Università Ca’ Foscari Venezia

Dottorato di Ricerca in Informatica, 24° Ciclo

(A.A. 2008/2009 – 2011/2012)

Sparse and Robust Matching Problem for 3D Shape Analysis

Settore scientifico disciplinare di afferenza: INF/01

Tesi di dottorato di Emanuele Rodolà
Matr. 955627

Tutore del dottorando
prof. Andrea Torsello

Coordinatore del dottorato
prof. Antonino Salibra

January, 2012

Author's Web Page: http://www.dsi.unive.it/~rodola

Author's e-mail: rodola@dsi.unive.it

Author's address:

Dipartimento di Scienze Ambientali, Informatica e Statistica
Università Ca' Foscari Venezia
Via Torino, 155
30172 Venezia Mestre – Italia
tel. +39 041 2348465
fax. +39 041 2348419
web: http://www.dais.unive.it

Abstract

In this thesis we approach different aspects of the all-pervasive correspondence problem in Computer Vision. Our main results take advantage of recent developments in the emerging field of game-theoretic methods for Machine Learning and Pattern Recognition, which we adapt and shape into a general framework that is flexible enough to accommodate rather specific and commonly encountered correspondence problems within the areas of 3D reconstruction and shape analysis. We apply said framework to a variety of matching scenarios and test its effectiveness over a wide selection of applicative domains, demonstrating and motivating its capability to deliver sparse, yet very robust solutions to domain-specific instances of the matching problem. Finally, we provide some theoretical insights that both confirm the validity of the method in a rigorous manner and foster new interesting directions of research.

Sommario

In questa tesi affrontiamo diversi aspetti del dilagante problema della corrispondenza nella Visione Artificiale. I nostri risultati principali traggono vantaggio da sviluppi recenti nel campo emergente dei metodi basati sulla Teoria dei Giochi in Machine Learning e Pattern Recognition, che adattiamo in un framework più generale. Tale framework è sufficientemente flessibile da gestire problemi di corrispondenza piuttosto specifici che comunemente si incontrano nelle aree della ricostruzione tridimensionale e di shape analysis. Il nostro metodo viene applicato a diversi scenari e altrettanti domini applicativi, dimostrando e motivandone l'efficacia nel fornire soluzioni sparse, ma al contempo molto robuste, a istanze specifiche del problema della corrispondenza. Diamo infine alcuni elementi teorici che non solo confermano in maniera rigorosa la validità del metodo, ma aprono anche nuove e interessanti direzioni di ricerca.

Contents

List of Figures

Preface

This thesis covers the work carried out during my post-graduate studies at the Department of Environmental Sciences, Informatics and Statistics of Ca' Foscari University of Venice, Italy. I worked in a small, yet tight-knit and enthusiast group comprising technically prepared, astonishingly creative, but above all very motivated young men who never feared undertaking research projects of any nature and proportion. While these three years as a PhD student undoubtedly proved to be essential for the shaping of both my technical and personal skills, it is with no second thoughts that I can safely confirm that passing this period in such a fervent environment contributed in fostering in me what I believe to be the most crucial of all achievements: the passion to do research.

Starting from a technology transfer project - the design and development of a full-fledged structured light 3D scanner for the eyewear industry - the main focus of my research became more and more apparent as I could discern a common denominator underlying the problems we encountered along the path. This cardinal point is the Matching Problem. While a major part of this thesis is dedicated to the question of correspondence, here, instead of providing an overview of the present treatise (which can indeed be found in the introductory chapters), I would like to start, perhaps provocatively, with a quote.

> With the first sentence, “Granted: I'm an inmate in a mental institution...”, the barriers fell, language surged forward, memory, imagination, the pleasure of invention, and an obsession with detail all flowed freely, chapter after chapter arose, history offered local examples, I took on a rapidly proliferating family, and contended with Oskar Matzerath and those around him over the simultaneity of events and the absurd constraints of chronology, over Oskar's right to speak in the first or third person, over his true transgressions and his feigned guilt.

Günter Grass, 2009.

Published Papers

[1] Albarelli, A., Rodolà, E., Rota Bulò, S., and Torsello, A. Fast 3d surface reconstruction by unambiguous compound phase coding. *IEEE International Workshop on 3D Digital Imaging and Modeling (3DIM2009)* (2009).

[2] Albarelli, A., Rodolà, E., and Torsello, A. Robust game-theoretic inlier selection for bundle adjustment. In *Fifth International Symposium on 3D Data Processing, Visualization and Transmission (3DPVT2010)* (2010).

[3] Albarelli, A., Rodolà, E., and Torsello, A. A game-theoretic approach to fine surface registration without initial motion estimation. In *The XXIII IEEE Conference on Computer Vision and Pattern Recognition (CVPR2010)* (2010).

[4] Rodolà, E., Albarelli, A., and Torsello, A. A game-theoretic approach to the enforcement of global consistency in multi-view feature matching. In *13th International Workshop on Structural and Syntactic Pattern Recognition (SSPR2010)* (2010).

[5] Albarelli, A., Rodolà, E., Cavallarin, A., and Torsello, A. Robust figure extraction on textured background: a game-theoretic approach. In *20th International Conference on Pattern Recognition (ICPR2010)* (2010).

[6] Rodolà, E., Albarelli, A., and Torsello, A. A game-theoretic approach to robust selection of multi-view point correspondence. In *20th International Conference on Pattern Recognition (ICPR2010)* (2010).

[7] Albarelli, A., Rodolà, E., and Torsello, A. Robust camera calibration using inaccurate targets. In *21st British Machine Vision Conference (BMVC2010)* (2010).

[8] Albarelli, A., Rodolà, E., and Torsello, A. Loosely distinctive features for robust surface alignment. In *11th European Conference on Computer Vision (ECCV2010)* (2010).

[9] Torsello, A., Rodolà, E., and Albarelli, A. Sampling relevant points for surface registration. In *2011 International Conference on 3D Imaging, Modeling, Processing, Visualization and Transmission (3DIMPVT2011)* (2011).

[10] Albarelli, A., Rodolà, E., and Torsello, A. A non-cooperative game for 3d object recognition in cluttered scenes. In *2011 International Conference on 3D Imaging, Modeling, Processing, Visualization and Transmission (3DIMPVT2011)* (2011).

[11] Bergamasco, F., Albarelli, A., Rodolà, E., and Torsello, A. RUNE-Tag: a high accuracy fiducial marker with strong occlusion resilience. In *The XXIV IEEE Conference on Computer Vision and Pattern Recognition (CVPR2011)* (2011).

[12] Torsello, A., Rodolà, E., and Albarelli, A. Multiview registration via graph diffusion of dual quaternions. In *The XXIV IEEE Conference on Computer Vision and Pattern Recognition (CVPR2011)* (2011).

[13] Albarelli, A., Rodolà, E., and Torsello, A. Imposing semi-local geometric constraints for accurate correspondences selection in structure from motion: a game-theoretic perspective. In *International Journal of Computer Vision (IJCV) - Special Issue* (2011).

[14] Torsello, A., Albarelli, A., and Rodolà, E. Stable and fast techniques for unambiguous compound phase coding. In *Image and Vision Computing (IVC)* (to appear).

[15] Albarelli, A., Rodolà, E., and Torsello, A. Fast and accurate surface alignment through an isometry-enforcing game. In *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)* (submitted).

[16] Rodolà, E., Albarelli, A., Bergamasco, F., and Torsello, A. A scale independent game for 3d object recognition. In *International Journal of Computer Vision (IJCV) - Special Issue* (submitted).

1
Introduction

This thesis is dedicated, for the most part, to the matching problem. Nevertheless, in an attempt to reflect the path followed by our research during these years, we devote the first part of this treatment to the general problem of 3D reconstruction, and present the research we carried out on the topic in two comprehensive chapters. However, it is in this first segment of our study that we will introduce our take on the matching problem. Making use of some notions from the mathematical field of Game Theory (Chapter 2), we will provide a first, if direct application of our matching framework to the widespread problem of feature detection (Chapter 4), and move on to the heart of our research starting from Chapter 5.

Surface Reconstruction

Surface reconstruction is the process by which the three-dimensional physical appearance of an object is captured and given a compact representation. Given the wide range of practical applications that could take advantage of a 3D reconstruction, it is not surprising that it has been a very active research topic during the last decades. In fact, many different approaches have been proposed in literature: some are aimed at solving the most general scenarios, others specialize to sub-domains, both in terms of the number of free parameters allowed and in terms of assumptions about some characteristic of the scene to be inferred. Here we will concentrate on two common and widespread (families of) methods, namely structured light scanning and Structure from Motion.

The main idea behind structured light reconstruction is that of assigning unique codes to surface points, according to a modulated intensity pattern which is projected over the object. Assuming the relative geometry between camera and light source is known, this projected information can be mapped to depth estimates for each point. Of course, there are many ways to design such a code, and all the methods come with their advantages and disadvantages depending on the specific task to solve. In case a two-camera setup is employed, reconstruction can be performed via triangulation of the (coded) image points depicted on the two cameras (here we are not interested in the technical details, for which we refer to the following chapters).

While structured light 3D reconstruction allows to obtain very accurate results and finds large application in many precision measurement tasks, there are also many scenarios in which it cannot be applied, say for the impossibility of controlling the scanning conditions, the lack of technological resources, or the physical characteristics of the object itself (e.g., translucent, transparent or simply huge objects such as entire buildings). In these cases, alternative scanning solutions have to be considered. Among these, the most widely utilized and studied, perhaps especially in these later years, are the so-called *Structure from Motion* (SfM) techniques. The common goal of SfM methods is to infer as many 3D clues as possible by analyzing a set of 2D images. In general, the 3D knowledge that can be obtained by such methods can be classified into two different (but related) classes: *scene* and *camera* information. Scene information is referred to the actual shape of the objects that are viewed in the images. This often boils down to assigning a plausible location in space to some significant subset of the acquired 2D points. These newly reconstructed 3D points are the "structure" part of SfM. By contrast, camera information includes all the parameters that characterize the abstract model of the image acquisition process. These can in turn be classified in intrinsic and extrinsic parameters. Intrinsic parameters are related to the physical characteristics of the camera itself, such as its focal length and principal point, while the extrinsic parameters define the camera pose, that is its position and rotation with respect to a conventional origin of the 3D space. Unlike the structure part, which is bound by physics of the object to a particular 3D position, the intrinsic and extrinsic parameter can vary in each shot; for this reason they are usually referred to as "motion".

Regardless of the technique adopted to obtain the reconstruction, there may be cases in which a single view of the object is not sufficient for the task. For instance, a single façade of a building might be enough for a photo insertion application, while objects should be scanned in their entirety in an archaelogical fragment reassembly scenario. Recomposing the partial views of an object together is a research topic in its own right, and we will return to this in the following section. Let us assume, for now, that we are given a method (for example a human) which, given a pair of surfaces representing different parts of the same object, aligns them with some residual error. Here by "alignment" we mean that the method gives an estimate of the rigid transformation linking the two partial views, and applies this relative motion to one of the surfaces. We can, in principle, achieve a global alignment of all the views by repeatedly invoking the motion oracle, and applying the transformation to a new view and the union resulting from the previous iterate. Nevertheless, this procedure is bound to induce an error accumulation, arising from both a not good enough pairwise alignment, and from surface noise that inevitably plagues the scanning process. First, we note that the initial approximate alignment can be refined by minimizing the overlap error between each pair of surfaces. This is usually done by means of some iterative technique bringing the two surfaces closer to one another at each successive step. This iterative procedure tends to be very susceptible to local minima; in our study, we isolated one of the major causes of this sensitivity in the specific choice of the surface samples used in the minimization process (typically, in order to reduce the size of the problem, not all the surface points are used). In Section 4.2, we propose a surface sampling technique that specifically tackles the point selection problem for rigid motion refinement in reconstruction pipelines. Even after applying the best possible refinement, error accumulation might still represent a problem, e.g. in measurement tasks. In literature there exist methods that try to alleviate this problem by moving the surfaces in such a way that the global overlap error is diffused evenly among all the views; however, all these methods act directly on the surface data and, for this reason, their performance both in terms of effectiveness and efficiency strongly depends on the specific choice of the surface samples on which they operate. Such methods have been in use for many years and are taken as the gold standard approach to surface alignment in most reconstruction pipelines. We offer a new solution to this problem, by introducing in Section 6.4 a novel multiview registration algorithm where the poses are estimated through a diffusion process on the view-adjacency graph. The diffusion process is over dual quaternions, a mathematical structure that is related to the group of 3D rigid transformations, leading to an approach that is both orders of magnitude faster than the state of the art, and much more robust to extreme positional noise and outliers.

The Matching Problem

We will start with a problem that we left open, specifically how to devise a method that computes the rigid transformation linking two 3D views of an object. Instead of explicitly looking for the best possible motion in the space of rigid transformations, we can translate this problem to a simpler one where we seek a group of corresponding points between the two shapes: given at least three such matches, it is in fact possible to estimate the rigid motion bringing one point set over the other. The search for such a correspondence constitutes an instance of the matching problem. Clearly, this kind of problem is extensively analyzed in literature, with methods taking a pointwise matching approach as above, and other approaches that rather explore the space of feasible motions in a RANSAC fashion (see Section 2.4 for a more detailed overview). All these methods share a common drawback: they attempt to obtain globally optimal solutions by taking local decisions, and only validate their matching hypotheses a posteriori. In our work, we take a totally different view and cast the matching problem in a novel framework (based on results from evolutionary game theory, Section 2.6), where the search for the best correspondence is regarded as an inlier selection problem; in this setting, the “strongest” possible group of matches that respect a measure of rigid compatibility is explicitly sought, and all incompatible matches are filtered out. The selection process operates in a globally aware manner, by making use of local information coming from pairs of points on the two surfaces. A thorough presentation of how we achieve this will be given in Chapter 6.

What we wrote above clearly represents only a particular instance of the matching problem, and more or less faithfully reflects the path of study we followed during our exploration of the somewhat more applicative reconstruction problem. We follow in fact a similar approach to the rigid registration case in a 2D matching setting for SfM applications (Chapter 5). In this case the matching problem bears similar characteristics to the 3D case, since it can be modelled again according to the same inlier selection principles:

one can define a measure of compatibility between pairs of image points, according to which the most coherent group of matches is extracted by means of some selection process operating over them. Clearly, in this case we need a different notion of compatibility, which also depends on the specific assumptions underlying the reconstruction task; to this end, we will rely on the common expectation (in SfM scenarios) that the two images are in fact separated by a small baseline, which in turn allows us to define a similarity measure between image patches in terms of some local affine transformation approximating the rigid motion in 3D space.

This second application of our matching framework allows us to inspect the general problem with more care, providing some interesting insights on its mathematical foundation and convergence properties. In the course of this thesis, we will investigate the possibilities of the framework in different areas, with applications to feature detection (Chapter 4), object-in-clutter recognition (Chapter 7) and deformable shape matching (Chapter 8). We will try to reconsider the matching problem under different points of view (Section 8.3), underlining the often overlooked fact that for many optimization problems there usually exist different, but equivalent mathematical formulations, which stress different structural characteristics of the problem, leading to different solution approaches. We will also try to establish a link between some of these formulations, although this remains an open direction of research that we are willing to explore.

2 Related Work

With this chapter we attempt to provide a comprehensive review of the existing literature, covering most of the topics approached in our work. What follows is by no means intended to be an exhaustive survey of existing techniques and state-of-the-art solutions, for which we refer to more thorough and rigorous treatments, such as [24, 117, 121, 120, 119]. We wish, instead, to give an untechnical (as possible), yet scientific and complete overview of recent research, considering both what is deemed to be seminal work and the current state of things. In doing so, we will also introduce much of the mathematical preliminaries that are required to go through the successive chapters, deferring the details and the specific techniques to appropriate sections, and referring to them directly when opportune.

We start off, maybe unsurprisingly, with an overview of the pinhole camera model (Section 2.1) and the epipolar geometry; this being a work focused for the major part on 3D reconstruction and the problems that frequently arise in this field, we soon shift our attention to camera calibration methods (Section 2.1.1) and structured light techniques (Section 2.2), to move rather directly to the Structure from Motion problem (Section 2.3). After introducing the all pervasive correspondence problem in computer vision, we move to the 3D surface domain by presenting the most common techniques for rigid surface alignment (Section 2.4), feature detection (Section 2.4.1) and object recognition (Section 2.5). We conclude with a section dedicated to Game Theory and its applications to computer vision (Section 2.6), the body of which is frequently referred in our work.

2.1 Camera Model and Epipolar Geometry

A camera is a mapping between the 3D world (object space) and a 2D image; while many different modelizations for this mapping have been proposed during the years in scientific literature, in this work we are primarily concerned with the simplest and most common camera model used in reconstruction frameworks, the *pinhole camera* (Figure 2.1a). Its wide adoption is due to its ability to approximate well the behaviour of many real cameras. In practical scenarios, radial and tangential lens distortions are the main sources of divergence from the pinole model, however it is easy to fit polynomial models to them and compensate for their effect [148, 156]. The most important parameters of this model are the pose of the camera with respect to the world (represented by a rotation matrix R and a translation vector T), the distance of the projection center from the image plane (the focal length f in Figure 2.1), and the coordinates on the image plane of the intersection between the optical axis and the plane itself (the principal point $c = (c_x, c_y)^T$ in Figure 2.1). The projection of a world point m on the image plane happens in two steps. The first step is a rigid body transformation from the world coordinate system to the camera system. This can be easily espressed (using homogeneous coordinate) as:

$$\begin{bmatrix} X_c \\ Y_c \\ Z_c \\ 1 \end{bmatrix} \sim \begin{bmatrix} \mathbf{R} & \mathbf{T} \\ 0 & 1 \end{bmatrix} \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix}$$

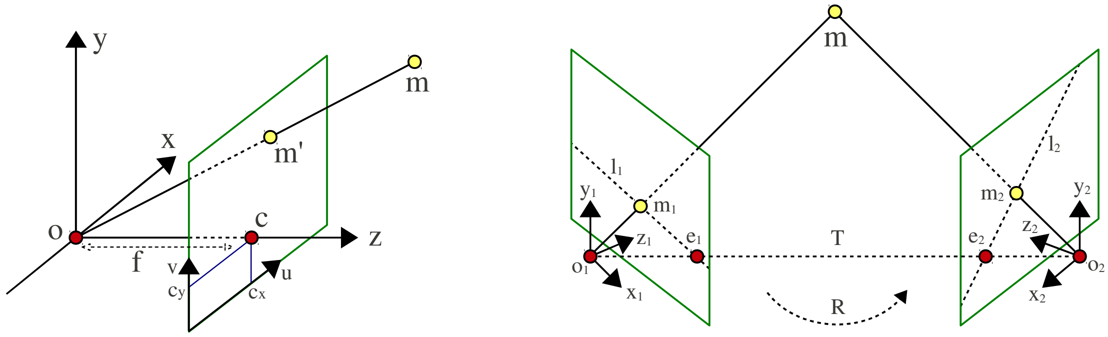
The second step is the projection of the point in camera coordiantes on the image planes that happens by applying a camera calibration matrix $\mathbf{K}$ that contains the intrinsic parameters of the model. The most general version of calibration matrix allows for a different vertical (f_y) and horizontal (f_x) focal length to accomodate for non-square pixels, and for a skewness factor (s) to account for non-rectangular pixels:

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

In practice, for most real cameras, pixels can be approximated by perfect squares, thus we can resort to the basic model of Figure 2.1 and assume $s = 0$ and $f_x = f_y = f$. Usually the camera pose and calibration matrices are combined in a single 3×4 projection matrix $\mathbf{P} = \mathbf{K}[\mathbf{R}\,\mathbf{T}]$. This projection matrix can be directly applied to a point in (homogeneous) world coordinate to obtain its corresponding 2D point in the image plane:

$$\mathbf{m}' = \mathbf{P}\mathbf{m} = \mathbf{K}[\mathbf{R}\,\mathbf{T}]\mathbf{m}\,.$$

When a point is observed by two cameras its projections on the respective image planes are not independent. In fact, given the projection m_1 of point m in the first camera, its projection m_2 on the second image plane must lie on the projection l_2 of the line that connects m_1 to m (see Figure 2.1b). This line is called the *epipolar line* and can be found for each point m_1 in the first image plane by intersecting the plane defined by o_1,o_2 and m_1 (the *epipolar plane*) with the second image plane. The epipolar constraint can be enforced exactly only if the position of m_1 and m_2 and the camera parameters are known without error. In practice, however, there will always be some distance between a projected point and the epipolar line it should belong to. This discrepancy is a useful measure for verification tasks such as the detection of outliers among alleged matching image points, or the evaluation of the quality of estimated camera parameters. The epipolar constraint can be expressed algebraically in a straightforward manner. If we know the rotation matrix and translation vector that move one camera reference system to the other (Figure 2.1b) we



(a) (b)

Figure 2.1: Illustration scheme of the pinhole camera model (a) and of the epipolar geometry (b). See text for details.

have that:

$$\mathbf{x_1^T E x_2} = \mathbf{x_1^T} \begin{bmatrix} 0 & -t_z & t_y \\ t_z & 0 & -t_y \\ -t_y & t_x & 0 \end{bmatrix} \mathbf{R x_2} = 0\,,$$

where the *essential matrix* $\mathbf{E}$ is the product between the cross product matrix of the translation vector $\mathbf{T}$ and the rotation matrix $\mathbf{R}$, and x_1 and x_2 are points expressed in the reference systems of the first and second camera respectively, belonging to the same epipolar plane. If the calibration matrices of both cameras are known this constraint can also be expressed in terms of image points by applying the inverse of the two calibration matrices to the image points:

$$(\mathbf{K_1}^{-1}\mathbf{m_1})^T\mathbf{E}(\mathbf{K_2}^{-1}\mathbf{m_2}^T) = \mathbf{m_1}^T(\mathbf{K_1}^{-1T}\mathbf{E}\mathbf{K_2}^{-1})\mathbf{m_2} = 0\,,$$

Where $\mathbf{F} = \mathbf{K_1}^{-1T}\mathbf{E}\mathbf{K_2}^{-1}$ is called the *fundamental matrix*. It is clear that if intrinsic camera parameters are known the epipolar constraint can be verified on image points by using just the essential matrix, which has only five degrees of freedom; otherwise, it is necessary to resort to the use of the fundamental matrix, which has seven degrees of freedom. Many algorithms are known to estimate both $\mathbf{E}$ or $\mathbf{F}$ from image point correspondences [71, 164, 144].

2.1.1 Camera Calibration

Accurate intrinsic camera calibration is essential to any computer vision task that involves image based measurements. Given its crucial role with respect to precision, a large number of approaches have been proposed over the last decades. Despite this rich literature, steady advancements in imaging hardware regularly push forward the need for even more accurate techniques. Some authors suggest generalizations of the camera model itself, others propose novel designs for calibration targets or different optimization schemas.

The growing availability of cheap and accurate digital imaging devices has led to the wide adoption of computer vision based methods in measurement applications. Typical scenarios range from the use of a single camera for quality check on planar profiles to the adoption of complex camera networks and structured light for precise 3D reconstructions. Crucial to these applications is the ability to know in a very accurate fashion the relation between physical points in the observed scene and the associated image points captured by the camera itself. This knowledge is usually attained by defining a parametric function from $\mathbb{R}^3$ to $\mathbb{R}^2$ and by supplying a technique for finding the most suitable set of parameters for a given camera. The function responsible for the transformation between world and image coordinates is usually (as anticipated) referred to as the (direct) *camera model*. The procedure used to search for the correct model parameters (with respect to some observations) is the *calibration method*. Obviously, camera models and calibration methods are strongly related: basic models require less parameters that can be estimated with fewer observations and simpler optimization procedures; by contrast, more sophisticated models describe more accurately the optical behaviour of real cameras and lenses, and thus are preferred when higher precision is required.

In this work we will focus on camera models based on perspective projection as they allow for a complete Euclidean scene reconstruction. The most basic model assumes an ideal pinhole camera characterized by a focal length f (the distance between the projective center and the image plane) and a principal point c (the intersection of the optical axis and the image plane). Often, the focal length is separated into a horizontal (fx) and vertical (fy) component in order to correct any possible non-squareness of the sensor elements. In real scenarios this simple model fails to deal properly with additional factors that contribute to the imaging process, such as distortions introduced by lenses or misalignment of the optical axis. In 1971 Brown [37] proposed a more sophisticated model that accounts for three radial and two decentering distortion coefficients (respectively $k1$, $k2$, $k3$, $t1$ and $t2$). While more complex models exist (accounting for affine and shear transformations of the image plane), the distortion coefficients introduced by Brown have shown to be adequate even for accurate metrology applications. Most calibration methods exploit the correspondences between 3D points of known geometry and 2D points on the image plane. Specifically, they seek a set of model parameters (intrinsic calibration) and camera orientation with respect to the world coordinate system (extrinsic calibration) that minimizes the distance from the measured 2D points and their respective projections obtained by applying the calibrated camera model to the 3D points (reprojection error). If the distortion parameters are not sought, it is possible to solve the calibration problem by using simple linear techniques, such as those suggested by Hall [69] and Faugeras [57]. Differently, it is necessary to introduce more complex approaches that usually alternate a linear technique to optimize a subset of the parameters and an iterative refinement step. One of the first calibration methods accounting for one radial distortion coefficient was proposed by Tsai [149]. Later approaches by Zhang [162] and Heikkilä [72] are able to deal respectively with two and three radial distortion coefficients; additionally, the latter technique also estimates two coefficients of tangential distortion. We refer to [119] for a comparative review of the most adopted calibration methods.

All the cited calibration methods rely both on the precise knowledge of the calibration object and on the accurate detection of its feature points on the image plane. The choice of an appropriate calibration object is often a compromise between the calibration accuracy required and the manufacturing complexity of the target itself. In general, 3D calibration objects (i.e., targets that are not composed of coplanar reference features) grant a more stable and precise calibration. This is mainly due to the lack of coupling between intrinsic and extrinsic parameters with respect to the minimization of the reprojection error. In addition, the presence of distinct depth information in the target limits the interplay between the estimation of the focal length and the lens distortion. Unfortunately, accurate 3D targets are very difficult to build, require an expensive tooling and need a remarkable amount of effort for maintenance and verification. While it is possible to use even one-dimensional objects [163], planar calibration targets are by far the most used both in experimental and industrial setups in virtue of their ease of construction. In principle, a single view of a planar target is not enough to determine all the calibration parameters at once [133]. Nevertheless, by exploiting a well planned network of independent views it is still possible to constrain the whole set of parameters and obtain an accurate estimation of the camera model [113].

Regardless of the target geometry chosen, it is equally important to select an appropriate feature pattern for the detection of the reference points. Zhang proposed to use a black square on a white background [162]; while this is an easily detectable pattern, it is biased by systematic localization errors caused by the inherent asimmetry of the marker (for instance, directional bleeding due to illumination). To reduce those biases, a better choice is to adopt symmetric markers such as circular points [72] or checkerboards [91]. Both these marker types have shown to be detectable with high subpixel accuracy, but the checkerboard schema deals better with the misplacement error introduced by radial distortion [92].

2.2 Structured Light

Assuming one or more calibrated cameras are available, it becomes possible to undertake a reconstruction process that allows to obtain a 3D point cloud (and eventually a connected surface) of the captured scene via triangulation [24]. The main challenge for any triangulation-based surface reconstruction technique is the assignment of reliable correspondences between features observed by two or more different points of view. Given the central role of this problem, many and diverse strategies have been proposed in literature over the past few decades [123]. When a sparse surface reconstruction is adequate, correspondences can be searched and tracked among repeatable features readily present in the scene, such as corners or edges. Unfortunately, in general it is not possible to guarantee that the same features are extracted from each image, or that the feature density is sufficient. Hence, complementary techniques, usually based on photometric correlation, are used to obtain an approximate reconstruction of the scene depth map. Structured light systems overcome these limitations as they do not rely on natural features, but instead use projected patterns of light in order to find correspondences that are usually as dense as the pixels of each image [24]. Among these, time-multiplexing strategies such as n-ary and Gray codes, as well as hybrid approaches, are by far the most utilized [121].

Simple binary coding assigns to every pixel a codeword retrieved from the digitized sequence over time of projected black and white stripes; binary coding methods require $log_2(t)$ pattern images to generate t code strings. Robustness of binary codes is improved by using Gray codes, where adjacent codes differ only in one bit. Both the techniques generate unique codes along each scanline, but at the same time are limited by their low resolution due to the inherently discrete nature of the coding. Also, the large number of projected patterns does not result in an increased accuracy. Generally, this class of measurements proves to be ineffective with objects having different reflective properties (such as slick metal parts or low reflective regions), thus they must rely on the assumption of uniform albedo [121].

Phase shifting methods, on the other hand, yield higher resolutions since they are based on the projection of periodic patterns with a given spatial period. Each projected pattern is obtained by spatially shifting the preceding one of a fraction of the period, and then captured by one or more cameras. The images are then elaborated and the phase information at each pixel determined by means of M-step relationships [138]. Since the phase is distributed continuously within its period, phase shifting techniques provide subpixel accuracy and achieve high measurement spatial resolution. Furthermore, the intensity change at each pixel for subsequent patterns is relative to the underlying color and reflectance, which makes phase shift locally insensitive to texture variance to a certain degree. A major drawback is that, in it basic formulation, phase shifted structured light renders only relative phase values and thus it is ambiguous. However, when both an extended measuring range and a high resolution are required, a combined approach proves to be very powerful. The integration of Gray code and phase shift brings together the advantages of both, providing disambiguation and high resolution, but the number of patterns to be projected increases considerably, and each strategy introduces a source of error [87].

Other high resolution shape measurement systems include optical profilometers. Non-contact phase profilometry techniques relate each surface point to three coordinates in a frame having the z axis orthogonal to a reference plane, which then represents the reference for the measured height [131, 158]. In classical phase measurement profilometry, gratings or sinusoidal patterns are projected and shifted first onto the plane and then over the object to be measured. Phase information from the deformed fringe pattern is then extracted by means of various techniques. Other, more effective profilometry techniques include the well-known Fourier Transform method [139] and other interesting derivatives [159, 67]. Fourier-based profilometry can require as few as one or two frames for depth estimation, which makes real-time reconstruction possible. Nevertheless, in profilometric methods phase variation caused by height modulation must be limited: ambiguity of the phase limits the measurement range, allowing for gauging of smooth-shaped objects only. Moreover, noise from camera, distortion of lens, difficulties of calibration, aliasing and imperfectness of the projecting unit influence the precision of most of these techniques

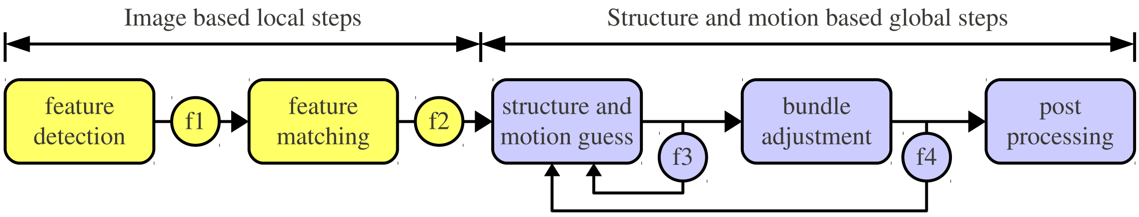


Figure 2.2: A simplified schema that captures the general steps found in many SfM approaches. The main loop is usually based on an iterative refinement of the candidate scene points based of their geometric consistency with respect to the estimated motion. Circles between steps represent the applied outlier filtering strategies.

[43, 135].

2.3 Structure from Motion

While the most relevant SfM approaches will be discussed with more detail in Section 2.3.2, in this section we will resort to the simplified general workflow presented in Figure 2.2 in order to introduce the key ideas and contributions of the proposed approach. To this end, the typical pipeline can be roughly split in two subsequent macro steps (respectively dubbed as *Image based* and *Structure and Motion based* in Figure 2.2). The first step deals with the localization on the source 2D images of salient feature points that are both distinctive and repeatable. Such points are meant to be tracked between different images, thus creating multiple sets of correspondences that will be used in the scene reconstruction step. The use of a reduced set of relevant points is crucial as their repeatable characterization allows us to minimize the chance of including wrong correspondences. Typically, filters are applied to the selection and matching phase in an attempt to make this phase more robust. In Figure 2.2 the extracted features are further culled by using filter *f1*, which eliminates point that exhibit very common descriptors or that are not distinctive or stable enough. A second refinement can be achieved after the matching: most implementations of filter *f2* remove correspondences that are not reliable enough, that is pairs where the second best match has a very similar score to the first one or that involve too different descriptors. Once a suitable set of point pairs has been found among all the images, the second macro step of the pipeline uses them to perform the actual structure and motion estimation. This happens by building a reasonable guess for both spatial locations of the found correspondences and the camera parameters and then, almost invariably, by adopting a bundle adjustment optimization to refine them. Also, at this stage, filtering techniques can be adopted in order to remove outliers from the initial set of matches. Specifically, a filter that removes pair that do not agree with the estimated epipolar constraints can be applied after combining some or all of the images into the initial guesses (*f3*), or after the bundle adjustment optimized the structure and motion estimates (*f4*). Depending on the result of the filtering a new initial estimation can be triggered, taking advantage of the (hopefully) more accurate selection of corresponding features. This kind of process leads to an iterative refinement that usually stops when the inlier set does not change or becomes stable enough. While this approach works well in many scenarios, it inherently contains a limitation that might drive it to poor results or even prevent it to converge at all: The main criterion for the elimination of erroneous matches is to exclude points that exhibit a large reprojection error or adhere poorly to the epipolar constraint after a first round of scene and pose estimation. Unfortunately this afterthought is based upon an error evaluation that depends on the point pairs chosen beforehand; this leads to a quandary that can only be solved by avoiding wrong matches from the start. This is indeed a difficult goal, mainly because the macro step from which the initial matches are generated is only able to exploit strictly local information, such as the appearance of a point or of its immediate surroundings. By contrast the following step would be able to take advantage of global knowledge, but this cannot be trusted enough to perform an extremely selective trimming and thus most methods settle with rather loose thresholds. In order to alleviate this limitation, in this work we introduce a robust matching technique that allows to operate a very accurate inlier selection at an early stage of the process and without any need to rely on preliminary structure and motion estimations. In the following we are going through a brief review of the most significant related contributions available in literature, and introduce some basic notions about the geometrical aspects of the SfM process.

2.3.1 Features Extraction and Matching

The selection of 2D point correspondences is arguably the most critical step in image based multi-view reconstruction and, unlike with techniques augmented by structured light or known markers, there is no guarantee that two pixel patches that exhibit strong photo-consinstecy are actually located on the projection of the same physical point. Further, even when correspondences are correctly assigned, the interest point detectors themselves introduce displacement errors that can be as large as several pixels. Such errors can easily lead to sub-optimal parameter estimation or, in the worst cases, to the inability of the optimization algorithm to obtain a feasible solution. For this reasons, reconstruction approaches adopt several specially crafted expedients to avoid the inclusion of outliers as much as possible. In the first place correspondences are not searched throughout all the image plane, but only points that are both repeatable and well characterized are considered. This selection is carried out by means of interest point detectors and feature descriptors. Salient points are localized with sub-pixel accuracy by general detectors, such as Harris Operator [70] and Difference of Gaussians [93], or by using techniques that are able to locate affine invariant regions, such as Maximally Stable Extremal Regions (MSER) [95] and Hessian-Affine [101]. This latter affine invariance property is desirable since the change in appearance of a scene region after a small camera motion can be locally approximated with an affine transformation. In Chapter 4 we will also introduce an outlier pruning technique that aims at isolating uncommon features, which are then treated as points of interest. Once interesting points are found, they must be matched to form the

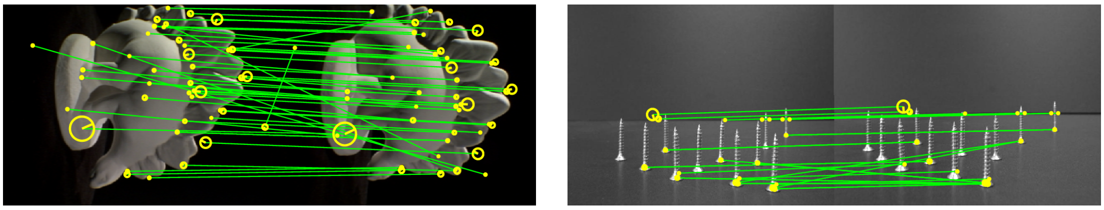


Figure 2.3: Example of SIFT features extracted and matched using the VLFeat package.

candidate pairs to be fed to the subsequent parameter optimization steps. Most of the currently used techniques for point matching are based on the computation of some affine invariant feature descriptor. Specifically, to each point is assigned a feature vector with tens to hundreds of dimensions, and a scale and a rotation value. Among the most used feature descriptor algorithms are the Scale-Invariant Feature Transform (SIFT) [90, 89], the Speeded Up Robust Features (SURF) [25], the Gradient Location and Orientation Histogram (GLOH) [102] and more recently the Local Energy based Shape Histogram (LESH) [122], the SIFT algorithm being the first of the lot and arguably the most widely adopted. For this reason several enhancements and specialization were introduced since the original paper by Lowe; for intance, PCA-SIFT [82] applies PCA to the normalized gradient patch to gain more distinctiveness, PHOW [31] makes the descriptor denser and allows to use color information, ASIFT [103] extends the method to cover the tilt of the camera in addition to scale, skew and rotation.

In all of these techniques, the descriptor vector is robust with respect to affine trasformations: i.e., similar image regions exhibit descriptor vectors with small mutual Euclidean distance. This property is used to match each point with the candidate with the nearest descriptor vector. However, if the descriptor is not distinctive enough this approach is prone to select many outliers. A common optimization involves the definition of a maximum threshold over the distance ratio between the first and the second nearest neighbors. In addition, points that are matched multiple times are deemed as ambiguous and discarded (i.e., one-to-one matching is enforced). Despite any effort made in this direction, any filter that operates at a local level is fated to fail when the matched regions are very similar or identical, a situation that is not uncommon as it happens every time an object is repeated multiple times in the scene or there is a repeated texture. In Figure 2.3 we show two examples of SIFT features extracted and matched by using the VLFeat [151] Matlab toolkit. In the first example almost all the correspondences are correct, still some clear mismatches are visible both between the plates of the saurus (which are similar in shape) and on the black background (which indeed contains some amount of noise). In the second example several instances of the same screw are matched and, as expected, features coming from different objects are confused and almost all the correspondences are wrong. It should be noted that such mismatches are not a fault of the descriptor itself as it performs exactly its duty by assigning similar description vectors to features that are almost identical from a photometric standpoint. In fact, this particular example is specially crafted to hinder traditional matchers that relies on local properties. In Chapter 5 we will show how introducing some level of global awareness in the process allows to deal well also with this cases that are indeed very common in the highly repetitive world of human-made objects and urban environments.

2.3.2 Parameters Estimation

The distinctive traits of every Structure from Motion technique proposed in literature are usually to be found in the approach used for the initial estimate and in the refinement technique adopted. In general this refinement happens by iteratively applying a bundle adjustment algorithm [147] to an initial set of correspondences, 3D points and motion hypoteses. This optimization is almost invariantly carried out by means of the Levenberg-Marquardt algorithm [86], which is very sensitive to the presence of outliers in the input data. For this reason any possible care should be taken in order to supply the optimizer with good hypoteses or at least a minimal number of outliers. When a reasonable subset of all the points is visible in all the images global methods can be used to obtain such initial hypothesis. This approach, commonly called *factorization*, was initially proposed only for simplified camera models that are not able to fully capture the pihole projection [142, 155]. More recently, similar approaches have been presented also for perspective cameras [134, 73]; however, their need to have each point visible in each camera severely reduces their usability in practical scenarios where occlusion is usually abundant. For this reason, incremental methods which allow to add one or a few images at a time are by far more popular in SfM applications. Usually, such methods start from a reliable image pair (for instance the pair with the higher number of good correspondeces), build an initial reconstruction by triangulation and thus extend it sequentially. The extension can happen by virtue of common 2D point between a new camera and one or more images already in the batch. If internal camera parameters are known (at least roughly), rotation and translation direction can be extracted from the essential matrix and translation magnitude can be found using the projection in the new image of an already reconstructed 3D point. In the more general case intrinsic parameters are not known and the new camera can be added exploiting the correspondences between its 2D features and previuosly triangulated 3D points to estimate the projection matrix [26, 110]. Finally, it is also possible to merge partial reconstructions by using corresponding 3D points [59]. Many modern SfM approaches iterate this process by including and excluding point correspondences or entire images by validating them with respect to the currently estimated structure and camera poses [38, 152, 130].

2.4 Fine and Coarse Registration Techniques

Once a collection of 3D surfaces is obtained by means of some reconstruction technique, dependently from the specific computer vision task it might be necessary to reassemble these surfaces together so as to compose the original object they represent. In a typical 3D object scanning setting, for instance, it is fairly common that the object be captured by different points of view and then these view-dependent scannings rigidly aligned together in a semi-supervised post-processing step. Note, however, that this is not always the case: a single 3D view might be sufficient in simple object recognition scenarios (e.g., in industrial inspection applications), or the scanning device might be equipped with a special optical apparatus that allows to capture more surface area than traditional pinhole models. Further, the last few months have seen the germination of depth-enabled motion sensing devices (especially for gaming purposes) capturing video data in 3D as time-sequenced depth maps. Nevertheless, alignment methods are not useful only for the assemblage of partial 3D scans. In fact, they find useful applications as tools for in-line quality control [104], 3D object recognition [99], advanced human interfaces [18], and SLAM [30], just to name a few. For this reason, surface registration is one of the most studied topics in the field of 3D data acquisition and processing. In this thesis we try to introduce a fresh view on the problem by proposing a novel approach that is very robust to noise and allows to attain a very accurate registration without needing an initial motion estimation.

The distinction between fine and coarse surface registration methods is mainly related to the different strategies adopted to find pairs of mating points to be used for the estimation of the rigid transformation. Almost invariably, fine registration algorithms exploit an initial guess in order to constrain the search area for compatible mates and minimize the risk of selecting outliers. On the other hand, coarse techniques, which cannot rely on any motion estimation, must adopt a mating strategy based on the similarity between surface-point descriptors or resort to random selection schemes. The tension between the precision required for fine alignment versus the recall needed for an initial motion estimation stands as the main hurdle to the unification of such approaches. The large majority of currently used fine alignment methods are modifications to the original ICP proposed by Zhang [161] and Besl and McKay [29]. These variants generally differ in the strategies used to sample points from the surfaces, reject incompatible pairs, or measure error. In general, the precision and convergence speed of these techniques is highly data-dependent and sensitive to the fine-tuning of the model parameters. Several approaches that combine these variants have been proposed in the literature in order to overcome these limitations (see [117] for a comparative review). No matter what variant is used, ICP, being an iterative algorithm based on local, step-by-step decisions, is susceptible to the presence of local minima. Some recent variants mitigate this problem by avoiding hard culling assigning a probability to each candidate pair by means of evolutionary techniques [88] or Expectation Maximization [66]. Other, non ICP-based, fine registration methods include the well-known method by Chen [44] and signed distance fields matching [94].

Coarse registration techniques can be roughly organized into three main classes: global methods, feature-based methods and technique based on RANSAC [58] or PROSAC [46] schemes. Global methods, such as PCA [47] or Algebraic Surface Model [140] exploit some global property of the surface and thus are very sensitive to occlusion. Feature-based approaches aim at the localization and matching of interesting points on the surfaces. They are more precise and can align surfaces that exhibit only partial overlap. Nevertheless, the unavoidable localization error of the feature points prevents them from obtaining accuracies on par with fine registration methods. A completely different coarse registration approach is the one taken by RANSAC-based techniques. DARCES [41] is based on the random extraction of sets of mates from the surfaces and their validation based on the accuracy of the estimated transformation. The more recent Four Points Congruent Sets method [19] follows a similar route, but filters the data to reduce noise and performs early check in order to reduce the number of trials. A recent and extensive review of many different methods can be found in [120].

Regardless of the criteria used to obtain pairs of mating points, the subsequent step in surface registration is to search for the rigid transformation that minimizes the squared distance between them. Many mature techniques are available to do this (see for instance [74]).

2.4.1 Feature Detection on 3D Objects

Feature detection and characterization is a key step in many tasks involving the recognition, registration or database search of 2D and 3D data. Specifically, when suitable interest points are available, all these problems can be tackled by working with the set of extracted features rather than dealing with the information carried by the whole data, which is less stable and noisier. For an interest point to be reliable it must exhibit two properties: repeatability and distinctiveness. A feature is highly *repeatable* if it can be detected with good positional accuracy over a wide range of noise levels and sampling conditions as well as different scales and transformations of the data itself. Further, description vectors calculated over interesting points are said to be *distinctive* if the descriptors related to different features lie far apart in feature space, while descriptor associated to multiple instances of the same point have lie within a small distance from one another. These properties are somewhat difficult to attain since they are subject to antithetical goals: In fact, to achieve good repeatability despite of noise, larger patches of data must be considered, which unfortunately leads to a lower positional precision and a less sharp culling of uninteresting points. Moreover, for descriptor vectors to be distinctive among different features, they need to adopt a large enough basis, which, owing to the well known "curse of dimensionality," also affects their coherence over perturbed versions of the same feature. In the last two decades these quandaries have been addressed with great success in the domain of 2D images, where salient points can be localized with sub-pixel accuracy using detectors exploiting strong local variation in intensity, such as Harris Operator [70] and Difference of Gaussians [93], or using techniques that are able to locate affine invariant regions, such as Maximally Stable Extremal Regions (MSER) [95] and Hessian-Affine [101]. Among the most used descriptors are the Scale-invariant feature transform (SIFT) [89], the Speeded Up Robust Features (SURF) [25] and Gradient Location and Orientation Histogram (GLOH) [102]. While these approaches work well with 2D intensity images, they cannot be easily extended to handle 3D surfaces since no intensity information is directly available. On the other hand, there has been huge effort to use other local measures, such as curvature or normals. One of the first descriptors to capture the structural neighborhood of a surface point was described by Chua and Jarvis, who with their Point Signatures [45] suggest both a rotation and translation invariant descriptor and a matching technique. Later on, Johnson and Hebert introduced Spin Images [76], a rich characterization obtained by binning the radial and planar distances of the surface samples respectively from the feature point and from the tangent plane. Given their ability to perform well with both surface registration and object recognition, Spin Images have become one of the most used 3D descriptors. More recently, Pottmann et al. proposed the use of Integral Invariants [111], stable multi-scale geometric measures related to the curvature of the surface and the properties of its intersection with spheres centered on the feature point. Zaharescu et al. [160] presented a comprehensive approach for interest point detection (MeshDOG) and description (MeshHOG), based on the value of any scalar function defined over the surface (i.e. curvature or texture, if available). MeshDOG localizes feature points by searching for scale-space extrema over progressive Gaussian convolutions of the scalar function and thus by applying proper thresholding and corner detection. MeshHOG calculates a histogram descriptor by binning gradient vectors with respect to a rotationally invariant local coordinate system. Finally, the recent SHOT descriptor [143], introduced by Tombari et al. exploits a novel 3D reference frame to offer enhanced descriptive power and robustness.

In the following chapters we introduce a novel pipeline that can be used to obtain an accurate surface registration without requiring an initial motion estimation. We propose very simple descriptors, named *Surface Hashes*, that span only 3 to 5 dimensions. As their name suggests, we expect Surface Hashes to be repeatable through the same feature point, yet to suffer from a high level of clashing due to their limited distinctiveness. In order to overcome this liability we also adopt a robust inlier selection approach which exploits rigidity constraints among surfaces to guarantee a global geometric consistency. The combination of these loosely distinctive features and our robust matcher leads to an effective and robust surface alignment approach.

2.5 Object Recognition

In the recent past, the acquisition of 3D data was only viable for research labs or professionals that could afford to invest in expensive and difficult to handle high-end hardware. However, due to both technological advances and increased market demand, this scenario has been altered significantly: Semi-professional range scanners can be found at the same price level of a standard workstation, widely available software stacks can be used to obtain reasonable results even with cheap webcams, and, finally, range imaging capabilities have been introduced even in very low-end devices such as game controllers. Given this trend, it is safe to forecast that range scans will be so easy to acquire that they will complement or even replace traditional intensity-based imaging in many computer vision applications. The added benefit of depth information can indeed enhance the reliability of most inspection and recognition tasks, as well as providing robust cues for scene understanding or pose estimation. Many of these activities include fitting a known model to a scene as a fundamental step. For instance, a setup for in-line quality control within a production line, could need to locate the manufactured objects that are meant to

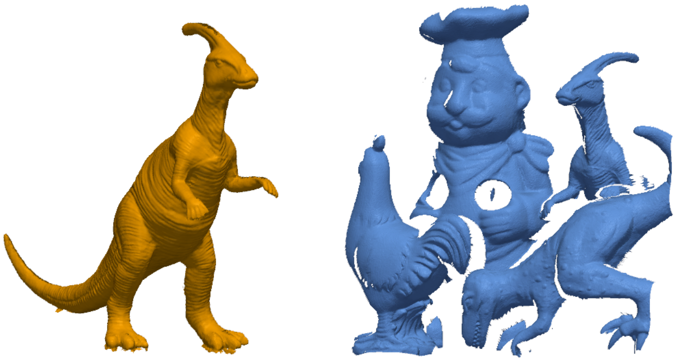


Figure 2.4: A typical 3D object recognition scenario. Clutter of the scene and occlusion due to the geometry of the ranging sensor seriously hinder the ability of both global and feature-based techniques to spot the model.

be measured [104]. Moreover, a range-based SLAM system can exploit the position of known 3D reference objects to achieve a more precise and robust robot localization [30]. Finally, non-rigid fitting could be used to recognize hand or whole-body gestures in next generation interactive games or novel man-machine interfaces [18].

The matching problem in 3D scenes shares many aspects with object recognition and location in 2D images: The common goal is to find the relation between a model and its transformed instance (if any) in the scene. In both cases, transformations may include uniform and non-uniform scaling, differences in pose or partial modification of the shape. In the 3D case, scenes can undergo a variety of non-rigid deformations such as topological noise (inherent in some measure to the very nature of the acquisition process), local scale variations and even global affine deformations or warping effects due, for instance, to miscalibration of the scanning device or to the action of natural forces on the objects in rather specific scenarios [64]. While in general severe deformations of the scene are unlikely to occur, they are commonly present in a measure and should be accounted for in matching applications. Other common hurdles, both in 2D and 3D, comprise measurement errors on intensities or point positions and indirect changes in the appearance due to occlusion or the simultaneous presence in the scene of extraneous objects that can act as distractions.

A similar problem to object recognition in the 3D domain is surface registration, which we briefly introduced in the previous sections. In this case, two surfaces representing different point of views (with unknown positions) of the same object are to be rigidly aligned one to another. While there may be apparently many aspects in common with this class of problems, adopting the same techniques to solve both can be far from beneficial. Most surface alignment methods like RANSAC-based DARCES [41] or 4-points Congruent Sets [19] (currently at the state-of-the-art for surface registration) first generate pseudo-random, not necessarily point-based matching hypotheses and then validate the match in an attempt to maximize the overall surface overlap. It is clear that in an object recognition scenario, where occlusion and clutter are present, and where the object itself cannot even be assumed to be in the scene, such approaches can give completely wrong results, fooled by the structured noise offered by the clutter. Thus, even though in a technical sense similar methods can be adopted for both surface registration and object recognition, the assumptions underlying the two problems, as well as the expected results, are very different.

Among the basic approaches to object recognition are feature-based techniques, which adopt descriptors that are associated to single points respectively on the image (in the 2D case) or on the object surface. In principle, each feature can be matched individually by comparing the descriptors, which of course decouples the effect of partial occlusion. In the 2D domain, intensity based descriptors such as SIFT [90] or affine-invariant alternatives such as MSER [95] have proven to be very distinctive and capable to perform very well even with naive matching methods that do not include any global information [89]. However, the problem of balancing local and global robustness is more binding with 3D scenes than with images, as no natural scalar field is available on surfaces and thus feature descriptors tend to be less distinctive. In practice, global or semi-global inlier selection techniques are often used to avoid wrong correspondences. This, while making the whole process more robust to a moderate number of outliers, can introduce additional weaknesses. As already said, for instance, if a RANSAC-like inlier selection is applied, occlusion coupled with the presence of clutter (*i.e.*, unrelated objects in the scene) can easily lower the probability for the process to find the correct match. The limited distinctiveness of surface features can be tackled by introducing scalar quantities computed over the local surface area. This is the case, for instance, with values such as mean curvature, Gaussian curvature or shape index and curvedness, which can be constructed in order to classify surface patches into types such as pits, peaks or saddles [20]. Unfortunately, this kind of characterization has proven to be not very selective for matching purposes, since it is frequent to obtain similar values in many different locations.

Another approach is to augment the point data with additional scalar values that can be obtained during the acquisition process. To this extent, the use of natural textures coming from the scanned object have shown to allow good performance since they show high variability, and can be used to compute descriptors similar to those usually adopted in the 2D domain. Examples of these augmented descriptors include Textured Spin Images [50], MeshDOG [160], which indeed can benefit from any scalar field defined over the surface, and more recently Color Cubic Higher-order Local Auto-Correlation Features [77], which measure the autocorrelation function of color 3D voxel data at specific points, so as to guarantee resilience to various forms of noise. However, not all the surface digitizing techniques allow to acquire texture information, and even when available, the usability of textures for descriptor extraction strongly depends on the appearance of the scanned object. To overcome the limitations of scalar descriptors, methods that gather information from the whole neighborhood of each point to characterize have been introduced. Such methods can be roughly classified in approaches that define a full reference frame for each point (for instance, by using PCA) and techniques that only need a reference axis (usually some kind of normal direction for the point). When a full reference frame is available it is possible to build very discriminative descriptors [45, 137]. Unfortunately, noise and differences in the mesh could lead to instabilities in the reference frame, and thus to a brittle descriptor. By converse, methods that just require a reference axis (and are thus invariant to the rotation of the frame) trade some descriptiveness to gain greater robustness. These latter techniques almost invariably build histograms based on some properties of points falling in a cylindrical volume centered and aligned to the reference axis. The most popular histogram-based approach is certainly Spin Images [76], but many others have been proposed in literature [42, 60, 111].

Lately, an approach that aims to retain the advantages of both full reference frames and histograms has been introduced in [143]. In their work, the authors introduce a robust method for the estimation of highly repeatable local reference frames, together with a novel representation that is designed to be efficient to compute, descriptive and robust to noise and clutter. The authors take the hint from SIFT descriptors in the 2D domain, to build a 3D descriptor that encodes histograms of basic differential entities which are further enhanced by introducing geometric information of the points within the given support. The set of local histograms is first computed over the 3D volumes defined by a spherical grid superimposed on the support; these local histograms are then grouped together to form the actual descriptor. Laying at the intersection between histograms and signatures, the descriptor is dubbed Signature of Histograms of Orientations (SHOT).

Any of the interest point descriptors above can be used to find correspondences between a model and a 3D scene that can contain it. Most of the cited papers, in addition to introducing the descriptor itself, propose some matching technique. These span from very direct approaches, such as associating each point in the model with the point in the scene having the most similar descriptor, to more advanced techniques such as customized flavors of PROSAC [46] and specialized keypoint matchers that exploit locally fitted surfaces for computing depth values to use as feature components [100].

Recent contributions providing both a descriptor and a matching technique for the specific problem of object recognition in clutter, include the works by [23] and [106]. The authors represent object and scene with a set of scale-dependent corners and associated scale-invariant descriptors. Recognition is performed via an interpretation tree whose nodes represent correspondences between a model feature and scene feature, with each branch representing a hypothesis about the possible presence (and pose) of that model in the scene. In the matching step, hypotheses are effectively culled from the tree by exploiting the intrinsic scale of each scale-dependent corner, whose possible correspondences are restricted to corners detected at the same intrinsic scale. In [23] the authors also claim to introduce the first method to attack scale-invariant recognition in cluttered scenarios. In [99], the authors presented an object recognition and segmentation algorithm based on a tensorial representation for descriptors. Pairs of vertices are randomly selected from the model to construct a tensor, then matched with the tensors of the scene by casting votes using a 4D hash table. Tuples receiving too few votes are discarded and a similarity measure is calculated between the scene tensor and the tensors of the remaining tuples. Hypotheses having highest similarities are verified by transforming the 3D model to the scene, and evaluating the surface overlap.

All the methods presented above share a common burden: given the highly dimensional space of matching hypotheses, they try to reduce the exploration set by adopting different kinds of heuristics and by validating their choices a posteriori. Further, there is almost invariably a random component underlying the search method, which renders these approaches non-deterministic by nature. While performance of these methods is generally more than acceptable, still, somehow surprisingly they seem to fail at apparently simple scenarios or, by converse, succeed in objectively more challenging cases. This behavior can be symptomatic of a number of reasons, among which we identify as a possible cause the fact that such approaches try to generate globally optimal solutions by operating exclusively at a local level, utilizing global information in ex-post validation steps only; this makes for a method that is not globally aware, and as such is bound to give unexpected results as the objects to be matched are incomplete, noisy or not even present in the scene.

2.6 Game Theory in Computer Vision

In this section we introduce some principles of Game Theory together with the mathematical preliminaries that will be used throughout the following chapters. This makes this last section rather technical, although we will limit ourselves to the general results and refer instead to the rest of the thesis for a more insightful treatment.

Originated in the early 40's, Game Theory was an attempt to formalize a system characterized by the actions of entities with competing objectives, which is thus hard to characterize with a single objective function [154]. According to this view, the emphasis shifts from the search of a local optimum to the definition of equilibria between opposing forces, providing an abstract theoretically-founded framework to model complex interactions. In this setting, multiple players have at their disposal a set of strategies and their goal is to maximize a payoff that depends also on the strategies adopted by other players. Here we will concentrate on symmetric two-player games, i.e., games between two players that have the same set of available strategies and that receive the same payoff when playing against the same strategy.

2.6.1 Non-cooperative Games

In [21] Albarelli et al. introduced an approach based on principles of Game Theory in an attempt to tackle the all-pervasive correspondence problems encountered in Computer Vision. Deriving their formulation from a clustering method [146], the authors made for a rather general approach which can be applied to many different settings, as long as an objective function and feasible set can be defined in terms of unary and pairwise interactions between players of some *matching game*.

More formally, let $O = \{1, \cdots, n\}$ be the set of available strategies (*pure strategies* in the language of game theory), and $\Pi = (\pi_{ij})$ be a matrix specifying the payoffs that an individual playing strategy i receives against someone playing strategy j. A *mixed* *strategy* $\mathbf{x}$ is a randomization of the available strategies, i.e., a probability distribution $\mathbf{x} = (x_1, \ldots, x_n)^T$ over the set O. As such, mixed strategies are constrained to lie in the n-dimensional standard simplex

$$\Delta^n = \left\{ \mathbf{x} \in \mathbb{R}^n \ : \ x_i \geq 0 \text{ for all } i \in 1 \ldots n, \ \sum_{i=1}^{n} x_i = 1 \right\} .$$

The *support* of a mixed strategy $\mathbf{x} \in \Delta^n$, denoted by $\sigma(\mathbf{x})$, is defined as the set of elements chosen with non-zero probability: $\sigma(\mathbf{x}) = \{i \in O \mid x_i > 0\}$. The expected payoff received by a player choosing element i when playing against a player adopting a mixed strategy $\mathbf{x}$ is $(\Pi\mathbf{x})_i = \sum_j \pi_{ij} x_j$, hence the expected payoff received by adopting the mixed strategy $\mathbf{y}$ against $\mathbf{x}$ is $\mathbf{y}^T \Pi \mathbf{x}$. The *best replies* against mixed strategy $\mathbf{x}$ is the set of mixed strategies maximizing the expected payoff against $\mathbf{x}$.

$$\beta(\mathbf{x}) = \{\mathbf{y} \in \Delta^n \mid \mathbf{y}^T \Pi \mathbf{x} = \max_{\mathbf{z}} (\mathbf{z}^T \Pi \mathbf{x})\} .$$

The best reply is not necessarily unique. Indeed, except in the extreme case in which there is a unique best reply that is a pure strategy, the number of best replies is always infinite.

A central notion of game theory is that of a Nash equilibrium. A strategy $\mathbf{x}$ is said to be a *Nash equilibrium* if it is a best reply to itself, i.e., $\forall \mathbf{y} \in \Delta^n, \ \mathbf{x}^T \Pi \mathbf{x} \geq \mathbf{y}^T \Pi \mathbf{x}$. This implies that $\forall i \in \sigma(\mathbf{x})$ we have $(\Pi\mathbf{x})_i = \mathbf{x}^T \Pi \mathbf{x}$; that is, the payoff of every strategy in the support of $\mathbf{x}$ is constant, while strategies outside the support of $\mathbf{x}$ earn a smaller or equal payoff. The idea underpinning the concept of Nash equilibrium is that a rational player will consider a strategy viable only if no player has an incentive to deviate from it.

2.6.2 Evolutionary Dynamics

We undertake an evolutionary approach to the computation of Nash equilibria. Evolutionary Game Theory originated in the early 70's as an attempt to apply the principles and tools of game theory to biological contexts. It considers an idealized scenario where pairs of individuals are repeatedly drawn at random from a large population to perform a two-player game. In contrast to traditional game-theoretic models, players are not supposed to behave rationally, but rather act according to a pre-programmed behavior, or mixed strategy. Further, it is supposed that some selection process operates over time on the distribution of behaviors favoring players that receive higher payoffs.

In this dynamic setting, the concept of stability, or resistance to invasion by new strategies, becomes central. A strategy $\mathbf{x}$ is said to be an *evolutionary stable strategy* (ESS) if it is a Nash equilibrium and

$$\forall \mathbf{y} \in \Delta^n \quad \mathbf{x}^T \Pi \mathbf{x} = \mathbf{y}^T \Pi \mathbf{x} \Rightarrow \mathbf{x}^T \Pi \mathbf{y} > \mathbf{y}^T \Pi \mathbf{y} . \tag{2.1}$$

This condition guarantees that any deviation from the stable strategies does not pay.

Interestingly, in the special case in which payoff matrix Π is symmetric, there is relationship with optimization theory [154]: Stable states correspond to the strict local maximizers of the average payoff $\pi(\mathbf{x}|\mathbf{x}) = \mathbf{x}^{\mathrm{T}} \Pi \mathbf{x}$ over Δ^n, whereas all critical points are related to Nash equilibria. Assuming a mechanism to reach a stable state is available, this interesting property provides us with a rather flexible and general tool that we can adapt and employ for our purposes.

The search for a stable state is performed by simulating the evolution of a natural selection process. A common way to characterize this process is via the *replicator dynamics* equation:

$$x_i(t+1) = x_i(t)\frac{(\Pi \mathbf{x}(t))_i}{\mathbf{x}(t)^{\mathrm{T}}\Pi \mathbf{x}(t)} \tag{2.2}$$

for $i = 1 \dots n$. It can be easily seen that the equation above guarantees $\mathbf{x}(t) \in \Delta^n$ for all $t \geq 0$. Under this dynamics, the average payoff of the population is also guaranteed to (strictly) increase (provided that Π is nonnegative and symmetric), that is for all $t \geq 0$ we have $\pi(\mathbf{x}(t+1)) \geq \pi(\mathbf{x}(t))$ and the equality holds only when $\mathbf{x}$ is a stationary point for the dynamics, i.e., $\mathbf{x}(t+1) = \mathbf{x}(t)$. According to equation (2.2), the fraction of individuals adopting strategy i will grow over time whenever their expected payoff exceeds the population average, decreasing otherwise. In addition, any such sequence will always converge to a unique solution (a Nash equilibrium).

The equation above can be conveniently interpreted as follows. It is supposed that, during a (non-cooperative) game, it is in each player's interest to pick strategies that are compatible with those the other players are likely to choose. In general, as the game is repeated, players will adapt their behavior to prefer strategies that yield larger payoffs, driving all inconsistent hypotheses to extinction, and settling for an equilibrium where the pool of strategies from which the players are still actively selecting their associations forms a cohesive set with high mutual support. In a sense, this iterative process can be seen as a contextual voting system, where each time the game is repeated the previous selections of the other players affect the future vote of each player in an attempt to reach consensus. This way, the evolving context brings global information into the selection process.

Unfortunately, replicator dynamics can be very inefficient for large problems, each iteration having $O(n^2)$ complexity. Faster alternatives have been introduced in literature [108]; in the following we focus on a new class of dynamics called *infection and immunization*, bringing per-iteration complexity to $O(n)$, recently introduced by Rota Bulò et al. [115]. For a more rigorous treatment we refer to the original paper.

Let $\mathbf{x} \in \Delta^n$ be an *incumbent* population state, and consider a *mutant* population $\mathbf{y} \in \Delta^n$ invading $\mathbf{x}$. Let $\mathbf{z} = (1-\varepsilon)\mathbf{x} + \varepsilon\mathbf{y}$ be the population state obtained after injecting a ε share of $\mathbf{y}$-strategists into $\mathbf{x}$. We define the *invasion barrier* of $\mathbf{x}$ against mutant $\mathbf{y}$ as

$$b_{\mathbf{x}}(\mathbf{y}) = \inf(\{\varepsilon \in (0,1) : h_{\mathbf{x}}(\mathbf{y},\varepsilon) > 0\} \cup \{1\}), \tag{2.3}$$

where $h_{\mathbf{x}}(\mathbf{y},\varepsilon)$ is a *score function* of $\mathbf{y}$ versus $\mathbf{x}$, defined as

$$h_{\mathbf{x}}(\mathbf{y},\varepsilon) = \pi(\mathbf{y} - \mathbf{x}|\mathbf{z}). \tag{2.4}$$

The barrier defines the largest population share of $\mathbf{y}$-strategists such that, for all smaller shares, $\mathbf{x}$ earns a greater or equal payoff than $\mathbf{y}$ in $\mathbf{z}$. Population $\mathbf{x}$ is said to be *immune* to

$\mathbf{y}$ if $b_{\mathbf{x}}(\mathbf{y}) > 0$, which happens whenever either $\pi(\mathbf{y}-\mathbf{x}|\mathbf{x}) < 0$, or $\pi(\mathbf{y}-\mathbf{x}|\mathbf{x}) = 0$ and $\pi(\mathbf{y}-\mathbf{x}) \leq 0$. Otherwise, $\mathbf{y}$ is *infective* for $\mathbf{x}$, which in turn implies that $b_{\mathbf{x}}(\mathbf{y}) = 0$. As long as the score of $\mathbf{y}$ versus $\mathbf{x}$ is positive, by allowing a small share of $\mathbf{y}$-strategists to invade $\mathbf{x}$, we are left with a share $\delta_{\mathbf{y}}(\mathbf{x})$ of mutants in $\mathbf{z}$, where

$$\delta_{\mathbf{y}}(\mathbf{x}) = \inf(\{\varepsilon \in (0,1) : h_{\mathbf{x}}(\mathbf{y},\varepsilon) \leq 0\} \cup \{1\}). \tag{2.5}$$

Then if $\mathbf{y}$ is infective for $\mathbf{x}$, $\delta_{\mathbf{y}}(\mathbf{x}) > 0$ and in case $\mathbf{x}$ is immune to $\mathbf{y}$ we have $\delta_{\mathbf{y}}(\mathbf{x}) = 0$. In [115] the authors prove that $\mathbf{x} \in \Delta^n$ is a Nash equilibrium if and only if there are no infective strategies for $\mathbf{x}$. Therefore, as long as we can find an infective strategy $\mathbf{y}$ for $\mathbf{x}$ we can go on updating the population state and obtain a new population $\mathbf{z}$ such that $\mathbf{y}$ is not among its infective strategies. Reiterating the process until no infective strategy is found allows us to reach an equilibrium. This is formalized by the equation

$$\mathbf{x}(t+1) = \delta_{S(\mathbf{x}(t))}(\mathbf{x}(t))[S(\mathbf{x}(t)) - \mathbf{x}(t)] + \mathbf{x}(t), \tag{2.6}$$

where $S : \Delta^n \to \Delta^n$ is a *strategy selection* function, returning an infective strategy for $\mathbf{x}$ if it exists and $\mathbf{x}$ otherwise. The equation above guarantees that $\mathbf{x}(t+1)$ does not include $S(\mathbf{x}(t))$ among its infecting strategies. With this formulation, $\mathbf{x}$ is a fixed point under dynamics (2.6) if and only if it represents a population state that cannot be infected by other strategies. Additionally, for any choice of selection function, $\pi(\mathbf{x}|\mathbf{x})$ is strictly increasing and always converges [115]. The specific choice of a selection function has an effect on efficiency of the dynamics. In [115] the authors introduce a selection function that seeks for an infective strategy among the set O of pure strategies. This restricts the search from Δ^n to a finite set, thus rendering (per-iteration) complexity linear while still retaining the guarantee of reaching a local optimum.

2.6.3 Matching Game

Central to our game-theoretic framework is the definition of a *matching game*, or, specifically, the definition of the strategies available to the players and of the payoffs related to these strategies. In the following we will refer to the Game-Theoretic Matching method as GTM.

Given a set of "model" points M and a set of "data" points D, we call a *matching strategy* any pair (a_1, a_2) with $a_1 \in M$ and $a_2 \in D$. We call the set of all the matching strategies $S \subseteq M \times D$. The total number of matching strategies in S can, in theory, be as large as the Cartesian product of the sets of features selected in the model and data objects (which may be images in the 2D case or triangulated meshes in the 3D case). This would, of course, lead to a very large sized problem even in the most simple cases, rendering the search for a local optimum unfeasible for the majority of real world scenarios. To this end, we can exploit unary (pointwise) information to reduce the feasible set of strategies by selecting, for instance, only pairs of points bearing similar descriptors or local image patches having good photo-consinstency, depending on the problem at hand.

Let us consider for a moment a 2D matching scenario. In this setting, we can generate for each source point k matching strategies that connect it to the k closest target features in the descriptor (e.g. SIFT) space. Since our game-theoretic approach operates an inlier selection regardless of the descriptor, we do not need to set any threshold with respect to the absolute descriptor distance or the distinctiveness between the first and the second nearest point [90]. In this sense, the only constraint that we need to impose over k is that it should be large enough that we can expect the correct correspondence to be among the candidates for a significant proportion of the source features. In the following chapters we will separately analyze the influence of paramater k over the quality of the final matches; yet, we anticipate that in common scenarios a very small amount of candidates (typically 3 or 4) are enough to guarantee a satisfactory performance, even though in the presence of highly repeating patterns, a larger value might be needed. Limiting the number of correspondences per source feature to a constant value, we limit the growth of number of strategies to be linear with the number of (source) features to be matched.

After defining a feasible set of strategies, we need to construct a proper payoff matrix specifying the reward that each strategy obtains when playing against all the others. Since the set of strategies S is built by proposing several attainable matches for each considered model point it is obvious that the number of outliers in S will be far superior to the number of correct correspondences. In fact, in the majority of cases we will expect the correspondences to be one-to-one, thus at most one match for each set of strategies that involve the same model or data point can be correct; this is a safe assumption for a wide range of matching scenarios, although we emphasize that imposing bijectivity is not a requirement of our method. In order to extract this minority of correct matches “buried” into S, the GTM framework must exploit the consistency of any pair of those strategies with respect to some property. By contrast, all the other wrong and thus randomly paired matches should not exhibit a wide agreement of the same property. This degree of compatibility is usually expressed through a (symmetric) real valued function $\pi : S \times S \to \mathbb{R}^+$ which can be (and usually is) materialized in a symmetric payoff matrix Π. Mapping constraints can be imposed directly on Π, e.g. by setting to 0 pairs of strategies not respecting the one-to-one condition.

This concept is somewhat similar to the basic idea that drives RANSAC; however, by casting the search in an evolutionary process, we are not relying on a one-shot vote counting validation, but rather we are using self-validation to drive to a consensus what can be seen as an iterated voting process. The very nature of the property to which pairs of correct matches must adhere depends on the exact scenario to which GTM is applied. For instance, in [21] adjacency relations of graphs built on segments extracted from the images are used; in Chapter 5 the payoff between pairs of potential matches is boosted by the compatibility of the locally affine transformation assigned to points by the SIFT descriptor. In Chapter 4 the payoff is used to cluster groups of non-distinctive features, while in Chapter 6 it is utilized to match the remaining ones.

Finally, the goal of the selection process is to evolve an initial mixed strategy $\mathbf{x}$ to a stable state from which the non-extinct matches will be extracted. Since $\mathbf{x}$ is associated to strategies in S, it can be seen as a point in $\Delta^{|S|}$ (each coordinate of such a point expresses the proportion of population playing a strategy, and for this reason we refer to $\mathbf{x}$ as a *population vector*). At the beginning of the process, $\mathbf{x}$ is initialized near the barycenter of $\Delta^{|S|}$. In practice this means that to each element of $\mathbf{x}$ is assigned the amount of population $\frac{1}{|S|}$ and then $\mathbf{x}$ is slightly perturbed. The perturbation serves two goals: First, to introduce some randomness in the process (which can be useful if different alternative configurations are sought); second, to avoid some rare stall conditions in evolutionary dynamics that could occur with some highly symmetric payoff configurations.

This initial population is evolved by means of equations 2.2 or 2.6 until convergence. In order to assess when this happens, a threshold on the speed of $\mathbf{x}$ in $\Delta^{|S|}$ or a maximum number of iterations can be set. After a stable state has been reached, all the matching strategies that thrive under the dynamics should be retained. Since no strategy will be completely extinct (as the dynamics do not allow to reach the faces of the simplex) another threshold is needed to select the strategies that succeeded; note that this not happens when using infection and immunization dynamics with pure selection: in this case, the equilibrium is guaranteed to be reached in finite time [115]. Strategy selection is not carried out via an absolute threshold, but rather by fixing a ratio over the maximum value in the population vector. In the following chapters we will show that the value of this threshold has little influence over the quality of the registration obtained. Further, we can weight the contribution of the surviving matches according to the corresponding value in $\mathbf{x}$. This has the consequence of reducing the impact of relatively low-quality points and was shown to allow a slight enhancement in the quality of the final solution [21]. For instance, in a rigid alignment setting, these population coefficients can be employed to compute the rigid motion between corresponding point sets in a weighted fashion (Chapter 6).

Being a rather general framework, the game-theoretic scheme we presented above will be frequently employed in the remainder of this thesis. While the basic formulation is the same for all considered scenarios, the specific definitions making up all its parts will be quite different, and will provide interesting insights emphasizing its flexibility in a variety of cases. In particular, in Chapter 8 we will establish a link with optimization theory, and point out some interesting analogies with more classical formulations of the matching problem.

3

The Reconstruction Process

Three-dimensional scanning is a wide and diverse research field that finds applications in many different areas ranging from measurement tasks, such as industrial monitoring, to artistically motivated reproductions, such as those that can be found in virtual museums. As such, scanning processes come in many variations reflecting both the motivations and the technical contexts in which they originated. Among the existing techniques, structured light scanning is to be regarded, without any doubt, as one of the most accurate measurement tools currently available, but is by no means to be considered the definitive instrument for 3D reconstruction suitable for all scenarios. Nevertheless, due to the degrees of accuracy it can attain, it is certainly a very useful research instrument and has been frequently adopted to produce real-life datasets under controlled conditions. In this Chapter we focus on a typical acquisition process composed of two fundamental steps, for which we provide effective contributions: camera calibration and projected light coding. Here, cameras are modeled according to the traditional pinhole model, and the correspondence problem is tackled via projection of phase-shift coded light. This research was motivated by the need to acquire complete, water-tight, yet very accurate 3D models for the manufacturing industry with low cost hardware requirements.

3.1 Robust Camera Calibration using Inaccurate Targets

Camera calibration consists in estimating the intrinsic parameters of a camera with acceptable accuracy, together with the distortion coefficients that characterize the lens distortion model when needed. To this end, many techniques have been proposed in literature (see Chapter 2 for a review on existing methods). Most, if not all, existing approaches work by taking several shots of some calibration pattern, setting up a 3D-to-2D reprojection system and then solving for the camera parameters, relying on the strong assumption that the calibration pattern be constructed with high accuracy. This might not be the case in a wide range of scenarios, where the calibration pattern is built using consumer technology that certainly cannot guarantee the high precision levels required by demanding computer vision tasks. In this work we take a completely different route from these techniques, by directly addressing one of the most overlooked problems in practical calibration scenarios. Specifically, we drop the assumption that the target is known with enough precision

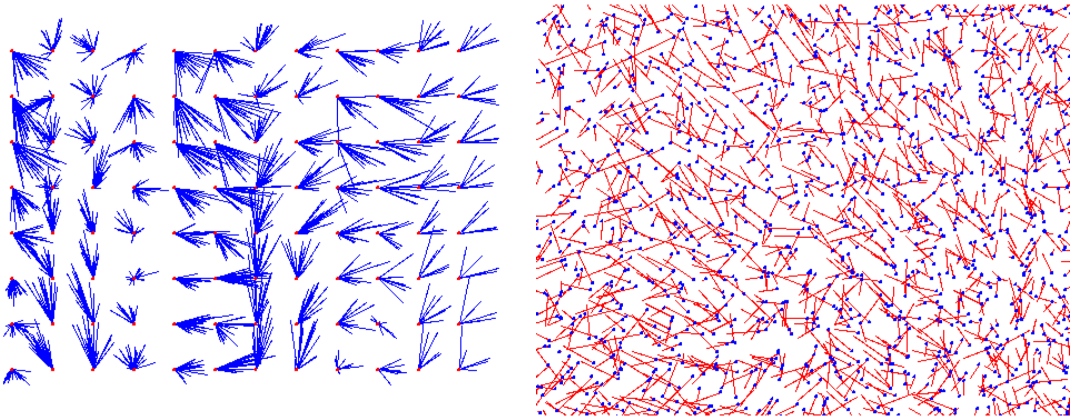


Figure 3.1: The strongly polarized distribution of the reprojection error with respect to target points obtained in our calibration experiments (left figure) suggests that the measurements on images were not subject to zero-mean error. On the other hand, the distribution of the error on the image plane (right figure) is isotropic, assessing the good compensation of radial and tangential distortions and the reliability of the corner detector. This supports our hypothesis that the discrepancies are due to systematic printing errors (errors magnified by a factor of 100).

and we adjust it in an iterative way as part of the whole process. This is in fact the case with the typical target used in most of the calibration literature, which is usually printed on paper and stitched on a flat surface. In the experimental section we show that even with such a cheaply crafted target it is possible to obtain a very accurate camera calibration that outperforms those obtained with well-known standard techniques.

3.1.1 Camera calibration with inaccurate targets

Throughout the literature the calibration quality obtained by different authors, even when using comparable techniques, is somewhat fluctuating. Given the complex nature of the image acquisition process and the uncertainties of the target manufacturing, it is often difficult to identify the exact sources of error and how they combine. The quality measure itself, which is usually the reprojection error, is not always a correct indicator (in section 3.1.2 we will show that it is possible to obtain a low reprojection error and still get a bad calibration). In [53] Douxchamps suggests that a higher accuracy could be attained by increasing the total interface length of the markers themselves and thus enhance their localization. A very in-depth theoretical analysis is validated by a large amount of experiments with synthetic images; however, real-world tests exhibit errors about an order of magnitude larger, which the author explains in part with the non-planarity of the target and in part with optical artifacts not corrected by the calibration model. Similar differences between results obtained with synthetic and real images have been previously observed by Heikkilä in [72]. While we agree that this discrepancy can be attributed to

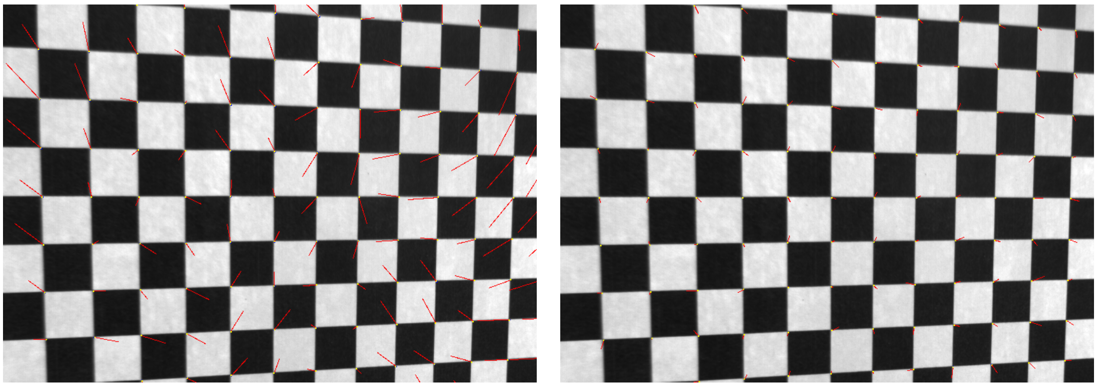


Figure 3.2: Reprojection errors superimposed over the corresponding image point obtained with a direct application of the calibration procedure with the theoretical checkerboard (left figure) and with the more accurate estimated model (right figure).

imperfection of the target manufacturing, we argue that the non-planarity is not the main error source. Specifically, we built a planar target by stitching a checkerboard printed on inextensible plotter paper onto a 6mm thick highly planar float glass. Using this target we intrinsically calibrated a test camera with the Zhang method and, beside studying the reprojection error with respect to image points, we also plotted the same oriented error bars applied to the theoretical model of the checkerboard. The results of this experiment are shown in Figure 3.1. It is immediate to observe that the measurements are not subject to an uniformly distributed error. From a general point of view this means that the optimization was not able to fit the theoretical checkerboard model with a zero-mean Gaussian error, which in turn highlights the presence of some systematic error source. Basically, this source can be correlated with three causes: a localization bias introduced by the subpixel corner detector, the inability of the adopted camera model to fully capture the real image formation model (namely lens distortions), or some unknown discrepancy between the theorical checkerboard model and the printed one. We are not inclined to think that the bias in corner detection is significant since, if this was the case, we should spot a higher coherence in error orientations as all the corner features present roughly the same orientation, scale and illumination conditions (see Figure 3.2). To rule out the deficiency of the camera model we plotted the same error measurements on the image plane space (i.e., the camera CCD sensor). By observing the right part of Figure 3.1 it is quite evident that the error distribution is isotropic and this is a strong indication that no systematic error is amendable by using a more sophisticated camera model. Those considerations suggest that the printing process itself could be inaccurate enough to represent a significant source of systematic error in the calibration process. For this reason, we decided to investigate the nature of such printing errors and to tailor a calibration procedure able to correct them. It should be noted that the influence of printing error was already observed in literature. Recently Strobl [132] described a similar scenario, albeit he suggested that the printing procedure introduces just two types of systematic biases: an error in the global scale of

Algorithm 3.1.1: Calibrate Camera-Target($target, images, minError, maxIter$)

```
$originalTarget \leftarrow target$
$reprojectionError \leftarrow +\infty$
$currentIteration \leftarrow 0$
while $reprojectionError > minError$ and $currentIteration < maxIter$
  do
    $cameraParameters \leftarrow calibrateCamera(target, images)$
    $reprojectionError \leftarrow bundleAdjust(target, cameraParameters, images)$
    $target \leftarrow hornAdjust(target, originalTarget)$
    $currentIteration++$
return $(cameraParameters)$
```

Figure 3.3: The simple, yet effective, algorithm proposed to iteratively estimate the intrinsic camera calibration and the real target geometry.

the checkerboard and a non predictable aspect ratio. While this simple formulation allows for an elegant calibration procedure that introduces just two additional parameters, we think that it is not general enough. In fact, the distribution of the error shown in Figure 3.1 can not be justified by scale and aspect ratio transformations. In order to deal with the most general scenario we propose to lift any direct constraint on the target geometry and iteratively run a three-step procedure (see Figure 3.3). In the first step we assume to know precisely the calibration target and we perform a standard calibration procedure with the currently trustable target model. Once plausible camera parameters are obtained, we assume them to be correct and we evaluate a more accurate target geometry by using a bundle adjustment technique to estimate only camera poses and scene (i.e., target). At this point a further step is needed, specifically we need to rescale the newly created target to fit the original one. This is necessary since the bundle adjustment step does not guarantee scale invariance. This adjustment step is performed by using the robust closed form point set alignment technique by Horn [74]. The procedure is stopped when the reprojection error falls below a fixed threshold or a maximum number of iterations is reached. Note that since the estimated target is scaled towards the theoretical one at each step, the final camera calibration could be subject to an absolute scaling error that is not avoidable as the real measures of the target are not known. Still, this error is averaged over the printing error of each corner and in practice this has shown to be very small (see Section 3.1.2).

3.1.2 Experimental results

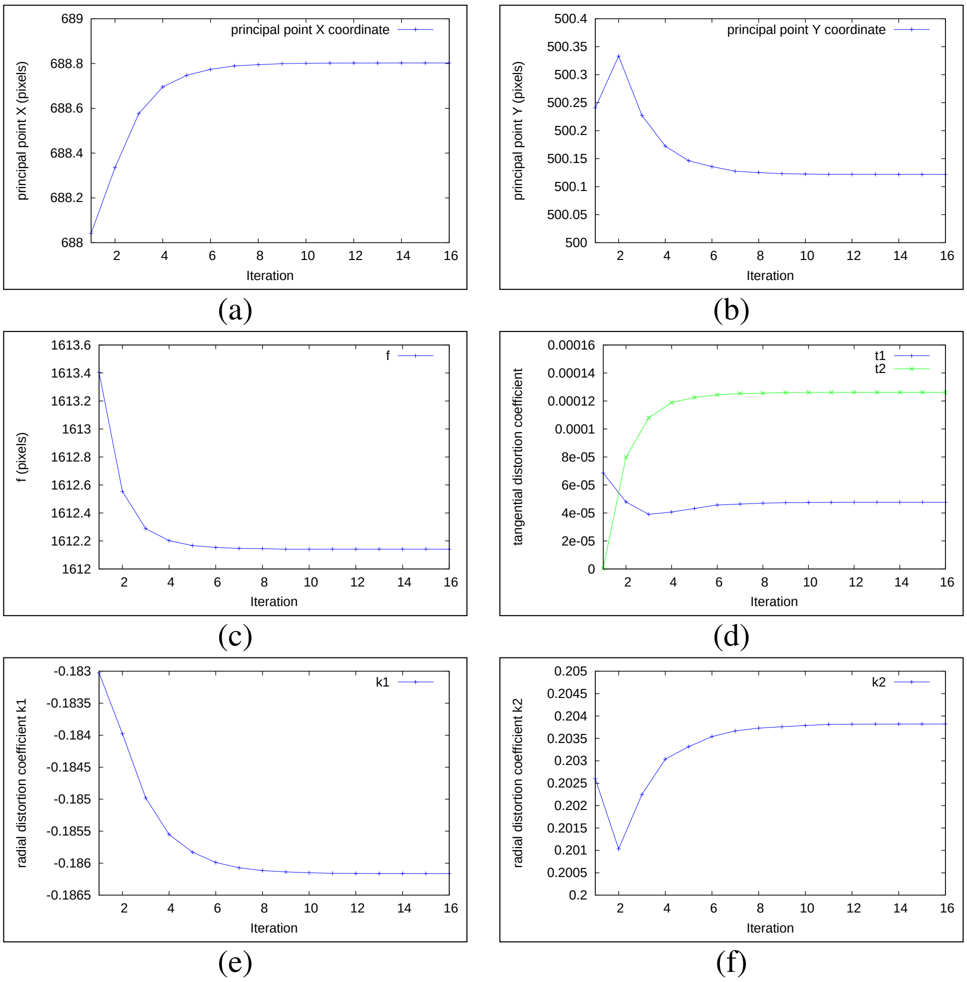
We performed all of our experiments using a pair of Basler Scout scA1300-32gc Gigabit Ethernet monochrome cameras with a resolution of 1280 x 960 pixels. We printed our checkerboard pattern with one laser and two different inkjet printers. Some rudimentary dimensional controls, made with a millimeter accurate ruler, showed that the laser printer is surprisingly inaccurate as it produces prints that suffer from a global error (measured between the first and last check of the A4 page) from one to three millimeters. This error was also very variable between subsequent prints and also the aspect ratio was not repeatable (as previously observed by Strobl [132]). The inkjet printers were more accurate, showing printing errors greatly below the measurement threshold of the ruler (i.e. submillimeter) both with respect to the global checkerboard size and its aspect ratio. Thus, we decided to build our test target by printing a large checkerboard on an inextensible A3 plotter sheet with the inkjet printer we felt to be more precise, and to stitch it on a thick planar glass with all the needed care to guarantee the best possible adhesion. The camera calibration step was performed by using the OpenCV [32] library, which implements the Zhang [162] method and allows to estimate, by non-linear optimization, three radial and two tangential distortion parameters. In our experiments we estimated only two radial distortion parameters since the used lens set was not wide-angle and a preliminary set of experiments showed that the third parameter was both irrelevant and unstable. Further, we fixed the camera aspect ratio to 1 (i.e. fx=fy), as guaranteed by the manufacturer. To perform the bundle adjustment a suitable sparse implementation of the Levenberg-Marquardt optimization algorithm was used. For each camera we took a set of 200 shots of the target: 100 shots were taken with small angles between the target and the optical axis (from 0 to 15 degrees), while the others exhibit a larger range of orientations (up to 30 degrees). These two sets have been used to study the influence of the target orientation in both the quality of the calibration and the reprojection error measure (see Section 3.1.2). Finally, the two cameras were mounted on a rigid support to form a stereo pair which was used to shot another set of 100 images of the same checkerboard, and thus perform a stereo calibration and target reconstruction as an additional validation step.

Effects of target estimation on camera calibration

In our first set of experiments we analyzed the convergence behaviour of the proposed algorithm. This was done by applying the calibration technique to the full set of shots taken by the first camera. By observing the plots presented in Figure 3.4, we can see that the position of the principal point (plots a and b) changes significantly during the first 8 iterations. In particular, its horizontal location moves by almost 1 pixel. In general, with a good quality calibration setup, this value is expected to be recovered with at most some fraction of pixel of estimation error.

We observe a similar behaviour for the focal length (plot c), which moves away from the initial approximation by more than 1 pixel. In the last three graphs of Figure 3.4 we see the distortion parameters variation. Again, the distances from the first rough estimate are similar in percentage to those observed for the internal parameters.

To give an idea of how these variations relate with the reprojection error, in Figure 3.5 we reported the RMS between observed and reprojected points for the same iterations shown in part (a) of Figure 3.4. It is interesting to note that the reprojection error dramatically decreases after just one iteration, then it becomes quite stable and after as few as 4



(a) (b)

(c) (d)

(e) (f)

Figure 3.4: Convergence of both internal and distortion parameters with respect to subsequent iterations of our method.

iterations it seems to have reached convergence. Indeed, this does not correspond to the behaviour observed for the camera parameters, which continues to move after 4 iterations. We think that this is due both to the misleading nature of the reprojection error (which we discuss in Section 3.1.2) and to the fact that most of such initial large error is to be attributed to the very rough first approximation of the checkerboard model.

To better explain the interplay between the estimation of the “true” calibration target and the associated error reduction, we plotted (in part (b) of Figure 3.4) the RMS error between the corresponding points of the theoretical and estimated target. We can see that the target estimation process requires a few more steps to stabilize, albeit the most significant variation happens during the first few bundle adjustments. The final distance from the theoretical model settles around one-tenth of a millimeter, which is compatible with the qualitative observation obtained by using a physical ruler. While this error magnitude can be considered small for a normal printing operation, it can be cumbersome when

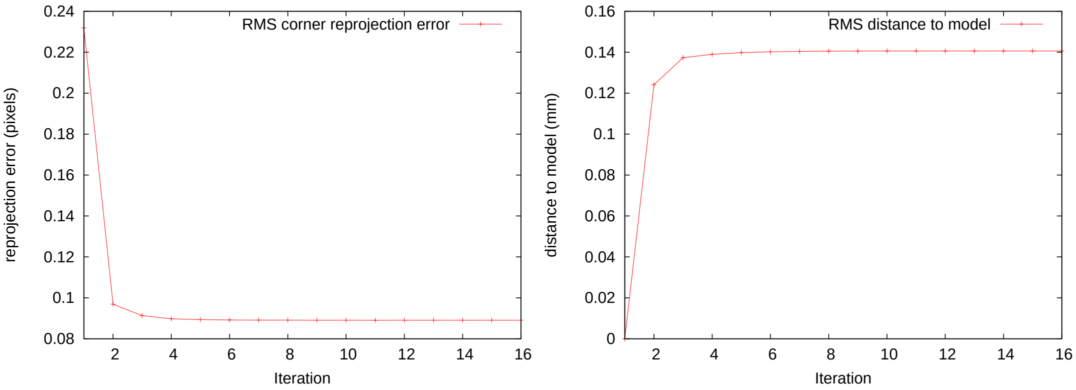


Figure 3.5: Convergence of reprojection error (left) and distance of the estimated calibration object with respect to the theoretical checkerboard (right) for subsequent iterations of the proposed method.

seeking for an accurate calibration, especially considering that high-quality specifically crafted industrial calibration targets guarantee a location error below one-hundredth or one-thousandth of a millimeter.

In Figure 3.6 we finally show the distribution of the reprojection error with respect to the target points (first image), and the image plane after the application of the proposed calibration method (second image). This figure is meant to be directly compared with Figure 3.1. Regarding the measured error applied to the target points, the large reduction in magnitude is very noticeable. In addition, the error sticks appear to exhibit an isotropic distribution. This suggests that most of the systematic error has been eliminated and only Gaussian noise (due to image noise and corner detection) is measured.

Validation of the calibration by stereo triangulation

The reprojection error obtained in Section 3.1.2 is below one-tenth of pixel. While this is not a bad result overall, it is a bit large with respect to common values declared in literature by recent methods, which usually reach a reprojection error as low as 0.05 pixels even using self-printed targets. To better understand the relation between the reprojection error and the calibration accuracy we performed and independent check based on stereo reconstruction. We calibrated a pair of cameras with the described technique and then we fixed them to a rigid bar and calibrated the stereo rig using a new set of stereo shots and the target geometry previously learnt. Since the correct geometry was assumed to be known, this latter step was performed simply by using the appropriate stereo calibration function of the OpenCV library. The stereo calibration was then used to recover, in the 3D space and for each stereo shot, the position of two checkerboard corners that are 10 cm apart. Our idea is that by analyzing the measured distance between the reference points under different conditions, we should be able to assess the overall quality of the two camera calibrations and of the target estimation (as the “learnt” target is used to perform the stereo

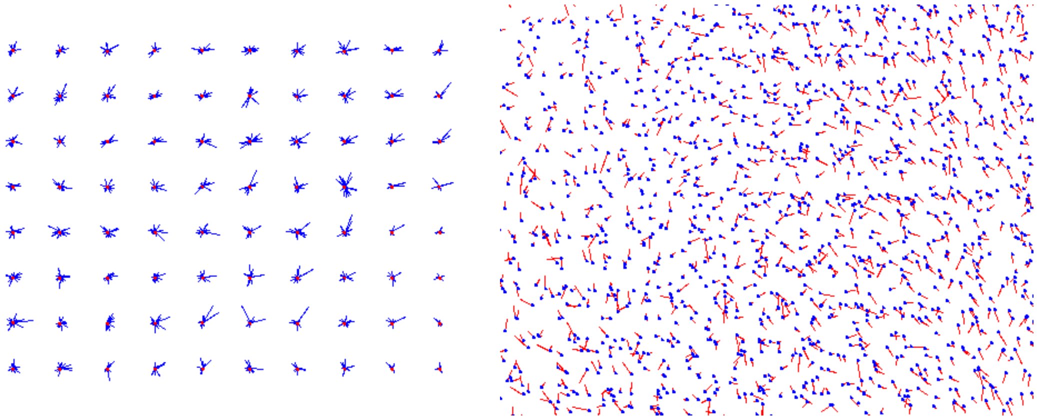
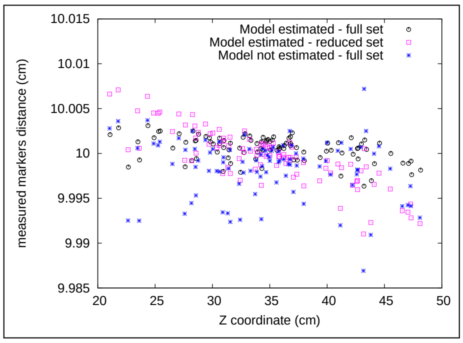


Figure 3.6: Reprojection error (magnified by a factor 100) with respect to the model points (first graph) and the sensor plane (second graph) after our method comes to convergence.

step). Specifically, we tested three calibration setups: the first is meant to be a comparison with standard calibration techniques and thus uses only the theoretical checkerboard, the second uses the estimated target and all the images, finally the last setup uses only a reduced set of images that feature angles with the optical axis lower than 15 degrees. In Figure 3.7 we present a scatter plot of the different measures obtained with the three setups described. Since the target is moved among the stereo shots, we reconstructed the pair of control points in different positions in space. Of course, the better the calibration is the more consistent is the measure between different shots, since if the calibration were perfect, the only measure error would be due to the subpixel corner localization. We can observe that by using our method with all the shots (thus with angles up to 30 degrees) we obtain the most reliable results: the measure has the lowest variance and it is quite uniform throughout the depth of the view field. By contrast, using the reduced set of images (those that are more fronto-parallel) not only leads to a larger variance, but in addition a clear gradient appears in the scatter plot; that is, the measure tends to decrease as the target gets far from the stereo pair. This effect is probably due to the fact that images with weak angulation are not able to fully constrain all the calibration parameters, especially the extrinsic ones (i.e. camera poses). As expected, when not estimating the target model we obtain a larger variance, albeit we do not get any gradient along the z coordinate as we use the full set of images. Note that the average in not exactly 10 cm even for the best technique: this is to be expected, as the real distance is not known and we just know that the two reference points are separated by 10 squares allegedly long 1 cm in the print. In table (b) of Figure 3.7 we show the different reprojection errors attained by the three setups. We can see that, notwithstanding the large average measurement error, the setup that uses the fronto-parallel images obtains a very low reprojection error. This apparent nonsense is indeed due to the lower error introduced by the corner detector as, when the model is correctly estimated, this zero-mean Gaussian error does not hinder a correct camera calibration, yet it contributes to the magnitude of the reprojection RMS error. On



(a)

	Reproject Error	Distance to Model	Measure Error
Model full set	0.08907	0.14062	0.00145
Model small set	0.05814	0.11645	0.00317
No model full set	0.23191	0	0.00362

(b)

Figure 3.7: Verification of the calibration quality by means of a stereo measuring check (see text for details).

the other hand, the use of weakly angulated images poses less constraints to parameters and thus leads to both a less accurate mono calibration and stereo reconstruction. In this sense, the reprojection error is not always a good indicator since, when the calibration setup is good, it tends to be dominated by localization errors.

3.1.3 Conclusions

In this Section we introduced a very simple but effective method to precisely calibrate a camera with a cheap self-built target. While this approach is not meant to substitute the use of professional-made calibration objects, it is very useful since printed targets are still widely used both by researchers and practitioners. The effects of an accurate learning of the real target geometry are shown with a set of experiments that explore both the convergence behavior of the algorithm and the quality of the obtained calibration.

In a typical stereo reconstruction setting, once a calibration is provided for the two cameras, and given that point-to-point matches are established between the two images, it is possible to triangulate object points in the 3D space; doing this for all the correspondences, one obtains the surface reconstruction of the captured object from that single point of view. In the next Section we will assume that a more or less accurate camera calibration is provided, and shift our focus to the correspondence problem itself. Specifically, we introduce a novel technique for phase coded light that aims at improving efficiency of the acquisition process by retaining a good level of accuracy.

3.2 Fast Reconstruction by Unambiguous Compound Phase Coding

The most common approach to structured light scanning is undoubtedly represented by binary Gray coding (Section 2.2). The method is of easy implementation and accurate enough for many applications, but is severely limited in resolution and is highly dependent on the albedo of the underlying surface. By contrast, phase shift methods have proven to be very robust and accurate for photometric 3D reconstruction. One problem of these approaches is the existence of ambiguities arising from the periodicity of the fringe patterns. While several techniques for disambiguation exist, all of them require the projection of a significant number of additional patterns. For instance, a global Gray coding sequence or several supplemental sinusoidal patterns of different periods are commonly used to complement the basic phase shift technique.

In this work we propose a new pattern strategy to reduce the total number of patterns projected by encoding multiple phases into a single sequence. This is obtained by mixing multiple equal-amplitude sinusoidal signals, which can be efficiently computed using inverse Fourier transformation. The initial phase for each fringe is then recovered independently through Fourier analysis and the unique projected coordinate is computed from the phase vectors using the disambiguation approach based on multiple periods fringes proposed by Lilienblum and Michaelis [87]. With respect to competing approaches, our method is simpler and requires fewer structured light patterns, thus reducing the measurement time, while retaining high level of accuracy.

3.2.1 Multi-Period Phase Coding

Recently, methods based on number theory have been proposed for disambiguation [165, 87]. In [87] the authors relate absolute, unambiguous phase values to projector coordinates $\xi \in \mathbb{R}$, and define $\xi(u, v)$ to be the projector coordinate at pixel (u, v). Then several phase shift sequences, each with a different local period λ_i, are projected onto the object to be measured. A phase image is obtained for each sequence through the computation of periodic phase values $\phi_i(\xi) \in [0, 1)$ at every pixel. In addition, the fringes of a pattern are assigned sequential natural numbers $\eta_i(\xi) \in \mathbb{N}$, which represent a simple counting of the fringes from left to right. A projector coordinate can then be directly obtained, for all $i = 1, 2, ..., n$, from a fringe number and a phase value:

$$\xi = (\eta_i(\xi) + \phi_i(\xi))\lambda_i \,. \tag{3.1}$$

Since the only available values during measurement are λ and ϕ, it is clear that the system of equations becomes ambiguous as the same value of ξ can be obtained for different values of η_i. This happens when two different projector coordinates yield the same phase values for all i. Therefore, the authors follow a number-theoretic approach and identify a general condition for generating unambiguous pattern sequences, by defining a maximum projector coordinate ξ_{max} up to which ambiguity can be excluded. Such a coordinate is defined as the least common multiple of relatively prime periods λ_i, and clearly for practical advantage it must entirely cover the projector range. An efficient method is given to calculate the fringe numbers from the ambiguous phase values at each pixel, given the local period lengths. This method takes advantage of a simple relationship between phase values and fringe numbers. Given any pair of pattern sequences, the following equivalence holds for each image pixel:

$$\lambda_i \phi_i(u,v) - \lambda_j \phi_j(u,v) = \lambda_j \eta_j(u,v) - \lambda_i \eta_i(u,v)\,. \tag{3.2}$$

This makes it possible to construct a theoretical phase difference vector beforehand, and then use it to retrieve the fringe numbers when real phase measurements become available. In addition to providing an efficient way to obtain the fringe numbers, this method allows to assign each point a reliability value related to the deviation between measured and expected values. The use of theoretical phase difference vectors makes for a powerful test, which allows to identify erroneous or weak measurements (such as mixed phase values) caused, for instance, by sharp edges, involuntary object movements and light reflections. Once the unknown fringe numbers are calculated, projector coordinates can be easily retrieved for each pattern sequence with equation 3.1. The independent measurements can then be averaged to obtain a unique and absolute phase value at every pixel in an efficient way, leading to an increase in accuracy of the measurements.

The big advantages of the multi-period method is its relative simplicity and high efficiency. The phase-coded images can be directly employed in general stereo reconstruction systems, ensuring high quality and density of the code. Specifically, the lack of surface points is mainly due to occlusions and camera disparity, and measurement errors are very low thanks to the averaging and validation procedures implicit to the approach, that exclude a large percentage of errors and outliers before the actual surface reconstruction takes place. The main drawback lies in the fact that, typically, three or more pattern sequences are needed to entirely cover the projector range (typical values are 800 or 1024 projector pixels). This requires the projection of as many as three times more patterns than required with classical phase shifting.

In the next section we introduce a novel coding strategy that retains the big advantages offered by the multi-period method, but requires a significantly lower number of structured light patterns while achieving comparable levels of accuracy.

3.2.2 Compound Phase Coding

The main idea behind the Compound Phase Coding strategy is to project several fringe patterns in a single spatio-temporal pattern. This is obtained by encoding the phases of the fringe vector as phases of a Fourier term at different frequencies. Each fringe is characterized by a different period and all of them are relatively prime. This way, once recovered the single initial phase shift for each fringe, we are able to build an unambiguous code with a numerical technique similar to the one suggested in [87].

Let $\lambda_1, \dots, \lambda_k$ be k periods and let ξ be the projector coordinate at some pixel. If the periods are coprime and the projector coordinates do not exceed $\prod_{j=1}^{k} \lambda_j$, then we

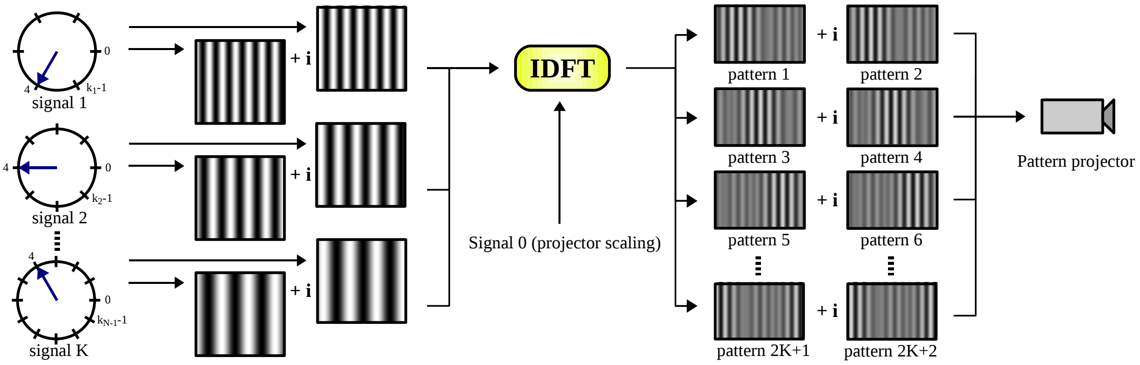


Figure 3.8: The composition of k fringe patterns, plus one unknown shift signal will produce a total of $2(k+1)$ image patterns that will be projected onto the surface to be reconstructed. The shift pattern (projector scaling) accounts for the unknown value of the albedo of the surface.

have a unique phase code for ξ [87]. Namely, this code is given by the vector $\boldsymbol{\phi} = (\phi_1 \quad \dots \quad \phi_k) \in [0,1)^k$, where $\phi_j = (\xi \mod \lambda_j)/\lambda_j$.

Our aim is to map $\boldsymbol{\phi}$ into a signal sequence of gray-scale values that can then be measured by the cameras to obtain the unique code of each fringe. We take the hint from phase shift methods that phase encodings are more robust than amplitude codings, but extend multi-period coding by decoupling the phase used to encode the message from the frequency of the sinusoidal signal used to transport it. This is done by projecting the sum of equal-amplitude sinusoidal signals at frequencies $\frac{1}{k+1}, \frac{2}{k+1}, \dots, \frac{k}{k+1}$, where k is the number of periods necessary to encode the coordinate x_i, and by encoding the phase parameters as the phases of the corresponding sinusoidal signal.

Given a phase code $\boldsymbol{\phi} \in [0,1)^k$, we create a $(k+1)$-dimensional complex vector $\mathbf{x} \in \mathbb{C}^{k+1}$, where

$$x_j = \begin{cases} 0\,, & \text{if } j = 0\,, \\ e^{-2\pi i \phi_j}\,, & \text{if } 1 \le j \le k\,. \end{cases}$$

Here, $i = \sqrt{-1}$. Note that given x_j for any $1 \le j \le k$, we can compute the phase ϕ_j as

$$\phi_j = \operatorname{frac}(1 + \frac{1}{2\pi} \arg(\Im(x_j), \Re(x_j)))\,, \tag{3.3}$$

for any value of s, where $\operatorname{frac}(\cdot)$ is the fractional part of the argument, and $\Im(z)$, $\Re(z)$ are the imaginary and real parts of $z \in \mathbb{C}$, respectively.

Each complex number x_j represents the amplitude and phase of a sinusoidal component with frequency $\frac{j}{k+1}$ cycles per sample. Hence we can reconstruct the intensity sequence of that coordinate by computing the Inverse Discrete Fourier Transform of $\mathbf{x}$, obtaining the vector $\mathbf{y} \in \mathbb{C}^{k+1}$, where

$$y_n = \frac{1}{k+1} \sum_{j=0}^{k} x_j e^{2\pi i \frac{j}{k+1} n}\,, \quad n = 0, \dots, k\,.$$

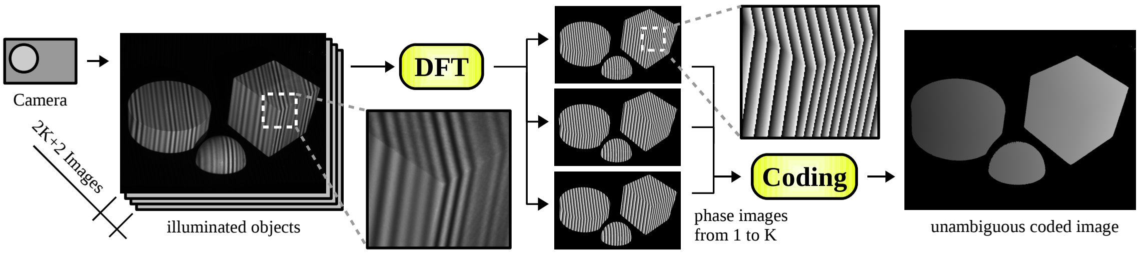


Figure 3.9: A total of 2K+2 images of illuminated objects are captured and single phase values are calculated for each composed fringe signal. Those values are subsequently used to get an unambiguous coding. Note that the intensity profile of each projected pattern is not sinusoidal.

We can then project separately the real and imaginary part of this vector as two time sequences obtaining a single set of $2(k+1)$ patterns to be projected to uniquely encode the x_i projector coordinate (see Figure 3.8).

Hence, given $\mathbf{y} \in \mathbb{C}^{k+1}$, we transform it into a real vector $\mathbf{z} \in \mathbb{R}^{2(k+1)}$, where $z_{2j-1} = \Re(y_j)$ and $z_{2j} = \Im(y_j)$ for $1 \leq j \leq k+1$. Afterwards we scale and shift each component of $\mathbf{z}$ in order to bound them within $[0, 255]$, obtaining the sequence of gray-scale values to project in correspondence to ξ. Note that by applying a shift we affect the information at frequency 0, while by scaling, we modify the amplitudes of all frequencies, without influencing the phase values.

The acquisition process introduces an additional linear deformation on $\mathbf{z}$, which depends on the physical properties of the object being scanned. Again this does not affect the phases.

Let $\overline{\mathbf{z}} \in \mathbb{R}^{2(k+1)}$ be the acquired gray-scale values and let $\overline{\mathbf{y}} \in \mathbb{C}^{k+1}$ be its representation into a complex vector. Then, the net effect of the projector scaling and the change in the reflectivity properties of the surface is a translation and scale of the observations, i.e. the relation beween the intended signal $\mathbf{y}$ and the observed signal $\overline{\mathbf{y}}$ is

$$\overline{\mathbf{y}} = \overline{\delta} + \overline{s}\mathbf{y}$$

for some real values $\overline{\delta}$ and $\overline{s}$. The phase code is finally recovered from $\overline{\mathbf{y}}$ by computing the Discrete Fourier Transform, namely

$$\overline{x}_j = \sum_{n=0}^{k+1} \overline{y}_n e^{-2\pi i \frac{n}{k+1} j} ,$$

and by applying (3.3) to the result $\overline{\mathbf{x}} \in \mathbb{C}^{k+1}$, obtaining $\boldsymbol{\phi}$ (see Figure 3.9).

This process allows to recover the phase code for each projector coordinate by taking only $2(k+1)$ measurements, where k is the number of signal periods. Nevertheless, one can also force a larger number of samples in order to increase accuracy, by appending null components to $\mathbf{x}$. More precisely, by appending M null components, we need $2(M+k+1)$ measurements in order to recall the phase code $\boldsymbol{\phi}$.

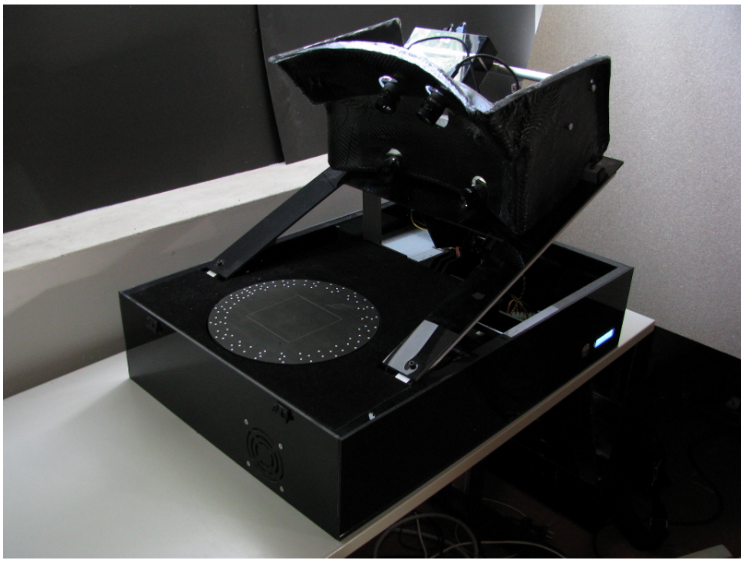


Figure 3.10: The general purpose structured light scanner used.

It should be noted that a drawback of this approach is that encoding multiple signals in a single pattern reduces the effective projector intensity range available to encode each phase, increasing the effects of the discretization error. However, experiments show that the error introduced by this is limited.

3.2.3 Experimental results

In order to validate the proposed technique we compare its accuracy to the results obtained with a state of the art phase shift technique. To this extent we choose to compare our measurements with those given by the Multi-Period Phase Shift proposed in [87]. The reasons for this choice are two-fold: Multi-Period Phase Shift is very accurate, as it uses information for each fringe projected in order to reduce the average error; in addition, once the phase vector from the composite signal is obtained, measurement quality is directly comparable as both techniques share the same numerical disambiguation step. In the following sections we will show both quantitative results, by evaluating the measurement error over a planar target, and qualitative results, by showing relative average distances between the measurements obtained by the two techniques with generic objects and by comparing the corresponding estimated quality.

Experimental setup

All the following experiments have been run on a test rig for structured light techniques that has been internally developed in our lab (figure 3.10). The rig is made up of a motorized plate for object positioning, four cameras and an illumination source mounted on a motorized liftable platform. Specifically the cameras are equipped with a 1/2 inch

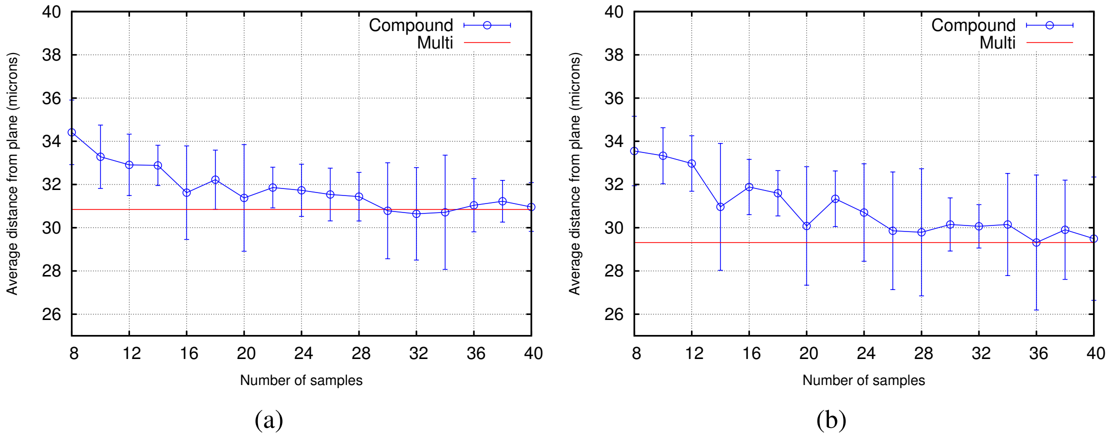


Figure 3.11: Accuracy comparison between the Compound Phase Coding method and the Multi-Period Phase Shift technique. In figure 3.11(a) we used periods of 7, 11 and 13 pixels (30 patterns for Multi-Period Phase Shift), and in figure 3.11(b) we used periods of length 9, 11 and 13 (34 patterns for Multi-Period Phase Shift). Note that the Multi-Period technique appears as a flat continuous red line and its standard deviation as dashed red lines. Vertical bars are standard deviations in the measurement of the error for the compound technique.

CMOS sensor which offers a full 1280x1024 resolution. The cameras are monochrome, thus no Bayer filters are placed over the sensor. While four cameras are available, in this experiment set we use only one pair of cameras to reconstruct the surfaces. Thus the system falls into the category of two calibrated cameras and one uncalibrated light source. The illumination source is a 800x600 color DLP projector which we use to project the monochromatic patterns. The system is controlled by a standard PC housed into the base of the rig. This PC is a 2.8 GHz AMD quad core system with 2 Gigabytes of ram.

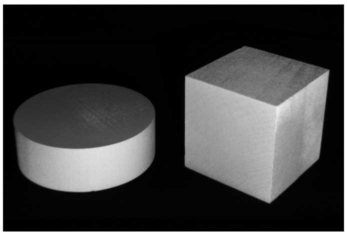
Intrinsic and extrinsic parameters of the cameras have been obtained through a standard calibration procedure using a planar checkerboard and OpenCV [32] calibration software. It must be noted that our system is not a full fledged production scanner, thus it does not guarantees extreme accuracy or resolution: in fact we estimate its precision in about $30\mu m$. This level of accuracy is adequate with respect to our experiments, as we are not interested in showing the absolute precision of our setup, but rather the relative performance of the coding schemes, showing that the proposed approach can be used as an effective replacement for slower approaches without suffering from a significant loss in accuracy.

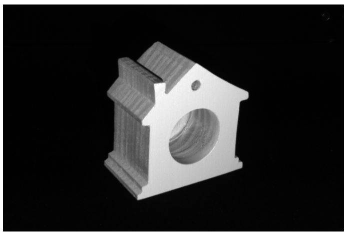
Planar target measurements

In this first set of experiments we measured the surface of a 200 by 200 mm squared piece of float glass which we previously sprayed with a very thin layer of acrylic paint. We made several sets of measurements with both the Compound Phase Coding technique and the Multi-Period Phase Shift technique. Since the exact pose of the test object is unknown and the surface cannot be perfectly flat, we approximated the ground truth with the best fitting plane (in the least squares sense) with respect to the measured points of the object. This way we estimate the expected measurement error of each technique as the average of the absolute value of the distance from the fitting plane of each measured point. We had a wide range of choices regarding the number of different signals to project and their periods. We chose to execute the test with two configurations: 3 signals of periods respectively 7, 11 and 13 pixels and other 3 signals of periods 9, 11 and 13 pixels. Since our projector has an horizontal resolution of 800 pixels both configurations allow to obtain a globally unambiguous coding of the object. Given those signal configurations, we projected respectively 30 and 34 patterns for testing the Multi-Period Phase Shift technique. Since we always used 3 signals, the Compound Phase Shift technique strictly requires only 8 patterns to be projected: nevertheless we repeated the measurement with a growing number of additional patterns in order to study the effect of the supplementary information on the final accuracy. Finally, each experimental measure was repeated for a total of 10 times. In figure 3.11 we compare the performance of the proposed approach against the baseline multi-period approach. Accuracy of the multi-phase approach is slightly better in 3.11(b), probably due to the higher number of patterns projected (34 instead of 30). The blue line shows the trend of the average error of the proposed Compound Phase Coding technique as the number of projected pattern is increased. It should be noted that even with the minimal number of projected patterns the accuracy is quite good: in fact, on average our distances from the fitting plane are only about three micrometers higher than those obtained using the Multi-Period technique, which requires almost four times as many patterns. Moreover, by projecting additional patterns, the quality of the measurements can be further enhanced: for instance, by doubling the number of projected patterns the distance from the Multi-Period approach is approximately halved. Finally, as we expected, using the same number of patterns projected by the Multi-Period technique, the two approaches yield equivalent performances. Of course, the whole point of the proposed technique is to reduce the number of projected patterns. In this sense, our approach allows to control the time/precision trade-off, obtaining good reconstructions even with just 8 patterns.

Generic objects measurements

In this set of experiments we measured the surface of several generic scenes built with simple wooden objects in a volume of about 200 mm of diameter in each direction. The set of objects was selected to maximize the range of surface orientations and conditions. Again we compared our results with those obtained by the Multi-Period approach: in this case we take the Multi-Period results as a direct comparison, as no ground-truth or knowledge of the surface of the objects (which are hand-made) is available. Compound Phase Coding was tested both with 8 and 16 projected patterns. Since we projected signals with periods of respectively 9, 11 and 13 pixels Multi-Period Phase Shift was tested with 34 samples. For each experiment we evaluated three quantities, namely:





	Cube and Disc			Toy House		
Technique	Multi	Comp 8	Comp 16	Multi	Comp 8	Comp 16
Points	109347	108306	109034	46749	46263	46505
Avg deviation	0.025	0.089	0.063	0.030	0.085	0.067
Avg distance	-	32.00	25.09	-	21.15	16.19

Figure 3.12: Comparison between reconstructions obtained with the Multi-Period Phase Shift and the Compound Phase Coding on several test objects. Compound Phase Coding has been tested using both 8 and 16 samples. Multi-Period Phase Shift has been tested with 34 samples in order to obtain the best quality. Distances are in microns and objects are 5 to 10 cm wide.

- The number of points acquired. That is the number of surface points that pass both the consistency and the quality check. The first verifies that the phase vector is consistent with the code assigned to the point, the latter ensures that the phase difference vector contains only integer values (as expected with respect to constraint 3.2). Specifically, for these experiments we rejected points associated to phase difference vectors where at least one entry deviates from the nearest integer by more than 0.2;

- The average deviation. That is the average distance from each entry in the phase difference vector and the nearest integer. Since each element of the phase difference vector should be integer a lower value suggests a higher quality in the measurement obtained. This parameter has been calculated over all the points obtained, even those filtered by the consistency and quality checks;

- The average distance. That is the average of the absolute distances between points obtained by the Multi-Period and Compound techniques. Note that this measure is possible because we implemented both algorithms in a way that allowed us to produce depth maps that are exactly overlapped along the x and y axis and differ only for the depth values assigned along the z axis.

In Figure 3.12 we report the values of those three quantities in different scenes. In general, the number of points acquired by the Compound Phase Coding with only 8 patterns is slightly smaller than the number of valid points given by the Multi-Period technique. This distance, while small, is almost completely eliminated when using 16 patterns. The

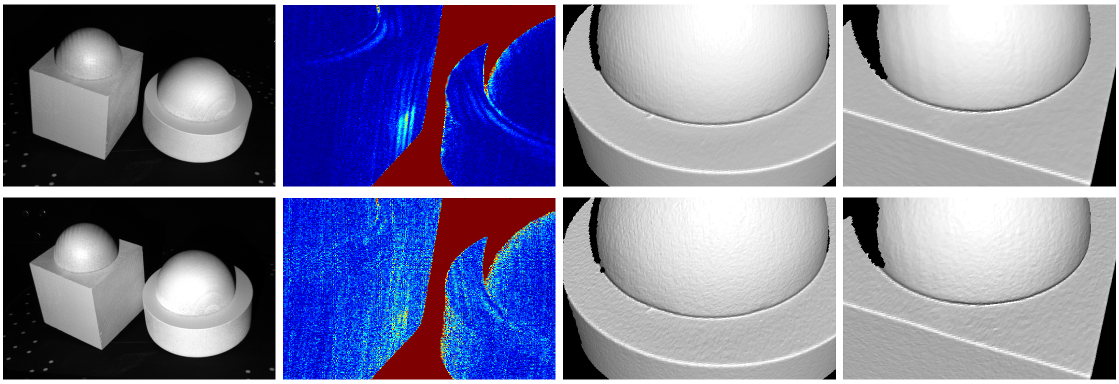


Figure 3.13: A surface reconstruction example complete with deviation assessment. In the first colum, the object as viewed by the left and right camera. In the second column, the deviance maps of Multi-Period Phase Shift (top) and Compound Phase Coding (bottom). In the last colums, two respective close-ups of the reconstructed surfaces.

average deviation resulting by applying our technique is always slightly higher, nevertheless it should be remarked that those values are all quite small: in fact, for any practical purpose a deviation smaller than 0.1 can always be associated with a valid measured point. Finally, the average distances between the two techniques also assess the suitability of our technique as a fast replacement for traditional Multi-Period: in fact, the results obtained are generally compatible with the measurement error estimated in the experiments done in section 3.2.3, a result that should be expected since both techniques are subject to measurement errors. It should also be noted that the higher average distance obtained with the scene “Cube and Disc” is probably due to the presence of sharp edges and of two large surfaces at a grazing angle with both the cameras and the light source. Conversely, the higher smoothness offered by the scene “Cone and Spheres” allows for a more precise reconstruction.

In Figure 3.13 we show a more qualitative example of the difference in accuracy of reconstruction between our method and a complete Multi-Period phase shift. In this experiment we used the less accurate instance of Compound Phase Coding, thus only 8 patterns were projected. In the first column we reproduced the scene viewed from the left and right camera. In the second column we calculated a map of the maximum deviation from integer values of the phase difference vectors of the Multi-period method (top) and the Compound method (bottom). In these maps we used the standard Matlab color scale, thus a full blue pixel indicates a low deviation and a bright red pixel a maximum deviation (in this case 0.5). Uncoded pixels are represented in dark red. Both techniques exhibit a fairly low deviation, with high values mainly on the borders, which is due to the camera integration of the projected signal and the background on boundary pixels. The other high deviation area on the side of the cube is probably due to inter-reflection between the two objects.

The last two columns show a close-up of the surfaces reconstructed with both methods. In general the surface reconstructed with the Compound technique is a bit more

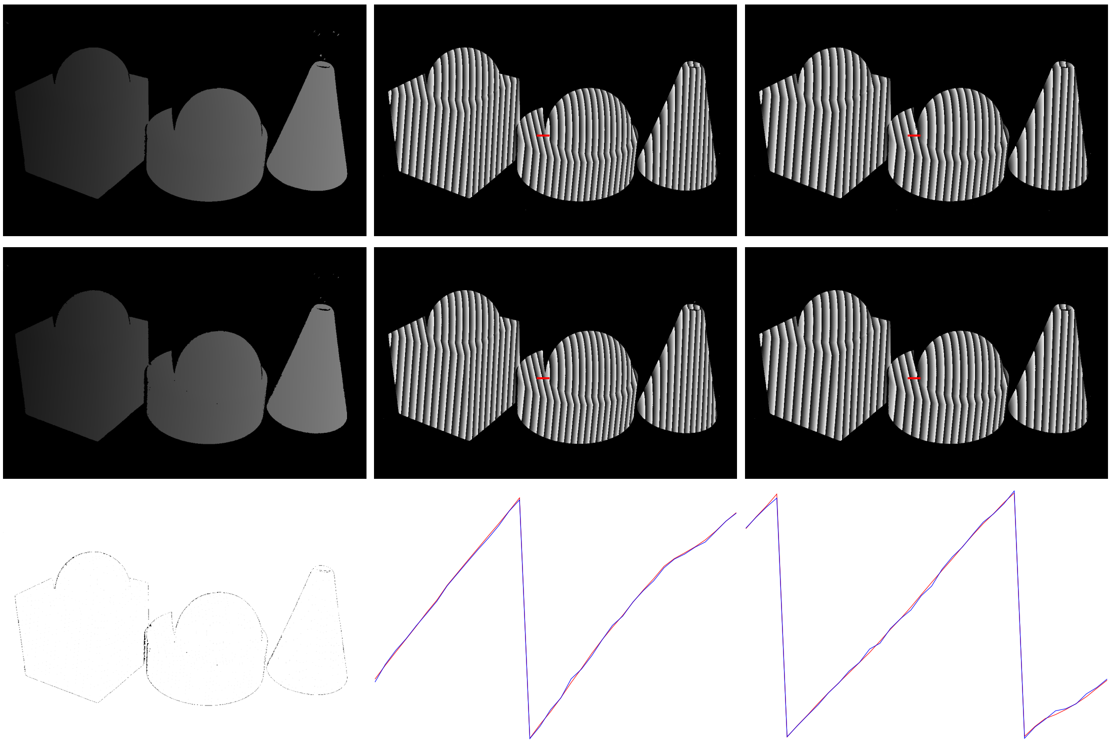


Figure 3.14: A comparison of the coding and of the phase values extracted by our technique and by Multi-Period phase shift. In the first colum we show the final codings and their difference. In the other columns we display the respective phase images for two reconstructed signals. The two magnified plots in the last row correspond to the small red segment. Note that the obtained phases are very similar even in the proximity of critical areas.

noisy, but the overall level of detail is maintained, even with details at a relatively high frequency being clearly visible on both reconstructions: for instance, in the close-up of the disc object (third column) a very small bulge is visible in both the reconstructions in proximity of the lower rim of the hemisphere.

Finally, in Figure 3.14, we compare the reconstructed phases obtained after the application of the Discrete Fourier Transform to the captured structured light images before and after the coding step. As in the previous experiments we choose periods of 9, 10 and 13 pixels and we projected 34 patterns for the Multi-Period method and 8 pattern for our approach. In the first column we show respectively the complete coding for the Multi-Period (top) and for the Compound (middle) methods. In order to highlight the differences between the two codings, we show in the third line of this column the difference image between them: as expected they are almost identical and the only slight differences are found in boundary regions, where the measurements are less precise.

In the other two columns we show the phase related to two different signals calculated by the two techniques (we show only two signals for space reasons). For the purpose of a simple comparison between them, we plotted a magnified graph of the small red lines in the phase images. In those graphs the red line represents the phase extracted by the Multi-Period method and the blue line the phase extracted by our technique. It can be seen that the values obtained are very similar, and that they diverge by a very small amount only in proximity of critical areas such as the interface between the disc and the hemisphere.

3.2.4 Conclusions

We have proposed a novel compound phase coding technique that is able to perform a complete and accurate surface reconstruction requiring the projection of as few as 8 patterns. This is obtained by encoding multiple phase information in a single pattern sequence as phases of sinusoidal signals at integer frequencies, thus encoding and decoding the compound signal using standard Fourier analysis. The comparison with another well known technique assesses the ability of the approach to obtain accurate reconstruction even with very few patterns. In addition, the time/quality trade-off can be easily controlled by adding more patterns. In fact, we show that measurement error decreases consistently by adding more information, to the point that our method reaches the performance of other state-of-the-art approaches when feeded with a comparable quantity of data.

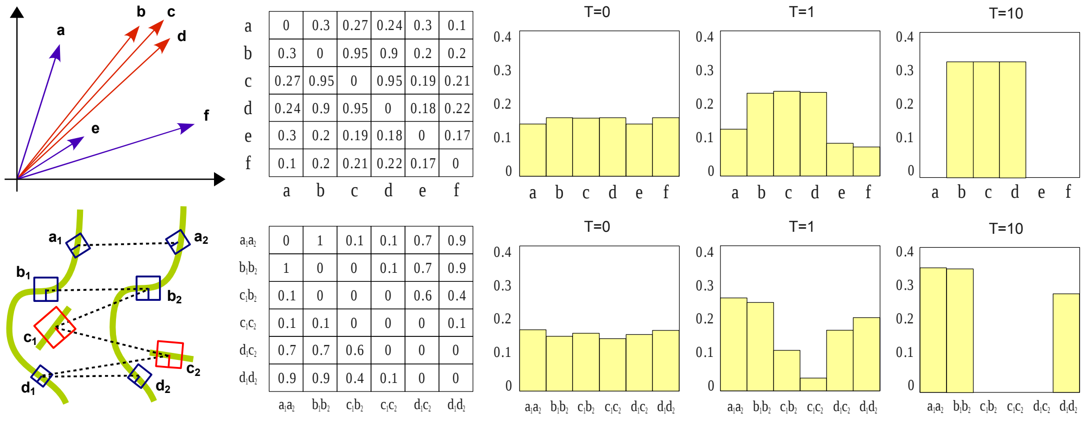
4
Features Selection

Feature detection is a pervasive problem in computer vision; as such, a large number of specialized techniques have been proposed during the years. An overview of the most common methods is given in Chapter 2. Unfortunately, the very concept of *feature* has a different meaning that depends not only on the domain of application (e.g. 2D vs 3D, points vs lines, et cetera), but also in regards to the specific task that is being tackled within a given problem. In the majority of cases, a feature is what can be considered an "interesting trait" of a given object; as such, distinctiveness comes as a natural requirement, whereas a feature can hardly be considered to be "interesting" when its characteristics are commonly encountered across the object [63]. Yet, this view of what constitutes a feature based on some notion of frequency of appearance is arguable, and does not take into account neither the transformations the object can undergo (e.g. additive noise), nor the problem at hand (as an example, this measure is highly sensitive to partiality or view transformations of the object). Another common desideratum is that the feature be robust to a variety of transformations, such as (in the 3D case) additive noise, local scale changes and topological deformations. Or again, repeatability under different instances of the same object might be preferable over robustness in a number of cases.

Given this somewhat vague and task-specific definition of the problem, it comes with no surprise that research dedicated to this topic has been carried out with constant vigor during the years in many different fields. In this Chapter we present some results that span three different areas within the Image Processing and Computer Vision fields: In Section 4.1, we introduce a method based on results from game theory (see Section 2.6) that tackles a 2D object recognition problem in highly noisy scenarios; then, Section 4.1.4 introduces the problem of interest point detection in a 3D setting, where features are extracted according to an outlier filtering principle similar to the one followed in the 2D case; finally, in Section 4.2 we present a sampling technique that attacks the specific problem of fine surface alignment in 3D reconstruction scenarios.

4.1 Figure Extraction from Textured Backgrounds

Feature-based image matching relies on the assumption that the features contained in the model are distinctive enough. When both model and data present a sizeable amount of



	a	b	c	d	e	f
a	0	0.3	0.27	0.24	0.3	0.1
b	0.3	0	0.95	0.9	0.2	0.2
c	0.27	0.95	0	0.95	0.19	0.21
d	0.24	0.9	0.95	0	0.18	0.22
e	0.3	0.2	0.19	0.18	0	0.17
f	0.1	0.2	0.21	0.22	0.17	0

	a_1a_2	b_1b_2	c_1b_2	c_1c_2	d_1c_2	d_1d_2
a_1a_2	0	1	0.1	0.1	0.7	0.9
b_1b_2	1	0	0	0.1	0.7	0.9
c_1b_2	0.1	0	0	0	0.6	0.4
c_1c_2	0.1	0.1	0	0	0	0.1
d_1c_2	0.7	0.7	0.6	0	0	0
d_1d_2	0.9	0.9	0.4	0.1	0	0

Figure 4.1: Examples of the two evolutionary matching games proposed.

clutter, the signal-to-noise ratio falls and the detection becomes more challenging. If such clutter exhibits a coherent structure, as it is the case for textured background, matching becomes even harder. In fact, the large amount of repeatable features extracted from the texture dims the strength of the relatively few interesting points of the object itself. It's exactly this property that we exploit in order to distinguish foreground features from background ones. In addition, the same technique can be used to deal with the object matching itself. The whole procedure is validated by applying it to a practical scenario and by comparing it with a standard point-pattern matching technique.

Several existing feature-based approaches operate by extracting attributed feature points from the images using detectors [126, 128, 114] and descriptors [89, 25] that are locally invariant to illumination, scale and rotation. Usually, the model features are matched with those obtained from the target image by means of some RANSAC-based approach that can exploit the prior given by the descriptors [46]. Critical to the success of this kind of techniques is of course the distinctiveness of the extracted features. Unfortunately, when dealing with textured clutter, this distinctiveness comes short and the number of very repeatable but irrelevant features overshadows those coming from the foreground object. To avoid false matches it is mandatory to recognize and ignore the background. In this work we cope with both the filtering of the background features and the recognition task by tailoring the matching framework introduced in the previous chapters to this specific context. Specifically, we model the filtering step as a self-matching game, where features that show high mutual similarity in the same image are deemed not distinctive enough and thus screened away. By converse, the recognition step is performed as a matching game between the model and a data image, where a set of highly coherent pairs of corresponding features is sought.

4.1.1 Filtering a Textured Background

When dealing with textures, we can expect a large number of features that exhibit very similar descriptors. This is a very unfortunate condition for matching: in fact, this high

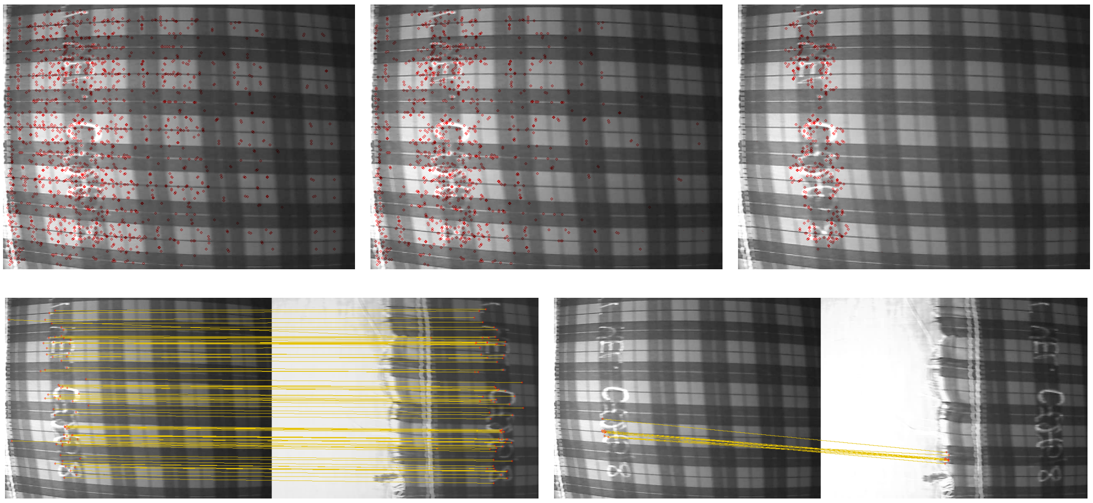


Figure 4.2: Background filtering and feature matching (best viewed in color).

level of congruence can easily distract any matcher from the foreground object. Paradoxically we use this property to screen out background features. Following Section 2.6, we model each feature as a strategy in a matching game where the payoff matrix is defined by:

$$C(i,j) = e^{-\alpha|d_i - d_j|} \tag{4.1}$$

where d_i and d_j are the descriptor vectors associated to features i and j, and α is a parameter that controls the level of selectivity. Clearly, features that are similar will get a large mutual payoff and thus are more likely to be selected by the evolutive process. A simplified (but numerically correct) example of such evolution is shown in the first row of Figure 4.1. Here, six descriptors of dimensionality 2 are labeled from a to f. Vectors b,c and d get high values in the payoff matrix since they are close in the descriptor space. Other descriptors get lower mutual payoffs, according to their respective distances. We start the replicator dynamics ($T = 0$) near the barycenter of Δ^6, which is sligthly perturbed to help avoiding local minima. After just one iteration ($T = 1$), strategies b,c and d get a significant evolutionary boost over the others, and after ten iterations ($T = 10$) they are the only strategies left in the support. We can then classify those features as background and filter them out.

4.1.2 Matching Model and Data

In order to match model and data points we need to define a slightly different matching game. We proceed again as in Section 2.6 by modelling strategies as pairs of features on model and data. We define a payoff measure among strategies proportional to the compatibility of the affine transformation estimated by the descriptor used. Since the main objective of this Chapter is to give a way to filter out "not interesting" features in an object recognition scenario, we restrict ourselves to a brief presentation and refer to

Chapter 5 for a more complete and rigorous treatment of the matching step. Thus, we are able to associate to each strategy (a_1, a_2) an affine transformation $T(a_1, a_2)$. When applied to a_1, the transformation produces the point a_2, but when it is applied to the model point b_1 it will give a point b'_2 that is near b_2 if $T(a_1, a_2)$ is similar to $T(b_1, b_2)$. Given two strategies (a_1, a_2) and (b_1, b_2) and their associated transformations $T(a_1, a_2)$ and $T(b_1, b_2)$ we calculate their reciprocal reprojected virtual points as: $a'_2 = T(b_1, b_2)a_1$ and $b'_2 = T(a_1, a_2)b_1$. Given virtual points a'_2 and b'_2 we are finally able to define the payoff between (a_1, a_2) and (b_1, b_2) as:

$$C((a_1, a_2), (b_1, b_2)) = e^{-\beta \max(|a_2 - a'_2|, |b_2 - b'_2|)} \tag{4.2}$$

where β is a selectivity parameter that allows to operate a more or less selective matching game. Large groups of point pairs that are coherent with respect to an affine transformation will receive a large payoff and thus an evolutive advantage. In the second row of Figure 4.1 we show an example of this matching game (with real data). Here, coherent strategies exhibit high payoff values (i.e., $C((a_1, a_2), (b_1, b_2)) = 1$), while less compatible pairs get lower scores (i.e., $C((a_1, a_2), (c_1, c_2)) = 0.1$). Note that strategies that share the same model or data point get payoff 0 to avoid one-to-many matching. Initially, the population is set to a slightly perturbed barycenter of Δ^6. After one iteration, (c_1, b_2) and (c_1, c_2) have lost a significant amount of support, while (d_1, c_2) and (d_1, d_2) are still played by a sizeable amount of population, despite being mutually exclusive. After ten iterations, (d_1, d_2) has finally prevailed over (d_1, c_2) and the final support has emerged.

4.1.3 Experimental Evaluation

The texture filtering method was tested by applying it to the detection of hand-written markers placed on textured fabric. This is a typical scenario for batch tracking in the textile industry, where barcodes or RFID tags are not viable solutions due to the harsh cloth processing conditions that would destroy them. The first three frames of Figure 4.2 show the background filtering performance of our method. The first frame contains all the original SIFT features extracted, the second one shows those survived after applying our filter with selectivity parameter $\alpha = 10^{-4}$. By using $\alpha = 10^{-3}$ all the background is screened in the third frame. We observed that a larger value of α does not affect much the result, as foreground features are quite disjointed. Matching performance was evaluated by comparing its precision-recall curve with those obtained by using an optimized RANSAC-based technique. Specifically, we implemented a PROSAC [46] variant by using descriptor vectors as hints for the selection of transformation candidates in an affine point-pattern matching. In order to assess the effect of the background elimination step, we applied this RANSAC schema to both filtered and unfiltered frames.

The trade-off between precision and recall was adjusted respectively by means of parameter β and by using different thresholds for the consensus. Tests were performed with 20 markers and 15 different fabric patterns. The markers were present in 59 frames of a 30,000 frames long video sequence. Given the constant presence of a textured background, the poor results obtained with RANSAC and the unfiltered video were expected.

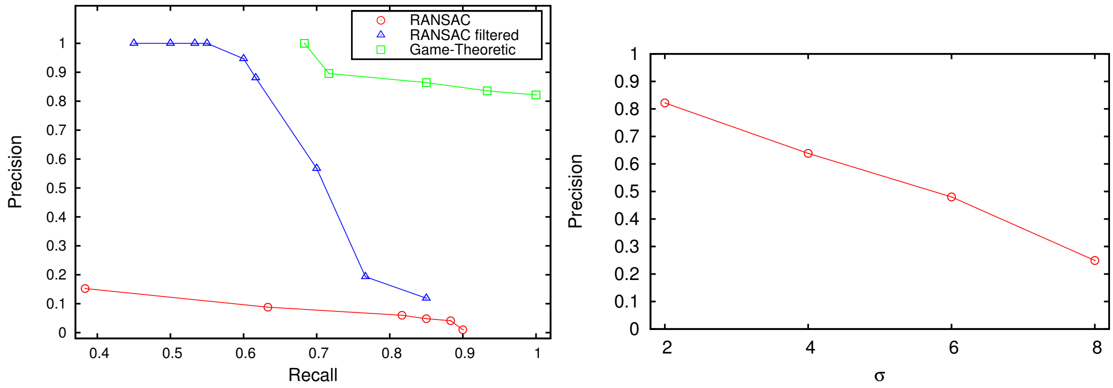


Figure 4.3: Comparison with RANSAC and effect of image noise.

Indeed, we were unable to reach a full recall without a complete loss of precision, and even when accepting a low recall most of the detected frames were false positives due to background matching. RANSAC performance increases dramatically after application of the filter. Nevertheless, it is not possible to obtain a high level of recall without losing precision. This is due to the presence of features that do not belong to the foreground marker and neither are part of a texture. This happens, for instance, with sewings, seams or dirt present in the fabric. In the right half of Figure 4.2 we show an instance where our method obtains the correct match, while RANSAC is distracted by a junction in the fabric. The game-theoretic matcher (applied over filtered frames) obtains by far the best results. In fact, a perfect recall is obtained with a precision value above 0.8 ($\beta = 10^{-3}$) and, by using a more selective parameter ($\beta = 10^{-2}$) all the false positives are avoided while still obtaining a recall just slightly below 0.7. In some practical applications it is more important to guarantee a recall of 1 since a moderate number of false positives can be tolerated (and filtered bottomward in the pipeline), while a miss in the detection is not allowed. To measure the loss in precision with respect to noise, we corrupted both data and model with additive Gaussian noise. At each noise level (expressed with the standard deviation in Figure 4.3) we tuned β to maintain a recall of 1 and measured the precision. While it was always possible to obtain a complete recall, we observed a linear decay of the precision. This is not a failure of the matcher itself, but an impaired effectiveness of the background filter due to the reduced similarity among the extracted descriptors. It should be noted, however, that in this experimental setup a precision of 0.3 with a recall of 1 corresponds to a fall-out of 0.006 (about 180 false positives over 30.000 tests).

Next, we are going to present a similar filtering method in a 3D setting, where common features are treated as structured outliers and filtered out in an iterative manner, much like we did in the previous sections for simple images.

4.1.4 Interest Points Detection

Given the large number of points contained in typical 3D objects, it is not practical for any matching algorithm to deal with all of them. In addition, the isolation of a relatively small

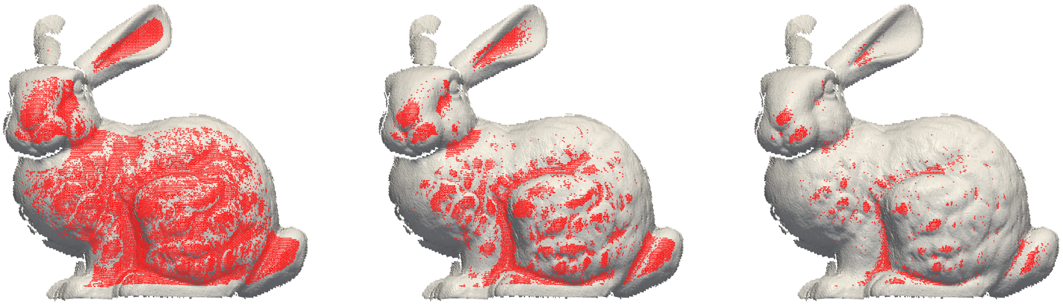

(a) First pass (b) Second pass (c) Third pass

Figure 4.4: Example of the interest points detection process.

number of interest points can enhance dramatically the ability of, for instance, a matching technique to avoid false correspondences, usually due to a large number of features with very common characterizations. This is particularly true when using loosely distinctive features. As in the previous section, we use exactly this property to screen out features exhibiting descriptors that are too common over the surface. The concept is similar to what Gelfand et al. proposed in [63]: feature points should come from regions with rare descriptor values. Again, this happens by defining a matching game where the strategy set S corresponds to the set of all the surface points and the payoff matrix is defined by:

$$\pi_{ij} = e^{-\alpha|d_i - d_j|}\,, \tag{4.3}$$

where d_i and d_j are the descriptor vectors associated to surface point i and j, and α is a parameter that controls the level of selectivity. Clearly, features that are close in the descriptor space will get a large mutual payoff and thus are more likely to be selected by the evolutive process. In this sense, our goal is to let the population evolve to an ESS and then remove from the set of interest points the features that survived the evolutive process. At the beginning we can initialize the set of retained features to the whole surface and run a sequence of matching games until the desired number of points are left. At this point, the remaining features are those characterized by less-common descriptors which are more likely to represent good cues for the matching. It should be noted that by choosing large values for α the payoff function decreases more rapidly with the growth of the distance between the descriptors, thus the matching game becomes more selective and less points survive after reaching an ESS. In the end this results in a blander decimation and thus in a larger ratio of retained interest points. By converse, a small value for α leads to a more greedy filtering and thus to a more selective interest point detector. In Figure 4.4 (from (a) to (c)) we show three steps of the evolutive interest point selection (with respect to a 3-dimensional Normal Hash, presented in Chapter 6 and shown in Figure 6.4). In Figure 4.4(a) we see that after a single pass of the matching game most of the surface points are still considered interesting, while after respectively two and three passes only very distinctive points (belonging to areas with less common curvature profile) are left.

4.1.5 Conclusions

The game-theoretic approach allows to perform a robust feature-based matching even when the foreground is absorbed in a highly textured background, or generally “plunged” among very common descriptors. This is done by playing two different non-cooperative games: a filtering game, that separates foreground from background (unstructured from structured features), and a matching game, that performs the actual point-pattern matching. The experimental validation shows that both the steps concur to the improvement of the whole matching task and the obtained results outperform in terms of precision and recall an optimized RANSAC-based approach. For a more thorough investigation of the 3D selection process, we refer to Chapter 6, where a complete 3D matching pipeline is also presented.

4.2 Sampling Relevant Points for Surface Registration

In this Section we attack a slightly different feature selection problem, which is commonly encountered in surface-based 3D reconstruction pipelines. Specifically, we focus on the alignment step, where after an approximate alignment of two 3D surfaces is obtained, this is subsequently refined by solving an error minimization problem on a subset of “relevant” points on the two shapes.

Surface registration is a fundamental step in the reconstruction of three-dimensional objects using range scanners since, due to occlusions and the limited field of view of the scanners, more range images are necessary to fully cover the object. Registration is typically a two-step process where an initial coarse motion estimation is followed by a refinement step on pairs of range images. Pairwise refinements are almost invariably performed using a variant of the Iterative Closest Point (ICP) algorithm [44, 29] that iteratively minimizes a distance function measured between pairs of selected neighboring points. The performance of ICP depends on several parameters such as the choice of the distance metric, the selection of points on one surface and the selection of mating points in the other. These parameters affect the amount of local minima in the error landscape, the speed of convergence and in general the precision of the resulting alignment in the presence of “complex” geometry. See [117] for a review of some of the variants of ICP.

In this Section we concentrate on the sampling process used to select points on one surface, with a view of correctly aligning “hard” surfaces that offer very few points that constrain the motion and large areas where the surface can slide. For this reason in all the experiments we will use the point-to-plane distance and the closest point will be used as mate, since these approaches, while not offering the fastest convergence, have been shown to be the most robust combination for “difficult” geometries .

The selection of relevant points on one surface to match against points on the other surface is an important issue in any efficient implementation of ICP with strong implications both on the convergence speed and on the quality of the final alignment. This is due to the fact that typically on a surface there are a lot of low-curvature points that scarcely

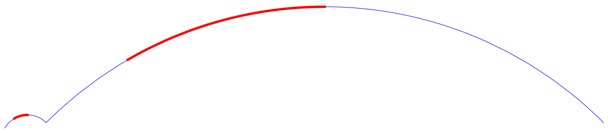


Figure 4.5: The two regions compete for the samples, resulting in a larger number of samples on the large low curvature area than on the smaller part. This results in a bias in the distance measure.

constrain the rigid transformation and an order of magnitude less descriptive points that are more relevant for finding the correct alignment. See Figure 4.5 for an illustration of the problem. There we can see a section of a surface composed of two parts with circular profile which in isolation allow the surfaces to slide. The transformation is fully constrained only if enough points are selected from both parts. However, since there is a large size difference between the two parts, uniform sampling will take proportionally more samples from the larger area than from the smaller one. The difference in sampling biases the error term towards fitting the larger region better than the smaller one, but since the region does not constrain the transformation fully, any sliding that would better fit local noise would be preferred to the correct alignment, resulting in a tendency to“overfit” noise on low-curvature areas. In order to better constrain the set of transformations Rusinkiewicz and Levoy [117] propose a normal space sampling approach that attempts to sample uniformly on the sphere of normal directions rather than on the surface. This, however, only partially solves the problem; to show why we refer again to Figure 4.5. There the thicker arcs on the surface section refer to points that fall in the same normal bin. Since points in the same bin are sampled uniformly, the points on the smaller arc must compete with the points on the larger arc resulting in the same, albeit a bit reduced, tendency of overfitting noise on the larger region that plagues uniform sampling. In effect, we would like to sample points from the two arcs with the same probability. Further, normal space sampling fully constrains only translational error, but in general cannot limit rotational sliding, and the binning interacts poorly with noise in the normal estimation.

An interesting approach to better constrain the transformation is to select points that best equalize the error covariance matrix. To this effect Guehring [68] proposes to weigh the samples based on their contribution to the covariance matrix, but since the analysis is performed after the sampling, the approach cannot constrain the transformation if too few samples were chosen in a relevant region. On the other extreme, Gelfand *et al.* [62] propose an approach that selects the points that constrain the transformation the most. However, the approach is deterministic and has a tendency of sampling very regularly on isolated regions. Since in general the range surfaces overlap only partially, the optimal constraining property only holds if the sampling is performed only on the overlapping part, which means that the analysis and the sampling cannot be performed only once for each surface, but has to be redone for each pair of surfaces. Further, high levels of noise can make some areas artificially strongly constraining, and for that reason the approach needs the surfaces to be smoothed before the points can be selected.

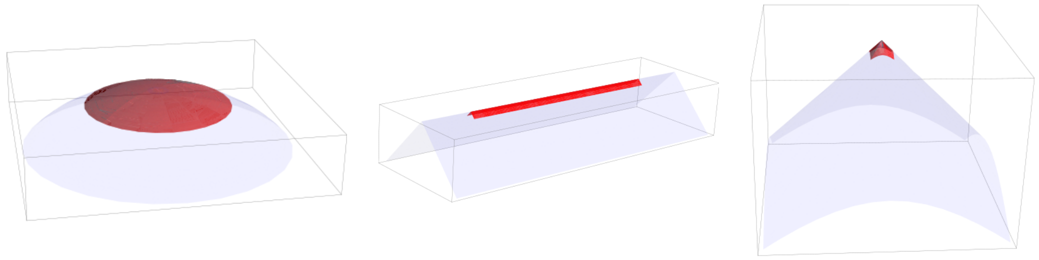


Figure 4.6: The region A_p grows in all directions in a "flat" part of the surface, in only one direction along edges and boundaries and does not grow much at all on vertices.

We propose a different approach to ensure that the rigid transformation is fully constrained, based on the relevance, or local distinctiveness, of points. The idea of point distinctiveness has been extensively used in image processing to develop interest point detectors such as the Harris Operator [70] and Difference of Gaussians [93]. While these approaches work well with 2D intensity images, they cannot be easily extended to handle 3D surfaces since no intensity information is directly available. Several efforts have been made to use other local measures, such as curvature or normals to find relevant points on a surface, but mostly with the end of finding repeatable associations for coarse registration or 3D object recognition. One of the first descriptors to capture the structural neighborhood of a surface point was described by Chua and Jarvis, who with their Point Signatures [45] suggest both a rotation and translation invariant descriptor and a matching technique. Later, Johnson and Hebert introduced Spin Images [76], a rich characterization obtained by a binning of the radial and planar distances of the surface samples respectively from the feature point and from the plane fitting its neighborhood. Given their ability to perform well with both surface registration and object recognition, Spin Images have become one of the most used 3D descriptors. More recently, Pottmann *et al.* proposed the use of Integral Invariants [111], stable multi-scale geometric measures related to the curvature of the surface and the properties of its intersection with spheres centered on the feature point. Finally, Zaharescu *et al.* [160] presented a comprehensive approach for interest point detection (MeshDOG) and description (MeshHOG), based on the value of any scalar function defined over the surface (i.e., curvature or texture, if available). MeshDOG localizes feature points by searching for scale-space extrema over progressive Gaussian convolutions of the scalar function and thus by applying proper thresholding and corner detection. MeshHOG calculates a histogram descriptor by binning gradient vectors with respect to a rotational invariant local coordinate system.

Here we propose a local distinctiveness measure that is associated with the average local radius of curvature, and a sampling strategy that samples points according with their distinctiveness. The distinctiveness is computed through an integral measure, and thus is robust with respect to noise.

4.2.1 Relevance-based Sampling

The relevance of a point p is related to how similar points around p are to it. The larger the number of similar points, the less distinctive, and thus the less relevant, p is. For this reason we formalize the idea of distinctiveness of point p in terms of the area of a surface patch around p where points are similar. More specifically, let p be a point of the surface S, we associate to it a connected region A_p such that

$$A_p = \{q \in S | N_p^T N_q > T \text{ and } p \sim q\} \tag{4.4}$$

where N_p and N_q are the normals of the surface S at points p and q, while $p \sim q$ means that there is a path in A_p connecting p to q, and the dot threshold T is a parameter of the approach. For small values of T the area of A_p is related to the average absolute radius of curvature

$$||A_p|| \approx \bar{r} = \frac{|r_1| + |r_2|}{2} = \frac{|1/k_1| + |1/k_2|}{2}, \tag{4.5}$$

where $||A_p||$ denotes the area of region A_p, k_1 and k_2 are the principal curvatures of S in p and $r_1 = 1/k_1$ and $r_2 = 1/k_2$ are the radii linked with the principal curvatures. Points within A_p have all the orientations similar to that of p and if the surface orientation varies quickly in one direction the growth of the region in that direction will be limited, thus the size of A_p is linked with the distinctiveness of p. The area will be inversely proportional to the curvature, along edges will extend only in one dimension attaining a size one order of magnitude smaller, and will be almost point-like on vertices, where the transformation is locally completely constrained with the exception of rotations along the point normal (see Figure 4.6). Hence, the area is inversely proportional to how much the surface is constraining the transformation locally.

With the patches A_p to hand, we can assign to each point p the measure of distinctiveness

$$f(p) = ||A_p||^{-k} \tag{4.6}$$

where k is an equalization parameter, changing the relative weight of "common" and "distinctive" point. In particular, the larger the value of k, the more the distinctiveness of points forming a small patch A_p is emphasized.

Moreover, since the region A_p is defined in terms of an angular threshold, $||A_p||$ is invariant with respect to resampling, up to the precision imposed by the new sampling resolution. Further, any scale change varies the areas proportionally, so the ratio between patch areas is scale-invariant.

Finally, the area of A_p is an integral measure, thus being less sensitive to noise, and varies continuously along the surface, with T being a smoothing factor.

When the surface is discretized into points and edges, A_p can be easily computed with a region growing approach starting from each point p. If the regions are big, one could use the continuity and locality of A_p to update the region from neighboring points, but in practice, we add a size threshold D limiting the growth of A_p to points whose distance from p is less than D. This way we limit the complexity of the region growing process to $O(D^2)$ for each point and we avoid the uncontrolled expansion of A_p on flat surfaces.

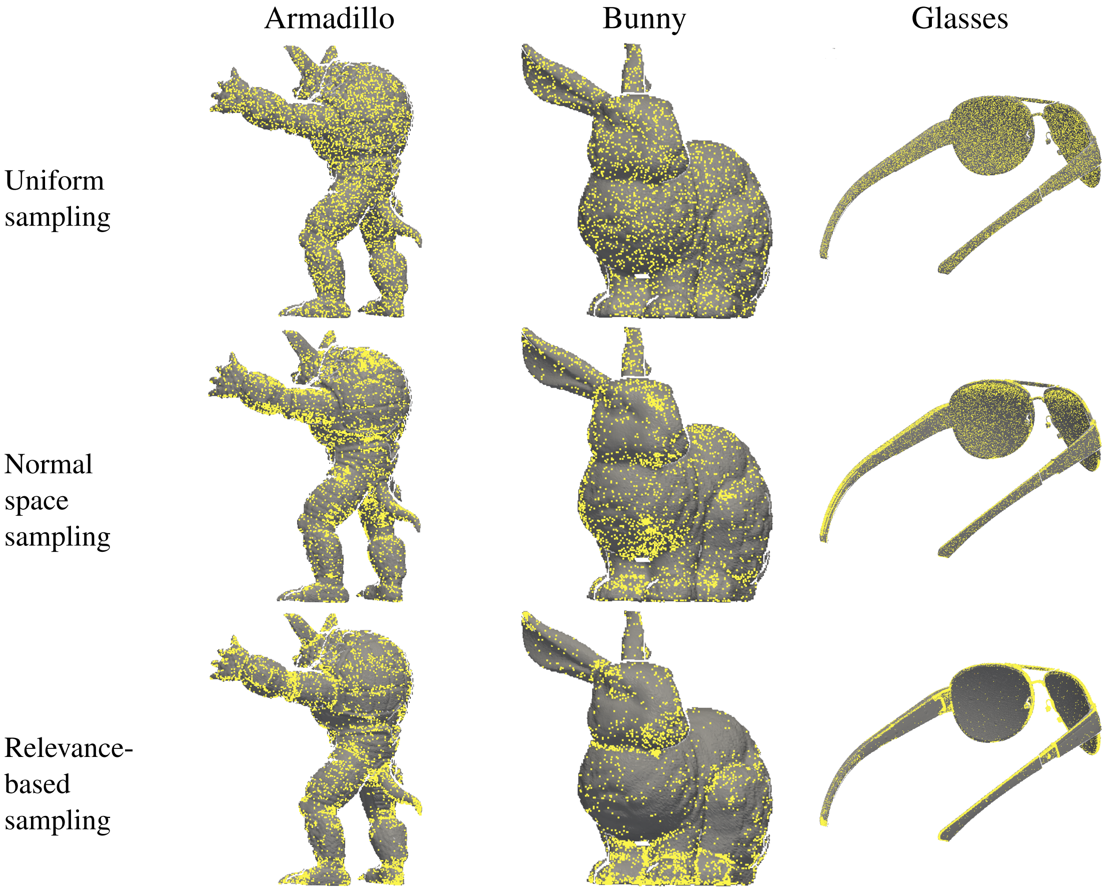


Figure 4.7: Examples of the different sampling approaches.

Once the areas A_p have been computed, we can assign to each point p the measure of distinctiveness

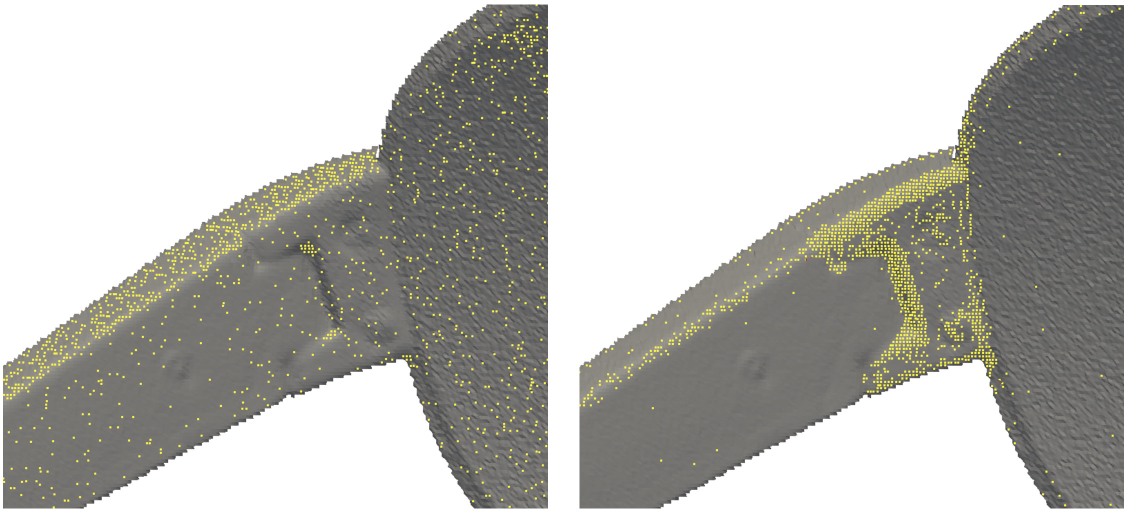
$$\hat{f}(p) = |A_p|^{-k} \tag{4.7}$$

where $|A_p|$ is the number of points in A_p. This approximation works under the assumption that the edge length is uniform through the discretization of surface S.

Once we have computed the distinctiveness of all points in the surface S, we can proceed to sample points from it with a density proportional to $\hat{f}$. To do this we select any order $p_1, \dots, p_n$ of the points in S and compute the cumulative distribution

$$\hat{F}(p_i) = \sum_{j=1}^{i} \hat{f}(j)\,, \tag{4.8}$$

then we sample a number x uniformly in $[0, \hat{F}(p_n)]$ and find the smallest index i such that $\hat{F}(p_i) > x$. To perform the search efficiently, we use interpolation search [109], a variant of binary search that instead of splitting the interval $[i, j]$ in half at each iteration, it splits it at point $i + \frac{x-\hat{F}(p_i)}{\hat{F}(p_j)-\hat{F}(p_i)}$. It is a well known result that interpolation search finds an



Normal space sampling

Relevance-based sampling

Figure 4.8: Closeup of the samples. Relevance-based sampling concentrates samples along the surfaces' fine structures.

element in a sorted array in $O(\log \log n)$ on average for near-uniformly distributed data, compared to the $O(\log n)$ complexity of binary search. Further, the search is faster the higher the entropy of f is. This results in an expected $O(m \log \log n)$ complexity when sampling m points from a surface containing n points.

4.2.2 Experimental Evaluation

To evaluate the performance of the sampling approach we created several range images with known ground-truth transformations. To this end we took the 3D models of the Bunny, the Armadillo, the Dragon, and the Buddha from the Stanford 3D scanning repository and range scans extracted from six sets of glasses scanned using a home-brew scanner built in our lab. The glasses were selected because they are a particularly hard real-world object since it is dominated by large perfectly spherical lenses, while the sliding is constrained only by a very thin rim around them. In the experiments we used 18 scans for each model. For the glasses we used directly the range images provided by the scanner, while for the models taken from the Stanford repository the range images were created by projecting the models onto virtual orthographic cameras placed on a ring around them. Once the range images were to hand, additive Gaussian noise was added along the z dimension to simulate measurement error. In order to avoid having perfect point correspondences, the virtual shots, and thus the points in the various range images, were obtained by projecting equally spaced points on the view-plane of the virtual cameras, and the depths were computed by finding the first intersection of the rays with the model.

All the measures of quality of the alignments are based on the ground-truth alignment, and not the usual Root Mean Square Error (RMSE) because the value of the RMSE de-

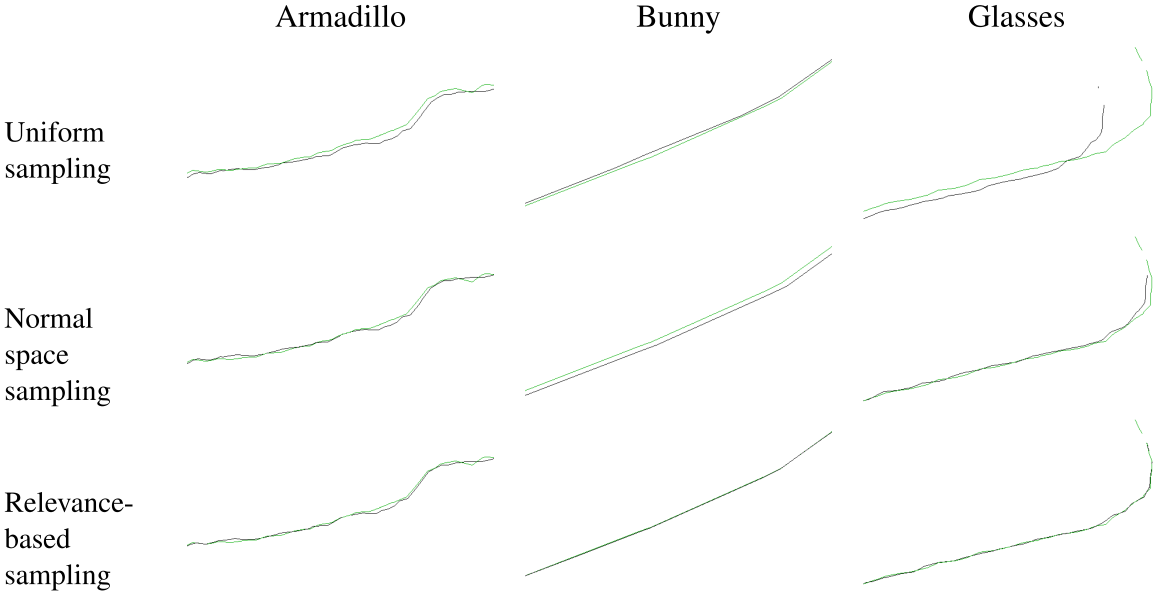


Figure 4.9: Slices of the surfaces aligned with the three sampling strategies.

pends heavily on the sampling strategy and it is completely blind with respect to the noise overfitting problem.

The proposed Relevance-based sampling (RBS) approach was compared against uniform sampling and normal space sampling (NSS).

Figure 4.7 shows an example of sampling range scans from the three models using the three sampling strategies. Uniform sampling does exactly what we expect, with the problems we have discussed. Normal space sampling is more selective of the points, but it still wastes quite a few samples on large low curvature areas like the back of the armadillo, the back and the chest of the bunny and the lenses of the spectacles. Relevance-based sampling, on the other hand, concentrates the samples along edges and feature discontinuities, which do a better job at locking the alignment. This difference in behavior can be seen clearly on the closeups in Figure 4.8. Here normal space sampling samples equally the three main faces of the spectacles, while Relevance-based sampling concentrates the samples along the fine structure details that limit the sliding along the surfaces.

Figure 4.9 Shows examples of the alignments obtained using the three sampling approaches. To better see the difference in alignment we only show a slice of the aligned surfaces cut approximately orthogonally to the surfaces. From the examples we can clearly see that the use of uniform sampling results on the surfaces sliding along large low-curvature areas. This is particularly evident on the surfaces taken from the spectacles, but it is also evident on the armadillo model. The bunny model is relatively simple and results in good alignments with all the methods, even though uniform sampling has a slightly worse performance here as well. Normal space sampling fares much better, but there is still some residual misalignment, especially on the spectacles model. Relevance-based sampling, on the other hand, results in an optimal alignment in all cases.

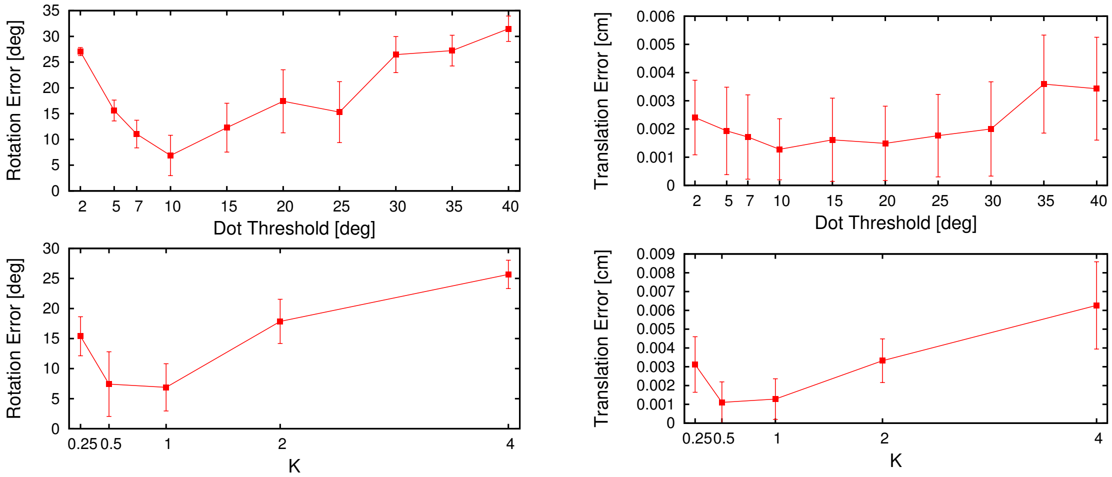


Figure 4.10: Effects of parameters on rotational and translational error.

The first set of quantitative experiments was a sensitivity analysis trying to assess the role of the dot threshold T and of the equalization parameter k. Figure 4.10 shows the angular error (in degrees) and the translation error (in centimeters) as a function of the two parameters. In all the cases the pairs of range images were selected randomly at a distance along the view-circle of at most 3 positions (90 degrees distance in view direction) and were perturbed with additive Gaussian noise along the z dimension with standard deviation equal to 0.4 times the average edge length. We can clearly see that there is an optimal value for the dot threshold at around 10 degrees, and it appears that the optimal value for the equalization parameter k is just slightly below 1. This is due to the fact that noise affects the size of small regions more than larger ones, keeping them smaller, resulting in over-inflated relevance values. A value of k smaller than 1 balances this effect by reducing the relative weight of the smaller regions with respect to larger ones.

Finally, Figure 4.11 plots the resulting rotation and translation error of the alignments obtained with the three sampling strategies as a function of the level of noise added to the range images. Here the range images were selected using the same strategy adopted for the previous set of experiments, allowing variations in the viewing directions of up to 90 degrees. We can see that the relevance-based sampling consistently outperforms uniform sampling by a large margin in both rotational and translational error. Normal space sampling, on the other hand, has the same performance as uniform sampling for rotational errors, while it exhibits the same low translational error obtained by the relevance-based sampling for noise levels smaller than 0.4 times the average edge length. This is consistent with the fact that normal space sampling constrains only the translational sliding. Note however, that with larger noise levels the translational error of normal space sampling breaks down to uniform sampling levels. This is probably due to the interaction between

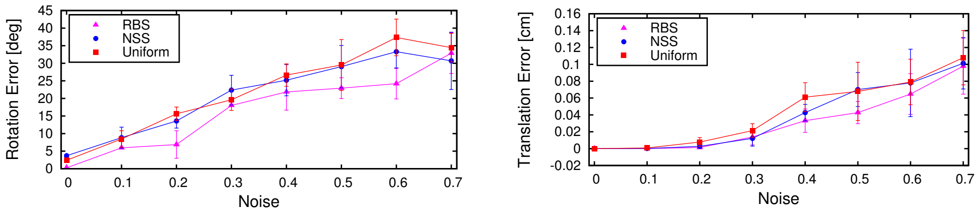


Figure 4.11: Comparison of rotational and translational error obtained with the three sampling strategies.

bin-size and noise, with high noise spreading neighboring points onto several bins.

4.2.3 Conclusions

We have presented a novel sampling strategy for ICP, based on the local distinctiveness of each point. The distinctiveness is gauged through an integral measure that is robust with respect to noise, and the points are then sampled with a density proportional to their distinctiveness. The sampling approach concentrates samples along the surfaces' fine structures, allowing to limit any sliding away from the ideal alignment. Experiments on range images with known ground-truth alignment show that the approach clearly outperforms the most commonly used sampling strategies.

5

Correspondence Selection in Structure from Motion

Our first detailed attempt at analyzing the effectiveness and robustness of the game-theoretic framework follows rather directly from [21]. In their paper, the authors tackle two different problems. The first considers a 2D object recognition scenario in which two images are first segmented and then a many-to-many correspondence is sought between the different segments; the method turns out to be fairly robust in case of relatively unstable segmentations, further confirming its effectiveness under generally not bijective mapping constratints. The second application is point-pattern matching, which considers a scenario in which one of the two images undergoes an affine transformation; here the method proves to be especially effective, allowing to perform very accurate parameter estimation and significantly outperforming other algorithms at the state of the art. In what follows, we introduce a correspondence selection scheme that attacks the specific problem of (2D) point-to-point matching in Structure from Motion scenarios.

5.1 Introduction

Similarly to [21], our method seeks to enforce geometric constraints that do not depend on the full knowledge motion parameters, but rather on some semi-local property that can be estimated from the local appearance of the image features. In this setting, most existing pipelines operate through the iterative refinement of an initial batch of feature correspondences. Typically this is performed by selecting a set of match candidates based on their photometric similarity; an initial estimate of camera intrinsic and extrinsic parameters is then computed by minimizing the reprojection error. Finally, outliers in the initial correspondences are filtered out by enforcing some global geometric property such as the epipolar constraint. In literature many different approaches have been proposed to deal with each of these three steps, but almost invariably they separate the first inlier selection step, which is based only on local image properties, from the enforcement of global geometric consistency. Unfortunately, these two steps are not independent since outliers can lead to inaccurate parameter estimation or even prevent convergence, leading to the well known sensitivity of all filtering approaches with respect to the number of outliers, especially in the presence of structured noise, as one would expect when the images present several repeated patterns. Conversely, our approach relies on a natural selection procedure where incompatible matches are filtered out during the matching process itself, thus removing completely the need for an ex-post verification of the matching consistency. The method operates by enforcing properties that are inferable on image regions at a local or semi-local scale and then extending their validation to a global scale. This is done by casting the selection process into a game-theoretic setting, where feature-correspondences are allowed to compete with one another, receiving support from correspondences that satisfy the same semi-local constraint, and competitive pressure from the rest. The surviving correspondences form a small cohesive set of mutually compatible correspondences, thus satisfying globally the semi-local constraint. In principle, our assumption is not that different from the one subtending to the very popular RANSAC inlier selection method, which assumes that a subset of large consensus exists in correspondence with the correct solution. However, the loose connection with RANSAC breaks as we analyze the selection process itself. Specifically, in our case there is no majority validation for a random subset; rather, the selection happens by letting the strategies compete in a non-cooperative game. The game starts with an initial population where each strategy is played by an equal percentage of players. Such population is then evolved through the action of discrete time dynamics until it reaches some stable state from which a (conceivably) correct matching set can be extracted. In order to assess the advantage provided by our approach, in the experimental section we compare our technique with a reference implementation of the structure-from-motion system presented in [129] and [130].

5.2 Non-Cooperative Games for Inlier Selection

The selection of matching points typically adopted in literature is based on local information provided by pointwise feature descriptors. This approach is limited in that it conflicts with the richness of information that is embedded in the scene structure. For instance, under the assumption of rigidity and small camera motion, intuition suggests that features that are close in one view cannot be too far apart in the other one. Further, if a pair of features exhibit a certain difference of angles or ratio of scales, this relation should be maintained among their respective matches. Following Section 2.6, we model the matching process in a (two-player) game-theoretic framework, where the two players select a pair of matching points from two images. Each player then receives a payoff proportional to how compatible his match is with respect to the other player's choice. We formalize this intuitive notion of consistency between pairs of feature matches into a real-valued compatibility function, and seek for a large set of matches that express a high level of mutual compatibility. Of course, the ability to define a meaningful pairwise compatibility function and a reliable technique for finding a consistent set is at the basis of the effectiveness of the approach.

In this work we will introduce two different payoff functions to address our multi-view point matching problem. In this section we will define a compatibility among pairs of

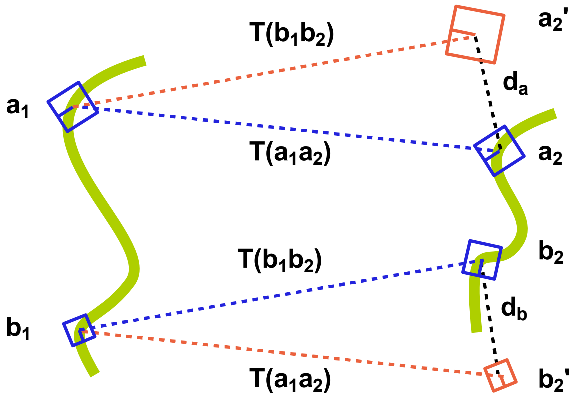


Figure 5.1: The payoff between two matching strategies is inversely proportional to the maximum reprojection error obtained by applying the affine transformation estimated by a match to the other.

correspondences that is proportional to the similarity of the affine transformation inferred from each match; this is done to exploit the expected local spatial and scale coherence among image patches. In Section 5.2.1, we will propose a refinement step that filters out groups of matches by letting them play an evolutionary game where the payoff is bound to their mutual ability to comply with the epipolar constraint.

After defining a set of candidate matches S (Section 2.6), our goal becomes to extract from it a large subset of correspondences that includes only correctly matched features: that is, strategies that associate a physical point in the source image with the same physical point (if visible) in the destination image. To this end, it is necessary to define a payoff function $\Pi : S \times S \rightarrow \mathbb{R}^+$ that exploits some pairwise information available at this early stage (i.e., before estimating camera and scene parameters) and that can be used to impose consistency globally.

We propose two different ways to attain this. Since location, scale, and rotation are associated to each feature, we can associate to each correspondence (a, b) between feature a in the source image and feature b in the target image a similarity transformation $T(a, b)$ that maps the neighborhood of a into the neighborhood of b, transforming the location, orientation and scale measured in the source image into the location, orientation, and scale observed in the target image. Under small motion assumptions, we can expect these similarity transformations to be very similar locally. Thus, imposing the conservation of the similarity transformation, we aim to extract clusters of feature matches that belong to the same region of the object and that tend to lie in the same level of depth. While this could seem to be an unsound assumption for general camera motion, in the experimental section we will show that it holds well with the typical disparity found in standard multiple view and stereo data sets. Further, it should be noted that with large camera motion, most, if not all, commonly used feature detectors fail, thus any inlier selection attempt becomes meaningless. We name this approach *affinity preserving game*.

The second approach enforces that all pairs of correspondences between 2D views be consistent with a common 3D rigid transformation. Here we assume that we have reasonable guesses for the intrinsic camera parameters and reduce the problem space to the search of a 3D rigid transformation from one image space to the other. This condition is in general underspecified, as a whole manifold of pairs of correspondences are consistent with a rigid 3D transformation. However, by accumulating mutual support through a large set of mutually compatible correspondences, one can expect to reduce the ambiguity to a single 3D rigid transformation. In the proposed approach, high order consistency constraints are reduced to a second order compatibility where sets of 2D point correspondences that can be interpreted as projections of rigidly-transformed 3D points all have high mutual support. The reduction is obtained by making use of the scale and orientation information linked with each feature point in the SIFT descriptor [89] and a further reprojection that can be considered a continuous form of hypergraph clique expansion [166]. In the following, we refer to this approach as *virtual point game*.

Affinity Preserving Game

In order to define the payoff function Π we need a way to measure the distance between similarity transformations. In order to avoid the problem of mixing incommensurable quantities, we compute the distance in terms of the reprojection error expressed in pixels. Specifically, given two matching strategies (a_1, a_2) and (b_1, b_2) and their respective associated similarities $T(a_1, a_2)$ and $T(b_1, b_2)$, we calculate tranformed points a'_2 and b'_2 by applying the other strategy's transformation to the source features a_1 and b_1 (see Figure 5.1). More formally,

$$\begin{aligned} a'_2 &= T(b_1, b_2)a_1 \\ b'_2 &= T(a_1, a_2)b_1 \,, \end{aligned}$$

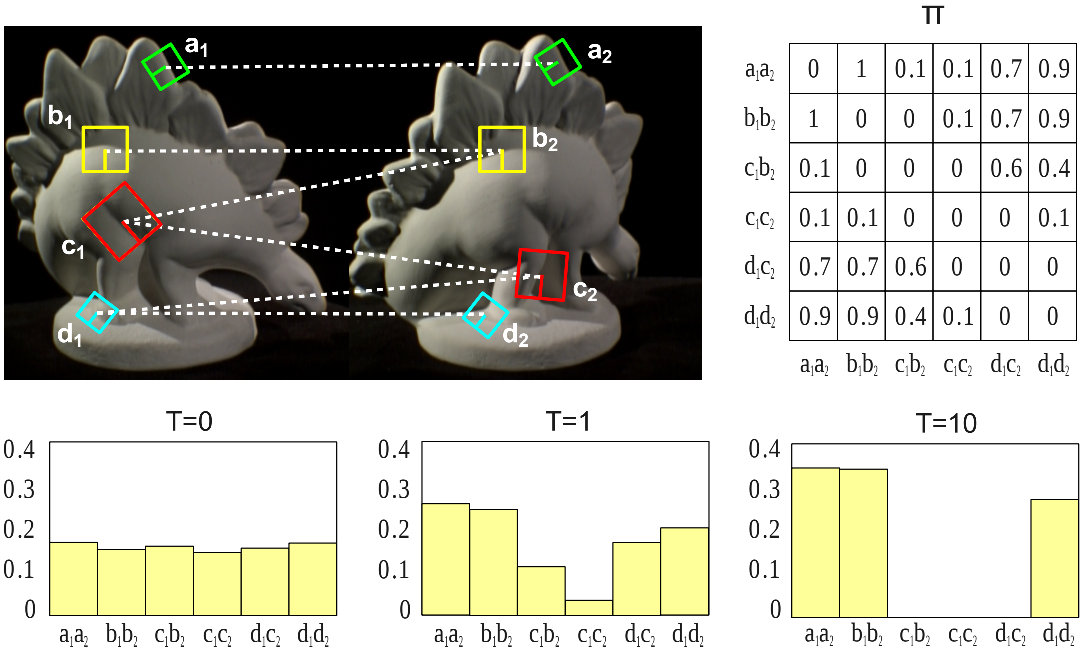
Given transformed points a'_2 and b'_2, we can measure the similarity between (a_1, a_2) and (b_1, b_2) as:

$$\text{sim}((a_1, a_2), (b_1, b_2)) = e^{-\lambda max(|a_2 - a'_2|, |b_2 - b'_2|)} \tag{5.1}$$

where λ is a selectivity parameter: If λ is small, then the similarity function (and thus the matching) is more tolerant with respect to deviation in the similarity transformations, becoming more selective as λ grows. Since each source feature can correspond with at most one destination point, it is desirable to avoid any kind of multiple match. It is easy to show that a pair of strategies with zero mutual payoff cannot belong to the support of an ESS (see [21]), thus any payoff function Π can be easily adapted to enforce one-to-one matching by defining:

$$\Pi((a_1, a_2), (b_1, b_2)) = \begin{cases} \text{sim}((a_1, a_2), (b_1, b_2)) & \text{if } a_1 \neq b_1 \\ & \text{and } a_2 \neq b_2 \\ 0 & \text{otherwise} \end{cases} \tag{5.2}$$

We define payoff (5.2) a *similarity enforcing payoff function* and we call an *affine matching game* any symmetric two player game that involves a matching strategies set S and a



	a_1a_2	b_1b_2	c_1b_2	c_1c_2	d_1c_2	d_1d_2
a_1a_2	0	1	0.1	0.1	0.7	0.9
b_1b_2	1	0	0	0.1	0.7	0.9
c_1b_2	0.1	0	0	0	0.6	0.4
c_1c_2	0.1	0.1	0	0	0	0.1
d_1c_2	0.7	0.7	0.6	0	0	0
d_1d_2	0.9	0.9	0.4	0.1	0	0

Figure 5.2: An example of the affine-based evolutionary process. Four feature points are extracted from two images and a total of six matching strategies are selected as initial hypotheses. The matrix Π shows the compatibilities between pairs of matching strategies according to a one-to-one similarity-enforcing payoff function. Each matching strategy got zero payoff with itself and with strategies that share the same source or destination point (i.e., $\Pi((b_1, b_2), (c_1, b_2)) = 0$). Strategies that are coherent with respect to the same similarity transformation exhibit high payoff values (i.e., $\Pi((a_1, a_2), (b_1, b_2)) = 1$ and $\Pi((a_1, a_2), (d_1, d_2)) = 0.9$)), while less compatible pairs get lower scores (i.e., $\Pi((a_1, a_2), (c_1, c_2)) = 0.1$). Initially (at T=0) the population is set to the barycenter of the simplex and slightly perturbed. After just one iteration, (c_1, b_2) and (c_1, c_2) have lost a significant amount of support, while (d_1, c_2) and (d_1, d_2) are still played by a sizable amount of population. After ten iterations (T=10) (d_1, d_2) has finally prevailed over (d_1, c_2) (note that the two are mutually exclusive). Note that in the final population $((a_1, a_2), (b_1, b_2))$ have a higher support than (d_1, d_2) since they are a little more coherent with respect to similarity.

similarity enforcing payoff function Π.

The main idea of the proposed approach is that by playing a matching game driven by a similarity enforcing payoff function such as (5.2), the strategies (i.e. correspondence candidates) that share a similar locally affine transformation are advantaged from an evolutionary point of view and shall emerge in the surving population. In Figure 5.2 we illustrate a simplified example of this process. Once the population has reached a local maximum, all the non-extinct mating strategies can be considered valid (even though technically strategies become truly extinct only after an infinite number of iterations). Since

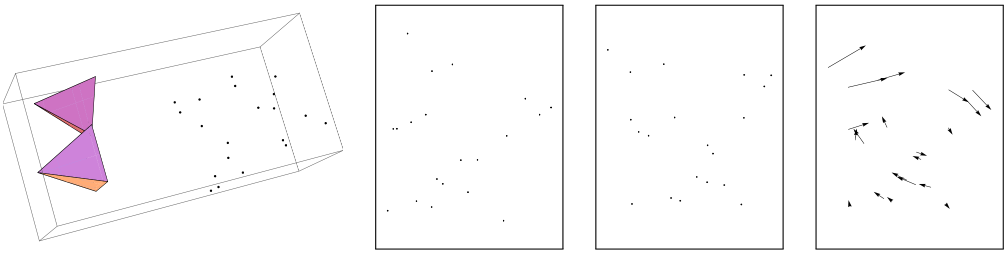


Figure 5.3: In the presence of strong parallax, locally uniform 3D motion does not result in a locally uniform 2D motion. From left to right: 3D scene, left and right views, and motion estimation (from right to left).

we halt the evolution when the population ceases to change significantly, it is necessary to introduce some criteria to distinguish correct from non-correct matches. To avoid a hard threshold we chose to keep as valid all the strategies played whose population size exceeds a percentage of the most popular strategy. We call this percentage *quality threshold* (q). This criterion further limits the number of selected strategies, but increases their consistency, since the population proportion is linked witht the coherence of the strategy with the other surviving strategies. Each evolution process selects only a single group of matching strategies that are mutually coherent with respect to a local similarity transformation. This means that if we want to cover a large portion of the image we need to iterate the process many times, pruning the previously selected matches at each new iteration.

Virtual Point Game

There are two fundamental hypotheses underlying the reduction to second order of a higher-order 3D geometric consistency. First, we assume that the views have the same set of camera parameters, that we have reasonable guesses for the intrinsic parameters, and we can ignore lens distortion. Thus, the geometric consistency is reduced to the compatibility of the projected points with a single 3D rigid transformation related to the relative positions of the cameras. Second, we assume that the feature descriptor provides scale and orientation information and that this is related to actual local information in the 3D objects present in the scene. The effect of the first assumption is that the geometric consistency is reduced to a rigidity constraint that can be cast as a conservation along views of the distances between the unknown 3D position of the feature points, while the effect of the second assumption is that we can recover the missing depth information as a variation in scale between two views of the same point and that this variation is inversely proportional to variation in projected size of the local patch around the 3D point and, thus, to the projected size of the feature descriptor.

More formally, assume that we have two points p_1 and p_2, which in one view have coordinates (u_1^1, v_1^1) and (u_2^1, v_2^1) respectively, while in a second image they have coordinates (u_1^2, v_1^2) and (u_2^2, v_2^2). These points, in the coordinate system of the first camera, have 3D coordinates $z_1^1(u_1^1, v_1^1, f)$ and $z_2^1(u_2^1, v_2^1, f)$ respectively, while in the reference frame of the second camera they have coordinates $z_1^2(u_1^2, v_1^2, f)$ and $z_2^2(u_2^2, v_2^2, f)$. Up to a change

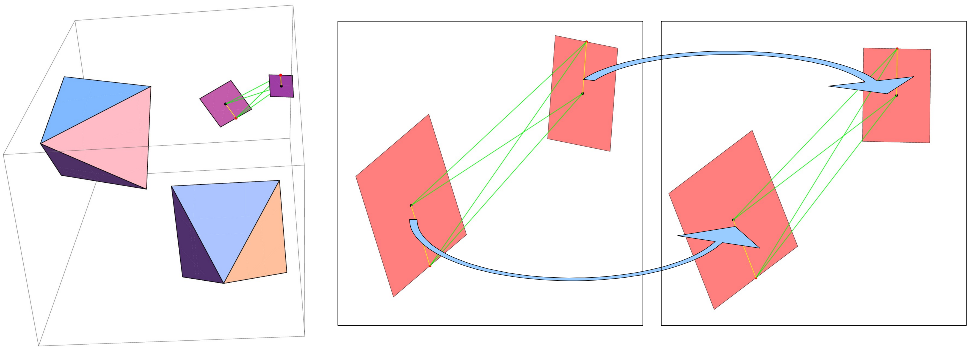


Figure 5.4: Scale and orientation offer depth information and a second virtual point. the conservation of the distances in green enforce consistency with a 3D rigid transformation.

in units, these coordinates can be re-written as

$$p_1^1 = \frac{1}{s_1^1}\begin{pmatrix} u_1^1 \\ v_1^1 \\ f \end{pmatrix},\ p_2^1 = \frac{a}{s_2^1}\begin{pmatrix} u_2^1 \\ v_2^1 \\ f \end{pmatrix},\ p_1^2 = \frac{1}{s_1^2}\begin{pmatrix} u_1^2 \\ v_1^2 \\ f \end{pmatrix},\ p_2^2 = \frac{a}{s_2^2}\begin{pmatrix} u_2^2 \\ v_2^2 \\ f \end{pmatrix},$$

where f is the focal lenght and a is the ratio between the actual scales of the local 3D patches around points p_1 and p_2, whose projections on the two views give the perceived scales s_1^1 and s_1^2 for point p_1 and s_2^1 and s_2^2 for point p_2.

The assumption that both scale and orientation are linked with actual properties of the local patch around each 3D point is equivalent to having 2 points for each feature correspondence: the actual location of the feature, plus a virtual point located along the axis of orientation of the feature at a distance proportional to the actual scale of the patch. These pairs of 3D points must move rigidly going from the coordinate system of one camera to the other, so that given any two sets of correspondences with 3D points p_1 and p_2 and their corresponding virtual points q_1 and q_2, the distances between these four points must be preserved in the reference frames of every view (see Figure 5.4).

Under a frontal-planar assumption for each local patch, or, less stringently, under small variation in viewpoints, we can assign 3D coordinates to the virtual points in the reference frames of the two images:

$$\begin{aligned} q_1^1 &= p_1^1 + \begin{pmatrix} \cos\theta_1^1 \\ \sin\theta_1^1 \\ 0 \end{pmatrix} & q_2^1 &= p_2^1 + a\begin{pmatrix} \cos\theta_2^1 \\ \sin\theta_2^1 \\ 0 \end{pmatrix} \\ q_1^2 &= p_1^2 + \begin{pmatrix} \cos\theta_1^2 \\ \sin\theta_1^2 \\ 0 \end{pmatrix} & q_2^2 &= p_2^2 + a\begin{pmatrix} \cos\theta_2^2 \\ \sin\theta_2^2 \\ 0 \end{pmatrix}, \end{aligned}$$

where θ_i^j is the perceived orientation of feature i in image j. At this point, given two sets of correspondences between points in two images, namely the correspondence m_1 between a feature point in the first image with coordinates, scale and orientation $(u_1^1, v_1^1, s_1^1, \theta_1^1)$ with the feature point in the second image $(u_1^2, v_1^2, s_1^2, \theta_1^2)$, and the correspondence m_2 between the points $(u_2^1, v_2^1, s_2^1, \theta_2^1)$ and $(u_2^2, v_2^2, s_2^2, \theta_2^2)$ in the first and second image respectively, we can compute a distance from the manifold of feature descriptors compatible with a single 3D rigid transformation as

$$\begin{aligned} d(m_1, m_2, a) = {} & (||p_1^1 - p_2^1||^2 - ||p_1^2 - p_2^2||^2)^2 + (||p_1^1 - q_2^1||^2 - ||p_1^2 - q_2^2||^2)^2 + \\ & (||q_1^1 - p_2^1||^2 - ||q_1^2 - p_2^2||^2)^2 + (||q_1^1 - q_2^1||^2 - ||q_1^2 - q_2^2||^2)^2 \,. \end{aligned}$$

From this we define the compatibility between correspondences as

$$C(m_1, m_2) = \max_a e^{-\gamma d(m_1, m_2, a)} \,, \tag{5.3}$$

where a is maximized over a reasonable range of ratio of scales of local 3D patches. In our experiments a was optimized in the interval $[0.5; 2]$.

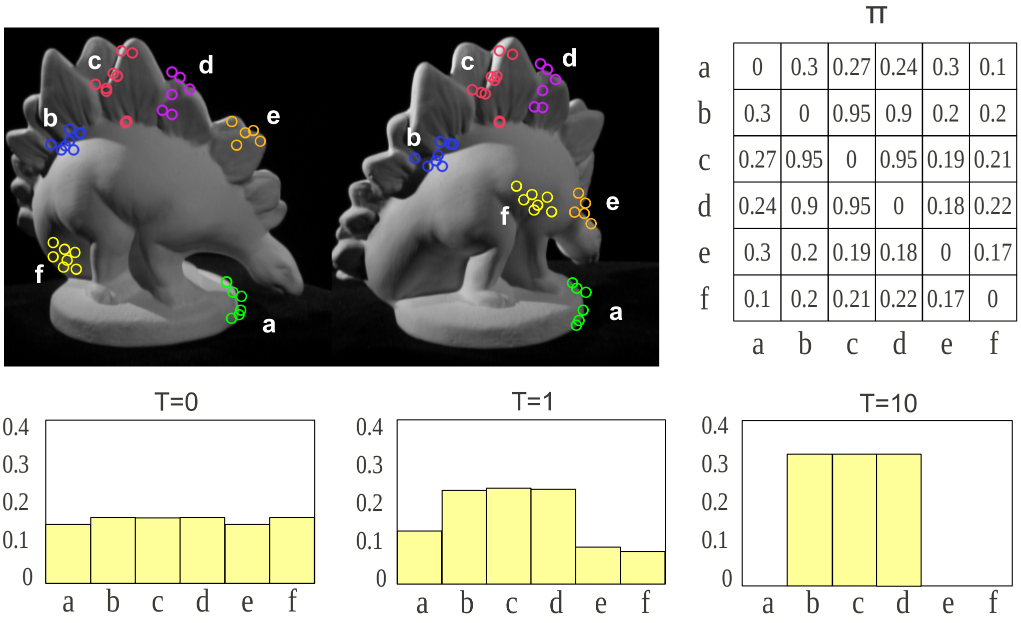
Each matching process selects a group of matching strategies that are coherent with respect to a local similarity transformation. This means that if we want to cover a large portion of the subject we need to iterate many times and prune the previously selected matches at each new start. Obviously, after all the depth levels have been swept, small and not significative residual groups start to emerge from the evolution. To avoid the selection of this spurious matches we fixed a minimum cardinality for each valid group.

5.2.1 Refinement by Epipolar Constraint Enforcement

The game formulations we just introduced shift the matching problem to a more global scope by producing a set of correspondences between groups of features. While the affine camera model extract very coherent groups, making such *macro features* more robust and descriptive than single points, in principle there is nothing that prevents the system to still produce wrong or weak matches. To reduce this chance we propose a different game setup that allows for a further refinement. In this game the strategies set S corresponds to the set of paired features groups extracted from the affine matching game and the payoff between them is related to the features' agreement to a common epipolar geometry. More specifically, given two pairs of matching groups $a \subseteq M \times D$ and $b \subseteq M \times D$, each one made up of model and data features, we estimate the epipolar geometry from $a \cup b$ and define the payoff among them as:

$$\Pi(a, b) = e^{-\lambda \sum_{(s,t) \in a \cup b} d(t, l(s))} \tag{5.4}$$

Where $l(p)$ is a function that gives the epipolar line in the data image from the feature point p in the model image, according to the estimated epipolar geometry, and $d(p, l)$ calculates the distance between point p and the epipolar line l. It is clear that this distance is low (and thus the payoff high) if the two groups share a common projective interpretation and high otherwise. Of course, different pairs of groups can agree on different epipolar geometry, but the transitive closure induced by the selection process ensures that the



	a	b	c	d	e	f
a	0	0.3	0.27	0.24	0.3	0.1
b	0.3	0	0.95	0.9	0.2	0.2
c	0.27	0.95	0	0.95	0.19	0.21
d	0.24	0.9	0.95	0	0.18	0.22
e	0.3	0.2	0.19	0.18	0	0.17
f	0.1	0.2	0.21	0.22	0.17	0

Figure 5.5: An example of the selection of groups of features that agree with respect to a common epipolar geometry. Six matching groups are selected by the affine matching step (labelled from a to f in figure). Each pair of feature sets is modeled as a matching strategy and the payoff among them is reported in matrix Π. Note that groups b,c and d are correctly matched and thus exhibit an high mutual payoff. By contrast, group a (which is consistent both in term of photometric and affine properties), e and f are clearly mismatched with respect to the overall scene geometry, which in turn leads to an high error on the epipolar check and thus to a low score in the payoff matrix. At the start of the evolutionary process each strategy obtains a fair amount of players (T=0). As expected, after just one iteration of the replicator dynamics the more consistent staregies (b,c and d) obtain a clear advantage. Finally, after ten iterations (T=10) the other groups have no more support in te population and only the correct matches survived.

strategies in the surviving population will agree on the same (or very similar) projective transformation (see Figure 5.5 for a complete example of this process). Regarding the estimation of the epipolar geometry this can be done in two different ways: if we know at least the intrinsic calibration of the camera we can estimate the essential matrix, by contrast, if we do not have any hint about the camera geometry, we must resort to a more relaxed set of constraints and use the fundamental matrix instead. In the experimental section we will test both scenarios.

5.3 Experimental Results

We performed an extensive set of tests in order to validate the proposed techniques and to explore their limits. Both quantitative and qualitative results are shown and the perfor-

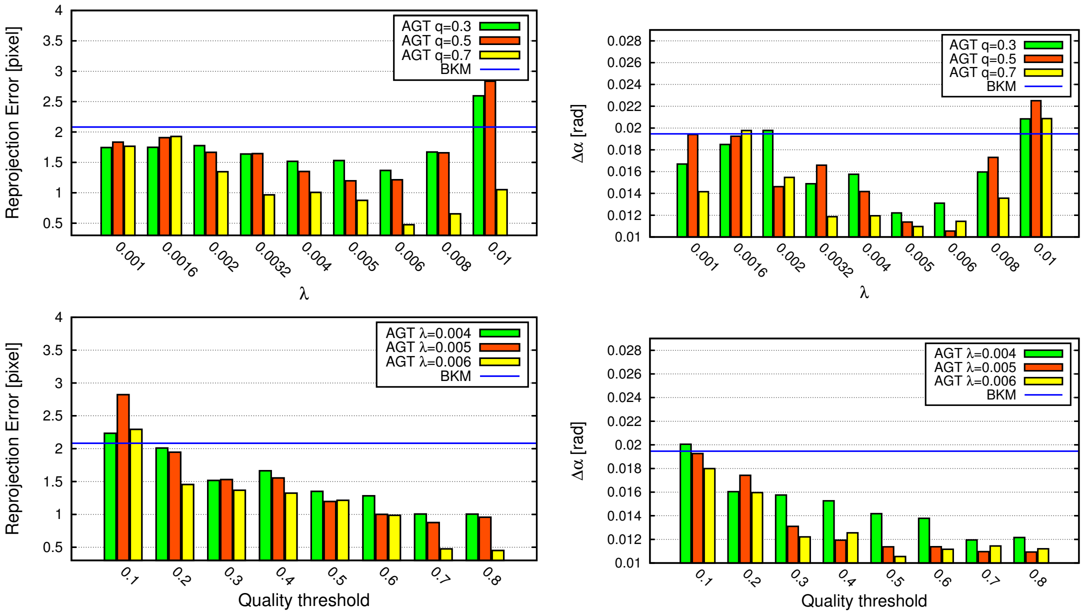


Figure 5.6: Analysis of the performance of the Affine Game-Theoretic approach with respect to variation of the parameters of the algorithm.

mances are compared with those achieved by a standard baseline method, i.e., the default feature matcher in the Bundler suite [130].

5.3.1 General Setup and Data sets

All the following experiments have been made by applying a common basic pattern: first a set of features is extracted from the images by using the SIFT keypoint detector made freely available in [89], then these interest points are paired using the matcher we want to test, finally scene and camera parameters are estimated by using the final portion of Bundler [130] pipeline (i.e. the part of the suite that applies Levenberg-Marquardt optimization to a set of proposed matches). Here we evalute three game-theoretic approaches: The first, referred to as Affine Game-Theoretic approach (AGT), uses the affine matching game without the further refinement provided by the enforcement of the epipolar geometry. In this case the iterative extraction and elimination of the groups is image-based, i.e., after a group of matches is selected, all matches that have souces or targets close to the source and target points of the extraced correspondences are eliminated, and then the evolutionary process is reiterated on the reduced set of strategies. The process is stopped when an extracted group is smaller than a given threshold or has average payoff smaller than a given threshold. The second and third approaches, referred to as Calibrated Projective Game-Theoretic approach (CPGT) and Uncalibrated Projective Game-Theoretic approach (UPGT) respectively, make use of the epipolar refinement. CPGT assumes that the camera intrinsic parameters are (approximately) known and estimate the epipolar geometry through the essential matrix, while UPGT uses the fundamental matrix. In both

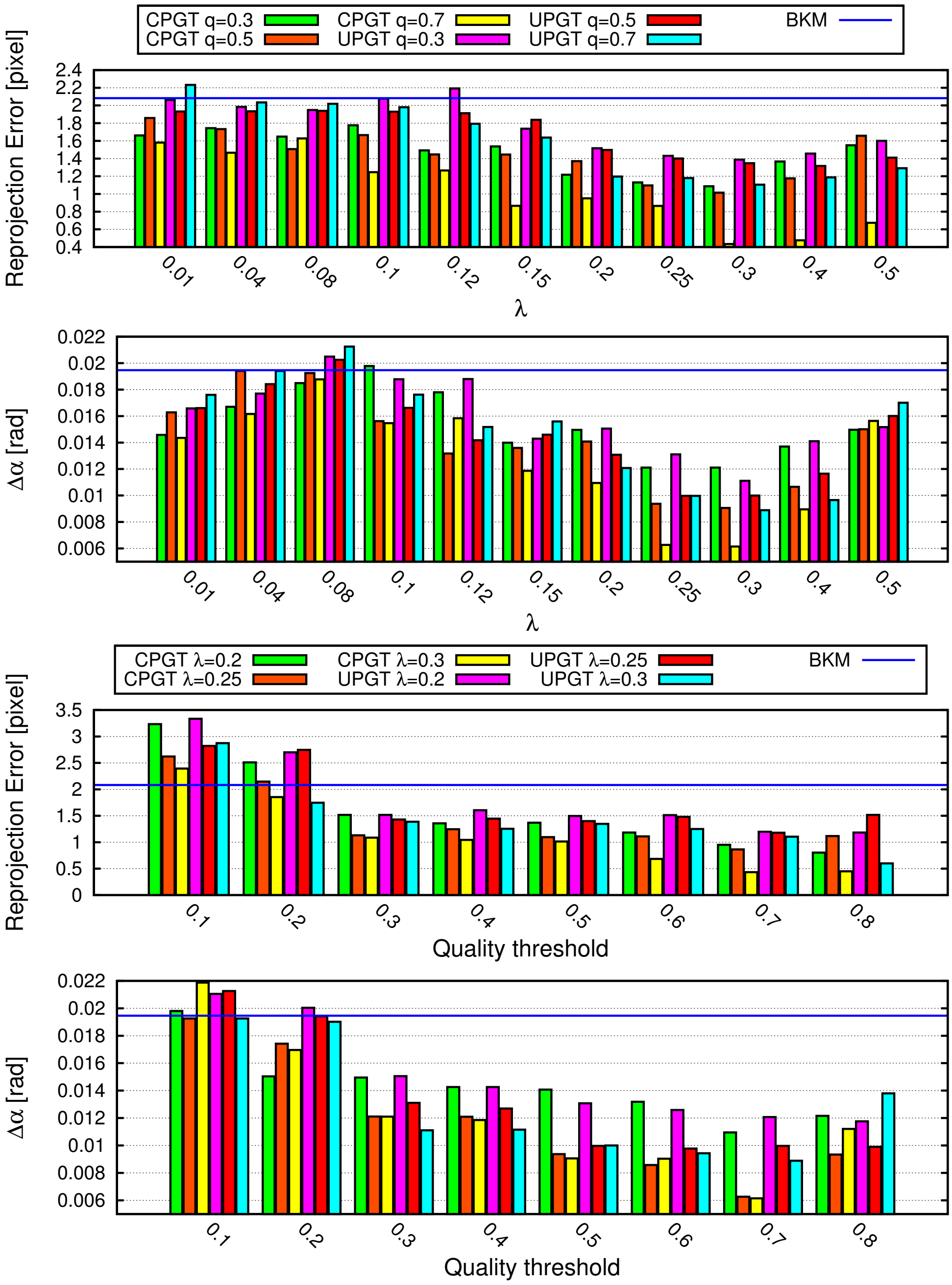


Figure 5.7: Analysis of the performance of the Calibrated and Uncalibrated Projective Game-Theoretic approaches with respect to variation of the parameters of the algorithm.

these appraoches the iterative extraction and elimination of the groups is strategy-based, i.e., after a group of matches is selected only those matches are eliminated from the strategy set, thus allowing for the same features to appear in several groups, while the stopping criterion here is the same as that of AGT. In our experiments the intrinsic parameters for

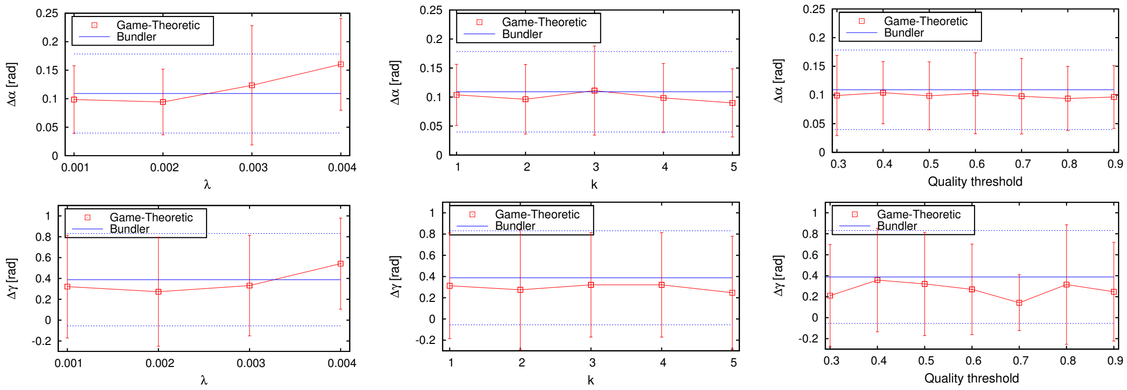
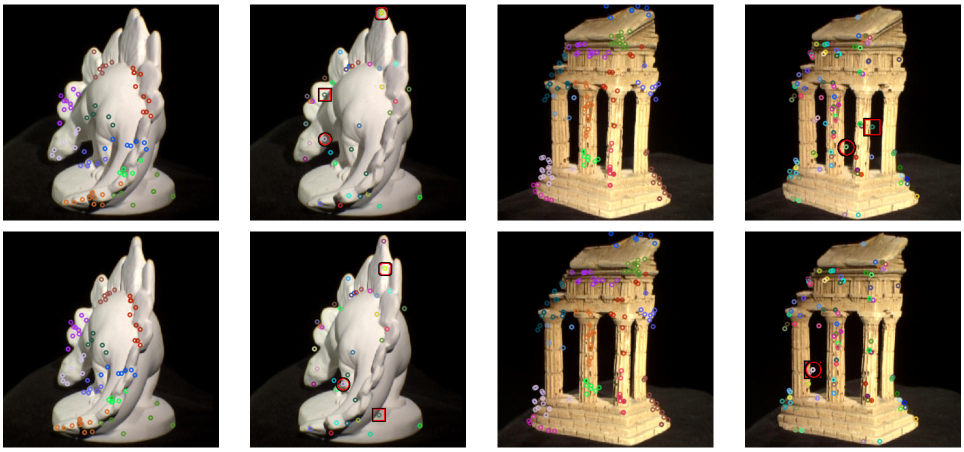


Figure 5.8: Performance analysis of the virtual points approach with respect to variation of its parameters.

CPGT have been estimated from the images' EXIF information. The three approaches are compared against the default feature matcher in the Bundler suite (BKM). This is a reasonable choice for several reasons: BKM is optimized to work with SIFT descriptors and, obviously, with the Bundler suite; in addition it is very popular in literature since Bundler itself has been used as the default matcher in many of the recent papers about SfM and dense stereo reconstruction. For each test we evaluated two quality measures: the average reprojection error (expressed in pixels) and the differences in radians between the ground-truth and the estimated rotation angle ($\Delta\alpha$). The first measure aims to capture the cumulative error made in the reconstruction of the structure and the estimation of the motion, while the second measure aims to decouple the error on the camera orientation from the one related to the scene reconstruction. This is possible since we used images pairs coming from a calibrated camera head or image sets with an available ground-truth. Specifically we used a pair of cameras previously calibrated through a standard procedure and took stereo pictures of 20 different, isolated objects; in addition we also included in the data set the shots coming from the "DinoRing" and "TempleRing" sequences from the Middlebury Multi-View Stereo dataset [124]. We conducted two main sets of experiments. The goal of the first set is to analyze the impact of the parameters, namely λ and *quality threshold* (q), over the accuracy of the results. Since AGT and CPGT/UPGT have different payoff functions and the selectivity λ is not directly comparable we investigate its influence separately. In addition, all the experiment regarding the refinement methods are made using very relaxed parameters for the AGT step. This is due to the fact that we are willing to accept a slightly higher number of outlier in the first step in exchange for an higher number of candidate groups in the hope that the refinement process is able to eliminate the spurious groups, but still resulting in a larger number of good correspondences from which to perform parameter estimation. In the second batch of experiments we compare our techniques with the default Bundler matcher. In these experiments the parameters are set to the optimal values estimated previously. We provide both quantitative and qualitative results from these comparisons: the quantitative analysis is based on



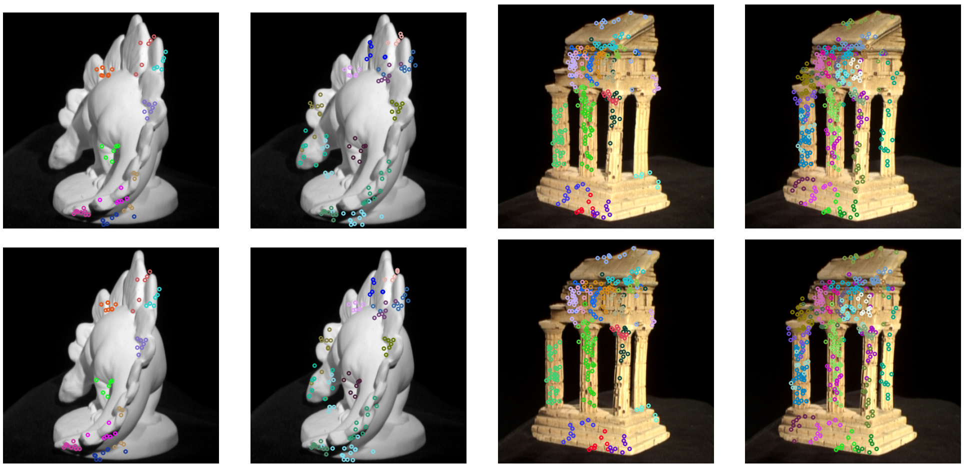
		Dino sequence		**Temple sequence**	
		AGT	BKM	AGT	BKM
Matches		14573	9245	25785	22317
ϵ	≤ 1 pix	24.83	6.49406	22.6049	24.6729
	≤ 5 pix	54.94	48.3659	62.7737	61.8957
	≥ 5 pix	20.21	45.1401	14.6214	13.4314
	Avg.	2.3086	4.5255	2.3577	2.3732
$\Delta\alpha$	Avg.	0.005751	0.005561	0.010514	0.009376
	S. dev.	0.003242	0.003184	0.005282	0.004646
	Max	0.012057	0.011475	0.021527	0.017016
Avg. levels		8.42	-	9.27	-

Figure 5.9: Results obtained with two multiple view data sets (image best viewed in color).

the errors in reprojection and motion estimation, while the qualitative results are based on a dense reconstruction obtained using the recovered parameters as input to the PMVS dense multiview suite [61].

5.3.2 Influence of Parameters

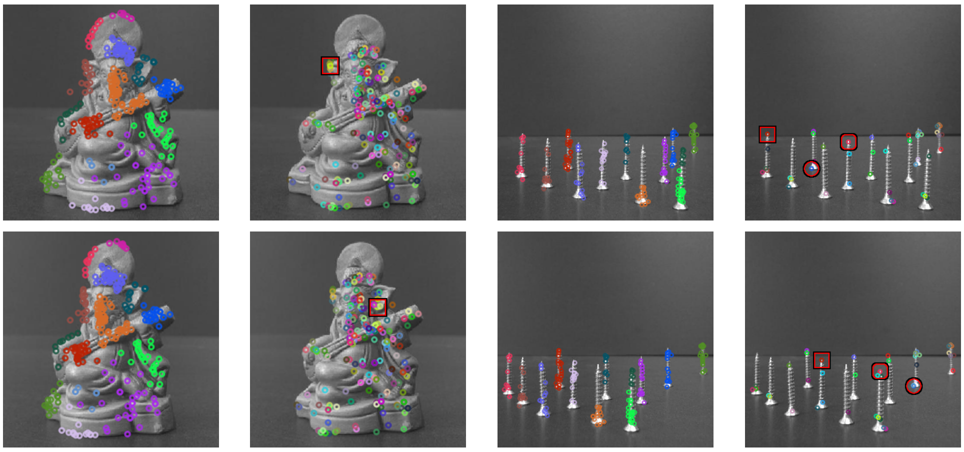
The AGT method depends on two explicit parameters: the sensitivity parameter λ, which modulates the steepness of the payoff function 5.2, and q, i.e. the percentage of the population density with respect to the most represented strategy that one match must obtain to be deemed not extinct. As stated in Section 5.2, λ controls the selectivity of teh selection process, while q allows to further filter the extracted group based on its cohesiveness. Higher values will lead to a more selective culling, while lower values will allow more strategies to pass the screening. Figure 5.6 reports the results of these



		Dino sequence		Temple sequence	
		CPGT	UPGT	CPGT	UPGT
Matches		15018	15231	28106	28407
ϵ	≤ 1 pix	32.1731	20.0126	25.7232	18.3715
	≤ 5 pix	61.4826	75.4671	64.5294	78.5347
	≥ 5 pix	6.3518	4.5203	9.7474	3.0938
	Avg.	1.7051	2.9841	2.1642	3.6713
$\Delta\alpha$	Avg.	0.004823	0.006437	0.009411	0.01328
	S. dev.	0.003671	0.004514	0.005143	0.006545
	Max	0.013147	0.017421	0.019725	0.027832
Avg. levels		17.21	18.34	20.13	22.05

Figure 5.10: Results obtained with two multiple view data sets (image best viewed in color).

experiments averaged over the full set of 20 stereo pairs taken with a previously calibrated camera pair. The first row shows the effect of the selectivity parameter λ. This is evaluated for three different q levels, from 0.3 to 0.7. As expected, both low and high values lead to higher errors, mainly with respect to the estimation of the angle between the two cameras. This is probably due to a too tight and a too relaxed enforcement of local coherence respectively. It could be argued that the estimation of the optimal λ can be tricky in practical situations; however, we must note that, whith a reasonable high q, it takes a very large sensitivity parameter to obtain a performance worst than that obtained with the default Bundler matcher. Regarding the quality threshold, we can see in the second row of Figure 5.6 that the best results are achieved by setting an high level of quality: this is clearly due to the fact that, in practice, the replicator dynamics have converged to a stable ESS and thus most of the non-zero strategies are indeed inliers and are mostly subject only to the (small) feature localization error, thus exhibiting all an equally high

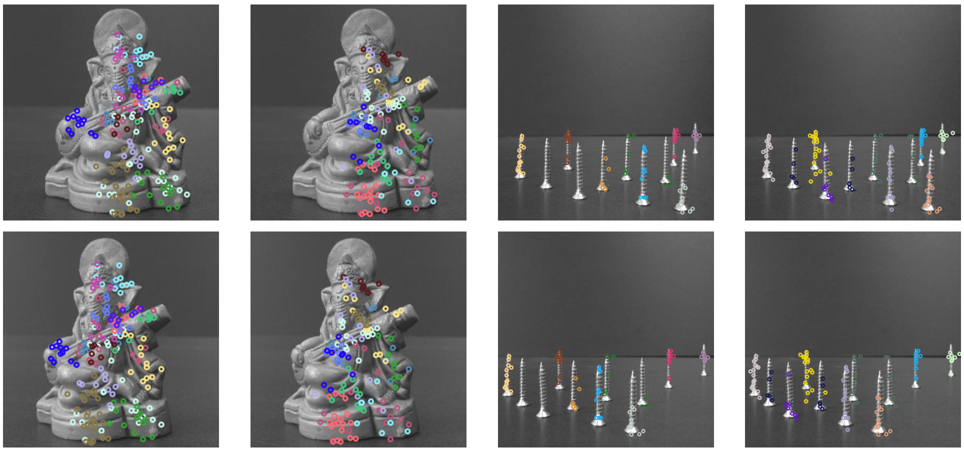


	Ganesha stereo		Screws stereo	
	AGT	BKM	AGT	BKM
Matches	280	200	211	46
ϵ ≤ 1 pix	98.2824	20	0	0
≤ 5 pix	1.7175	80	34.7716	6.75676
≥ 5 pix	0	0	65.2284	93.2432
Avg.	0.321248	1.67583	5.86237	10.2208
$\Delta\alpha$	0.001014	0.007424	0.020822	0.030995
Levels	14	-	12	-

Figure 5.11: Results obtained with two stereo view data sets (image best viewed in color).

density. In Figure 5.7 we show the result obtained by trying different parameters with CPGT and UPGT. As previously stated these experiments were made by performing an affine matching step with relaxed parameters: namely a value of λ of 0.09 and a q of 0.6. The overall behaviour with respects to parameters is similar to what observed for AGT: very low and very high values for λ lead to less satisfactory results (whereas in general better than those obtained with the Bundler key matcher) and high q seems to guarantee good estimates. Overall it seems that CPGT always gives better results than UPGT. We will analyze this behaviour with more detail in the next section.

Finally, we also analyzed the impact of parameters of the virtual points algorithm over the quality of the final results. To this end we investigated three parameters: the similarity decay λ, the number k of candidate mates per features, and the *quality threshold*, that is the minimum support for a correspondence to be considered non-extinct, divided by the maximum support in the population. Figure 5.8 reports the results of these experiments. Overall, these experiments suggest that those parameters have little influence over the quality of the result. However the game-theoretic approach achieves better average results and smaller standard deviations for almost all reasonable values of the parameters.



	Ganesha stereo		Screws stereo	
	CPGT	UPGT	CPGT	UPGT
Matches	315	282	72	108
ϵ ≤ 1 pix	99.0017	83.4812	2.1637	0
≤ 5 pix	0.9983	16.5188	37.5721	26.3417
≥ 5 pix	0	0	60.2642	73.6583
Avg.	0.300272	1.2311	3.92133	4.6379
$\Delta\alpha$	0.001623	0.00466	0.025341	0.03945
Levels	15	13	8	9

Figure 5.12: Results obtained with two stereo view data sets (image best viewed in color).

5.3.3 Comparisons between Approaches

To further explore the differences among the proposed techniques and the Bundler matcher, we executed two sets of experiments. The first set applies the approaches to unordered images coming from the DinoRing and TempleRing sequences from the Middlebury Multi-View Stereo dataset [124]; for these models, the camera extrinsic parameters are provided and used as a ground-truth. The rationale for using these sets (in opposite to simple stereo pairs) is to allow Bundler to optimize the parameters and correspondences over the complete sequence. The second set is composed of two calibrated stereo scenes selected from the previously acquired collection of 20 items, specifically a statue of Ganesha and a handful of screws placed on a table. For all the sets of experiments we evaluated both the rotation error of all the cameras and the reprojection error of the detected feature points. In the Middlebury sets the results are presented as averages. The Dino model is a difficult case in general, as it provides very few distinctive features; Figure 5.9 shows the correspondences produced by AGT (left column) in comparison with BKM (right column). The parameters where set to the optimal values estimated in the previous experiments ($\lambda = 0.06$ and $q = 0.8$). This resulted in the detection of many correct matches organized

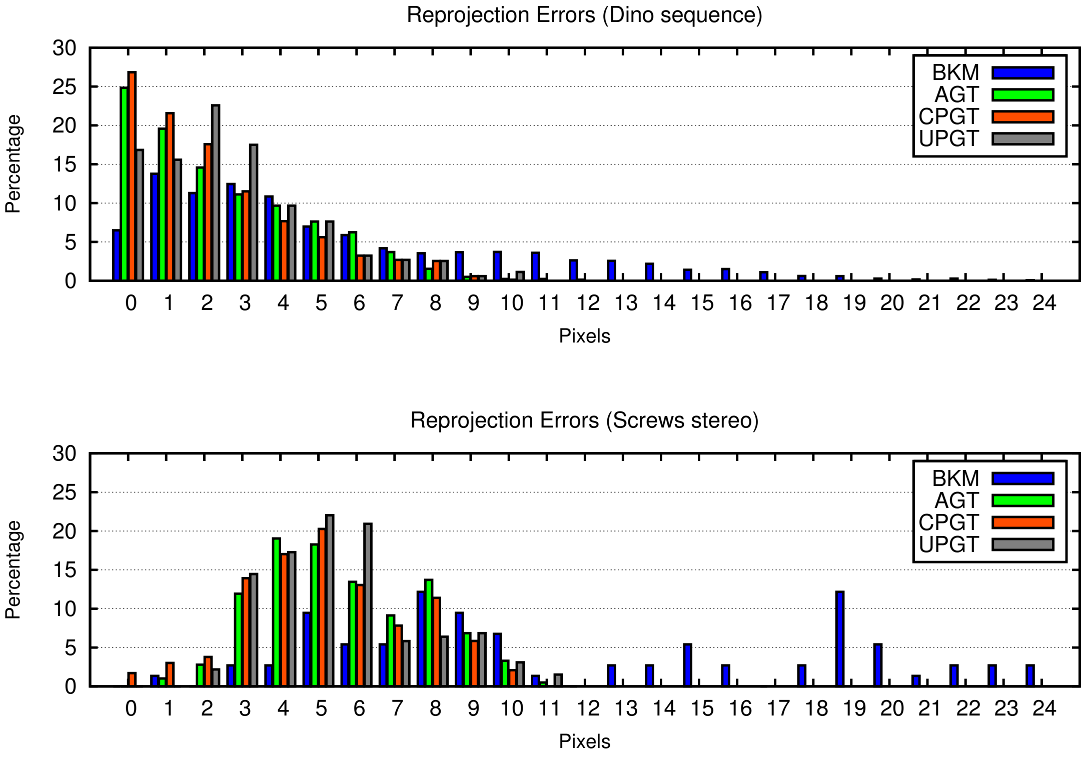
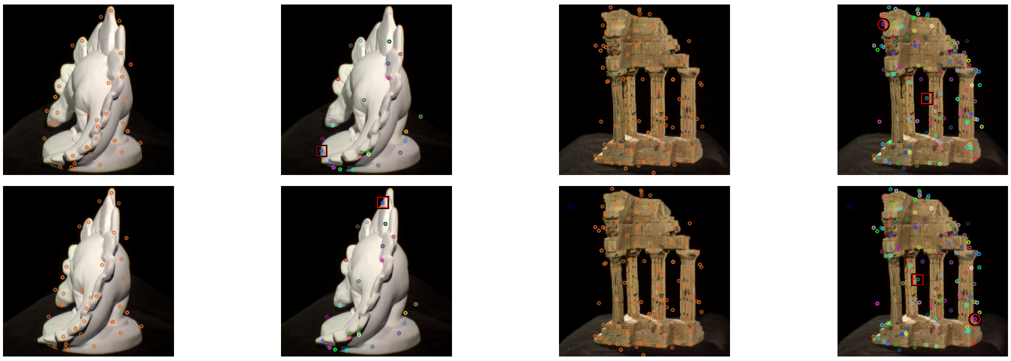


Figure 5.13: Distribution of the reprojection error on one multiple view (top) and one stereo pair (bottom) example.

in groups, each corresponding to a different depth level, and visualized with a unique color in the figure. As can be seen, the different depth levels are properly estimated; this is particularly evident throughout the arched back going from the tail (in foreground) to the head of the model (in background), where clustered sets of feature points follow one after the other. Furthermore, these sets of interest points maintain the right correspondences within the pair of images. The Bundler matcher on the other hand, while still achieving good results in the whole process, also outputs erroneous correspondences (marked in the figure). In Figure 5.10 we can see the results obtained with CPGT and UPGT with $\lambda = 0.3$ and $q = 0.7$ after an affine matching step performed with $\lambda = 0.09$ and $q = 0.9$. We can observe that CPGT gives a significant boost to all the statistics. By contrast UPGT performed worse than AGT (albeit still better than BKM). This is probably due to the higher number of degrees of freedom in the estimation of the fundamental matrix and, thus, to the reduced ability to discriminate incompatible groups. In fact, we can see that the size of the groups obtained with AGT is generally rather small (from 4 to about 10 points), and it is easy to justify such a small number of correspondences under a common fundamental matrix. The quality of reconstruction following the application of all methods can be compared visually by looking at the distribution of the reprojection error in the top row of Figure 5.13. While most reprojections fall within 1-3 pixels for the Game-Theoretic approaches, the Bundler matcher exhibits a long-tailed trend, with repro-



Game-Theoretic (VP) Bundler matcher Game-Theoretic (VP) Bundler matcher

Dino sequence Temple sequence

	Dino sequence		Temple sequence	
	Game-Theoretic (VP)	Bundler matcher	Game-Theoretic (VP)	Bundler matcher
Matches	262.5 ± 61.4	172.4 ± 79.5	535.7 ± 38.7	349.3 ± 36.2
$\Delta\alpha$	0.0668 ± 0.0777	0.0767 ± 0.1172	0.1326 ± 0.0399	0.1414 ± 0.0215
$\Delta\gamma$	0.4393 ± 0.4963	0.6912 ± 0.8793	0.0809 ± 0.0144	0.0850 ± 0.0065

Figure 5.14: Results obtained with the virtual points approach on the Dino and Temple data sets.

jection errors reaching 20 pixels. Unlike the Dino model, the Temple model is quite rich of features: for visualization purposes we only show a subset of the detected matches for all the techniques. While the effectiveness of our approaches is not negatively impacted by the model characteristics, several mismatches are extracted by BKM. In particular, the symmetric parts of the object (mainly the pillars) result in very similar features and this causes the matcher to establish one-to-many correspondences over them. In the calibrated stereo scenario, the Ganesha images are rich of distinctive features and pose no particular difficulty to any of the methods. The Bundler matcher provides very good results, with only one evident false match out of a total of 200 matches (see Figure 5.11). The resulting bundle adjustment is quite accurate, giving very small rotation errors and reprojection distances. Nevertheless, our methods performs considerably better: reprojection errors dramatically decrease, with around 98 percent of the feature points falling below one pixel of reprojection error for AGT and 99 percent for CPGT (Figure 5.12). Unfortunately UPGT is still unable to refine the results obtained with AGT, but still achieves smaller errors than BKM. The second calibrated stereo scene, "Screws stereo", is an emblematic case and provides some meaningful insight. The images depict a dozen of screws standing on a table, placed by hand at different depth levels. This configuration, together with the abundance of features, should provide enough information for the algorithms to extract significant matches. However, the scene is a difficult one due to the very nature of the objects depicted, which are all identical and highly symmetric, resulting in several

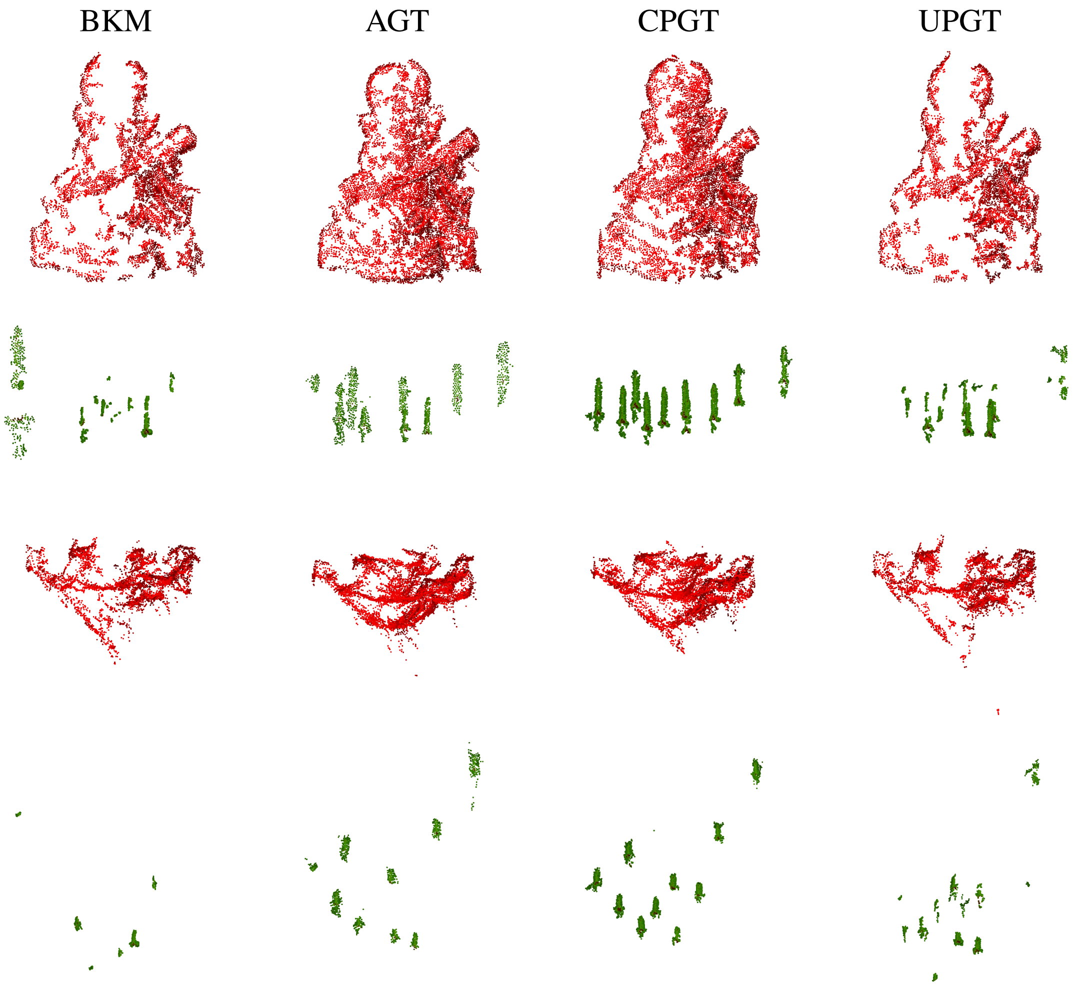


Figure 5.15: Comparisons of the point clouds produced by PMVS using the motion estimated using different matching methods. Respectively the Bundler default keymatcher (BKM), the affine game-theoretic technique (AGT) and the calibrated and uncalibrated projective techniques (CPGT and UPGT).

features with very similar descriptors and a difficulty in extracting good matches based only on photometric information. Indeed, several false matches are established by the Bundler matcher (see the last column of Figure 5.11). Still, BKM results in a reasonable estimation of the rigid transformation linking the two cameras, as erroneous pairings are removed *a posteriori* during the subsequent phases of bundle adjustment. By contrast, the AGT approach outputs large and accurate sets of matches, roughly one per object, and even difficult cases, such as the left-right parallactic swaps taking place at the borders are correctly dealt with. It is interesting to note that in this case the boost given by CPGT is even more significant than in the previous experiments, with a lower average reprojection error and an overall better error distribution. Unlike with the previous cases, this happens by reducing the number of total matches rather than increasing it, as the refinement pro-

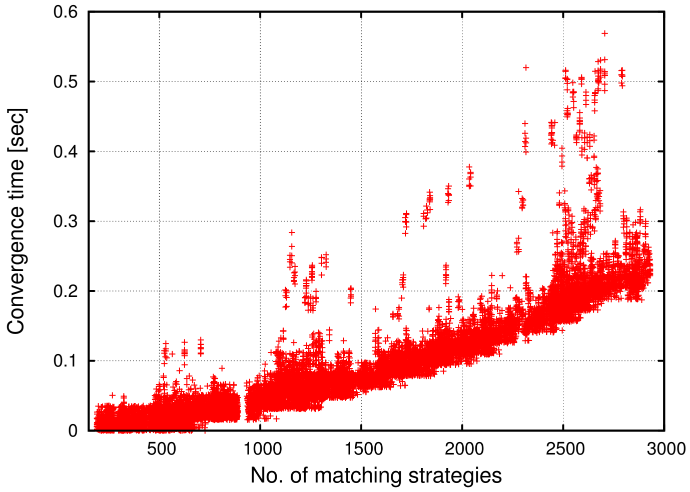


Figure 5.16: Plot of the convergence time of the replicator dynamics with respect to the number of matching strategies.

cess eliminates correspondences that are not globally consistent. In addition this time even the UPGT gives better results than AGT: a histogram of the reprojection errors for this object is shown in Figure 5.13. Finally, a qualitative analysis of the different approaches is shown in Figure 5.15, where the estimated parameters and correspondences are fed to the PMVS [61] dense multiview stereo reconstruction tool. The first and the second rows show the Ganesha and screws scenes from a frontal view, while the other two show a top view of the same scenes. AGT and CPGT give the best results for Ganesha with CPGT providing a denser reconstruction with a more circular halo over the head. With the screws scene CPGT allows by far the more consistent reconstruction, while BKM is substantially unable to offer to PMVS a satisfactory pose estimation.

The final set of experiments analyzes the relative performance of the virtual points approach in relation to the Bundler baseline. In particular we analuzyed the differences in radians between the (calibrated) ground-truth and respectively the estimated rotation angle ($\Delta\alpha$) and rotation axis ($\Delta\gamma$). Figure 5.14 shows the correspondences produced by our method (left column) in comparison with the other matcher (right column). Again, the "Temple" model is richer in features and for visualization purposes we only show a subset of the detected matches for both techniques. Our method, by enforcing global 3D consistency, can effectively disambiguate the matches. Looking at the results we can see that our approach extracts around 50% more correspondences than BKM, providing a slight increase in precision and reduction in variance of the estimates. Note that the selected measures evaluate the quality of the underlying least square estimates of the motion parameters after a reprojection step, thus small variations are expected.

5.3.4 Complexity and Running Time

With respect to the complexity all the game-theoretic approaches are dominated by the steps of the replicator dynamics. Each step is quadratic in the number of strategies, but there is no guarantee about the total number of step that are needed to reach an ESS. We chose to stop the iterations when the variation of the population was below a minimum threshold. Execution times for the matching steps of our technique are plotted in Figure 5.16; the scatter plot shows a weak quadratic growth of convergence time as the number of matching strategies increases with a very small constant in the quadratic term, resulting in computation times below half a second even with a large number of strategies.

5.4 Conclusions

We introduced a game-theoretic technique that performs an accurate feature matching as a preliminary step for multi-view 3D reconstruction using Structure from Motion techniques. Unlike other approaches, we do not rely on a first estimation of scene and camera parameters in order to obtain a robust inlier selection, but rather, we enforce geometric constraints based only on semi-local properties that can be estimated from the images. In particular, we define two selection games and one consolidation game that filters out groups of matches by considering their compliance with the epipolar constraint. The first matching game selects local groups of compatible correspondences enforcing a weak affine camera model, whereas the second game projects what is left of a high-order compatibility problem into a pairwise compatibility measure, by enforcing the conservation of distances between the unknown 3D positions of the points. Experimental comparisons with a widely used technique show the ability of our approach to obtain a tighter inlier selection and thus a more accurate estimation of the scene parameters.

6
Surface Registration and Multiview Error Diffusion

Surface alignment (also commonly found as *surface registration* in literature) is a fundamental step in many 3D computer vision tasks such as 3D reconstruction and shape analysis, and is generally approached with two-step techniques that first aim at obtaining an approximate ("coarse") alignment of the surfaces, then refine the relative motion via error-minimizing procedures. Since most reconstruction pipelines operate in a pairwise manner, in this Section we will first concentrate on the problem of (rigidly) aligning two surfaces with one another. Then, in Section 6.4, we will give a method to compose these pairwise registrations together in an attempt to globally optimize on the registration error of all the views simultaneously.

Most coarse registration algorithms exploit local point descriptors that are intrinsic to the shape and do not depend on the relative position of the surfaces. On the other hand, refinement techniques iteratively minimize a distance function measured between pairs of selected neighboring points and are thus strongly dependent on initial alignment. In this work we propose a novel technique that allows to obtain an accurate surface registration in a single step, without the need for an initial motion estimation. Following the previous chapters, we cast the selection of correspondences between points on the surfaces in a game-theoretic framework, where a natural selection process allows mating points that satisfy a mutual rigidity constraint to thrive, eliminating all the other correspondences. This process yields a very robust inlier selection scheme that does not depend on any particular technique for selecting the initial strategies as it relies only on the global geometric compatibility between correspondences. The practical effectiveness of the proposed approach is confirmed by an extensive set of experiments and comparisons with state-of-the-art techniques.

6.1 Surface Alignment Through an Isometry-Enforcing Game

We follow a similar approach to the one presented in Chapter 5; specifically, after defining an appropriate set of strategies (candidate matches), we let them compete with one

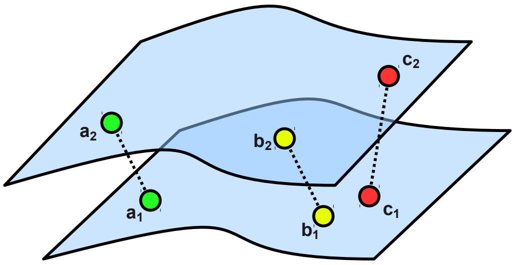


Figure 6.1: Example of three matching strategies.

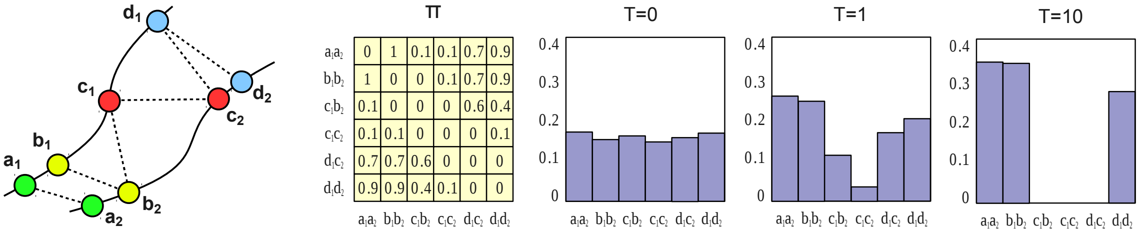
another, each obtaining support from compatible associations and competitive pressure from all the others. At the equilibrium, only pairings that are mutually compatible should survive and are then taken to be inliers. In practice, the full process happens through the following steps: First, we extract a set of candidate matches; then, we define a payoff function between such candidates; finally, we use evolutionary dynamics to evolve towards an equilibrium state. These steps will be detailed in the following sections.

6.1.1 Matches as Strategies

Since we will deal with the registration of two different surfaces we will refer to the points belonging to the first surface with the term *model points*, while we will use the term *data points* with respect to the second surface. This distinction is captious since there is no actual difference in role between the two surfaces, however it is consistent with the current registration literature and helps in defining an order within matches.

Here we follow the same notation of Section 2.6 and define M to be set of model points, D the set of data points, and S the set of available matching strategies. Our goal is of course to extract from S the subset of correct matches, that is, strategies that associate a point in the model surface with the same point in the data surface. Since in this context we are dealing with rigid alignment of surfaces, it is quite natural to exploit the rigidity constraint to measure the feasibility of a pair of matches. In fact (Section 6.2), we relax the rigidity assumption to an isometry assumption, assigning a high payoff to pairs of matching strategies that preserve the Euclidean distance between the corresponding points on model and data (see Figure 6.2). Also, we will assign a payoff equal to zero to pairs that share the same source or destination point, so as to enforce a one-to-one matching.

In order to reduce the number of the initial candidates, we subsample model points by keeping only points that are deemed to be interesting, and we use a semi-local surface descriptor (Section 6.1.2) to assign to each of them a small set of feasible matches. This is done simply by choosing the mates with the nearest descriptor in the Euclidean sense. The amount of model subsampling, the level of distinctiveness required, and the number of matching candidates to select for each model point are parameters of the method and can be modulated to balance speed and accuracy. Of course, the distinctiveness of the descriptor used to characterize the data points has a big influence in fixing a reasonable



	a_1a_2	b_1b_2	c_1b_2	c_1c_2	d_1c_2	d_1d_2
a_1a_2	0	1	0.1	0.1	0.7	0.9
b_1b_2	1	0	0	0.1	0.7	0.9
c_1b_2	0.1	0	0	0	0.6	0.4
c_1c_2	0.1	0.1	0	0	0	0.1
d_1c_2	0.7	0.7	0.6	0	0	0
d_1d_2	0.9	0.9	0.4	0.1	0	0

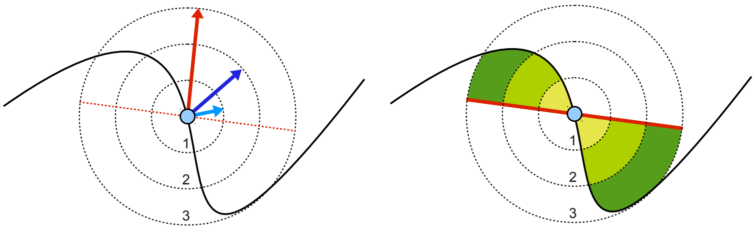
Figure 6.2: An example of the evolutionary process. Four points are sampled from the two surfaces and a total of six mating strategies are selected as initial hypotheses. Matrix Π shows the compatibilities between pairs of mating strategies according to a one-to-one rigidity-enforcing payoff function. Each mating strategy got zero payoff with itself and with strategies that share the same source or destination point (e.g., $\pi((b_1, b_2), (c_1, b_2)) = 0$). Strategies that are coherent with respect to a rigid transformation exhibit high payoff values (e.g., $\pi((a_1, a_2), (b_1, b_2)) = 1$ and $\pi((a_1, a_2), (d_1, d_2)) = 0.9)$), while less compatible pairs get lower scores (e.g., $\pi((a_1, a_2), (c_1, c_2)) = 0.1$). Initially (at T=0) the population is set to the barycenter of the simplex and slightly perturbed (3-5%). After just one iteration, (c_1, b_2) and (c_1, c_2) have lost a significant amount of support, while (d_1, c_2) and (d_1, d_2) are still played by a sizable amount of population. After ten iterations (T=10), (d_1, d_2) has finally prevailed over (d_1, c_2) (note that the two are mutually exclusive). Note that in the final population $(a_1, a_2), (b_1, b_2)$ have a larger support than (d_1, d_2) since they are a little more coherent with respect to rigidity.

number of candidates for each match. This observation, in turn, raises the quandary between the repeatability and the distinctiveness of feature descriptors. In fact, while a high distinctiveness is always desirable, this often comes at the price of much more instability with respect to noise and thus can lead to a poor repeatability. Nevertheless, with our approach, the descriptor itself is only used to construct the set S and has no role in the evolutionary selection, which is purely driven by the payoff function. For this reason, we find it reasonable to resort to a feature characterization that is scarcely distinctive and to allow for several candidate matches, letting the game-theoretic selection process to operate a severe culling. The loose descriptor that we are introducing has many other advantages, such as being easy to implement and fast to compute, and it will be described in full depth in Section 6.1.2.

In Figure 6.2 we show a complete example of the matching process. We refer to Section 2.6 for details. While the example is kept simple on purpose and the data does not come from real surfaces the illustrated evolution is computed exactly with the payoff matrix Π using equation 2.2.

6.1.2 Surface Hashes

Since the proposed framework relies on geometric consistency, we choose to adopt a very loose feature descriptor that enhances the probability for a feature point to be repeatable,

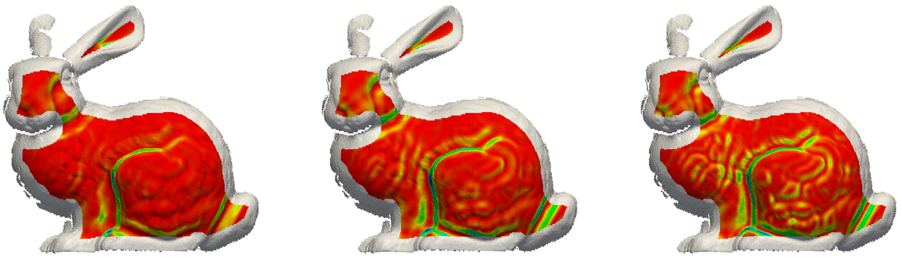


(a) Normal Hash (b) Integral Hash

Figure 6.3: Example of the two basic Surface Hashes.

albeit allowing a much higher number of outliers to get into the set of initially proposed matches. In this context, in order to avoid feature points that carry little useful information for registration purposes (such as flat areas or regions of constant curvature), a minimal matching game as the one described in Chapter 4 is also carried out among the features associated to the model points.

Intuitively, a Surface Hash is a concise point feature descriptor that exhibits the property of being highly repeatable at the cost of a relatively high probability of clashing. In practice this happens with any low-dimensional descriptor, such as the Gaussian or Mean Curvature (1 dimension), the first two Principal Components of a patch (2 dimensions), or the normal vector associated to a point (2 dimensions). While those descriptors could be used with our registration pipeline, we prefer to introduce two multiscale Surface Hashes based respectively on the dot product between normals and a local surface integral. Each of our descriptors corresponds to a vector of scalar measures evaluated at different scales. By increasing or reducing the number of scales, we are able to obtain vectors of different length, thus being more or less distinctive. The *Normal Hash* (Figure 6.3(a)) is obtained by setting as a reference the average surface normal over a patch that extends to the largest scale (red arrow in figure) and then, for each smaller scale, calculate the dot product between the reference and the average normal over the reduced patches (blue arrows in figure). This measure finds its rationale in the observation that at the largest scale the average normal is more stable with respect to noise and that the dot product offers a concise representation of the relation between the vectors obtained at various scales. The *Integral Hash* is similar in spirit to the Normal Hash (see Figure 6.3(b)). In this case, we search for the best fitting plane (in the least squares sense) with respect to the surface patch associated to the largest scale. Then we calculate the volume enclosed between the surface and such a plane. In practice, it is not necessary to evaluate this volume accurately: even naive approximations, such as the sum of the distances of the surface points from the plane, have shown to provide a reasonable approximation in all the empirical tests. Note that Normal Hashes evaluated over n scales yield descriptor vectors of length $n - 1$ (since the larger scale is used only to calculate the reference normal), while Integral Hashes provide n-dimensional vectors. In Figure 6.4 a Normal Hash of dimension 3 (respectively from (a) to (c)) evaluated over 4 scales is shown. Note that the descriptor is not defined on the points for which the larger support is not fully contained in the surface,



(a) First dimension (b) Second dimension (c) Third dimension

Figure 6.4: Example of a 3-dimensional Normal Hash.

i.e., points close to the surface boundary.

6.2 Isometry-Enforcing Payoff

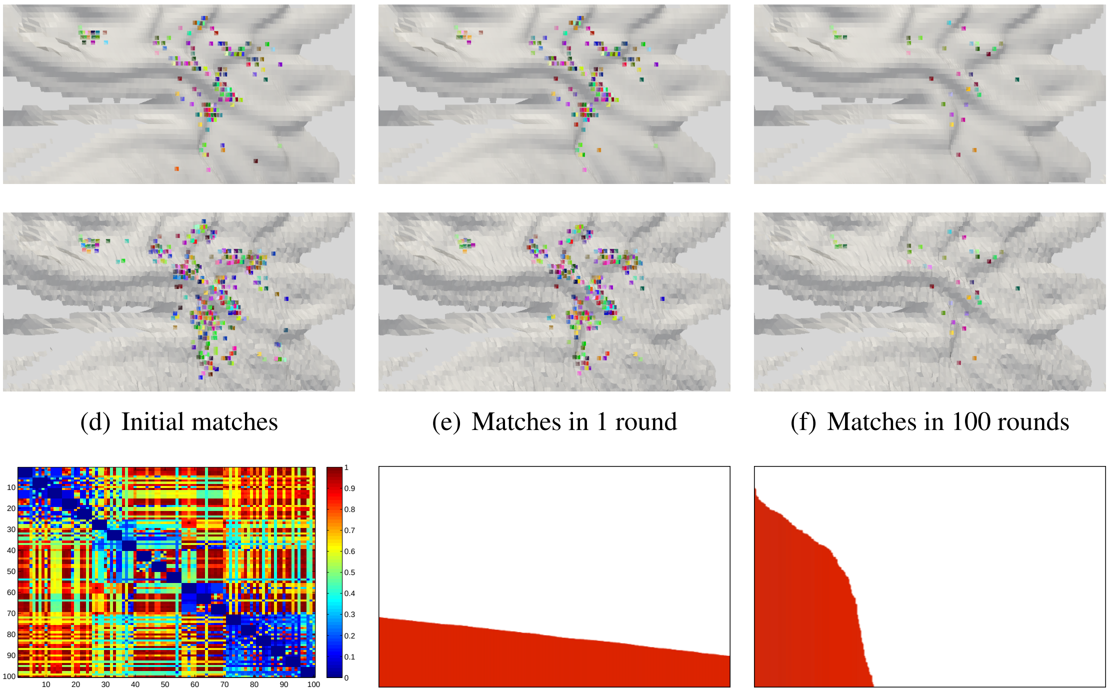
As already stated, for this particular application of the GTM framework we decided to assign to each pair of matching strategies a payoff that is inversely proportional to a measure of violation of the surface rigidity constraint. This violation can be expressed in several ways, but since all the rigid transformations preserve Euclidean distances, we choose this property to express the coherence between matching strategies. Clearly this isometry constraint is looser than the rigidity constraint as it cannot prevent specular flips of the surfaces, but the global consistency provided by the game-theoretic framework ensures that only rigid alignments will prevail.

Definition 1. *Given a function $\pi : S \times S \to \mathbb{R}^+$, we call it an* isometry-enforcing payoff function *if for any $((a_1, a_2), (b_1, b_2))$ and $((c_1, c_2), (d_1, d_2)) \in S \times S$ we have that $||a_1 - b_1| - |a_2 - b_2|| > ||c_1 - d_1| - |c_2 - d_2||$ implies $\pi((a_1, a_2), (b_1, b_2)) < \pi((c_1, c_2), (d_1, d_2))$. In addition, if $\pi((a_1, a_2), (b_1, b_2)) = \pi((b_1, b_2), (a_1, a_2))$, π is said to be* symmetric.

An isometry-enforcing payoff function is a function that is monotonically decreasing with the absolute difference of the Euclidean distances between respectively the model and data points of the matching strategies compared. In other words, given two matching strategies, their payoff should be high if the distance between the model points is equal to the distance between the data points and it should decrease as the difference between such distances increases. In the example of Figure 6.1, matching strategies (a_1, a_2) and (b_1, b_2) are coherent with respect to the rigidity constraint, whereas (b_1, b_2) and (c_1, c_2) are not, thus it is expected that $\pi((a_1, a_2), (b_1, b_2)) > \pi((b_1, b_2), (c_1, c_2))$.

Further, if we want mating to be one-to-one, we must put an additional constraint on the payoffs, namely that mates sharing a point are incompatible.

Definition 2. *An isometry-enforcing payoff function π is said to be* one-to-one *if $a_1 = b_1$ or $a_2 = b_2$ implies $\pi((a_1, a_2), (b_1, b_2)) = 0$.*



(d) Initial matches (e) Matches in 1 round (f) Matches in 100 rounds

(g) Payoff matrix (h) Population in 1 round (i) Population in 100 rounds

Figure 6.5: Example of a rigidity enforcing payoff and of the evolution of the matching process.

Given a set of matching strategies S and an enumeration $O = \{1, ..., |S|\}$ over it, a *matching game* is a non-cooperative game where the population is defined as a vector $\mathbf{x} \in \Delta^{|S|}$ and the payoff matrix $\Pi = (\pi_{ij})$ is defined as $\pi_{ij} = \pi(s_i, s_j)$, where $s_i, s_j \in S$ are enumerated by O and π is a symmetric one-to-one isometry-enforcing payoff function. Intuitively, $\mathbf{x}_i$ accounts for the percentage of the population that plays the i-th matching strategy. By using a symmetric one-to-one payoff function in a matching game we are guaranteed that ESS's will not include mates sharing either model or data nodes. In fact, given a non-negative payoff function, a stable state cannot have in its support pairs of strategies with payoff 0 [21]. Moreover, a matching game exhibits some additional interesting properties.

Theorem 1. *Given a set of model points M, a set of data points $D = TM$ that are exact rigid transformations of the points in M, a set of matching strategies $S \subseteq M \times D$ with $(m, Tm) \in S$ for all $m \in M$, and a matching game over them with a payoff function π, the vector $\hat{x} \in \Delta^{|S|}$ defined as*

$$\hat{x}_i = \begin{cases} 1/|M| & \textit{if } s_i = (m, Tm) \textit{ for some } m \in M\,; \\ 0 & \textit{otherwise,} \end{cases}$$

is an ESS and obtains the global maximum average payoff.

Proof. Let $\hat{S} \subseteq S$ be the set of mates that match a point to its copy, clearly for all $s, q \in \hat{S}, s \neq q$ we have $\pi(s, q) = 1$, while for $s \in \hat{S}$ and $q \in S \setminus \hat{S}$, we have $\pi(s, q) < 1$. For all $s \in \hat{S}$ we have that $\pi(\hat{x}, \hat{x}) = \frac{|M|-1}{|M|}$ while, since π is one-to-one, for any $q \in S \setminus \hat{S}$ there must be at least one $s_q \in \hat{S}$ with $\pi(q, s_q) = 0$, thus $\pi(q, \hat{x}) < \frac{|M|-1}{|M|}$, thus $\hat{x}$ is a Nash equilibrium. Further, since the inequality is strict, it is an ESS. Finally, $\hat{x}$ is a global maximizer of π since $\frac{|M|-1}{|M|}$ is the maximum value that a one-to-one normalized payoff function over $|M|$ points can attain. □

This theorem states that when matching a surface with a rigidly transformed copy of itself the optimal solution (i.e., the population configuration that selects all the matching strategies assigning each point to its copy) is the stable state of maximum payoff. Since well established algorithms to evolve a population to such a state exist, this provides us with an effective mating approach. Clearly, aligning a surface to an identical copy is not very useful in practical scenarios, where occlusion and measurement noise come into play. While the quality of the solution in presence of noise will be assessed experimentally, we can give some theoretical results regarding occlusions.

Theorem 2. *Let M be a set of points with $M_a \subseteq M$ and $D = TM_b$ a rigid transformation of $M_b \subseteq M$ such that $|M_a \cap M_b| \geq 3$, and $S \subseteq M_a \times D$ be a set of matching strategies over M_a and D with $(m, Tm) \in S$ for all $m \in M_a \cap M_b$. Further, assume that the points that are not in the overlap, that is the points in $E_a = M_a \backslash (M_a \cap M_b)$ and $E_b = M_b \backslash (M_a \cap M_b)$, are sufficiently far away such that for every $s \in S, s = (m, Tm)$ with $m \in M_a \cap M_b$ and every $q \in S, q = (m_a, Tm_b)$ with $m_a \in E_a$ and $m_b \in E_b$, we have $\pi(q, s) < \frac{|M_a \cap M_b|-1}{|M_a \cap M_b|}$, then, the vector $\hat{x} \in \Delta^{|S|}$ defined as*

$$\hat{x}_i = \begin{cases} 1/|M| & \text{if } s_i = (m, Tm) \text{ for some } m \in M_a \cap M_b\,; \\ 0 & \text{otherwise,} \end{cases}$$

is an ESS.

Proof. We have $\pi(\hat{x}, \hat{x}) = \frac{|M_a \cap M_b|-1}{|M_a \cap M_b|}$. Let $q \in S$ be a strategy not in the support of $\hat{x}$, then, either it maps a point in M_a or M_b, thus receiving payoff $\pi(q, \hat{x}) < \frac{|M_a \cap M_b|-1}{|M_a \cap M_b|}$ because of the one-to-one condition, or it maps a point in E_a to a point in E_b, receiving, by hypothesis, a payoff $\pi(q, \hat{x}) < \frac{|M_a \cap M_b|-1}{|M_a \cap M_b|}$. Hence, $\hat{x}$ is an ESS. □

The result of theorem 2 is slightly weaker than theorem 1, as the face of the simplex corresponding to the "correct" overlap, while being an evolutionary stable state, is not guaranteed to obtain the overall highest average payoff. This is not a limitation of the framework as this weakening is actually due to the very nature of the alignment problem itself. The inability to guarantee the maximality of the average payoff is due to the fact that the original object (M) could contain large areas outside the overlapping subset that are perfectly identical. Further, objects that are able to slide (for instance a plane or a sphere) could allow to move between different mixed strategies without penalty. These situations cannot be addressed by any algorithm without relying on supplementary information. However, in practice, they are quite unlikely, exceptional cases. In the experimental section we will show that our approach can effectively register a wide range of surface types.

In theory, any rigidity-enforcing payoff function can be used to perform surface registration. Throughout the experimental section we adopted:

$$\pi((a_1, b_1), (a_2, b_2)) = \left(\frac{min(|a_1 - a_2|, |b_1 - b_2|)}{max(|a_1 - a_2|, |b_1 - b_2|)}\right)^{\lambda} \tag{6.1}$$

where a_1, a_2, b_1 and b_2 are respectively the two model (source) and data (destination) points in the compared matching strategies. Parameter λ allows to make the enforcement of the Euclidean distance more or less strict.

In Figure 6.5 we show a complete example of the evolutionary matching process. In order to make the example easy to understand we restricted our focus to a detail of a range scan of the Stanford dragon [51]. In this example (and throughout all the experimental section) S is built by including all the strategy pairs composed by a feature point in the model and the 5 nearest feature points in the data in terms of Surface Hash (in this example we used an Integral Hash with 3 scales). In Figure 6.5(g) we show, on a colored scale from 0 to 1, the payoff matrix of the rigidity enforcing function 6.1. Note that in the diagonal area of the matrix blocks of five strategies with reciprocal 0 payoff can be found: this is related to the way we built S. In fact we chose to include for each model point 5 candidates in the data which are mutually non compatible as they share the same source point and we are looking for a one-to-one match. In the first and second row of Figure 6.5(d) we can see respectively model and data feature points at the beginning of the matching process. After just one round of replicator dynamics we see that many outliers have been eliminated from the initial set S, but still some wrong matches are present. After 100 iterations only a few matches are retained, but it is easy to see that they are extremely coherent. Finally, in Figure 6.5(h) and Figure 6.5(i) we show the (sorted) population histogram respectively after 1 and 100 iterations. The first histogram shows that all the strategies are still played by a sizeable amount of the population, while after 100 iterations most of the consensus is held by the few surviving matches.

6.3 Experimental Results

We introduced a Game-Theoretic Registration approach (GTR) that is based both on a feature detector/descriptor and on a matching technique. To better explore the role of both, we designed a wide range of experimental validations. First, we analyzed the sensitivity of the descriptor to several sources of noise and the influence of the number of scales (and thus of the size of the descriptor vector). Further, we studied the sensitivity of the matching algorithm to its parameters, with the goal of identifying an optimal

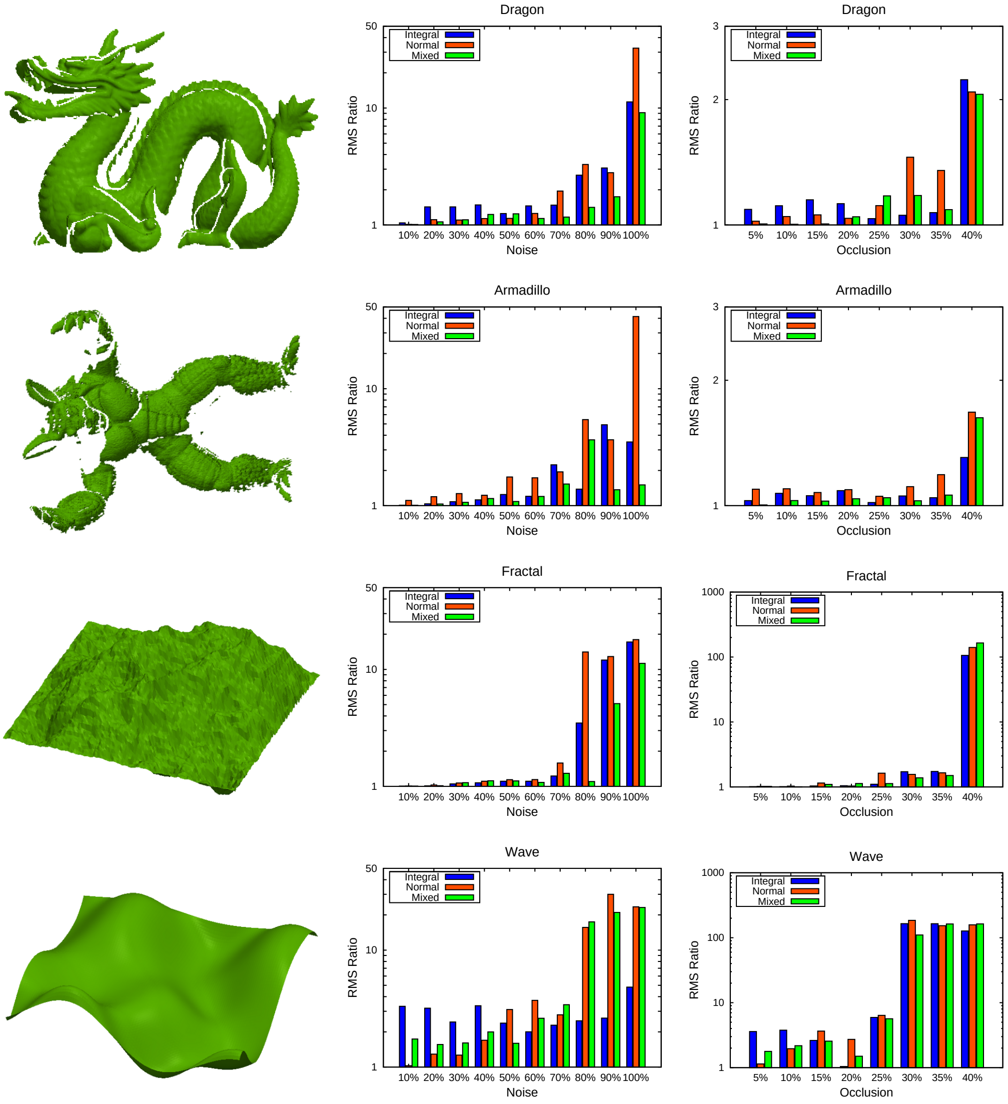


Figure 6.6: Comparison of different descriptors using real and synthetic objects.

parameterization (if any) and assess the stability of the method. Also a number of comparative tests were made. Specifically, we analyzed the performance obtained by using our matcher with different feature detectors and the overall comparison with respect to other well-know registration pipelines.

All the experiments were performed on a modern personal computer equipped with a Core i7 Intel processor and 8 GB of memory. The dataset used, where not differently stated, was built upon publicly available models; specifically the Bunny [150], the Armadillo [84] and the Dragon [51] from the Stanford 3D scanning repository. To further assess the shortcomings of the various approaches, we used two synthetic surfaces representative of as many difficult classes of objects: a wave surface and a fractal landscape (see Figure 6.6). Since a ground truth was needed for an accurate quantitative comparison,

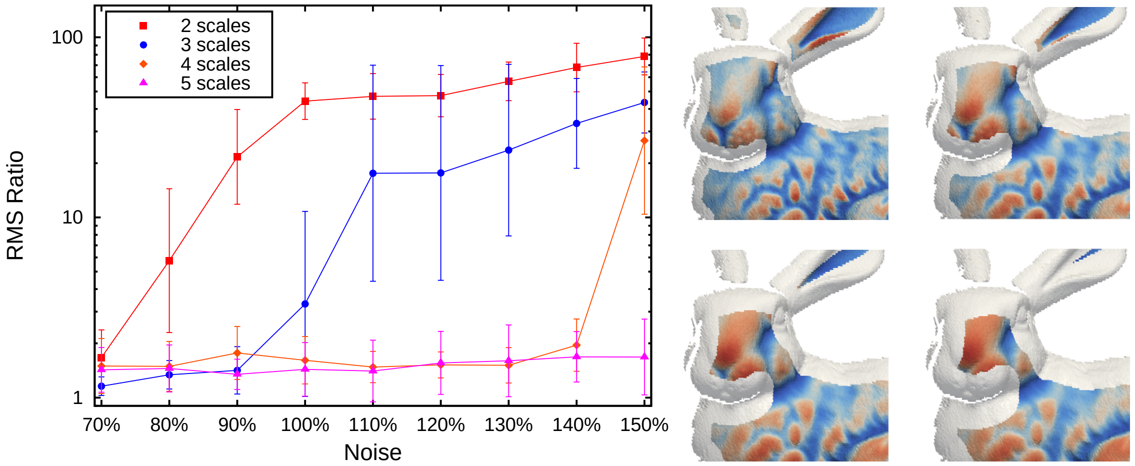


Figure 6.7: Effect of scale on the matching accuracy.

we generated virtual range images from the models and then applied additive Gaussian noise to them. All the registration experiments adopted the payoff function 6.1 with the additional one-to-one constraint. The descriptor used is a mixed Surface Hash with 3 scales.

6.3.1 Sensitivity Analysis of the Descriptor

The performance of different descriptors was tested for various levels of noise and occlusion applied to two surfaces obtained from real range scans ("armadillo" and "dragon" from Stanford) and two synthetic surfaces designed to be challenging for coarse and fine registration techniques ("fractal" and "wave"). The noise is a positional Gaussian perturbation on the point coordinates with its level (σ) expressed in terms of the percentage of the average edge length, while occlusion denotes the percentage of data and model surfaces removed. The RMS Ratio in the charts is the ratio of the root mean square error (RMS) obtained after registration and the RMS of ground truth alignment. The Normal and Integral Hashes were calculated over 3 levels of scale and the "Mixed" Hash is simply the juxtaposition of the previous two.

In Figure 6.6 we see that all the descriptors obtain good results with real range images and the registration "breaks" only with very high levels of noise (on the same order of magnitude of the edge length). It is interesting to observe that the Mixed Hash always obtains the best performance, even with high level of noise: This higher robustness is probably due to the orthogonality between the Normal and Integral Hashes. The behavior with the "fractal" synthetic surface is quite similar, by contrast all the descriptors seem to perform less well with the "wave" surface. This is due to the lack of distinctive features on the model itself, which indeed represents a challenge for any feature based registration technique [117]. The performance obtained with respect to occlusion is similar: all the descriptors achieve fairly good results and are resilient to high levels of occlusion (note that 40 percent occlusion is applied both to data and model). Overall the Mixed Hash appears to be consistently more robust. Since we found that the descriptors calculated over 3 levels of scale break at a certain level of noise, we were interested in evaluating if their performance can be improved by increasing their dimension.

In Figure 6.7 we present the results obtained with different levels of scale for the Mixed Hash. The graphs show the average over all the surfaces and the associated RMS. It is interesting to observe that by reducing the scale level the technique becomes less robust, whereas its performance increases dramatically when the number of scales increases. With a scale level of 5 our approach can deal even with surfaces subject to Gaussian positional noise of σ greater than the edge length. Unfortunately, this enhanced reliability comes with a drawback: by using larger levels of scale the portion of boundary that cannot be characterized grows. In the right half of Figure 6.7 the shrinking effect is shown for scale levels from 2 to 5.

6.3.2 Sensitivity to Parameters of the Matcher

The game-theoretic matching technique presented basically depends on four parameters:

- The number of points sampled from the model object;
- The number of neighbors considered when building the initial set of candidates;
- The selectivity λ for the rigidity-enforcing payoff 6.1;
- The quality threshold used to deem a strategy as non-extincted upon convergence.

The first two parameters are related to the building of the set of strategies S. From a performance point of view, adopting a strong subsampling will produce a smaller set of strategies and a smaller payoff matrix, thus each recurrence of the evolutionary dynamics will be faster. However, a too sparse sampling can lead to groups of mutually compatible matches that are too small and are unable to thrive during the evolutionary process. In a similar manner, a small number of neighbors will reduce the number of strategies. However, this will give fewer chances of capturing the correct pairings since such tightening would require the descriptor to always give a high rank to the correct pairing. In Figure 6.8 it can be seen that optimal results can be achieved with less than 1000 samples and that there is virtually no gain in using more than 6 neighbors. Later in this chapter we will show that on the test system used these conditions allow to perform an alignment in less than one second.

The third parameter (λ) is related to the level of strictness with respect to the enforcement of the rigidity constraint: Higher values for λ will make the payoff function more steep, thus making the selection process more picky. By contrast, lowering λ will yield a payoff matrix with smaller variance, up to the limit value of 0, when the matrix assumes value 1.0 for all the strategies pairs that do not break the one-to-one constraint and 0 otherwise. As expected, our experiments show that very low or very high values for λ deliver

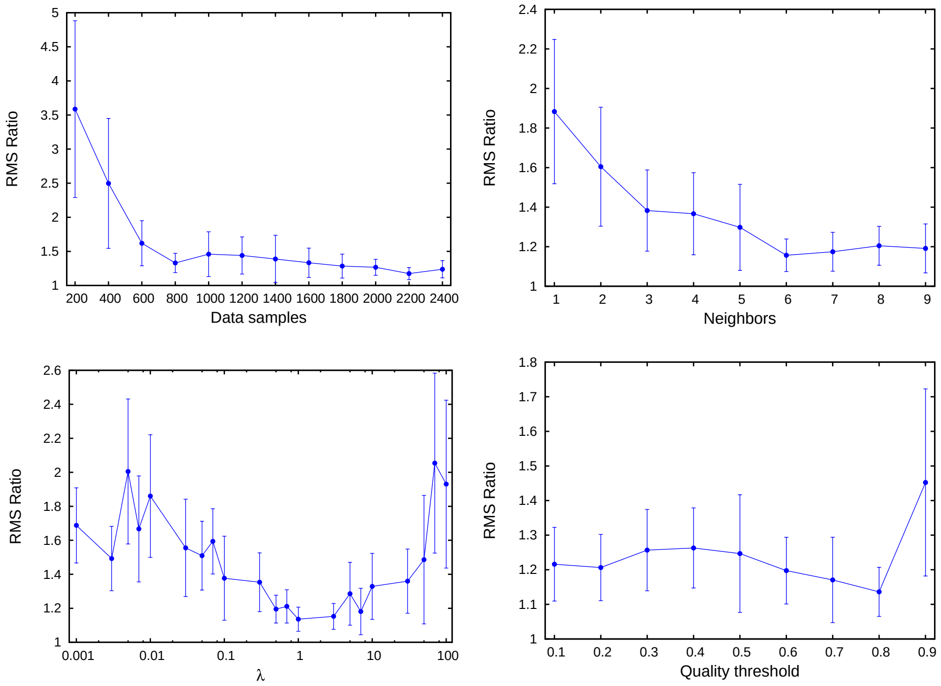


Figure 6.8: Analysis of the sensitivity of the Game-Theoretic Matcher with respect to the parameters of the algorithm.

poor results and, while there is clearly a larger variance than what has been captured by the experiments, the optimal value seems to be around 1.

Finally, the fourth parameter sets the ratio (with respect to the most successful match) used to classify a strategy as surviving or extinct. The last experiment of Figure 6.8 shows that all the tested values below 0.8 give similarly good results. This simply means that there is good separability between extinct and non-extinct strategies, the former being very close to 0.

Overall, we can assess that the matching method has a very limited dependency on its parameters, which can be easily fixed at values that are both safe and efficient. The most influent parameter is probably λ, however a value of 1.0 (that indeed simplifies equation 6.1 to a simple ratio) appears to be optimal for our test set.

6.3.3 Comparison with Full Pipelines

The whole registration algorithm we introduced can be classified as a coarse method, since it does not require initialization. For this reason we compared it with several other coarse techniques. Specifically, we implemented the whole Spin Images pipeline [76] and used the implementation supplied by the authors respectively for the MeshHOG/MeshDOG [160] and the Four Points Congruent Sets [19] methods. The latter method was initialized both with the parameters suggested by the authors and also with values for t and s that we

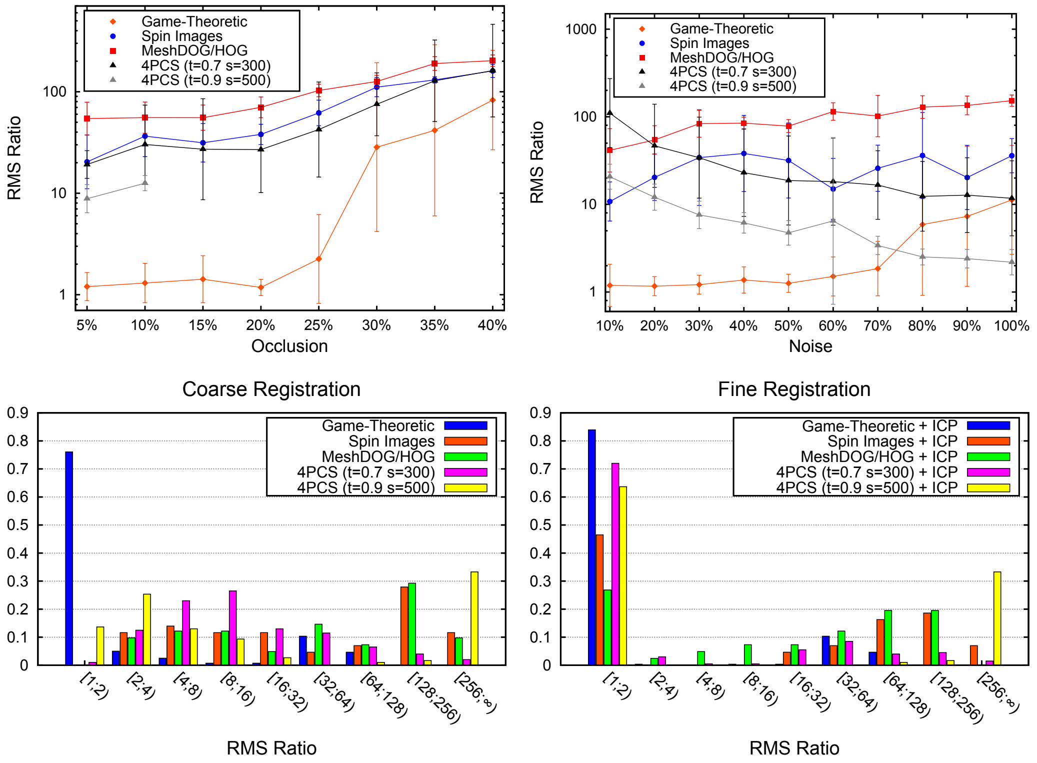


Figure 6.9: Comparisons between our Game-Theoretic Registration technique and other widely used surface registration pipelines.

manually optimized to get the best possible results from our dataset.

In the first row of Figure 6.9 we present the results of this comparison. In these experiments the occlusion is measured with respect to each range image and is applied in opposite directions of the overlapped area. That means that with an occlusion of 10% the actual ovelap is reduced by 20%. The noise is an additive Gaussian noise with a standard error expressed as a percentage over the average edge length. The occlusion test has been made with noise at level 10% and the noise test was performed with no occlusion. From the tests our method exhibits better results in both scenarios and breaks only with high levels of occlusion and noise. Note that the 4PCS method with parameters $t = 0.9$ and $s = 500$ does not always give a feasible solution with any occlusion greater than 10%. With extreme levels of noise the 4PCS seems to get better and obtains lower RMS ratios than our method. The reduction in performance of our method is related to the breaking of the descriptors, that at such high levels of noise do not carry sufficient information any more. A clarification should finally be made about the apparent improvement that 4PCS seems to exhibit as noise increases. In fact, at high noise levels the RMS associated to ground-truth motion is also high. In such conditions the additional error due to misalignment becomes less relevant in terms of contribution to the overall RMS ratio, which is dominated by random noise. Since 4PCS explores thoroughly the set of feasible motions until a solution with RMS low enough is found (depending on the stop criteria), it is expected to test more

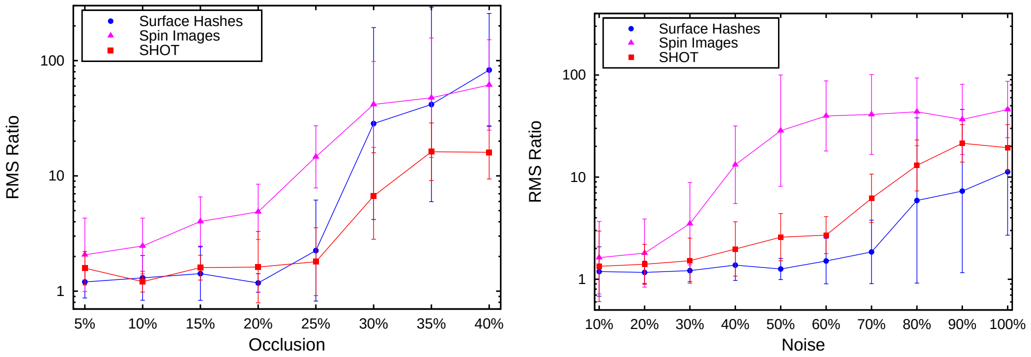


Figure 6.10: Performance of the Game-Theoretic Registration method using feature descriptors different than Surface Hashes.

alignments when surfaces are noisier and thus yield lower RMS ratio values. However, it is easy to build simple examples where a solution can obtain a low RMS ratio (even lower than one) and still being far from the correct alignment.

These results only indicate that GTR gives a better coarse registration, however to seek a perfectly fair comparison it is also needed to measure how much enhancement can be obtained by performing a fine registration step starting from the obtained coarse initialization. To this end we applied the ICP algorithm starting from the initial motion estimated with the different methods with no occlusion and random noise values below 60%. The results are shown in the bottom row of Figure 6.9 with histograms obtained by binning the distance between model points and data surface along the normal vector. Normals that do not intersect the data surface are discarded. The size of the bins grows exponentially. The first histogram shows the distribution obtained from the coarse registration and the second reports the enhancement obtained by applying ICP. Again, the results are favorable to our method, with very few points exhibiting large errors after refinement.

6.3.4 Influence of different Feature Descriptors

In principle, there is nothing that binds the proposed method to the Surface Hashes descriptors. Actually, the game-theoretic step does not use the descriptor at all and any other interest point characterization could be used as a drop-in replacement in order to build the initial matching strategies. To show the generality of the technique and to investigate the robustness of GTR we swapped the Surface Hashes descriptor with a dense variation of Spin Images [40] and the more recent SHOT 3D feature [143].

In Figure 6.10 the results obtained performing the same experiments designed to compare different methods are shown. It is apparent that Spin Images do not work very well with our method. By contrast, the SHOT descriptor behaves well with respect to high noise levels and is even a little more tolerant to occlusion than Surface Hashes. This is mainly due to the large support needed by the latter. On the other hand, Surface Hashes offer better alignment under noise, and are also much faster to compute.

6.3.5 Quality of Fine Registration

In addition to the full pipeline comparisons we also investigated how reliable the proposed approach would be if directly used as a fine registration technique. The goal of this test is two-fold: we want to evaluate our quality as a complete alignment tool and, at the same time, find the breaking point of traditional fine registration techniques.

The method we used for comparison is a best-of-breed ICP variant, similar to the one proposed in [150]. Point selection is based on Normal Space Sampling [117], and point-surface normal shooting is adopted for finding correspondences; distant mates or candidates with back-facing normals are rejected. To minimize the influence of incorrect normal estimates, matings established on the boundary of the mesh are also removed. The resulting pairings are weighted with a coefficient based on compatibility of normals, and finally a 5%-trimming is used. Each test was performed by applying a random rotation and translation to different range images selected from the Stanford 3D scanning repository. Additionally, each range image was perturbed with a constant level of Gaussian noise with standard deviation equal to 12% of the average edge length. We completed 100 independent tests and for each of them we measured the initial RMS error between the ground-truth corresponding points and the resulting error after performing a full round of ICP (ICP) and a single run of our registration method (GTR). In addition, we applied a step of ICP to the registration obtained with our method (GTR + ICP) in order to assess how much the solution extracted using our approach was further refinable.

A scatter plot of the obtained errors before and after registration is shown in Figure 6.11. The final error i on a log scale, so the dotted curve represent the points with identical initial and final error. We observe that ICP reaches its breaking point quite early; in fact with an initial error above the threshold of about 20mm it is unable to find a cor-

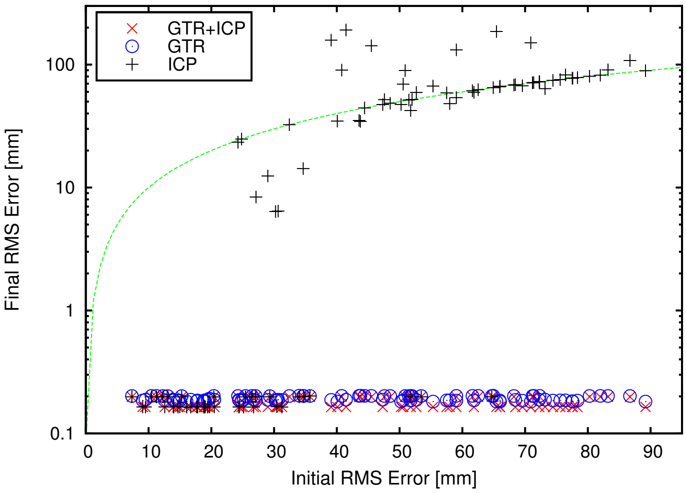


Figure 6.11: Comparison of fine registration accuracies (the green dashed line represents y=x).

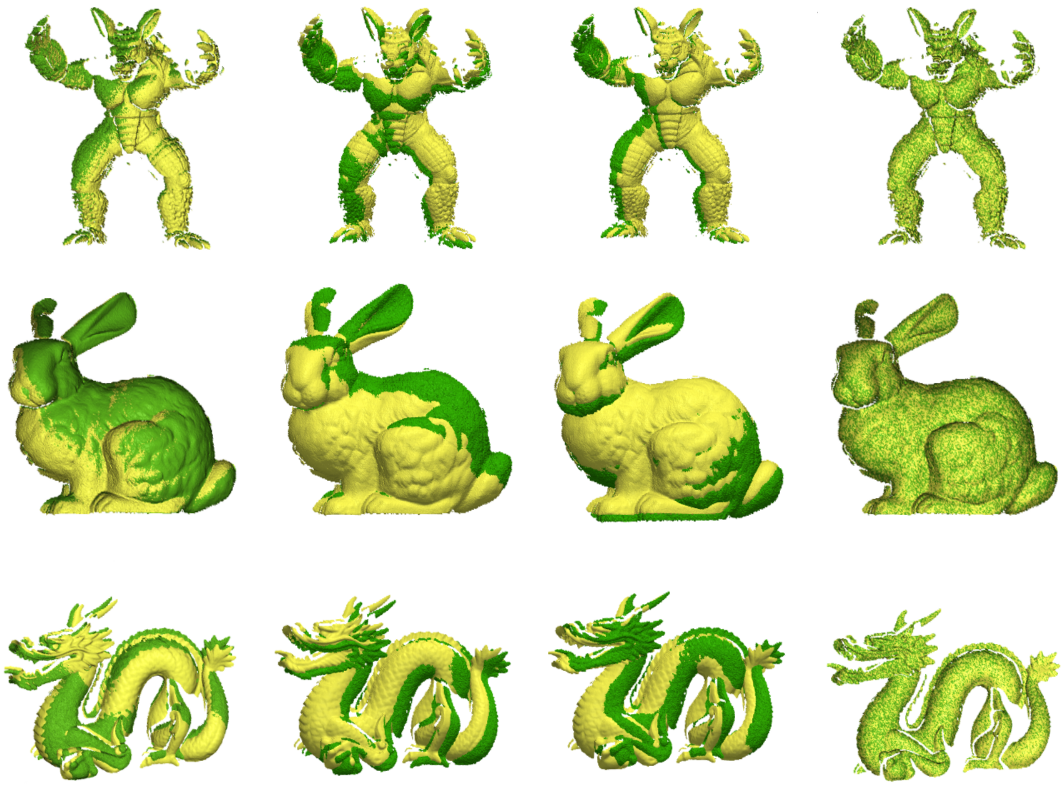


Figure 6.12: Examples of surface registration obtained respectively with Spin Images (first column), MeshDOG (second column), 4PCS (third column) and our Game-Theoretic Registration technique (last column).

rect registration. By contrast, GTR is able to obtain excellent alignment regardless of the initial motion perturbation. Finally, applying ICP to GTR decreases the RMS only by a very small amount.

6.3.6 Some Qualitative Results

In addition to the quantitative experiments presented, we also performed some qualitative tests. While these tests do not offer a measurable comparison between the results obtained by different methods, it certainly helps in putting the number presented so far in perspective. It is in fact sometimes harder to tell how much the RMS ratio affects the registration than judging some anecdotal alignment.

In Figure 6.12 we show some coarse alignments obtained with the four methods under comparison. In this particular example we can see that Spin Images fails on the Dragon, MeshDOG does not performs very well with any of the meshes, while the Four Points Congruent Sets method obtains good results with all the three meshes. On the other hand, GTR exhibits by far the best alignment in every example.

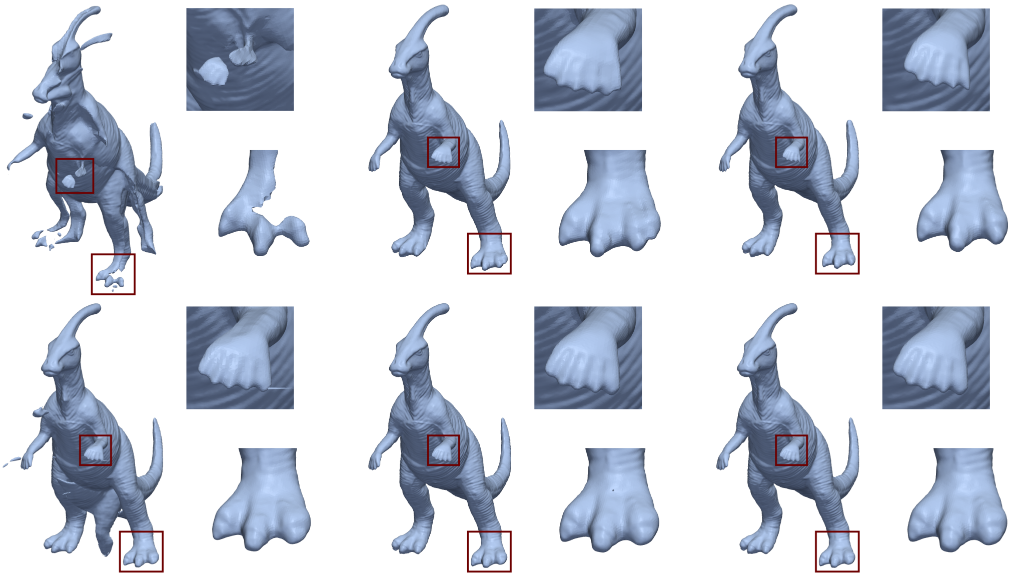


Figure 6.13: Comparison of surface reconstruction using different descriptors before (first row) and after (second row) ICP enhancement of the coarse registration obtained respectively using Spin Images (first column), SHOT (second column) and Surface Hashes (third column).

Figure 6.13 illustrates qualitative differences with the registration obtained by using different descriptors with and without ICP refinement. Since it is very difficult to spot defects on registration after the refinement we decided to perform a full registration of ranges acquired from a laser scanner contained in a publicly available database [99] and to build a closed surface using the Poisson Surface Reconstruction technique [81]. It is easy to note that for the same detail Surface Hashes and SHOT allow to build a smoother surface with no artifacts even before applying the refinement step. However a small defect can be spotted on the third finger on the foot reconstructed using SHOT without refinement. By contrast, Spin Images cannot be used to obtain reliable alignment even after applying ICP (the big artifacts that can be observed are mainly due to a few grossly misaligned ranges).

Finally, in Figure 6.14 we show the result obtained using the GTR pipeline as a fully automatic tool for aligning 12 range images acquired by a laser scanner.

6.3.7 Memory and Execution Time

Finally, we analyze the memory and cpu time requirements for our method.

The memory needed depends on the number of strategies in the initial set S. Since the payoff matrix sets a compatibility between each pair of strategies, the memory required is quadratic with $|S|$. To give a rough figure, with 1000 points sampled from the model and 6 neighbors for the initial matches, a single-precision matrix would require a little

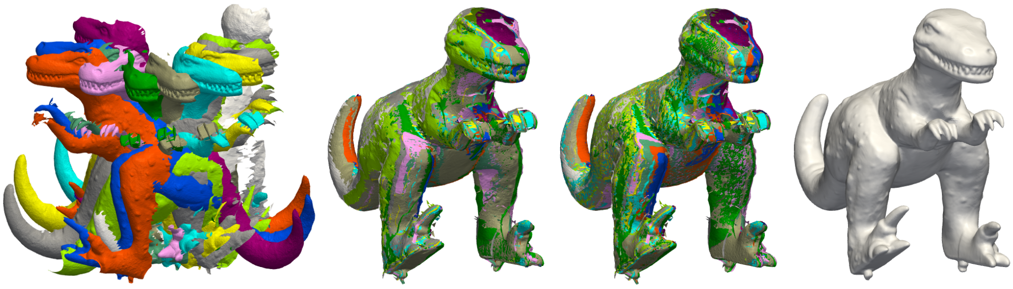


Figure 6.14: Fully automatic registration of 12 views of the t-rex model from [99] using the GTR pipeline with Surface Hashes. A set of range images (first column) with unknown initial positions is given; coarse alignment is then performed with our technique and refined with a few steps of ICP (second and third column). The last column shows the final model obtained by Poisson reconstruction [81].

more than 64MB. Of course, it is not really needed to materialize the payoff matrix. In fact, if the payoff function is simple enough, it could be advantageous to compute it on the fly during the iteration of the replicator dynamics. Actually, given that memory access is often the bottleneck on modern architectures, this could even speed-up computation (especially for GPU-based implementations of the replicator dynamics).

Regarding the execution time, each iteration of the replicator dynamics is quadratic with $|S|$. However, it is not easy to state how many iterations are required to converge as it depends on many factors and parameters. In Figure 6.15 we plotted a point cloud that relates the number of strategies with the convergence time (on our setup) for a large number of trials generated with the database of meshes adopted in the previous experiments. When dealing with some thousands of strategies (which is the common case) the evolution happens in about one second, which is reasonable for most non-real time applications.

6.4 Multiview Registration via Graph Diffusion of Dual Quaternions

While there are several fast and reliable methods to align two surfaces, the tools available to align multiple surfaces are relatively limited. In this Section we propose a novel multiview registration algorithm that projects several pairwise alignments onto a common reference frame. The projection is performed by representing the motions as dual quaternions, an algebraic structure that is related to the group of 3D rigid transformations, and by performing a diffusion along the graph of adjacent (i.e., pairwise alignable) views. The approach allows for a completely generic topology with which the pairwise motions are diffused. An extensive set of experiments shows that the proposed approach is both orders of magnitude faster than the state of the art, and more robust to extreme positional noise and outliers. The dramatic speedup of the approach allows it to be alternated with

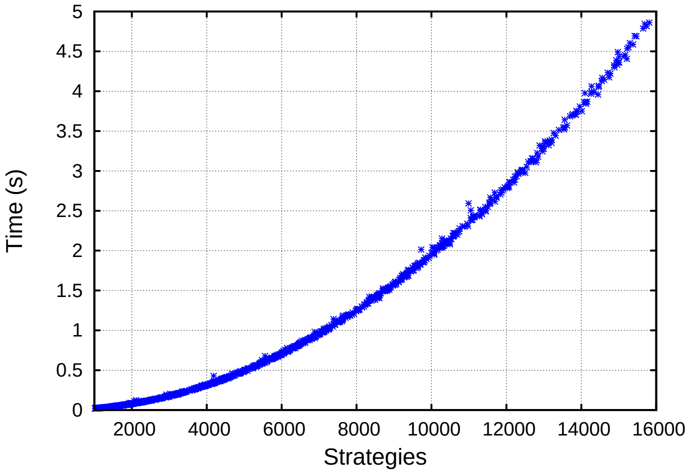


Figure 6.15: Execution time of the game-theoretic registration method with respect to the number of initial strategies.

pairwise alignment resulting in a smoother energy profile, reducing the risk of getting stuck at local minima.

6.4.1 Introduction

Full-object surface registration is typically a two step process where all the views are first registered against each other, and then all the pairwise transformations are lowered to a common coordinate frame through a process commonly referred to as multiview registration.

The literature on pairwise registration is quite ample, with modifications to the original ICP proposed by Zhang [161] and Besl and McKay [29] taking the lion's share. A more comprehensive review was given in the previous sections. ICP-based methods start from an initial pose estimate and iteratively refine it by minimizing a distance function measured between pairs of selected neighboring points. The variants generally differ in the strategies used to sample points from the surfaces, reject incompatible pairs, or measure error. In general, the precision and convergence speed of these techniques is highly data-dependent and very sensitive to the fine-tuning of the model parameters. Several approaches that combine these variants have been proposed in the literature in order to overcome these limitations (see [117] for a comparative review). Some recent variants avoid hard culling by assigning a probability to each candidate pair by means of evolutionary techniques [88] or Expectation Maximization [66]. ICP variants, being iterative algorithms based on local, step-by-step decisions, are very susceptible to the presence of local minima. Other fine registration methods include the well-known approach by Chen [44] and signed distance fields matching [94].

By contrast, the literature of multiview registration is more diverse. In [44] Chen and

Medioni propose to iteratively merge new views into a single metaview: The registration of a new view against the metaview is obtained with a common pairwise registration technique, such as ICP, and then the points of the new registered view are merged to the metaview; the approach is iterated until all the range images are merged. This metaview approach has problems since registration errors are accumulated rather than mediated. To solve this problem Bergevin *et al.* [28] match points in every view with all the views overlapping with it, and calculate a transformation that registers the first view using all the mating points. This process is iterated to convergence, thus diffusing the errors among all views. This is implicitly a diffusion process where the random walk in the transformation space is governed by the constraints offered by nearby views in the view-graph; however, convergence toward the steady-state is extremely slow and computationally demanding. Eggert *et al.* [55] constrain the pairings so that the points of each scan map with exactly one other point and then minimize the total distance between the paired points. This speeds up convergence, but can prevent the algorithm from converging to a correct solution as the views may cluster into groups that are well registered, without improving inter-group registration. With these iterative algorithms based on global point correspondences, how and when to apply the transformation remains an open issue: For example, Bergevin *et al.* [28] calculate a transformation for each view separately and then apply them simultaneously before the next round of matchings, while Benjemaa and Schmitt [27] apply the new transformations independently as soon as they are calculated, and Eggert *et al.* [55] solve for the update by simulating a spring model. An alternative was explored in [75] where an approximate surface model is created and the view are registered against the model. The surface model is then iteratively refined using the new registrations.

In [112] Pulli takes a simplifying view that pairwise registrations are "as good as it gets" and that the role of multiview registration is only to project the transformation into a common reference frame in such a way as to limit the accumulation of registration errors. To this end, he proposes a greedy approach that tries to limit the difference between the position of point sets as positioned in two frames and transformed by the pairwise registration of the two frames. More formally, he tries to keep the distortion $\mathcal{D}(S)$ of the points from a set S within a given tolerance ϵ, where

$$\mathcal{D}(S) = \sum_{s \in S} \sum_{(i,j) \in \mathcal{V}} ||P_i(s) - T_{ij}\big(P_j(s)\big)||^2 .$$

Here P_i is the transformation that maps a point into the coordinate system of view i, T_{ij} is the transformation that maps the coordinate frame j into the coordinate frame i obtained through pairwise registration, and $\mathcal{V}$ is the set of pairs of neighboring views for which pairwise registration is performed. Pulli suggests to sample the set of fiduciary points S from the surface of the object. Interestingly, by working only on the space of transformations, this approach limits the memory requirements since it does not need to keep all the points from all the views in memory at once. Note, however, that the approach cannot guarantee that an optimal solution will be found, nor that any solution within the given tolerance will be found.

More recently, Williams and Bennamoun [157] adopted a similar view, posing the problem as the minimization of the distortion on a set of fiduciary points and computing the minimization by an iterative approach optimizing each rotation via singular value decomposition.

In this work we propose a novel multiview registration algorithm where the poses are estimated through a diffusion process on the view-adjacency graph. The diffusion process is over dual quaternions [48], a non-commutative and non-associative algebraic structure that is related to the group $SE(3)$ of 3D rigid transformations, leading to an approach that is both orders of magnitude faster than the state of the art, and more robust to extreme positional noise and outliers.

6.4.2 Dual Quaternions and 3D Transformations

Quaternions have been a popular geometrical tool for more than 20 years as they represent 3D rotations in a way that is arguably more efficient and robust than 3×3 rotation matrices [127]. Quaternions are an algebraic extension of complex numbers with 3 imaginary bases i, j, and k, thus a quaternion is a number of the form

$$q = a + ix + jy + kz. \tag{6.2}$$

The multiplication of two quaternions is defined through the following multiplication rules for the three imaginary bases:

$$i^2 = j^2 = k^2 = -1 \tag{6.3}$$
$$ij = k = -ji \tag{6.4}$$
$$jk = i = -kj \tag{6.5}$$
$$ki = j = -ik. \tag{6.6}$$

The *conjugate* of a quaternion $q = a+ix+jy+kz$ is the quaternion $q^* = a - ix - jy - kz$, while the *norm* of a quaternion is the quantity

$$\|q\| = \sqrt{qq^*} = \sqrt{q^*q} = \sqrt{a^2 + x^2 + y^2 + z^2}. \tag{6.7}$$

Quaternions with unitary norm are called *unit quaternions*. In the following we will use the vectorial representation of quaternions: Let $\mathbf{i} = (i, j, k)$ be the row-vector of the imaginary bases, we can write the quaternion q as $a + \mathbf{iv}$, where $\mathbf{v} = (x, y, z)^T$ is a 3D vector. A right-handed 3D rotation of angle θ around the axis of unit vector $\mathbf{v}$ is in relation with the unit quaternion $q = \cos(\theta/2) + \sin(\theta/2)\mathbf{iv}$. In fact, let $\mathbf{p} = (p_x, p_y, p_z)^T$ be a 3D point and $\mathbf{p}_r$ its rotation, we have $\mathbf{ip}_r = q(\mathbf{ip})q^*$. The ring of quaternions, however, is a dual cover of the group $SO(3)$ of 3D rotations, as q and $-q$ represent the same rotation. Quaternions are particularly interesting since they allow for optimal interpolation between rotations. The famous Spherical Linear Interpolation (SLERP) algorithm [127] interpolates quaternions on the unit hypersphere and exhibits the following useful properties:

Table 6.1: Multiplicative table of dual quaternions.

	1	i	j	k	ϵ	ϵi	ϵj	ϵk
1	1	i	j	k	ϵ	ϵi	ϵj	ϵk
i	i	-1	k	$-j$	ϵi	$-\epsilon$	ϵk	$-\epsilon j$
j	j	$-k$	-1	i	ϵj	$-\epsilon k$	$-\epsilon$	ϵi
k	k	j	$-i$	-1	ϵk	ϵj	$-\epsilon i$	$-\epsilon$
ϵ	ϵ	ϵi	ϵj	ϵk	0	0	0	0
ϵi	ϵi	$-\epsilon$	ϵk	$-\epsilon j$	0	0	0	0
ϵj	ϵj	$-\epsilon k$	$-\epsilon$	ϵi	0	0	0	0
ϵk	ϵk	ϵj	$-\epsilon i$	$-\epsilon$	0	0	0	0

- **Shortest path:** the motion between the initial rotation R_0 and the final rotation R_1 is a rotation about a fixed axis with the smallest angle.

- **Constant speed:** the angle of the interpolated rotation varies linearly with respect to parameter t.

- **Coordinate system invariance:** the interpolation path does not change if we change the coordinate system.

On the other hand, linear interpolation followed by reprojection onto the unit hypersphere guarantees the first and third properties, but exhibits changes in speed, in particular it accelerates around the middle of the interpolation. This implies that a linear averaging of quaternions does not minimize the squared geodesic distance in the unit quaternion manifold in the same way that the mean of a set of points minimizes the squared Euclidean distances to the points. Unfortunately, SLERP does not generalize to the blending of several rotations. Buss and Fillmore [39] provide an iterative algorithm to find the proper (weighted) mean in the unit-quaternion manifold, while Kavan and Žára [80] show that the difference between spherical and linear interpolation is always less than 0.071 radians.

Dual quaternions are less known than quaternions, but their ability to efficiently represent rigid transformations has been successfully adopted in 3D animation and skinning [78], robot control and registration [17, 52], and have been used in theoretical kinematics for a long time [96]. Dual quaternions are an algebraic extension of quaternions much like complex numbers are an extension of the reals. They are defined in terms of a dual basis ϵ that commutes with the imaginary bases; thus, a dual quaternion is a number of the form $q + \epsilon r$, where q and r are quaternions, and the product follows the multiplicative rule $\epsilon^2 = 0$, yielding

$$(q + \epsilon r)(s + \epsilon t) = qs + \epsilon(qt + rs)\,.$$

This results in the multiplicative table shown in Table 6.1. Dual quaternions have three different conjugates:

$$(q + \epsilon r)^* = q^* + \epsilon r^* \quad (q + \epsilon r)^\dagger = q^* - \epsilon r^* \quad (q + \epsilon r)^+ = q - \epsilon r$$

The norm of a dual quaternion dq is $||dq|| = \sqrt{dq^*dq} = \sqrt{dq\,dq^*} = ||dq^*||$ and the inverse of a dual quaternion dq is $dq^{-1} = \frac{dq^*}{||dq||^2}$. The rigid transformation obtained by a rotation defined by the unit quaternion r and then a translation by $\mathbf{t} = (t_x, t_y, t_z)^T$, is represented by the dual quaternion

$$dq = r + \tfrac{1}{2}\epsilon\mathbf{it}\,r\,.$$

In fact, if we represent a 3D point $\mathbf{p}$ as $1 + \epsilon\mathbf{ip}$, the following holds:

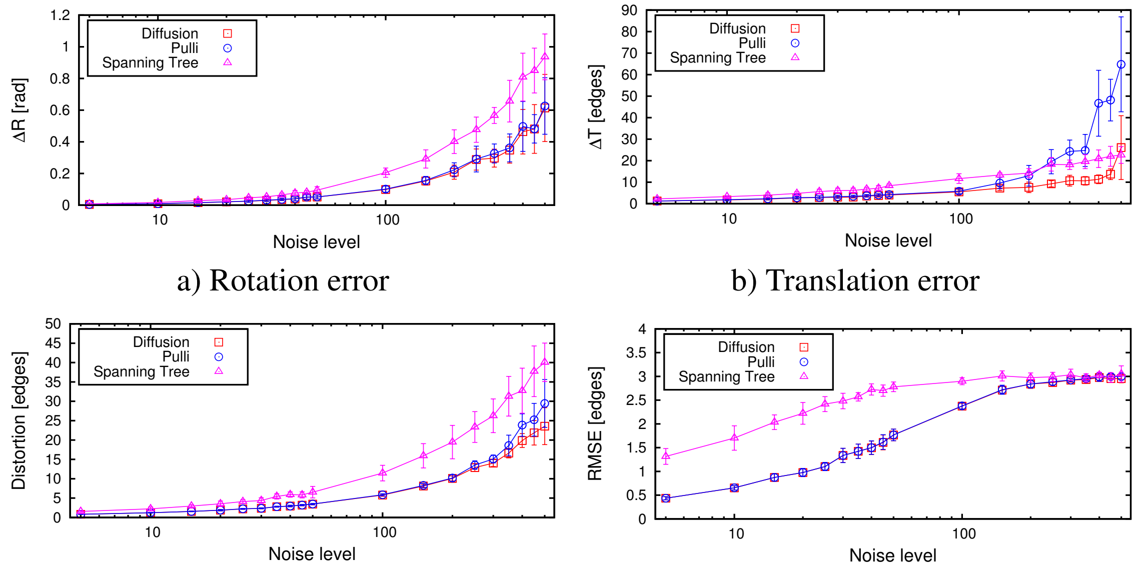
$$\begin{aligned}(r + \tfrac{1}{2}\epsilon\mathbf{it}rt)(1 + \epsilon\mathbf{ip})(r + \tfrac{1}{2}\epsilon\mathbf{it}r)^\dagger = \\ (r + \tfrac{1}{2}\epsilon\mathbf{it}r + \epsilon r\mathbf{ip})(r^* - \tfrac{1}{2}\epsilon(\mathbf{it}r)^*) = \\ (r + \tfrac{1}{2}\epsilon\mathbf{it}r + \epsilon r\mathbf{ip})(r^* + \tfrac{1}{2}\epsilon r^*\mathbf{it}) = 1 + \epsilon(r\mathbf{ip}r^* + \mathbf{it})\,.\end{aligned}$$

Thus, $1 + \epsilon\mathbf{ip}$, i.e., the dual quaternion representation of the point $\mathbf{p}$, gets mapped into the dual quaternion $1 + \epsilon(r\mathbf{ip}r^* + \mathbf{it})$, which is the representation of $\mathbf{p}$ after the rotation and translation have been applied. In fact, $r\mathbf{ip}r^*$ represents the rotation by r with the usual quaternion notation, while the addition of $\mathbf{it}$ takes care of the subsequent translation. Further, any dual quaternion $q + \epsilon r$ with $||q|| = 1$ and $q \cdot r = 0$ represents a rigid transformation. Here $\cdot$ represents the standard dot product in the quaternion viewed as a four-dimensional vector space over $\mathbb{R}$. However, the dual quaternion representation is not unique since, as with the normal quaternions, dq and $-dq$ represent the same transformation.

In [79, 78] the authors present ScLERP, a generalization of the SLERP interpolation algorithm for dual quaternions. Let $\alpha(t)$, $\mathbf{a}(t)$, $\delta(t)$, and $\mathbf{d}(t)$ be respectively the rotation angle and axis, and the translation magnitude and direction, then ScLERP was shown to have the following properties: a) $\mathbf{a}(t)$, and $\mathbf{d}(t)$ are constant and $\alpha(t) \in [-\pi; \pi]$ (shortest path); b) $\frac{d}{dt}\alpha(t) = 0$ and $\frac{d}{dt}\delta(t) = 0$ (constant speed). Further, it is invariant to changes in the coordinate system. In [79] was presented an iterative algorithm called Dual quaternion Iterative Blending (DIB) for averaging dual quaternions in a way that minimizes the (weighted) squared geodesic distances between the target mean and the input quaternions in the Riemannian manifold of unit dual quaternions, and it was shown that the variation between the proper geodesic average and a linear blending followed by a reprojection has an upper bound of 0.143 radians in rotation and a relative variation of 15% in translation, while in general the differences remain much smaller. In particular, if the set of dual quaternions we want to blend has small variance, the linear average and the geodesic average converge rapidly, since the difference between the geodesic and Euclidean distance is $O(\theta)$ where θ is the angle of rotation. In the following we will use the notation $\mathrm{ScAVG}(q_1, \ldots, q_n)$ to refer to geodesic mean of the quaternions $(q_1, \ldots, q_n)$ as obtained by applying DIB with uniform weights.

6.4.3 View-Graph Diffusion

We cast the multiview registration problem into a diffusion of rigid transformations over the view-graph, i.e., a graph in which nodes correspond to the range images and the edges



a) Rotation error

b) Translation error

c) Euclidean distortion

d) RMS error

Figure 6.16: Comparison of the different methods in the synthetic experiments at various levels of noise.

reflect the adjacency relation between views, or, equivalently, the existence of an overlap between the scans. Let V_i be a transformation taking the coordinate frame of view i into a global coordinate frame, and T_{ij} the result of the pairwise registration taking the coordinate frame of view j into the frame of view i. Then, if the pairwise registration was noise-free, we would have $T_{ji} * V_i = V_j$ for all adjacent views i and j. In this setting the problem of multiview registration is that of finding a set of rigid motions from each view to a common frame of reference, say that of view 0, such that a measure of distortion between the position V_i of view i and the position $T_{ij}V_j$ obtained from the composition of position V_j and the pairwise registration T_{ij} is minimized, i.e., we seek to minimize the functional

$$D = \sum_i \sum_{j \in N(i)} d(T_{ij}V_j, V_i)$$

for an appropriate distortion function d. Here $N(i)$ is the set of neighbors of view i. This is in spirit similar to the approach taken by Pulli [112], where the distortion function is the (squared) Euclidean distance between the final position of a set of points S transformed with motions V_i and $T_{ij}V_j$:

$$D_P = \sum_i \sum_{j \in N(i)} \sum_{\mathbf{p} \in S} ||T_{ij}V_j\mathbf{p} - V_i\mathbf{p}||^2.$$

We adopt a different measure of distortion that derives from the fact that any rigid transformation is in fact a screw motion, i.e., a rotation around an axis placed anywhere in the 3D space, and a translation along the direction of the axis. We define the *screw dis-*

tance $d_{SC}(q_i, q_2, \mathbf{p})$ as the length of the screw path of the point $\mathbf{p}$ along the transformation $\hat{q} = q_i^\dagger q_2$, i.e.,

$$d_{SC}(q_i, q_2, \mathbf{p}) = \sqrt{t^2 + \alpha^2 r_{\mathbf{p}}^2}\,, \tag{6.8}$$

where t is the length of the translation along the axis of $\hat{q}$, α is the rotation angle, and $r_{\mathbf{p}}$ is the distance of $\mathbf{p}$ from the rotation axis. The *screw distortion* is then defined as

$$D_{SC} = \sum_i \sum_{j \in N(i)} \sum_{\mathbf{p} \in S} d_{SC}(T_{ij}V_j, V_i, \mathbf{p})^2\,.$$

A direct consequence of the fact that ScLERP interpolation is both shortest path and constant speed is that, given a set of dual quaternions $Q = q_i, \dots, q_n$, their screw average $\hat{q} = \mathrm{ScAVG}(q_i, \dots, q_n)$ minimizes the sum of squared screw distances $\sum_i d_{SC}(q_i, \hat{q}, \mathbf{p})$ for any point $\mathbf{p} \in \mathbb{R}^3$ [145]. That is, by measuring along the curved screw path, the transformation that minimizes the distortion does not depend on the points selected, which was arguably the most problematic aspect with the Euclidean distortion adopted by Pulli. Further, when the variation in orientation among the dual quaternions $q_i, \dots, q_n$ is very small, we have that $d_{SC}(T_{ij}V_j, V_i, \mathbf{p}) \approx ||T_{ij}V_j\mathbf{p} - V_i\mathbf{p}||$ for any point $\mathbf{p} \in \mathbb{R}^3$. In fact, we have $d_{SC}(T_{ij}V_j, V_i, \mathbf{p})^2 = {d_E^{||}}^2 + \frac{\theta/2}{\sin(\theta/2)} {d_E^{\perp}}^2$ where θ is the rotation angle, and $d_E^{||}$ and $d_E^{\perp}$ are respectively the components of the Euclidean distance parallel and orthogonal to the axis.

The optimal multiview alignment is thus obtained by computing the steady-state of the following process

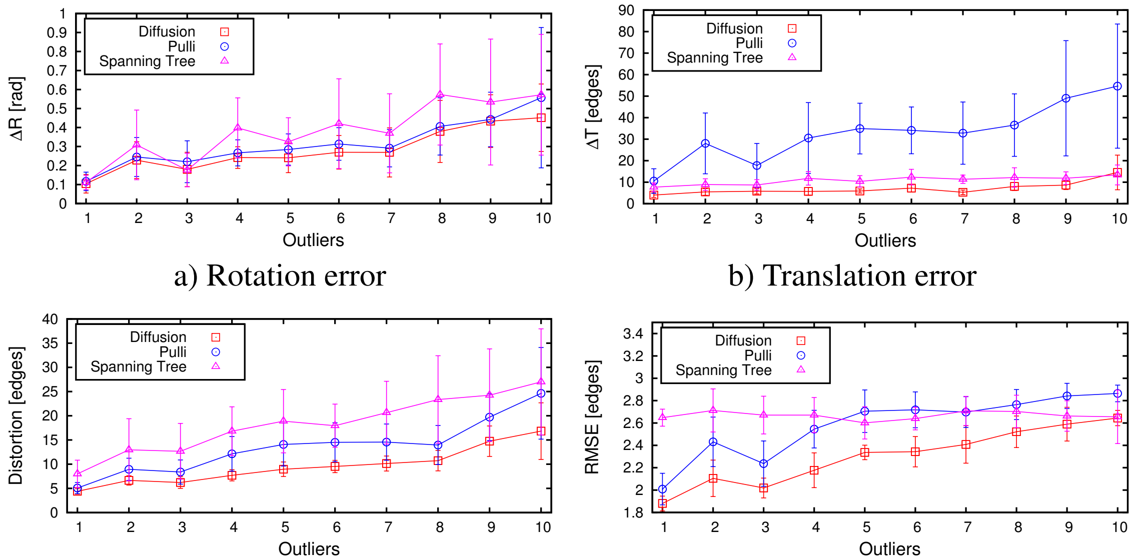
$$V_i^{t+1} = \underset{j \in N(i)}{\mathrm{ScAVG}}(\pm T_{ij} V_j^t)$$

where the sign uncertainty is a consequence of the sign uncertainty in the dual quaternion representation, and is chosen so that $\pm T_{ij} V_j^t \cdot V_i^t > 0$. Further, since for small and moderate rotational variability in the vectors the linear average approximates well the screw average, while being much faster, in all our experiments we are using linear averages.

Finally, the proposed approach is a refinement method that requires initial motion estimates, but these can be computed by simple composition of the transformations along adjacent views with a breadth-first visit starting from the view 0.

6.4.4 Experimental Evaluation

Multiview registration techniques have both a sparse and diverse coverage in literature, and as such they suffer from the lack of a robust and fair methodology for performance assessment and comparison. Specifically, in real scenarios, where ground-truth data is not available, it can be very hard to evaluate and quantify the results of a global alignment and settle for a solution, without resorting to a thorough and time-consuming analysis of the registered views.



a) Rotation error

b) Translation error

c) Euclidean distortion

d) RMS error

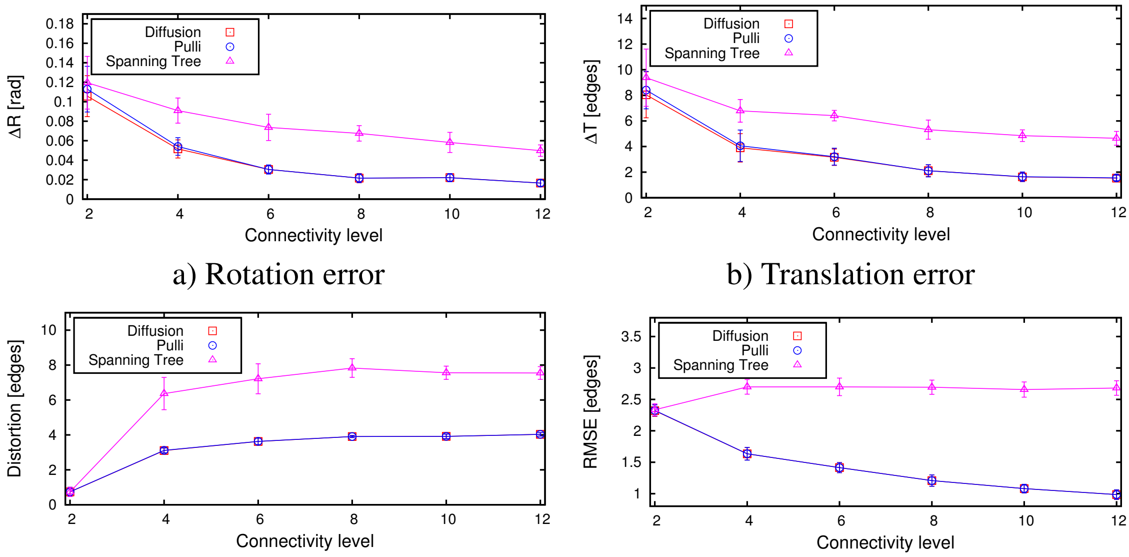
Figure 6.17: Comparison of the different methods in the synthetic experiments with different numbers of outliers.

To this end, we performed a wide range of experiments with both synthetic and real-world data. For each complete 3D model (from Georgia Tech's large geometric models archive[1]), a total of 36 orthographic snapshots were taken, each at a different angle of view; these, together with the ground-truth rigid motions used to produce the range images, constitute the dataset over which synthetic experiments were performed. In all the experiments we compare our method against Pulli's algorithm (as implemented by the author in the Scanalyze[2] software package), currently the method of choice in many applications. We evaluated the performance of the two algorithms, together with the initialization results obtained through a breadth-first coverage of the view graph (indicated as *Spanning Tree*), under different noise conditions and connectivity levels. For all the experiments we show the relative displacement with respect to ground-truth motion (with ΔR being the angle between the two unit quaternions, and ΔT the translation error expressed in median edge length), the RMS error among all the ranges and range 0 (point pairs were obtained through normal-shooting in both directions), and the Euclidean distortion metric adopted by Pulli (indicated as *Distortion*).

In Figure 6.16 we show the results at different levels of initial displacement, where every view in the graph is connected with the next two in a ring topology. *Noise level* refers to a quantity which is proportional to the amount of Gaussian noise applied to the ground-truth pairwise motions, and ranges from a few units of edges and radians to tens of units; for each noise level, 10 independent runs of each method were performed. Both the

[1] `http://www.cc.gatech.edu/projects/large_models`

[2] `http://www.graphics.stanford.edu/software/scanalyze`



a) Rotation error

b) Translation error

c) Euclidean distortion

d) RMS error

Figure 6.18: Comparison of the different methods in the synthetic experiments at various connectivity levels.

Diffusion and Pulli methods compensate well rotational errors and are comparably good at low levels of noise, whereas the latter is outperformed when positional noise increases, both in terms of translation error and Euclidean distortion. It is worth noting that when we perturbed either the rotational or translational part of the rigid motion, keeping the other fixed, Spanning Tree and Pulli's methods performed a joint optimization modifying both components, while our method never changes the already optimal part of the motion, yielding better results at all levels of noise.

In Figure 6.17 we assess the resilience of the tested methods to the presence of outliers: starting from a close-to-optimal initialization, we introduced strong pairwise misalignments so as to simulate a realistic scenario in which pairwise registrations get stuck at local minima. Connectivity is the same as in the previous experiments. In these graphs, the x-axis grows with the number of such mis-registrations; here, 20 runs were performed at each level, and for each run random pairs were picked from a uniform distribution and perturbed strongly. It can be seen that all the methods handle well rotational errors, with the Diffusion method giving particularly good performance constantly, even when the number of outlying pairs becomes large. Figure 6.17b) and Figure 6.17d) interestingly show the inability of Pulli's method to deliver good results in such situations: this is an inherent weakness of the method, since it acts in such a way to minimize all motion, relying on the assumption that pairwise alignments are nearly perfect.

The next set of experiments (Figure 6.18) is aimed at studying the effect of view-graph connectivity on the registration results. In these figures, *Connectivity level* refers to the number of links per view. As it can be readily seen, performance tends to increase with the number of edges in the view graph, as all three methods greatly benefit from more

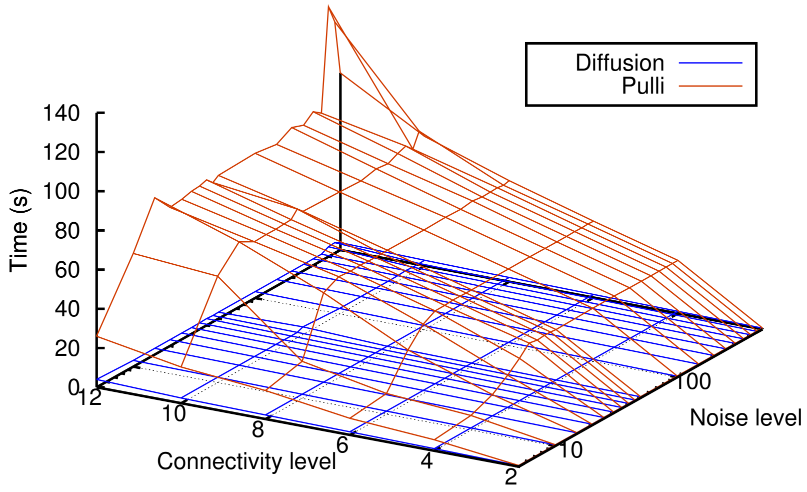


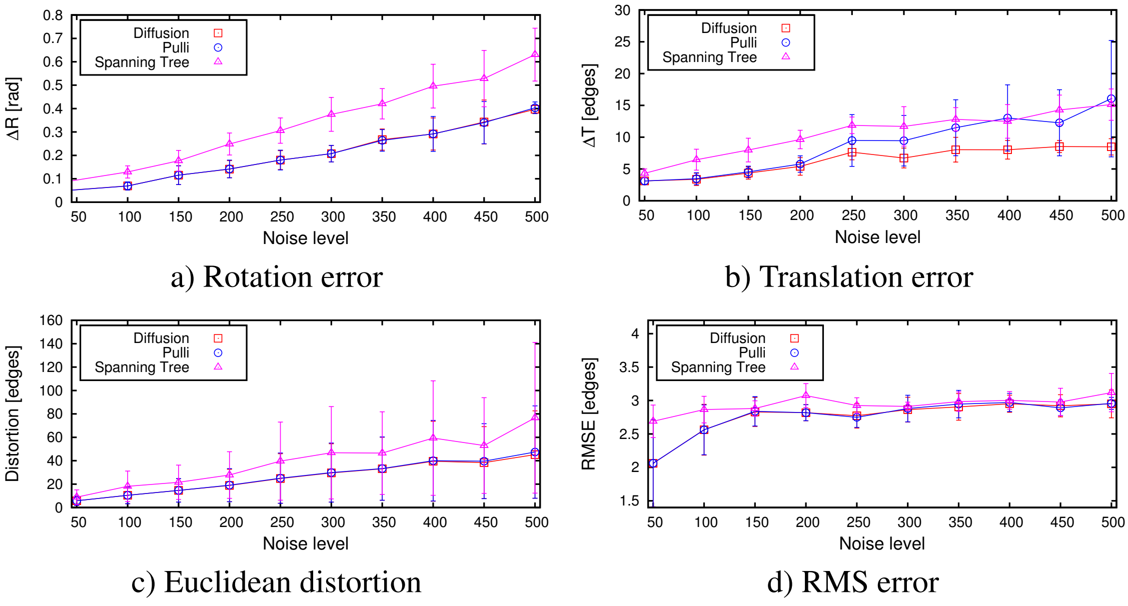
Figure 6.19: Computation times versus noise and connectivity levels.

structure being brought in. It must be noted, however, that connectivity augmentation dramatically increases the time requirements of Pulli's method, bringing up convergence times by orders of magnitude (see Figure 6.19).

Figure 6.19 compares the average convergence time of the proposed approach with the time required by Pulli's method. The times are shown as a function of noise level and connectivity. We can see that, with the exception of extreme values, the times required by Pulli's approach are independent from the level of noise, but grow linearly with the connectivity level and are in almost all tested situations in the order of 10 to 100 seconds. The proposed approach exhibits the same independence with respect to noise and linear growth with respect to connectivity level, but it is always around 2 to 3 orders of magnitude faster.

In order to simulate a more realistic type of noise, we also tested the following setup: we perturbed the initial pairwise motion as for the first set of experiments, and then applied ICP to refine the alignment. This setup simulates a normal registration process with increasingly bad coarse registration, with the possibility that ICP gets stuck on local minima providing more structured outliers than the previous experiments. The results of this set of experiments can be seen in Figure 6.20. Note that our approach and Pulli's method yield very similar results in all the metrics except for ΔT. This can be justified by the fact that, when caught in local minima, ICP slides the surface one over the other, resulting in strong translational outliers which Pulli's method cannot deal with. The very low RMSE derives from the fact that sliding along the surface each point still finds close-by mates on the other mesh. The proposed method, on the other hand, manages to smooth all the outliers effectively, yielding low errors in all the metrics.

Figure 6.21 shows an example of a real set of range images acquired with a scanner and aligned using Pulli's method and the proposed approach. While at a large scale the overall alignment appears similar, by examining closeups of various sections of the glasses we see that Pulli's method provides a slightly worse motion estimate, resulting in



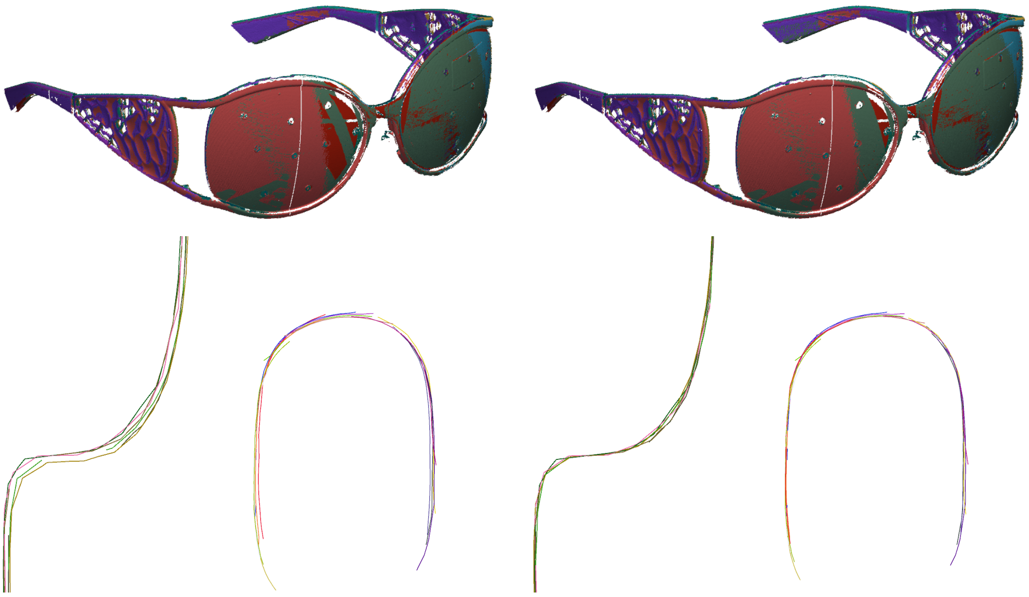
a) Rotation error

b) Translation error

c) Euclidean distortion

d) RMS error

Figure 6.20: Comparison of the different methods in the experiments with motion refined by ICP at various levels of noise.



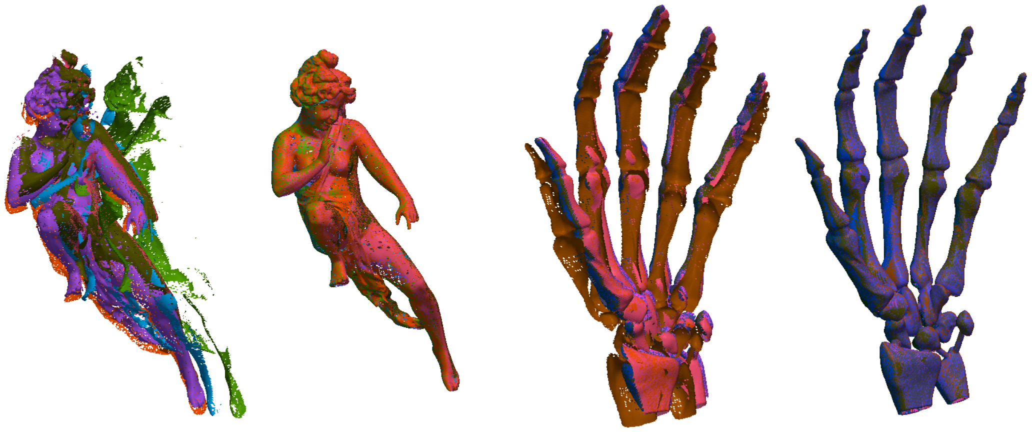
Pulli's method

Dual quaternion diffusion

Figure 6.21: Global registration and closeup of slices of Pulli's method and our approach.

a wider stratification of the meshes.

It is worth noting that the orders of magnitude speedup provided by our approach makes it possible to run it several times combining it with pairwise registration. The idea



262 sec. Trailing diffusion

233 sec. Alternating diffusion

115 sec. Trailing diffusion

117 sec. Alternating diffusion

Figure 6.22: Alignments obtained with the trailing and alternating approaches, and respective timings.

is to alternate a few steps of ICP performed on all adjacent views with the diffusion. This way the diffusion process can be seen as a projection operator taking the incremental pairwise motion onto a set of consistent motion estimates. The advantage of this projection is that the constrained motion space smooths the energy profile of the resulting "global" ICP, reducing the risk of getting stuck in local minima. Figure 6.22 shows two examples of alignments obtained by performing ICP from bad initial motion estimates and then performing diffusion at the end (Trailing diffusion) and alternating between 10 steps of ICP and a diffusion process until convergence (Alternating diffusion). Clearly alternating pairwise registration allows to avoid local minima in these examples without incurring in any noticeable penalty in running times.

6.5 Conclusions

In this Chapter we adopted the game-theoretic framework introduced in the previous chapters to tackle the commonly encountered matching problem of 3D surface registration. We tried to attack both the coarse and fine registration problems at once. Our approach has several advantages over the state-of-the-art: it does not require any kind of initial motion estimation, as it does not rely on spatial relationships between model and data points, and, unlike most inlier selection techniques, it is not affected by a large number of outliers since it operates an explicit selection of good inliers rather than using random selection or vote counting for validation. The approach has also shown to be general enough to accept different feature descriptors. From a theoretical point of view, a sound correspondence between optimal alignments and evolutionary equilibria has been presented and a wide range of experiments validated both the robustness of the approach with respect to noise and its performance in comparison with other well-known techniques.

Having approached the registration problem in a pairwise setting, we successively shifted our efforts towards the “derivative” problem of recomposing the pairwise motions together, in an attempt to minimize the global registration error over the view topology. To do this, we proposed a novel multiview registration algorithm that projects several pairwise alignments onto a common reference frame. The projection is performed by representing the motions as dual quaternions which are then diffused along the graph of adjacent (i.e., pairwise alignable) views. The approach is general allowing for any topology of the view-adjacency graph. An extensive set of experiments has shown that the proposed approach is both orders of magnitude faster than the state of the art, and more robust to extreme positional noise and outliers. Finally, the dramatic speedup of the approach allows it to be alternated with the pairwise alignment process resulting in a “global” ICP that exhibits a smoother energy profile, reducing the risk of getting stuck at local minima.

7

A Scale Independent Game for 3D Object Recognition

During the last few years, a wide range of algorithms and devices have been made available to easily acquire range images. To this extent, the increasing abundance of depth data boosts the need for reliable and unsupervised analysis techniques, spanning from part registration to automated segmentation. In this context, we focus on the recognition of known objects in cluttered and incomplete 3D scans. Locating and fitting a model to a scene are very important tasks in many scenarios such as industrial inspection, scene understanding, medical imaging and even gaming. For this reason, this problem has been addressed extensively in the literature. Nevertheless, while many descriptor-based approaches have been proposed, a number of hurdles still hinder the use of global techniques. In this work we offer a different perspective on the topic. Specifically, we adopt an evolutionary selection algorithm that seeks to attain a global pairwise agreement among surface points, while operating at a local level. The approach effectively extends the scope of local descriptors by actively selecting correspondences that satisfy global geometric consistency constraints, allowing us to attack a more challenging scenario where model and scene have different, unknown scales. This leads to a novel and very effective pipeline for 3D object recognition, which is validated with an extensive set of experiments and comparisons with recent techniques at the state of the art.

7.1 Introduction

In this Chapter we propose a feature-based 3D object recognition pipeline that deals in a robust manner with strong occlusion and clutter. This happens by adopting a recent local surface descriptor to find a set of matching candidates among a selection of relevant points on model and scene. Acting as distinctive priors, the introduced descriptors allow to reduce the problem size and to gain in robustness. These candidates are then let to compete in a non-cooperative game, where payoff values are proportional to the degree of Euclidean compatibility between them. The competition induces a selection process in which incompatible matches are driven to extinction whereas a set of sparse, yet very reliable correspondences survive. To attain scale invariance, we devise another game

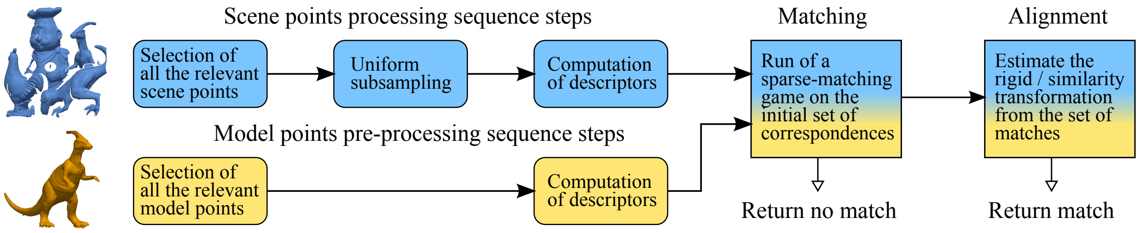

Scene points processing sequence steps
Selection of all the relevant scene points
Uniform subsampling
Computation of descriptors
Model points pre-processing sequence steps
Selection of all the relevant model points
Computation of descriptors
Matching
Run of a sparse-matching game on the initial set of correspondences
Return no match
Alignment
Estimate the rigid / similarity transformation from the set of matches
Return match

Figure 7.1: An overview of the object recognition pipeline presented.

where the change in scale is accounted for by considering geometric information along the paths connecting pairs of points. Specific scale-invariant descriptors are not needed for this game. Rather, we take a different approach by computing scale-dependent descriptors at different scales and then let the selection process extract the correct matches from the generated multi-scale pool of hypotheses. This new pipeline is tested in a wide range of experiments and is shown to outperform the state-of-the-art for 3D object recognition in clutter.

7.2 Feature Detection and Description

Much of this section follows rather directly from the formulation given in Chapter 6. For both efficiency and robustness reasons, our matching technique works on a subset of model and scene vertices. Interest point selection is performed by computing for each point a single-component *Integral Hash* (Section 6.1.2) at a given support scale σ, and retaining only those samples that obtain a negative value. Note that in this case, for the sake of efficiency, we don't carry out any feature clustering for the feature detection step. Being designed as a simple approximation to the integral invariant [111], calculation of the Integral Hash is very fast and the selection step is roughly equivalent to extracting points that belong to concave surface areas, where the measure of concavity is proportional to the absolute value of the Integral Hash at that point. Keeping only negative values means, in practice, that we are avoiding flat and convex areas which, empirically, we have seen to be less distinctive in a large variety of cases. By modulating the value of σ, a more or less inclusive sample selection can be carried out (see Figure 7.2). All the relevant points extracted from the model surface are kept. By contrast, uniform subsampling is optionally performed on the set of relevant points in the scene. Although more sophisticated detection algorithms could be used for this step (see [100], or [118] for a recent survey), we favored efficiency over repeatability since the game-theoretic selection mechanism is very effective at eliminating wrong guesses. Finally, a descriptor vector is computed for each vertex. To this extent, any of the descriptors discussed in Chapters 2 and 6 may be used; however, after an initial round of experiments, the SHOT descriptor [143] was chosen as it obtains the best performance overall.

Again, these steps are not strictly necessary, but introducing such priors proves to be beneficial both for reducing the problem size (which is proportional to the cardinality of

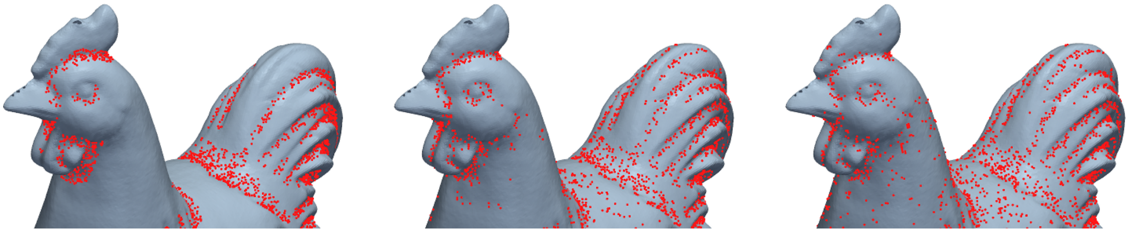


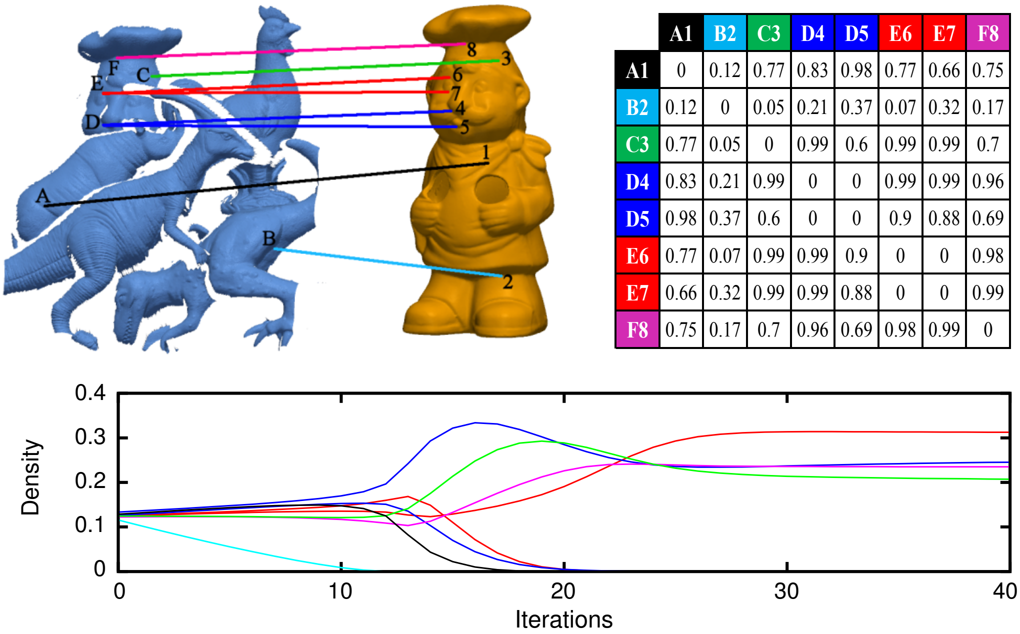
Figure 7.2: In order to avoid mismatches and reduce the convergence time it is important to use only relevant points. Model vertices selected with a σ respectively equal to 8, 5 and 2 times the median model edge are shown from left to right.

the set of matching strategies) and in terms of inlier ratio, which increases with rejection of unlikely hypotheses. In this regard, we remind that our method acts as an inlier selector whereas no ex-post verification is performed to validate the matches, and that this inlier selection behavior is put under considerable strain in the specific case of object-in-clutter scenarios, where strong groups of structured outliers can divert the selection process towards the wrong solution. We also note that existing techniques usually tend to forge ad-hoc matching methods for the specific descriptors they propose [76, 99, 100, 106, 23], while our method is general in this respect. In the experimental section we investigate both the influence of the relevant point selection and of the descriptor adopted.

7.3 Sparse Matching Game

We derive our matching approach directly from the technique we introduced in Chapter 6. The complete pipeline we are proposing is made up of a preprocessing step and two non-cooperative games (see Figure 7.1). The preprocessing is performed both on the model and on the scene. This step involves an initial selection of relevant points on the respective surfaces. The relevance criteria are explained in the next section, however, in this context the general meaning of this culling step is to avoid surface patches that are not significant for matching, such as flat areas. All the interest points on the model are kept, while those on the scene are uniformly subsampled. This makes sense for many reasons. In many applications the set of models does not change in time, and thus descriptors must be computed just once. In addition, as explained in the following sections, the direction of the matching is from the scene to the model and having less source than target points allows the game to proceed faster without compromising accuracy. Finally, the model tends to be measured with greater accuracy (either because more time can be spent on it or because it comes from a CAD model). A descriptor is computed for all the retained points, and these are used to build the initial candidates that, in turn, are fed to a matching game.

The remainder of this section details the (sparse) matching game that we adopt for the 3D object recognition scenario. We assume that relevant points were previously extracted from model and scene, and that every point of interest has a descriptor vector associated



	A1	B2	C3	D4	D5	E6	E7	F8
A1	0	0.12	0.77	0.83	0.98	0.77	0.66	0.75
B2	0.12	0	0.05	0.21	0.37	0.07	0.32	0.17
C3	0.77	0.05	0	0.99	0.6	0.99	0.99	0.7
D4	0.83	0.21	0.99	0	0	0.99	0.99	0.96
D5	0.98	0.37	0.6	0	0	0.9	0.88	0.69
E6	0.77	0.07	0.99	0.99	0.9	0	0	0.98
E7	0.66	0.32	0.99	0.99	0.88	0	0	0.99
F8	0.75	0.17	0.7	0.96	0.69	0.98	0.99	0

Figure 7.3: An example of the evolutionary process (with real data). Here we use exponential replicator dynamics for faster convergence [108]. A set of 8 matching candidates is chosen (upper left), a payoff matrix is built to enforce their respective Euclidean constraints (upper right, note that cells associated to many-to-many matches are set to 0) and the replicator dynamics are executed (bottom graph). At the start of the process the population is set around the barycenter (at 0 iterations). This means that initially the vector $\mathbf{x}$ represents a quasi-uniform probability distribution. After a few evolutionary iterations the matching candidate B2 (cyan) is extinct. This is to be expected since it is a clearly wrong correspondence and its payoff with respect to the other strategies is very low (see the payoff matrix). After a few more iterations, strategy A1 vanishes as well. It should be noted that strategies D4/D5 and E6/E7 are mutually exclusive, since they share the same scene vertex. In fact, after an initial plateau, the demise of A1 breaks the tie and finally E6 prevails over E7 and D4 over D5. After just 30 iterations the process stabilizes and only 4 strategies (corresponding to the correct matches) survive.

to it. We take a correspondence-based approach in that a match, if present, is established by means of point-wise correspondences between the two surfaces. Again, this matching process is similar to the surface registration technique presented in Chapter 6. However, both the scope of the methods and their underlying assumptions are quite different; in fact, preliminary experiments demonstrated the inability of the "pure" surface registration algorithm to deal with the strong structured outliers due to clutter, strong occlusions and possible absence of the object from the scene, which are characteristic of the object recognition scenario.

We start by defining the initial set of strategies S, where each reference point in the scene is associated with the k-nearest model points in the descriptor space:

$$S = \{(a,b) \in D \times M | b \in dn_k(a)\}, \tag{7.1}$$

where $dn_k(a)$ is the set of model vertices associated to the k-neighbors of the descriptor of a. This means that each (relevant) sample in the scene is considered to be a possible match with samples in the model that exhibit similar surface characteristics, and we limit the number of "attempts" to k. If the closest model descriptor is deemed too far apart from the data query, the corresponding scene point can be excluded altogether from the matching, so as to operate a form of clutter pre-filtering (although in our experiments we did not perform any filtering of this kind). If the chosen descriptor allows it, using fast search structures such as kd-tree can be beneficial for this step. Note that the direction of the matching is from scene to model; this is motivated by the fact that the scene likely contains only a partial view of the model object, and that originating candidate matches from the scene helps to reduce the false positive rate for equal number of strategies.

Next we define a pairwise compatibility function among the strategies. Since we are interested in finding a correspondence between the model and part of the surface in the scene, we are looking for a subset of candidates that enforce an isometric transformation among the two sets of vertices. Even though we discard connectivity information at this point, we argue that strategies enforcing this isometry constraint are likely to lay on the same surface both in the scene and in the model, and thus to be a viable solution. We define the payoff function $\delta : S \times S \to [0,1]$ as

$$\delta\big({\scriptstyle (a_1,b_1),\,(a_2,b_2)}\big) = \frac{min(\|a_1 - a_2\|, \|b_1 - b_2\|)}{max(\|a_1 - a_2\|, \|b_1 - b_2\|)}. \tag{7.2}$$

This function takes pairs of strategies $(a_1, b_1), (a_2, b_2) \in D \times M$ and gives a reward (a value close to 1) if the corresponding source and destination points are separated by the same Euclidean distance up to positional noise. By contrast, the value of δ will be small when the two strategies exhibit very different distances. This kind of check will succeed with correct pairs and will give false positives only for a small amount of cases, those preserving the isometry constraint by chance. However, since our game is seeking a large group of candidates with large mutual payoff, such outliers will be filtered out with high probability by the other strategies that participate to the Nash equilibrium. This makes for a semi-local approach that guarantees a robust global agreement among mating strategies, while operating at a local level.

Again, equation 7.2 does not guarantee injectivity of the solution; to avoid possible many-to-many matches, we impose a hard constraint by setting to 0 the compatibility between candidates that share the same source or destination vertex. Additionally, we require that the variation in orientation between each pair of data points be maintained on the model. In order to obtain a higher stability in the measurement, we characterize the variation in orientation as the angle between the principal axes of the descriptor frames rather than between the normal vectors computed from the mesh. Thus, the final payoff for the sparse matching game that we are defining is

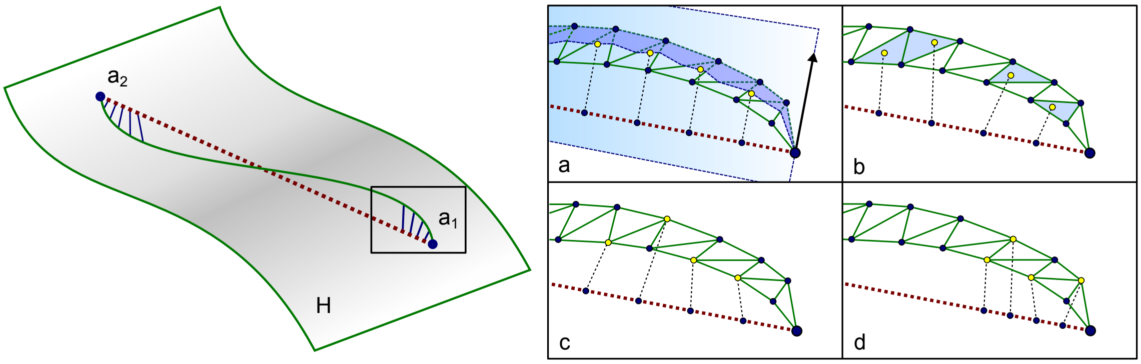


Figure 7.4: Definition of a binary descriptor between mesh points a_1 and a_2 for scale-invariant recognition. The n equally-spaced samples along the (dotted) segment separating the two points can be projected onto the mesh along a specific direction, such as the normal vector at a_1 (a); a minimum-distance projection can instead be computed so as to avoid the choice of a possibly unstable direction and for increased accuracy (b); efficiency can be attained by approximating minimum projections with closest points, which is appropriate in the majority of real cases where sampling density is consistent between model and scene (c); in order to be robust to occlusions, only the first $m \leq n/2$ samples are considered (d).

$$\Pi = \begin{cases} \delta\big((a_1,b_1),\,(a_2,b_2)\big) & \text{if } a_1 \neq a_2 \text{ and } b_1 \neq b_2 \\ 0 & \text{otherwise.} \end{cases} \tag{7.3}$$

Once the candidate set and the payoff matrix are built, the game is started from the barycenter of the simplex. When a stable state is reached, all the strategies supported by a large percentage of the population are considered non-extinct and retained as correct matches (see Figure 7.3). Since convergence is only reached in infinite time, we cannot expect the weakest strategies to be completely extinct at the equilibrium. We address the resulting thresholding problem by selecting only strategies whose population is within a fixed proportion of that of the best strategy. Then, if the total number of surviving matches is more than a fixed minimum (set to 8 in our experiments), the object is recognized and, optionally, its pose computed. Note that, unlike other approaches, we do not run any costly hypothesis verification step by making considerations on the resulting surface overlap. Finally, similarly to Section 6.2, we note that equation 7.2 admits symmetric groups of matches and that reflections are not accounted for a posteriori in the pipeline. Nevertheless, probably due to the strong inlier selection nature of the method and to the lack of perfectly symmetric shapes in the dataset, we never observed mismatches of this sort in all of our experiments. However, in Chapter 8 we will introduce a method that tries to tackle this problem.

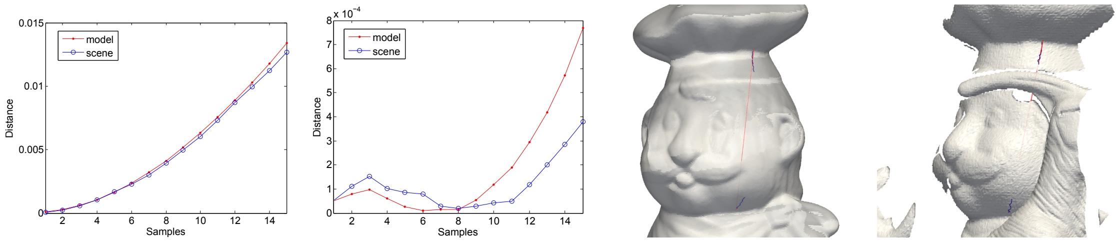


Figure 7.5: Example of two (correct) matching distance sequences as extracted by the scale-independent selection process. The path (red segment in the images on the right) joins the neck of the chef to the hat and has been sampled at a resolution of $n = 100$ and $m = 15$ samples are considered from each endpoint, allowing to mitigate the influence of clutter on the match. The two graphs plot model and scene descriptors originating at the neck and hat endpoints respectively (corresponding to the blue paths in the images). The scene (last image) is rescaled to the same scale of the model for visualization purposes.

7.4 Scale-Invariant Matching

The matching scheme presented in section 7.3 assumes a scenario where model and scene have the same scale (although sampling may be different). This allowed us to devise a game that explicitly enforces pairwise isometries between the two surfaces. In this section we tackle a more general setting by allowing the model object and the scene to have different scales.

It is clear that in an object-in-clutter setting there is no simple way to give a model-data relative scale estimate. For example, considerations on the bounding boxes of the two surfaces have little significance. In fact, basic assumptions on the location and possible presence of the object in the scene involve solving the recognition problem itself. In principle, the game-theoretic framework can be adapted so as to consider triplets rather than pairs of points from model and scene. The change in scale can then be accounted for by introducing triangular distance ratios in the formulation of the payoff function. Using ratios in place of plain distances would eliminate the effect of the scale and then allow to extract isometric-compatible groups in a similar way to the previous game. The resulting 3-way payoff tensor can in fact be used in a higher-order selection process via generalization of the replicator dynamics [116]. Similar techniques have been applied in hypergraph and probabilistic clustering [125]. A major problem with this approach is in its computational complexity, which grows with the third power of the total number of strategies, rendering the game infeasible for medium to large-scale problems; looking at memory usage, an unrealistically simple example with 50 points on both model and scene would produce a 58.2 GB single-precision payoff tensor.

Instead, we wish to fit the scale-invariance property within the current scheme. Our insight is to consider the straight line connecting points in each pair (from the same mesh) and then segment this line into a predetermined number of parts. Being in a rigid setting (up to scale), taking a Euclidean path is appropriate. Then, we construct a pairwise descriptor by enriching each pair of points with geometric information at fixed, equally-spaced steps along the line; since the number of segments into which lines are divided is the same for both model and scene, this effectively removes dependence from scale. There are many ways to characterize the surface along the edge, but thanks to the strongly selective behavior of our framework, we can restrict ourselves to simple measures, without taking into account any additional information that may increase distinctiveness at a computational cost. In Figure 7.4 we illustrate three possible choices. Given n ordered line samples $s_{i=1\dots n}$, we take their projection over the mesh H (Figure 7.4b) and build the sequence of minimum (normalized) distances

$$D_n = (d_1, d_2, \dots, d_n)\,, \tag{7.4}$$

$$d_i = \|s_{i\perp}^H - s_i\| / \|s_1 - s_n\|, \quad i = 1 \dots n\,, \tag{7.5}$$

where $x_\perp^H$ is the minimum-distance projection of point x onto surface H and the denominator acts as a scale normalization term. For efficiency reasons we avoid computing the actual projection and approximate it by taking the nearest mesh point to the given line sample (Figure 7.4c). We can then associate a descriptor vector $P_{ab} \in \mathbb{R}_+^n$ to each pair of points $a, b \in H$, representing the corresponding distance sequence of length n between them. To be robust against clutter and deal with boundary conditions, we only consider the first $m \leq n/2$ samples from each endpoint (Figure 7.4d). The set of strategies S' can be built in a similar way to the isometry-enforcing game, although in this case candidate matches cannot be directly constructed as per equation 7.1, since the descriptors we use are not invariant to scale. To this end, instead of introducing new descriptors, we prefer to rely on gameplay and compute, for each (relevant) point, fixed-scale descriptors at multiple scales; when the game is run, the selection process will operate on the pool of multiple scales and hopefully extract the most (scale-)compatible pairs of strategies. Thus, similarly to equation 7.1, we define the set of strategies as

$$S' = \{(a, b) \in D \times M | a \in dn_k(b)\}. \tag{7.6}$$

Here, descriptors at many different scales are associated with each $a \in D$ and the set of matching strategies will thus include mixed-scale associations between model and scene. We remind that the size of S' depends on parameter k and thus the problem does not necessarily grow in size with respect to the first game; in fact, in Section 7.5 we will use the same k in both games, making the matching step equally efficient in both cases. Next, we define the new payoff function to be

$$\begin{aligned} \rho\big((a_1,b_1),\,(a_2,b_2)\big) &= \frac{1}{2} + \frac{A^T B}{2\|A\|\|B\|}\,, \\ A &= P_{a_1a_2} - \bar{P}_{a_1a_2}\,, \\ B &= P_{b_1b_2} - \bar{P}_{b_1b_2}\,, \end{aligned} \tag{7.7}$$

where $\bar{X}$ denotes the sample mean of X. Although we don't repeat them here, the same hard constraints from the previous game are also applied in this case. Again, the payoff function takes pairs in $S' \times S'$ and gives values in $[0, 1]$; the payoff in this case is a

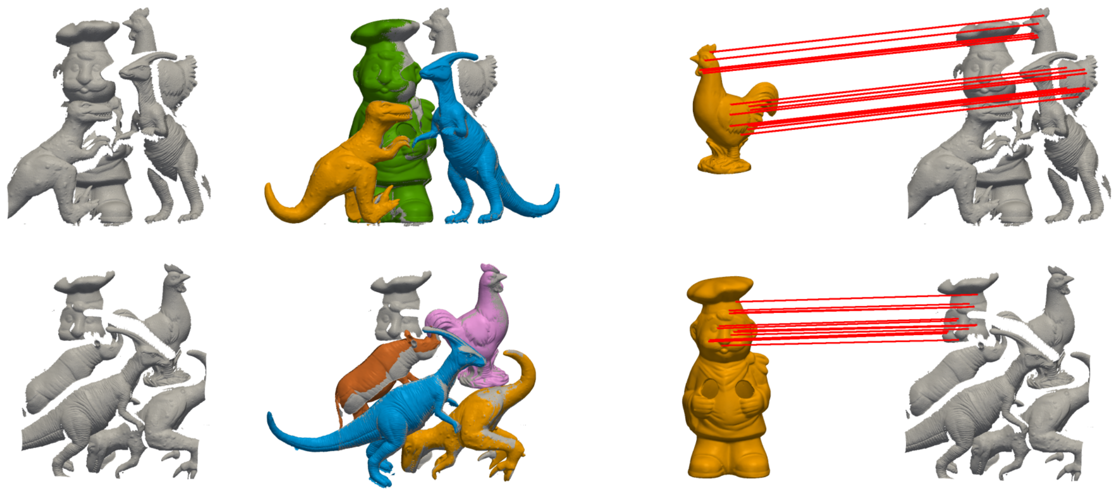


Figure 7.6: Two mismatches generated with the method by Mian et al (first row) and the method by Bariya and Nishino (second row), which are corrected by our technique. The first column shows the range image of the scene, onto which the matched models are successively registered (second column). The chicken and chef models have been missed respectively in the first and second scene, while our method is able to extract correct matches in both cases (third column). The figure is best viewed in color.

normalized inner product reflecting the degree of similarity of the distance sequence for a pair in the model with one in the scene. This new formulation is quite different from equation 7.2 in that we choose to avoid enforcing isometries explicitly, and use instead information coming from the pairwise descriptors alone. Of course, there are other ways in which this information can be used in the definition of a payoff function. In the experimental section we present three alternatives and compare the results obtained with each of them. Figure 7.5 shows an example of two correctly matched strategies and the corresponding distance descriptors.

The scale-invariant game we have just defined favors pairs of matches having compatible distance sequences on each surface, and similar descriptors between the two. While this works well in practice, it can be easily improved by imposing the additional requirement that consistent matches should also give similar estimates of the relative scale between model and scene. We do this by multiplying the payoff function by an additional term favoring pairs with similar scale ratio, expressed as the ratio of the descriptors support radii:

$$\rho'\big((a_1,b_1),(a_2,b_2)\big) = \rho\big((a_1,b_1),(a_2,b_2)\big)e^{-\lambda\left|\frac{\sigma(a_1)}{\sigma(b_1)}-\frac{\sigma(a_2)}{\sigma(b_2)}\right|}, \tag{7.8}$$

with $\sigma(x)$ indicating the support radius of the descriptor at point x, and λ is a parameter regulating the tolerance level for different scale estimates (a small value for λ indicates high tolerance to different scales). This definition enforces the final group of matches to map the model to the scene at consistent scale. In the experimental section we give a quantitative evaluation of what can be obtained with and without the newly introduced

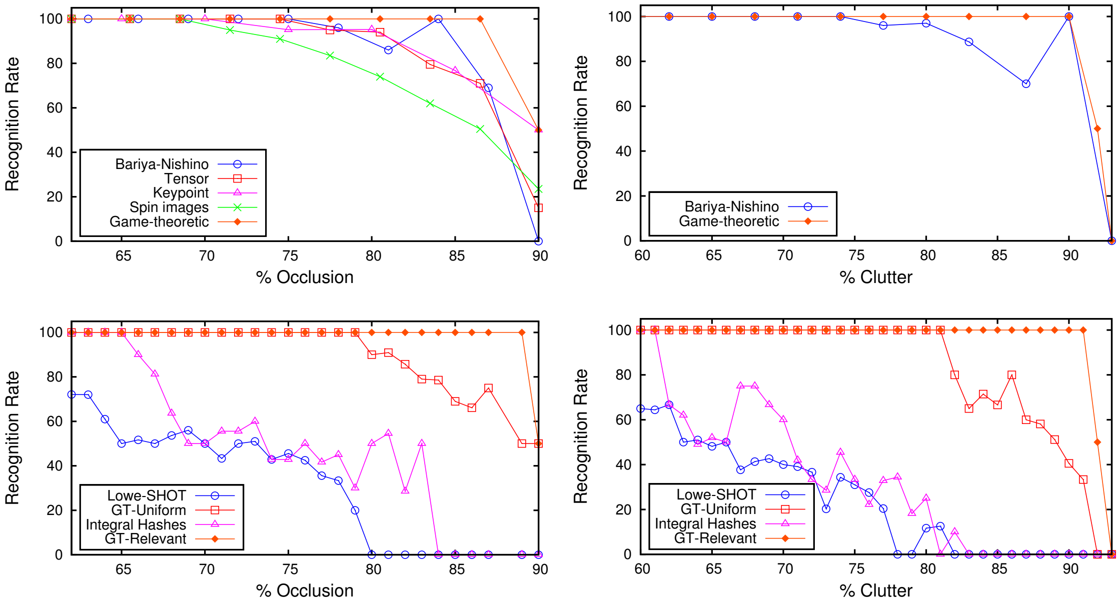


Figure 7.7: In the top row the recognition rate of our pipeline is compared with state-of-the-art techniques, which are outperformed with respect to both occlusion and clutter. In the bottom row the contribution of each part of the overall approach is tested separately (see text for details).

unary term.

Finally, while in our experiments we found that the resolution of the pairwise descriptors (proportional to the number of line samples n) has no significant influence on the matching results, the actual number of samples (from each endpoint) m that take part to the game has a more direct impact. This value can be set to a fixed percentage of n, but this would bring to an imbalance between strategies where spatially close groups of matches become favored. While this could be desirable in certain applications, we aim at a sparse match covering the target object as much as possible. To do this, we first note that, in general, it is not required that P_{xy} has the same number of components for every pair $x, y \in H$. The number of samples m may be different among pairs of points as long as it is the same on model and scene for each pair of strategies. That is, when calculating the payoff between two pure strategies $(a_1, b_1), (a_2, b_2) \in S'$, it must be $P_{a_1 a_2}, P_{b_1 b_2} \in \mathbb{R}^{2m}$, but any such pair may have a different value for m. We determine this value dynamically for each pair of strategies as the number of steps required to reach a fixed distance d (equal for all pairs) on the model mesh, that is, $m = \lceil d/(\|b_1 - b_2\|/n) \rceil$ with $b_1, b_2 \in M$. This allows to obtain spatially sparse correspondences more easily, and thus increase robustness in presence of occlusions and a more stable pose estimate after a solution is found. Quantitative results comparing the adoption of fixed versus adaptive sampling are presented in the experimental section.

After the payoff matrix is constructed, the game is started from the barycenter as in section 7.3 and the final group of matches, if any, is extracted. Each of these correspondences has associated a value representing its relative degree of participation to the final equilibrium, and can be used to compute the similarity transformation linking model and scene in a weighted fashion [74].

7.5 Experimental Results

In order to evaluate the performance of the proposed pipeline we performed a wide range of tests and comparisons with recent techniques. To offer a fair comparison we used the model/scene dataset adopted in [23] and [99, 100]. This dataset is composed of five high resolution models scanned from real objects (chef, dino1, dino2, chicken and rhino), plus 50 range scans of these objects under various conditions of occlusion (due to the overlap of objects and limits on the field of view of the sensor) and clutter (due to the presence of many objects in the scene). The minimum number of matches to assume the model as recognized in the scene was set to 8 for both (fixed scale and scale-invariant) matching games. This value is rather conservative as in general it is very unlikely that outliers form consistent groups of more than a few elements. Also, in both games a value of descriptor neighbors of $k = 5$ was used to build the strategy set; relevant points were detected via Integral Hashes with a scale of $\sigma = 8$ edges and then uniformly subsampled to 3000 points in the data surfaces, while retaining all relevant points in the model surfaces; 10-bins SHOT descriptors were computed at each relevant point ($\sigma = 8$ edges); the angle separating reference axes at scene points was considered maintained in the model with up to 15 degrees of difference; the final solution at the equilibrium was obtained by thresholding the population vector at 50% with respect to the most played strategy. Again, this is very conservative as in theory all the matches having a population share as low as 5% can contribute to the final solution.

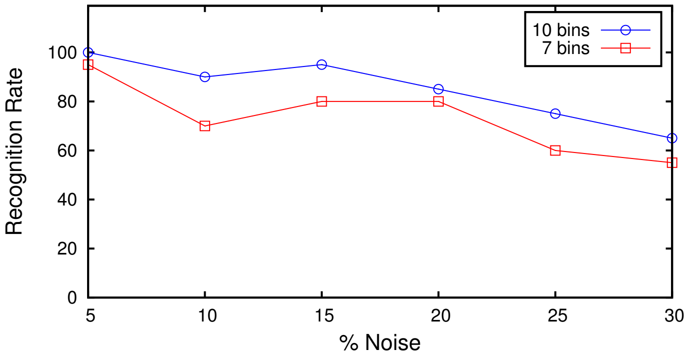


Figure 7.8: Evaluation of the robustness of the proposed pipeline with respect to increasing positional noise applied to the scene.

7.5.1 Comparison with the State-of-the-Art

In Figure 7.7 we compare our results with recent state-of-the-art algorithms (respectively [23] and [99, 100]) and with the well-known 3D Spin Image matching technique [76], which is often used as a baseline in literature. Looking at the recognition rate (defined as the ratio of models correctly labeled as matched/absent over the total number of match scenes) with respect to model occlusion, the proposed pipeline outperforms even the most recent techniques. Regarding the evaluation of the effects of clutter we were only able to compare our algorithm with [23], since an implementation for the other approaches and the data they used were not available. Still, it is apparent that the game-theoretic approach obtains good recognition with uniform performance.

Some examples of critical scenes where the proposed technique fixes matches missed by the other methods are shown in Figure 7.6. The behavior with respect to false positives has not been plotted since the proposed pipeline does not get any in the whole dataset.

In the second row of Figure 7.7 several combinations of components of the pipeline are evaluated one at a time in order to highlight their respective contributions. Note that the plots in these experiments are more dense than before as we have full control over all the algorithms. Specifically, we show the results obtained using the same descriptor [143] with the classical matcher proposed in [89] (Lowe-SHOT), the game-theoretic matcher without operating the initial relevance-based sampling (GT-Uniform), the descriptors and matching presented in Chapter 6 (Integral-Hashes) and finally the full proposed pipeline (GT-Relevant). It is apparent that the proposed pipeline gives its best with all the components in place. This experiment gives us further insight on the specific setting of object recognition as opposed to other matching scenarios, and confirms some expectations anticipated in the previous sections. First, it is clear that descriptors alone, as robust and descriptive as they may be, are hardly sufficient to guarantee correct matches at moderate levels of occlusion and clutter; they are, in fact, surprisingly good at some challenging scenes while they can fail at very simple ones. This can be caused, for instance, by the presence of repeated structure or featureless objects. The matching process presented in the previous chapters proves to be very effective in a wide range of scenes, but its performance worsens rapidly with increasing levels of clutter. This is symptomatic of the different problem scope of the method, which is tailored to a rigid alignment scenario with symmetric assumptions on the roles of model and data meshes. Skipping only the relevant point selection step yields very high performance, on par with the state of the art (compare with the top row). Nevertheless, at severe levels of occlusion and clutter uniform sampling ceases to be effective as it blindly gives equal importance to all surface regions; this has the effect of drastically reducing the inlier ratio in the construction of hypotheses S, which in turn leads to equilibria where wrong correspondences form larger and stronger groups than the (few) correct ones. Applying relevant sampling to the scene is a simple and fast step, and allows us to obtain excellent results without resorting to more sophisticated interest point detection techniques.

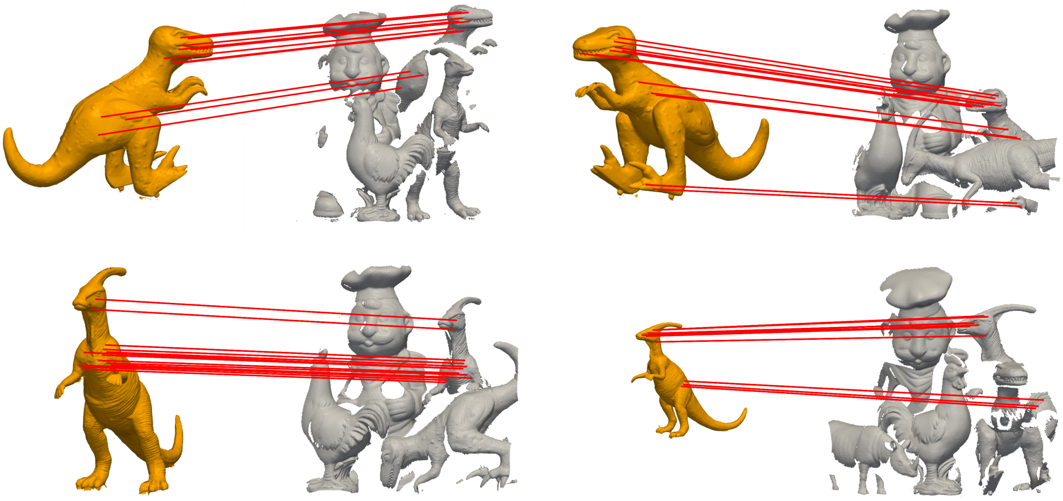


Figure 7.9: Examples of scale-invariant object retrieval from cluttered scenes. The foot of t-rex in the top-right image demonstrates the capability of the method to deal with strong occlusions.

7.5.2 Resilience to Noise

The dataset used in our experiments is made of dense models (300-400k triangles) and slightly less dense scenes produced with a range scanner. Although there is not an exact correspondence between models and scenes, they are rather similar by construction. With the next set of experiments we tried to characterize the performance of the proposed method in presence of positional noise. To do so, we added Gaussian displacement of varying intensity to each vertex in the scenes, and ran the recognition experiments again with the same framework parameters used in the previous evaluations. In order to assess the relative contribution given by descriptors under noisy conditions, we performed this test with two different SHOT parameterizations (the number of bins has a direct effect on resilience to noise, see [143] for details). Figure 7.8 reports the results of this test. As expected, performance gets lower as the noise level increases; still, reasonable recognition rates are maintained also with a moderate amount of noise (with standard deviation equal to 30% the median edge length). Further, the descriptors do not seem to have a significant impact over the results obtained with additional noise, thus suggesting that robustness to noise is for the major part a result of the inlier selection method itself, rather than the specific descriptors used.

7.5.3 Scale Invariance

In this section we evaluate the effectiveness of the scale-invariant scheme using different definitions for the payoff function and under different parameterizations of the pairwise descriptor. We used the same dataset from the previous experiments, where each

	Dot product	L1-norm	L2-norm
None	88.73%	83.10%	84.51%
Exp	92.96%	89.75%	90.14%
Cut-off	**97.18%**	87.32%	91.55%

Table 7.1: Scale-invariant recognition rates under different combinations of payoff functions with a scale term.

scene was randomly scaled from 0.5 to 2.5 times the original scale, and model descriptors spanned over 20 different support radii at each relevant point. All the parameters in common with the isometry-enforcing game are kept at the same values.

The first set of experiments is aimed at determining the best choice for a payoff function. First, we introduce two alternative definitions to ρ (equation 7.7), giving again a similarity measure based solely on pairwise distance descriptors:

$$\rho_{l2}\big((a_1,b_1),\,(a_2,b_2)\big) = e^{-\beta\left\|P_{a_1a_2}-P_{b_1b_2}\right\|_2} \tag{7.9}$$

$$\rho_{l1}\big((a_1,b_1),\,(a_2,b_2)\big) = e^{-\gamma\left\|P_{a_1a_2}-P_{b_1b_2}\right\|_1}\,, \tag{7.10}$$

where $\|\cdot\|_2$ denotes the standard Euclidean norm, $\|\cdot\|_1$ is the L1-norm and parameters β and γ make the functions more or less selective. In our experiments we set $\beta = 1000$ and $\gamma = 1$, values that where empirically seen to yield good results. Note that, after building the strategies set, we do not take into account descriptor information at points a_1, b_1, a_2, b_2 in the definition of the payoff function, although it is certainly possible to introduce another term accounting for their similarity. As in equation 7.8, we wish instead to enforce a common scale mapping by multiplying each payoff function ρ_* by a compatibility term μ based on the local scale of the descriptors:

$$\mu\big((a_1,b_1),\,(a_2,b_2)\big) = e^{-\lambda\left|\frac{\sigma(a_1)}{\sigma(b_1)}-\frac{\sigma(a_2)}{\sigma(b_2)}\right|}. \tag{7.11}$$

Introduction of this term helps the selection process by giving small payoff to unlikely hypotheses and thus bring more stable matches in difficult scenarios, as well as increased efficiency. In the experiments we evaluated all possible combinations of the payoff functions with three variations on the usage of μ. First, we consider no scale enforcement at all ($\lambda = 0$). Then, we increase its steepness ($\lambda = 30$) so as to make μ very selective and give high values to similar relative scales and very small payoffs to different scales. Finally, we use μ as a cut-off function by putting a hard threshold on the value obtained with $\lambda = 30$; in this case, pairs of strategies receiving $\mu < 0.8$ have the corresponding value of the payoff function set to 0. Table 7.1 reports the recognition rates obtained with these different combinations on a reduced dataset spanning many levels of occlusion and clutter. We evaluate payoff functions ρ (Dot product), ρ_{l1} (L1-norm) and ρ_{l2} (L2-norm) with no scale consistency (None), large λ (Exp) and hard thresholding (Cut-off). The best results by far are obtained by cutting-off dissimilar relative scales and weighting the result

	Dot product	L1-norm	L2-norm
None	31.70	42.08	43.41
Exp	19.11	12.45	13.54
Cut-off	23.21	12.93	14.10

Table 7.2: Average number of matches obtained with different payoff functions, using the same parameters as in Table 7.1.

with the inner product of the pairwise descriptors. Scale enforcement is beneficial in all the cases, while L1-norm always gives the worst results. Figure 7.9 shows some examples of matches obtained with the Dot-Cutoff combination. It should be noted, however, that all the reported recognition rates are rather good considering the scale-invariant setting. In fact, we could gain robustness to mesh sampling and thus achieve better results, on average, by computing the pairwise descriptors P_{xy} more accurately and not using the closest mesh point in place of projections, as described in Section 7.4. The average number of matches for each payoff function on the same dataset are reported in Table 7.2. Looking at the reported values, it is apparent that the increased selectivity brought by the additional scale constraints (second and third rows) has a direct influence on the size of the solution at the equilibrium.

The second set of experiments analyzes sensitivity to parameters of the pairwise descriptor, namely its resolution (expressed as the total number of line samples n) and the actual number of samples used in the descriptor ($2m$). For these experiments we used the best payoff function as evaluated in Table 7.1 (ρ' with cut-off). We observed that, while in principle a higher resolution should give better results, in practice the recognition rate is not affected by this parameter: using a small or large value for n (from 10 to 2000 samples) gives the same results on the whole dataset. This is probably due to the fact that, after removing the effect of scale, the game operates a very robust inlier selection in a rigid setting, where a few good hints are sufficient for extracting a consistent group of matches. As a reference, we used $n = 100$ in the following experiments. As described in Section 7.4, we analyzed two different approaches to determine a value for m, and thus the size of the descriptor. In Figure 7.10 we plot their recognition rate against clutter and occlusion on the full dataset. The first approach (Fixed) takes a fixed number of samples for all the pairs, set in our experiments to the first 7% samples from each end. As a result, any P_{xy} on model and scene has only 14 components; this value was determined empirically as the smallest number for which performance does not start to decrease. The second approach (Adaptive) is dynamic and each pair of strategies has the value of m set to the number of steps required on the model mesh to reach a (Euclidean) distance of 8 times the model resolution (calculated as its median edge length). Both methods exhibit remarkable performance at high levels of scene noise, with adaptive sampling giving better results on average. Comparing the results with those in Figure 7.7, we observe that performance of the scale-invariant pipeline is at least as good as the state of the art for fixed-scale recognition on the same dataset.

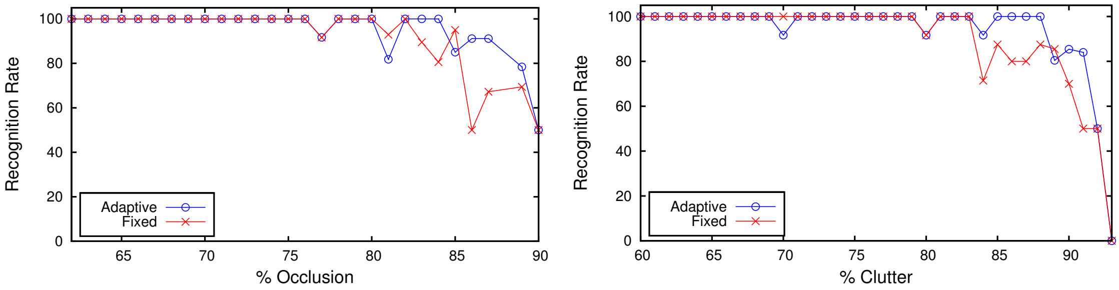


Figure 7.10: Recognition rate of the scale-invariant matching game against occlusion and clutter. The two curves correspond to different ways of determining the descriptor size.

7.5.4 Performance Considerations

In this section we carry out a performance evaluation of the proposed pipeline. We remind that in a typical matching scenario, only a subset of interesting points from model and scene take part to the matching game (see Section 7.3).

Given a payoff matrix and an initial set of candidate correspondences, the selection process is executed by means of evolutionary dynamics. This evolutionary process is iterative in nature and, as such, it is difficult to give an upper bound for its convergence time. In the case of standard, first-order replicator dynamics (as per equation 2.2), the computational complexity of each step is $O(N^2)$, with N being the total number of strategies (*i.e.*, the candidate correspondences). For this reason replicator dynamics are rarely used in practice, even more so for large-scale problems, where the cardinality of the set of strategies can be in the order of thousands even after strong candidate rejection via descriptor priors. A simple alternative to the standard replicator equations is the adaptive exponential replicator model [108], which can be employed in order to drastically reduce the number of iterations for converging to a solution, but still suffers from a per-step quadratic complexity. An even faster alternative is provided by the infection-immunization dynamics [115] (equation 2.6), which has an $O(N)$ complexity for each step; under this model, the time per iteration is only quadratic with respect to the number of mesh points, allowing to reach convergence in 4-5 seconds (around 15,000 iterations) with tens of thousands of strategies. Figure 7.11 reports computational times of the pipeline for the scale-invariant game, using infection-immunization to compute the equilibria. The computation is dominated by the construction of the payoff matrix Π, while the matching step takes only a small fraction of time. It can be seen that the selection process attains an equilibrium within seconds even with thousands of strategies. The experiments were written in C++ and run on a Core i7 machine with 12 GB of memory. Again, we would like to stress that the choice of specific dynamics is driven by efficiency considerations only, and that in general different evolutionary dynamics should converge to similar solutions.

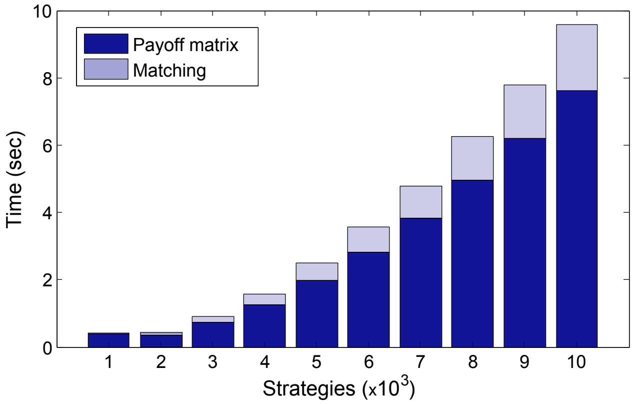


Figure 7.11: Time versus the number of strategies in the scale-invariant setting.

7.6 Conclusions

We presented a novel pipeline for model-based 3D object recognition in cluttered scenes obtained with a range scanner. The pipeline starts with the detection of distinctive keypoints in the scene, which in turn is composed of a relevance filter, a subsampling step and the calculation of a descriptor for each sample kept. These relevant points are then matched pairwise with all the model keypoints and a set of candidate pairings is obtained. Finally, a non-cooperative, isometry-enforcing game is played. The gameplay performs the actual recognition step and returns a sparse set of reliable matches. An additional game is then introduced to tackle the more challenging recognition problem where model and scene are allowed to take different scales. To this end, a novel pairwise strategy descriptor utilizing geometric information along the Euclidean path linking surface points is adopted. The scale mapping is further enforced by computing local descriptors at different scales and putting them in the pool of candidate matches, thus letting the selection process extract the most compatible group of correspondences. The two matching approaches fit within the same general framework, and are extensively evaluated through a wide range of experiments under different conditions. The results demonstrate that the proposed pipeline outperforms the most recent state-of-the-art techniques on the same dataset.

8
Non-rigid Shape Matching

In this Chapter we consider the problem of minimum distortion intrinsic correspondence between deformable shapes, many useful formulations of which give rise to the NP-hard quadratic assignment problem (QAP). Previous attempts to use the spectral relaxation have had limited success due to the lack of sparsity of the obtained “fuzzy” solution. Following the previous chapters, we adopt the game-theoretic framework but take a different point of view in which we regard it as a L^1 relaxation of the QAP. We relate it to the Gromov and Lipschitz metrics between metric spaces and demonstrate on state-of-the-art benchmarks that the proposed approach is capable of finding very accurate sparse correspondences between deformable shapes.

8.1 Introduction

A particularly challenging family of matching problems is *deformable* shape correspondence, in which shapes may undergo non-rigid deformations under which the correspondence has to be invariant. In the past decade, significant attention has been devoted to problems related to deformable shape correspondence. A large corpus of research makes use of the notion of intrinsic geometry – an umbrella term referring to geometric structures that remain invariant under non-rigid bendings and other types of transformations. In [56, 98, 34, 97, 141] and followup studies, it was proposed to use the distortion of intrinsic metrics as a measure of the correspondence quality. Finding a minimum distortion correspondence can be rigorously formulated in geometric terms and cast as an optimization problem. Several particularly useful instances of minimum distortion correspondence problems can be reduced to quadratic assignment problems (QAP). However, the combinatorial nature of QAPs makes them challenging computationally.

Different relaxations of the NP-hard QAP have been explored in the computer vision literature. In their seminal work, Gold and Rangarajan [65] relax the assignment and solve the optimization problem through a gradient method over the set of bistochastic matrices. In [85], a spectral approach to correspondence finding was presented. The authors regard pointwise assignments as nodes of an undirected graph, whose edges represent the agreement between pairs of such potential matches and cast the correspondence problem into an eigenvector problem over the (weighted) adjacency matrix. Since the cor-

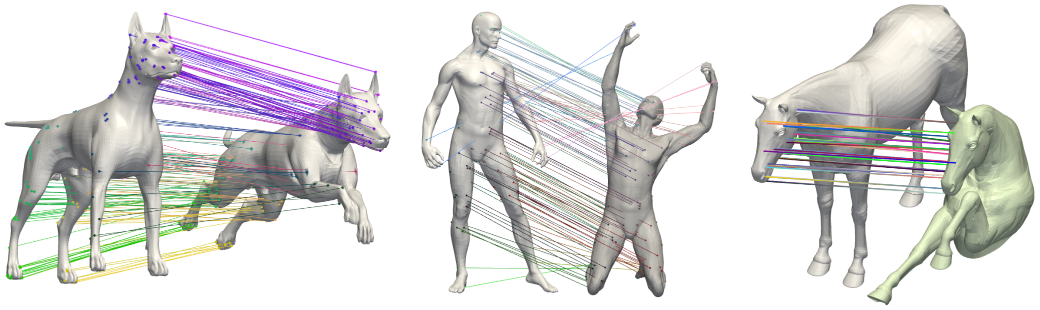


Figure 8.1: Examples of correspondences obtained with our method. The game-theoretic approach produces a sparse (around 1% of the shape is matched), yet very accurate correspondence which can be used as a robust initialization for subsequent refinement (first two images). The last image presents a case of partial matching, where the second shape additionally underwent a local scale deformation. In this case we applied the merging approach on 5 groups, resulting in 51 matches with an average ground-truth error of 2.57 (see Section 8.4.1).

rect correspondences are likely to form a strongly connected cluster, the authors build a symmetric weighted adjacency matrix $\mathbf{A}$ of the graph and try to solve for the set of assignments (represented by an indicator vector $\mathbf{u}$) maximizing the quadratic inter-cluster score $\mathbf{u}^{\mathrm{T}}\mathbf{A}\mathbf{u}$. Optimization is performed by relaxing the binary constraint $\mathbf{u} \in \{0,1\}^n$ and allowing $\mathbf{u}$ to take continuous values; then, by Rayleigh's quotient theorem, the solution $\mathbf{u}^*$ is the principal eigenvector of $\mathbf{A}$ and has unit L^2 norm. Mapping constraints are finally met by iteratively removing inconsistent or weak assignments until an optimum binarized solution is found. The procedure has been successfully applied to 2D matching and recognition, and subsequently extended to other contexts such as isometry-enforcing 3D nonrigid matching [107]; more recently, the technique has been generalized to higher order hypergraph matching [54].

In Chapters 5, 6 and 7 we have been considering a similar setup, with a cardinal difference of replacing the L^2 constraint $\mathbf{u}^{\mathrm{T}}\mathbf{u} = 1$ by $\mathbf{u}^{\mathrm{T}}\mathbf{1} = 1$, $\mathbf{u} \geq \mathbf{0}$. This modification makes the assignment problem more combinatorial in nature, and, like most types of L^1 constraints, promotes sparsity of the solution. In what follows, we show an interpretation of the QAPs commonly used in shape matching from the point of view of Gromov and Lipschitz distances between metric spaces. Then, we adapt the game-theoretic framework to efficiently solve the resulting optimization problems, and again show the relation between different heuristics used by this framework to distances between metric spaces. Finally, we propose a method to aggregate multiple sparse solutions obtained using the game-theoretic solver into a denser correspondence. Though the proposed approaches are general and work with any intrinsic distances, in this study we focus on the family of diffusion distances that has a natural scale-space interpretation, and show how to aggregate information from different scales into a single distortion functional.

8.2 Intrinsic geometries

We model shapes as compact smooth Riemannian manifolds equipped with an *intrinsic metric* d and the standard measure induced by the volume form. By the term intrinsic metric we refer to a distance function on the manifold that depends only on the Riemannian structure and is independent of the way it is embedded in the ambient space. One of the straightforward constructions of an intrinsic metric is the geodesic metric measuring the length of the shortest path (minimal geodesic) connecting a pair of points on the surface. Such a metric is invariant to inelastic bendings, that is, such deformations that do not stretch or tear the shape. A serious disadvantage of the geodesic geometry is its extreme sensitivity to topological noise. In fact, even a point topological change has a great influence on the length of the shortest path. Generally, this influence does not decay as one goes away from the affected point, limiting the practical applicability of geodesic distances.

A partial remedy to this problem has been found in another family of intrinsic geometries introduced by Coifman and Lafon [49] under the name of *diffusion geometry*. Diffusion geometry is an umbrella term referring to intrinsic distances and other geometric quantities based on the properties of diffusion processes on the surface. Diffusion processes are described by the heat equation

$$\Delta f(x,t) + \tfrac{\partial}{\partial t} f(x,t) \;=\; 0, \tag{8.1}$$

with $f(x,t)$ denoting the distribution of heat on the surface at point x at time t, and Δ being the Laplace-Beltrami operator. The equation has the initial condition $f(x,t=0)$ describing the initial heat distribution; boundary conditions apply in case the manifold has a boundary.

The solution of the heat equation at point x at time t initialized with a point distribution at x' is called the *heat kernel* and is denoted by $h_t(x,x')$. The heat kernel describes the proximity of two points x and x' at different scales t. This notion of proximity can be used to define a family of intrinsic metrics

$$d_t^2(x,x') \;=\; \int_X (h_t(x,y) - h_t(x',y))^2 dy \tag{8.2}$$

called the *diffusion metrics*. The family is parameterized by the scale parameter t and naturally forms a scale space: diffusion metric with small t are sensitive to small features while being rather indiscriminative at larger scale; on the contrary, d_t with large values of t is insensitive to small features, yet captures the global geometry of the shape.

In order to make diffusion distances commensurable and comparable across different scales, they are often normalized by the trace of the heat kernel (the heat trace),

$$H_{X,t} \;=\; \frac{1}{\mathrm{Vol}(X)} \int_X h_t(x,x)dx, \tag{8.3}$$

where $\mathrm{Vol}(X)$ stands for the total area of X. This results in the family of normalized metrics,

$$\hat{d}^2_{X,t}(x,x') = \frac{d^2_{X,t}(x,x')}{H_{X,t}}. \tag{8.4}$$

The framework of diffusion geometry also allows to define intrinsic point-wise feature descriptors (or signatures) on the surface. In [136], it was shown that under mild technical assumptions, the diagonal $\{h_t(x,x)\}_{t>0}$ of the heat kernel contains full information about the shape's intrinsic geometry (i.e., fully describes the underlying Riemannian structure). The authors proposed to associate each point of the surface with a vector-valued descriptor $\mathbf{h}(x) = (h_{t_1}(x,x),\ldots,h_{t_k}(x,x))$, dubbed as the *heat kernel signature* (HKS). A scale-invariant version of the HKS (SIHKS) was consequently introduced in [36]. In [22], the authors proposed to study the solutions of the Schrödinger equation in lieu of the heat equation arriving at the *wave kernel signature* (WKS) claiming better feature localization. Since the Laplace-Beltrami operator is an intrinsic property of the shape, quantities associated with it such as the heat kernel and descriptors based on it are also intrinsic. Being constructed from the same geometric quantities, both diffusion metrics and corresponding signatures are related in that nearly isometric shapes in the sense of the diffusion metrics will also be described by similar HKS and vice versa.

In what follows, we are going to use diffusion geometric quantities to formalize the notion of correspondence between shapes. Most of the presented discussion is however valid for any type of intrinsic metrics.

8.3 Intrinsic shape correspondence

We define a *correspondence* between two shapes X and Y as the subset $U \subset X \times Y$ satisfying: 1) for every $x \in X$, there exists (at least one) $y \in Y$ such that $(x,y) \in U$; and, vice versa, 2) for every $y \in Y$, there exists $x \in X$ such that $(x,y) \in U$. This relation can be thought of as a generalization of the notion of a function, and can be alternatively formulated as the binary function $u : X \times Y \to \{0,1\}$ satisfying for every $x \in X$ and $y \in Y$,

$$\max_{y\in Y} u(x,y) = \max_{x\in X} u(x,y) = 1. \tag{8.5}$$

Suppose two pairs of points (x,y) and (x',y') are in correspondence. Then, we can quantify the quality of the correspondence by measuring to which extent the distance between x and x' measured on X using d_X matches the distance between the corresponding points y and y' measured on Y using d_Y,

$$\epsilon(x,y,x',y') = |d_X(x,x') - d_Y(y,y')|. \tag{8.6}$$

The worst-case distortion of the metric caused by the correspondence U is given by

$$\|\epsilon\|_{L^\infty(U\times U)} = \sup_{(x,y),(x',y')\in U} \epsilon(x,y,x',y'). \tag{8.7}$$

Minimizing the distortion over all possible correspondences between X and Y yields a distance

$$D(X,Y) = \frac{1}{2}\inf_U \|\epsilon\|_{L^\infty(U\times U)} \tag{8.8}$$

between X and Y called the *Gromov-Hausdorff distance*. If the infimum is realized by some U^*, the latter is called a *minimum distortion correspondence* (note that more than one minimum distortion correspondence might exist if the shape possesses intrinsic symmetries). By using intrinsic metrics d_X and d_Y, the obtained correspondence is also intrinsic. In particular, this implies invariance to inelastic bending of the shapes.

It is worthwhile noting that taking the logarithm of the metrics d_X, d_Y, one can replace the absolute distortion (8.6) with a relative counterpart

$$\begin{aligned}\epsilon(x,y,x',y') &= |\log d_X(x,x') - \log d_Y(y,y')| \\ &= \log\max\left\{\frac{d_X(x,x')}{d_Y(y,y')}, \frac{d_Y(y,y')}{d_X(x,x')}\right\}.\end{aligned} \tag{8.9}$$

The resulting distance (8.8) is called the *Lipschitz distance*. Note that $\epsilon(x,y,x',y') = \infty$ whenever $x = x'$ or $y = y'$, requiring the correspondence u to be bijective. For this reason, the Lipschitz distance is only applicable to topologically equivalent shapes.

Both the Gromov-Hausdorff and the Lipschitz distances constitute a metric on the space of all (homeomorphic in case of the Lipschitz metric) shapes modulo their d-isometries. They naturally express the similarity relation of two shapes being "approximately isometric", and can be consistently discretized [34]. However, the L^∞ formulation makes the Gromov-Hausdorff and the Lipschitz distances of little practical use due to their sensitivity to noise and outliers.

While an L^p relaxation of the distortion (8.7) would theoretically yield a more robust distance, its direct introduction into (8.8) results in a distance inconsistent to sampling. A way to overcome this difficulty was proposed by [97]. We first relax the binary notion of correspondence into a *fuzzy* notion allowing the function u to assume a continuum of values between 0 and 1, $u : X \times Y \to [0,1]$. Condition (8.5) is relaxed by demanding for every measurable subsets $A \subseteq X$ and $B \subseteq Y$,

$$\begin{aligned}\int_A\int_Y u(x,y)dydx &= \int_A dx; \\ \int_B\int_X u(x,y)dxdy &= \int_B dy.\end{aligned} \tag{8.10}$$

In other words, $u(x,y)dxdy$ defines a weighted product measure on $X\times Y$ whose marginals are the measures dx and dy on X and Y, respectively. The quantity $u(x,y)dx$ can be thought of as the infinitesimal amount of mass transported from point x on X to point y on Y, while ϵ^p quantifies the cost of the transport.

Using this relaxed notion of correspondence, a new family of distances can be defined as

$$D(X,Y) = \frac{1}{2}\inf_u \|\epsilon\|_{L^p(u\times u)}, \tag{8.11}$$

where $1 \leq p \leq \infty$, and

$$\|\epsilon\|^p_{L^p(u\times u)} = \int_{(X\times Y)^2} \epsilon^p(x,y,x',y')u(x,y)u(x',y')dxdydx'dy'. \tag{8.12}$$

$D(X,Y)$ constitute metrics on the space of equivalence classes of shapes under the isomorphism relation of metric-measure spaces (i.e., measure-preserving isometries). In literature, this class of metrics is usually referred to as Wasserstein or earth mover's distances. Here, following [97] we will refer to them as the Gromov-Wasserstein metrics to emphasize the relation to the Gromov-Hausdorff distances. We note, however, that the two metrics are not equivalent, for the very same reasons the Hausdorff and the earth mover's metrics are not equivalent.

8.3.1 Multi-scale distortion

In the particular case where diffusion metrics are used to measure distances on X and Y, the selection of the scale parameter is crucial. Small scales alone give excellent feature localization (and hence accurate correspondence), but are not robust globally; on the other hand, large scales alone do not give accurate correspondences, while stabilize the global matching problem. Here, instead of selecting a single scale, we propose to combine several scales into a single distortion criterion,

$$\epsilon^p(x,y,x',y') = \int_{T_1}^{T_2} \left(\hat{d}_{X,t}(x,x') - \hat{d}_{Y,t}(y,y')\right)^p dt, \tag{8.13}$$

where T_1 and T_2 are parameters determining the range of scales, and $\hat{d}_{X,t}$ are the scale-normalized diffusion distances. Aggregation of multiple scales of spectral distances has been previously successfully used in shape retrieval applications [97, 35].

8.3.2 Discretization

In the discrete setting, let us assume the shapes X and Y to be represented by m and n points, respectively, with the vectors $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$ discretizing the corresponding area elements. The Gromov-Wasserstein metric assumes the form

$$D(X,Y) = \frac{1}{2}\min_{\mathbf{U}} \sum_{i,j,i',j'} \epsilon^p_{iji'j'}\mu_i\nu_j\mu_{i'}\nu_{j'}u_{ij}u_{i'j'}. \tag{8.14}$$

Absorbing the area elements into the cost term and using matrix notation, we arrive at the following quadratic program

$$\min_{\mathbf{U}\geq\mathbf{0}} \mathrm{vec}\{\mathbf{U}\}^{\mathrm{T}}\mathbf{B}\mathrm{vec}\{\mathbf{U}\} \quad \text{s.t} \quad \begin{cases} \mathbf{U1} = \mathbf{1} \\ \mathbf{U}^{\mathrm{T}}\mathbf{1} = \mathbf{1} \end{cases} \tag{8.15}$$

where $\text{vec}\{\mathbf{U}\}$ stands for the mn-dimensional column-stack vector representation of the $m \times n$ correspondence matrix $\mathbf{U}$, $\mathbf{1}$ is a vector of ones of appropriate dimensions, and $\mathbf{B}$ is the $mn \times mn$ cost matrix containing the elements $\epsilon^p_{iji'j'}\mu_i\nu_j\mu_{i'}\nu_{j'}$.

Since our final goal lies in finding the minimum-distortion correspondence rather than computing the Gromov-Wasserstein metric, we are interested in a minimizer rather than a minimum of the above problem. We observe that while the L^1-type constraints are known to favor a sparse solution (i.e., $\mathbf{U}$ will have few strong non-zero elements), it is still a fuzzy correspondence matrix. This may be disadvantageous in matching applications, where usually bijectivity is required. In order to impose bijectivity of the solution, we modify the cost function by setting $\epsilon_{iji'j'} = \infty$ for every $i = i'$ or $j = j'$, exactly as we did in the case of the Lipschitz metric. We denote the modified cost matrix by $\tilde{\mathbf{B}}$.

Finally, observe that the constraints on row and column sums of $\mathbf{U}$ in (8.15) require it to be a full correspondence (i.e., each point in X corresponds to a point in Y). This is rather a restrictive setting for many applications where a *partial* rather than full correspondence is sought. In order to allow for some points on X to have no corresponding points on Y and vice versa, one has to allow some of the rows or columns of $\mathbf{U}$ to sum to zero. We propose to replace problem (8.15) by an under-constrained counterpart

$$\min_{\mathbf{U}\geq\mathbf{0}} \text{vec}\{\mathbf{U}\}^{\mathrm{T}}\tilde{\mathbf{B}}\text{vec}\{\mathbf{U}\} \quad \text{s.t} \quad \mathbf{1}^{\mathrm{T}}\mathbf{U}\mathbf{1} = 1. \tag{8.16}$$

Note that the obtained partial correspondence is bijective by virtue of the modified cost matrix $\tilde{\mathbf{B}}$. In what follows, we show how to efficiently solve the above optimization problem using tools from game theory.

8.4 Game-theoretic matching

Following the previous chapters, we cast the optimization problem in an evolutionary game-theoretic framework. We start by modeling strategies as candidate assignments $(x, y) \in X \times Y$ based on some measure of pointwise similarity among the surface points. Here we use SIHKS [36] descriptors with the standard Euclidean distance since they demonstrate good resilience to a variety of deformations. We emphasize though that the descriptor need not be extremely robust given the strongly selective behavior of the method, and that this step has the intended effect of reducing the size of the problem and increase the inlier ratio.

We simplify the notation and formulate program (8.16) as a maximization problem

$$\max \mathbf{u}^{\mathrm{T}}\mathbf{A}\mathbf{u} \quad \text{s.t} \quad \mathbf{u} \in \Delta \tag{8.17}$$

where $\mathbf{u} \equiv \text{vec}\{\mathbf{U}\}$ is the correspondence vector, constrained to lie in the standard mn-simplex

$$\Delta = \{\mathbf{u} \in \mathbb{R}^{mn} \ : \ \mathbf{u}^{\mathrm{T}}\mathbf{1} = 1 \text{ and } \mathbf{u} \geq 0\}$$

and $\mathbf{A}$ is a $mn \times mn$ matrix whose elements represent the *similarity* between corresponding pairs of correspondences. Such a quantity is inversely related to distortion and can be defined in a variety of ways. In this work we follow [65] and adopt a softmax ansatz to the (relaxed) QAP, which is known to improve the convergence properties of gradient methods. Choosing $p = 2$ in the distortion term, we set $a_{(ij)(i'j')} = \exp(-\alpha\epsilon^2_{iji'j'})$, which incidentally gives $a_{(ij)(i'j')} = 0$ whenever either $i = i'$ or $j = j'$. While here we are using a Gromov-Wasserstein metric, it is worth noting that in Chapter 6 the rigid correspondence problem was solved using the equivalent of a Lipschitz metric.

We remind from the previous chapters that the bijectivity constraints imposed on $\mathbf{A}$ are guaranteed since a stable state cannot have in its support pairs of strategies with zero payoff. Thus, the matching game can be initialized by putting $\mathbf{u}(0)$ on the barycenter of Δ, and then iteratively updated via evolutionary dynamics until convergence. The final iterate $\mathbf{u}(t^*)$ at the equilibrium constitutes a L^1 solution to (8.17). We note that the final values $u_i = u(x, y)$ can be interpreted as the relative contribution of each strategy to the global coherence of the correspondence, in terms of the distortion measure ϵ^2.

The correspondence function u can then be binarized by keeping only the fittest strategies, e.g. by setting $u(x, y) = 1$ for the top 80% strategies (with respect to the maximum u_i), and putting the others to zero. In the experimental section, a specific set of experiments analyzing the influence of this parameter on the quality of the final match is presented.

8.4.1 Merging correspondences

The final correspondence resulting from the local maximization of (8.17) is characterized by a very strong internal coherence, and typically includes only a small percent (around 5-10%) of matches selected from the initial set of candidates. There exist effective methods to render correspondences denser [153]. Here we repeat application of the game-theoretic scheme in an attempt to "densify" the initial correspondence. This iterative approach is justified by the fact that the extinct strategies of a single game (those not supported by $\mathbf{u}^*$) do not necessarily have a smaller payoff than the extracted (local) maximum, thus motivating the interest to explore the solution space further. A similar approach was followed in Chapter 4 for iteratively removing clusters of similar features in an attempt to select interesting points on a shape.

After an initial solution is obtained, we proceed by invalidating the selected strategies from the set of candidates and play a new (smaller) game with the remaining matches. Once several sets of correspondences are extracted, we need a way to merge correspondences in a manner coherent with the possible intrinsic symmetries. We take the hint from spectral clustering [105] and blended intrinsic maps [83], and formalize this notion of coherence by defining a pairwise measure of distortion between groups of matches, and successively operate on the resulting affinity matrix. Let G and H be two correspondence groups $(g_i, g'_i) \in G$ and $(h_j, h'_j) \in H$. We define distortion ζ as:

$$\zeta(G, H) = \frac{1}{mn} \sum_{i,j} w_{g_i g'_i} w_{h_j h'_j} (d_X(g_i, h_j) - d_Y(g'_i, h'_j))^2 \,, \tag{8.18}$$

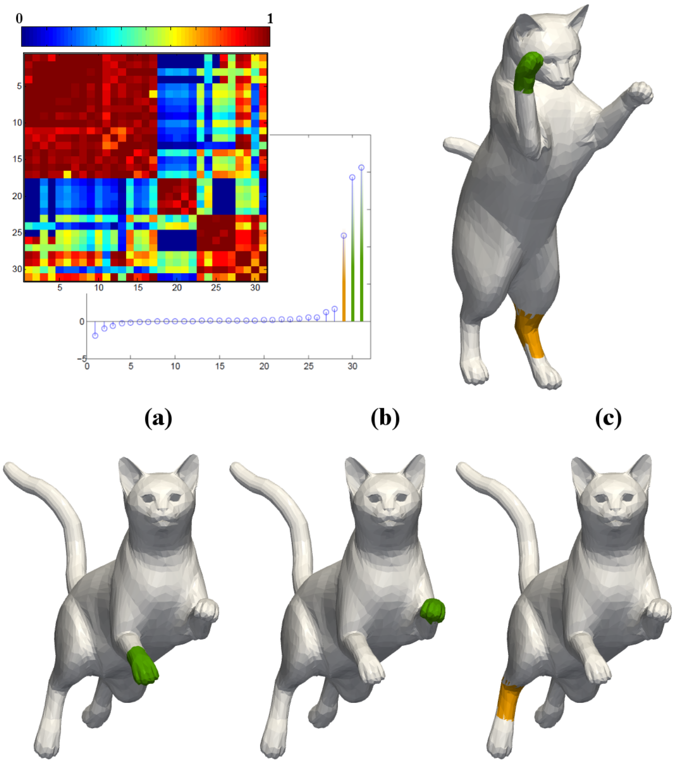


Figure 8.2: An example of the merging process (with real data) between two isometric shapes. After obtaining 30 correspondences, we compute the spectrum of matrix $\mathbf{S}$ (top left). The dominant eigenvector allows to retrieve the most consistent cluster of correspondences, matching the right paw of the cat (in green) (a); the next eigenvalue is only separated by a very small spectral gap, and the corresponding matches associate the right paw again with a symmetric patch (b); finally, the maximum gap eigenvector represents a reflected correspondence (in orange) with larger error (c). Figure best viewed in color.

where the w_{xy} are proper weights proportional to the point-to-point matching confidence between $x \in X$ and $y \in Y$ (for instance, $w_{xy} = u(x, y)$ before binarization). From this we define the corresponding similarity $\Gamma(G, H) = \exp\left(-\gamma\zeta(G, H)\right)$, where γ is a scale parameter.

If we play the game k times, we get to the definition of a (non-negative) similarity matrix $\mathbf{S} \in \mathbb{R}^{k \times k}$. The best group separation can then be represented by a selection vector $\mathbf{y}$, which (similarly to [85, 83]) we relax to take continuous values and constrain to have unitary L^2-norm. We get to the quadratic program

$$\max\ \mathbf{y}^{\mathrm{T}}\mathbf{S}\mathbf{y} \quad \text{s.t} \quad \|\mathbf{y}\|_{L^2}^2 = 1, \tag{8.19}$$

which is maximized by the leading eigenvector of $\mathbf{S}$. In presence of intrinsic symmetries, program (8.19) will yield a large energy value for more than one choice of $\mathbf{y}$, corresponding to different groups of coherent matches separated by a small spectral gap

Transform.	**1**	**≤2**	**≤3**	**≤4**	**≤5**
Isometry	1.47	1.73	6.83	1.77	0.68
Topology	2.45	1.05	3.29	14.70	11.64
Holes	3.93	3.87	3.88	7.44	22.69
Micro holes	1.09	2.59	3.70	2.34	2.87
Scale	4.01	0.81	2.11	9.54	47.99
Local scale	2.64	9.12	8.50	8.15	8.57
Sampling	1.19	2.56	11.84	8.72	20.25
Noise	3.74	4.34	8.63	10.72	12.22
Shot noise	1.46	1.06	1.09	2.06	14.43
Average	2.44	3.01	5.54	7.27	15.70

Table 8.1: Performance of the game-theoretic method using SIHKS and the diffusion metric. Average number of corresponding points is 10.

$|\mathbf{y}^{\mathrm{T}}\mathbf{S}\mathbf{y} - \tilde{\mathbf{y}}^{\mathrm{T}}\mathbf{S}\tilde{\mathbf{y}}|$ (see Figure 8.2). This provides us with a robust means to separate symmetric solutions into distinct consistent sets, while at the same time helps to filter out distorted matches that might occur as the game is repeated.

8.5 Experimental results

We performed a wide range of experiments on the SHREC'10 correspondence dataset [33], for which ground-truth assignments were made available by the authors. The dataset consists of 3 high-resolution (10K-50K vertices) shape classes (human, dog, horse) with simulated transformations, which are split into 9 classes: isometry, topology, small and big holes, global and local scaling, noise, shot noise, sampling. Each transformation class appears in five different strength levels, making for a total of 45 transformations per shape class. When we compute the ground-truth error of correspondence U, we take into consideration reflection intrinsic symmetries by evaluating both the direct and symmetric errors [33]:

$$D(U,U_g) = \frac{1}{|U|}\min\left\{\sum_{k=1}^{|U|} d_X(x_k, x'_k), \sum_{k=1}^{|U|} d_X(x_k, x''_k)\right\},$$

where d_X is a geodesic metric on X and $x'_k, x''_k \in U_g$ are, respectively, the direct and symmetric ground-truth positions of point $x_k \in U$.

8.5.1 Comparisons

We evaluate the performance of the game-theoretic method in relation to existing techniques. Table 8.1 reports per-deformation results at all strengths, which can be directly

Transform.	**1**	**≤2**	**≤3**	**≤4**	**≤5**
Isometry	9.82	15.97	3.28	7.52	3.26
Topology	3.44	3.80	3.03	8.81	4.73
Holes	31.80	18.13	13.49	8.07	49.88
Micro holes	8.61	4.90	3.44	6.98	3.38
Scale	11.76	6.53	8.75	8.70	3.17
Local scale	6.89	15.11	13.00	58.76	50.50
Sampling	6.93	26.55	40.81	13.20	16.06
Noise	6.46	7.81	9.47	11.06	18.34
Shot noise	6.77	13.82	10.28	6.06	15.03
Average	10.28	12.51	11.73	14.35	18.26

Table 8.2: Results obtained after merging the correspondences gathered from 25 games. Average number of corresponding points is 50.

compared with state-of-the-art methods in [33]. Here we used the best parameters determined through the sensitivity analysis that will be presented in the section. The table shows that the proposed approach provides better accuracy than all of the sparse approaches reported in [33], regardless of transformation class. Further, we achieve near-ideal performance in a number of cases. An interesting instance of surprisingly good behavior is represented by the local scale class, which seems to perform equally well at increasing intensities. This is due to the selective nature of the evolutionary process, which explicitly seeks for the most compatible group of matches in terms of preservation of the metric; in this case, the parts of shape that undergo a local change in scale are filtered out by the selection process, naturally favoring those portions of surface that are mostly left untouched by the transformation. By contrast, due to the multi-scale approach followed by our method, global rescaling of the shapes can easily pose problems (compare also with the “scale” curve in Figure 8.4).

The only approach that provides better accuracy in some instances is the spectral matching algorithm, which also provides a dense correspondence. Note, however, that this approach completely breaks for all topology-modifying transformation classes, i.e., topology, holes, and sampling. On the other hand, our performance is close to that of the spectral matching algorithm for the topology-preserving transformation classes, but is also robust with respect to topology-modifying classes.

We also investigated the effectiveness of the correspondence merging approach presented in section 8.4.1. For this test, we iteratively generated 25 groups of matches (for each pair of shapes), built the similarity matrix with $\gamma = 10^8$ and kept the principal eigenvector by thresholding it at 60% of its maximum value. Again, the experiments were carried out on the whole dataset and can be compared with the results published in [33] (a direct comparison can be made with GMDS since it gives the same average number of matches).

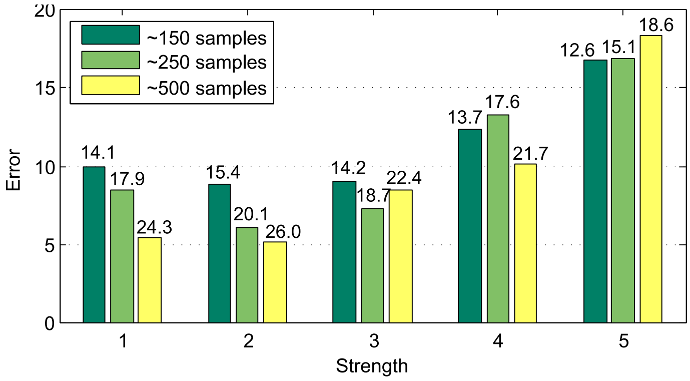


Figure 8.3: Evaluation of the results obtained under different initial samplings of the transformed mesh, averaged over all deformations of every shape. The initial number of samples has a direct and consistent influence on the final size of the correspondence (noted above each bar), whereas its quality does not appear to be affected at all deformation strengths.

8.5.2 Sensitivity analysis

The next set of experiments is aimed at analyzing performance of the game-theoretic method under different parameterizations. Similarly to Chapters 6 and 7, in order to limit the size of the problem, we only consider a subset of points from the deformed mesh X. Feature points are detected by computing for all $x \in X$ the HKS function $h_t(x, x)$ for 3 values of t, and keeping points that are 2-ring local maxima across all time scales [136]. The set of strategies is finally built by generating 5 candidate matches per feature point, based on the vicinity of the associated descriptors with points from the model mesh. Finally, diffusion distances in equation (8.13) were calculated at time scales $(2^7, 2^8, \ldots, 2^{16})$.

Figure 8.4 shows the results obtained by our method with different choices of payoff coefficient α and of the selection threshold used to determine the final set of matches. We used a threshold of 0.8 for the former experiment, and $\alpha = 10^3$ for the latter. The value of α in these graphs ranges over 50 equally spaced values from 10^3 to 36×10^3. Next, since both the size and quality of the correspondence also depend on the specific set of strategies used, we performed some additional tests with a progressively less aggressive feature detection on the data meshes (Figure 8.3). The outcome of this experiment suggests that increasing the number of initial samples can be beneficial to the matching process; indeed, settling for a selectivity level in the feature detection step is more a matter of memory consumption, while the algorithm is able to extract correspondences in 0.5-4 seconds even with large games with tens of thousands of strategies.

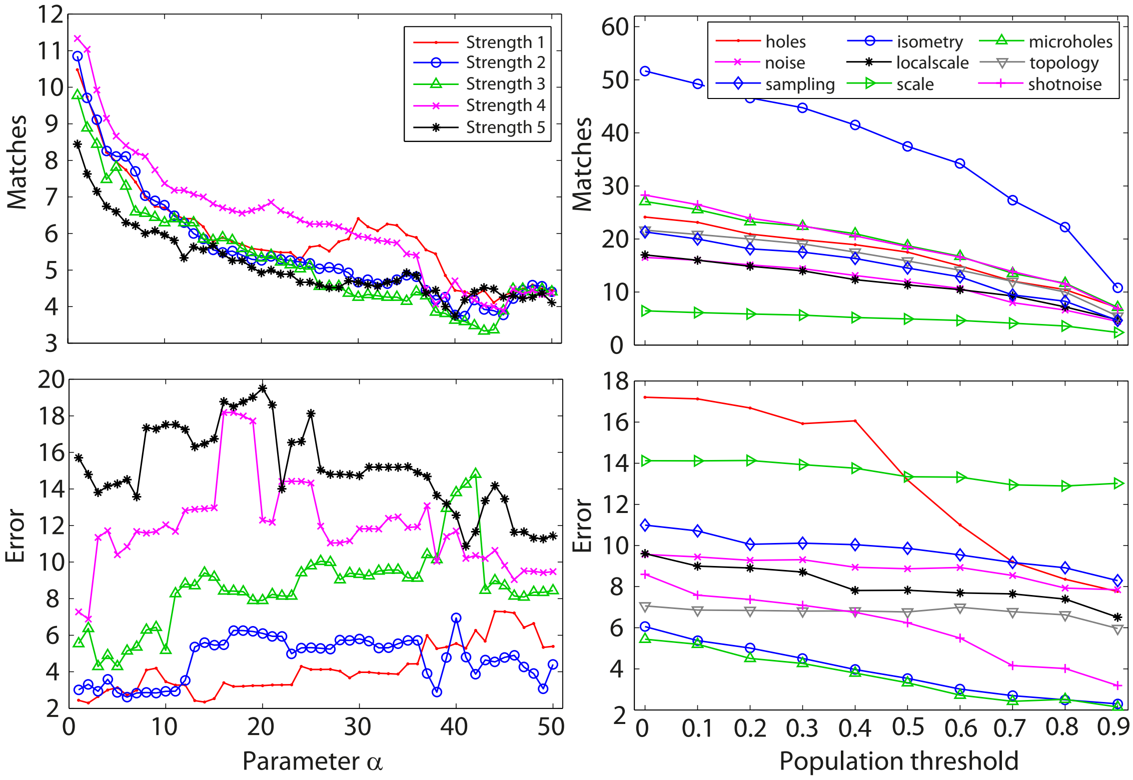


Figure 8.4: Sensitivity of our method to payoff coefficient α (first column) and the selection threshold used on the final population (second column). Increasing the α parameter reduces the average match distortion at the cost of a smaller correspondence. On the other hand, the population threshold has a more definite effect on size rather than quality of the final correspondence. In particular, while most transformations behave similarly, the “isometry” and “holes” classes appear to be more sensitive to this parameter.

8.6 Conclusions

We showed an application of the game-theoretic approach presented in the first chapters to the solution of intrinsic correspondence problems arising in deformable shape analysis. Through the use of multi-scale diffusion metrics, we showed how to fuse information from different scales into a single distortion criterion minimized in search of a minimum distortion correspondence. Evaluation on the SHREC’10 non-rigid shape correspondence benchmark demonstrated that the proposed approach is capable of recovering accurate sparse correspondences between shapes and is robust under a variety of strong deformations. Although these are mostly preliminary results, the game-theoretic framework seems to provide a good basis for attacking problems in non-rigid settings.

9
Conclusions

In this thesis we approached different aspects of the all-pervasive correspondence problem in Computer Vision. Our main results took advantage of recent developments in the emerging field of game-theoretic methods for Machine Learning and Pattern Recognition, which we adapted and shaped into a general framework that is flexible enough to accommodate rather specific and commonly encountered correspondence problems within the areas of 3D reconstruction and shape analysis. We were able to apply said framework to a variety of matching scenarios and tested its effectiveness over a wide selection of applicative domains. To this end, we proposed domain-specific instances of the framework and provided some theoretical insights that confirm the validity of the method and foster new interesting directions of research.

Contributions and novelty of the work

We started our presentation with an overview of the existing literature, and gave some mathematical preliminaries covering most of the topics successively explored throughout our study. Specifically, the final part of this preliminary section was dedicated to introducing the matching framework in its main components: we defined the concept of matching game, and elaborated on the notion of payoff and matching strategy, which constitute the principal material of our formulation for the matching problem. We showed how to impose mapping constraints, and how to evolve an initial state to a stable solution through the use of evolutionary dynamics.

In Chapter 3 we concentrated briefly on the reconstruction problem, specifically analyzing two fundamental components of a typical pipeline: camera calibration (Section 3.1) and coded light projection (Section 3.2). In the first case, we pointed out the frequently overlooked necessity of utilizing accurate calibration patterns, and proposed a method that allows to obtain an accurate calibration in a robust manner even with non-professionally crafted targets. This happens by modelling the calibration process as an iterative procedure in which the pattern structure is allowed to change so as to reflect the actual geometry of the calibration object, in what can be regarded as a way of "re-calibrating" the calibration item itself. The approach is novel to our knowledge, and allows to attain very accurate calibrations even under severe distortions of the model.

The second contribution regards the projection of intensity patterns in a structured light scanning scenario. While research in this field is rather active and spans many different communities (Optical Engineering and Photogrammetry among the others), relatively little interest is dedicated to the design of efficient scanning procedures, whereas many of the current solutions avoid the problem by exploiting technological advancement. Here we presented a modification to a state-of-the-art phase shift technique that allows to reduce the number of projected patterns to the theoretical minimum, while maintaining comparable levels of accuracy.

Chapter 4 introduced the general issue of feature detection in computer vision applications, and offered three different pictures of the problem. In the specific, in Section 4.1 we modelled the detection problem as an outlier selection task in which features that do not occur frequently are treated as interesting points, while common descriptors are regarded as structured noise and filtered out as such. The filtering step is performed in an iterative manner by casting the selection process into the game-theoretic framework introduced in Section 2.6. This first application of the framework proves to be extremely effective and is applied to both 2D and 3D settings. Section 4.1.4 tackles a different feature detection scenario which is specific to the surface registration problem: given two surfaces in approximate alignment, the problem is to select a limited number of relevant points such that the application of a motion refinement algorithm over these points gives the best possible results. To do this, we proposed an approach defining a local distinctiveness measure to each point that is associated with the average local radius of curvature, and a sampling strategy that samples points according with their distinctiveness. Experiments on range images with known ground-truth alignment showed the effectiveness of the approach.

Even if the game-theoretic framework was introduced in this chapter as a practical tool for clustering (not differently from its first use in the original paper), it is in Chapter 5 that it finds its first application in a matching scenario. We considered the Structure from Motion problem, an image-based reconstruction setup where typical assumptions such as a controlled environment and calibrated cameras do not always hold. Feature matching is an essential step in this context, and can be rendered particularly difficult by a variety of factors (changing illumination, reflections and moving objects are just a few). To tackle the matching problem, we decided to rely on a simple, yet very common assumption: the different views are linked by (sufficiently) small motions, in such a way that the 3D rigid motions connecting the single shots can be locally approximated by affine transformations on the image plane. Assuming this assumption holds, we devised a matching game in which pairs of candidate matches (generated by taking directly into account the descriptor distance of extracted features from the two images) receive a payoff that is proportional to the mutual fitness to a common similarity transformation, which can be calculated from the scale and orientation estimates given by the associated 2D descriptors. After computing this payoff value for all possible pairs of candidate matchings, we were able to extract the (globally) most consistent group of pairings with respect to the affinity/similarity constraint enforced by the proposed payoff function. Correspondences were densified with repeated applications of this matching scheme; the procedure emphasized the selective behavior of the game-theoretic approach, exhibiting an interesting effect of separating the “parallax groups” present in the scene according to their average (similarity) distortion. Application of this method to a standard dataset demonstrated its superior performance in relation to another state-of-the-art technique.

In subsequent chapters we proposed several other applications to a variety of matching problems. Chapter 6 tackles the surface registration problem frequently encountered in 3D reconstruction pipelines, where a collection of rangemaps, e.g. produced by a depth scanner, are to be rigidly aligned to one another so as to obtain an accurate as possible reconstruction of the scanned object. This problem has been frequently tackled in literature, with existing methods separating the matching process into a first (even supervised) approximate alignment of the surfaces, to which an automatic refinement (usually through variations of the ICP algorithm) follows. We took a different view of the problem by casting it into a correspondence search problem, which is solved via the game-theoretic inlier selection framework. Similarly to the SfM case, we modelled matching strategies as candidate matches and defined the payoff to be a measure of consistency between matches respecting the same rigid transformation. For increased efficiency, we relaxed rigidity to an isometry constraint and proposed a compatibility measure that reflects the extent to which the Euclidean distance among strategies is maintained by the correspondence map. The method exhibited excellent performance with both real-world and synthetic data, and demonstrated its ability to attack the surface registration problem efficiently and in a rigorous manner without the need of ex-post validation of the generated transformations. Additionally, a specific set of experiments showed an outstanding capability of the method to provide really accurate alignments in one shot, without the need of further refinement. To our knowledge, this is the first completely automatic method that attempts to solve the coarse and fine registration problems in a single step.

We followed a similar approach in Chapter 7, where we concentrated on the apparently similar, yet much more challenging scenario of 3D object recognition in cluttered scenes. While the general problem is treated similarly to what we did in the surface alignment case, by enforcing an isometry constraint on the generated set of matching hypotheses, in this chapter we also attacked the more general case in which model and scene may have a different scale. This adds a degree of freedom to the parameter estimation problem, and would in principle require a higher order approach in order to estimate the similarity (view) transformation connecting the two surfaces. Instead of resorting to high order dynamics, we decided to stick to a pairwise setting; our insight here is that the paths connecting pairs of points on the respective surfaces contain all the necessary information to obtain invariance to global changes in scale. We proposed many different ways to exploit this information, by defining a pairwise descriptor and different variations of the payoff function taking into account both unary and binary information of the matching strategies. We performed a wide range of experiments evaluating different combinations of the method, and compared the results with state-of-the-art techniques, demonstrating far superior performance on a standard dataset. We also presented, for the first time to our knowledge, a thorough quantitative analysis of scale-invariant object recognition in cluttered scenes. The research carried out in this chapter confirmed the effectiveness of the method to deal with generally complex scenarios (differently from the simple surface alignment case, here the data surface may undergo rigid, clutter, topological, partiality and noise transformations as well as total absence from the scene), emphasizing its strongly selective behavior as an inlier selection method. Further, it demonstrated the flexibility of the game-theoretic framework as a general tool to model a variety situations that might arise in common computer vision tasks, provided that they can be formulated in terms of some pairwise notion of compatibility among the available hypotheses.

The final Chapter considered the problem of deformable shape matching. Again, we tried to formulate the search for a correspondence in game-theoretic terms, but in doing this we offered a different perspective to the problem. Specifically, we established a clear link with optimization theory and Gromov-Lipschitz metrics, and regarded the selection process as a local maximizer of a (relaxed) quadratic assignment problem. Additionally, we proposed a way to integrate intrinsic information at different time scales in an attempt to gain in robustness under global scale transformations. The method was tested against a standard dataset for minimum distortion correspondence problems, placing our approach among the state-of-the-art for all considered deformations. Finally, we introduced a spectral technique that tries to take benefit from the repeated application of the game-theoretic method in order to merge together compatible groups of nonrigid matches. This way, one can obtain a denser correspondence by simply repeating the matching steps many times, producing strong clusters of matches that are separated by an intrinsic measure of consistency in a robust and coherent manner. Again, the method was tested against a standard dataset obtaining results on par with the state of the art.

Future directions

Of course, the game-theoretic method we presented doesn't come without drawbacks, the first and most obvious of which is the fact that it is not able to model all the matching problems. Indeed, with the current formulation we are only able to attack problems where a global property is verifiable over a subset of two or more matching hypotheses. While this does not represent the entire set of scenarios, it is nevertheless the case for most parameter estimation problems, as well as structural constrained problems. Throughout our treatment we successfully adapted the framework to a variety of scenarios, demonstrating its effectiveness in delivering robust and sparse solutions to the modelled problems. Yet, as we proceeded with our research, we gained some interesting insights that we intend to explore in subsequent studies. Chapter 8, in particular, represents preliminary research on a wide and complex topic. While we could observe good performance of the game-theoretic method as is in a nonrigid matching scenario, we feel that the approach needs a more rigorous treatment, especially in relation to the mathematical structures partaking in the selection process. This particular application leads us to reconsider the whole framework in mathematical optimization terms, whereas the correspondence problem can be formulated in other interesting ways, and the game-theoretic tools regarded as a robust means to seek for critical points over different relaxations of the assignment problem under sparse constraints. In particular, our ongoing research is directed towards three main topics. First, the population vector we introduced in the presentation of the game-theoretic framework finds a natural interpretation in the matching problem: since each strategy represents a putative assignment, the vector can be regarded (and reshaped) as a correspondence matrix where each element represents some confidence measure of the match among the two shape points. With this view, we can reconsider the L1 constraint on the matrix by requiring, instead, that some measure associated to each point be maintained by the mapping over the second shape; this way, the feasible set of solutions is included in the space of doubly stochastic matrices and the assignment problem finds a better interpretation in optimal transport theory. Of course, in doing this we assume that the two shapes have the same number of points; this suggests a second direction of research where the correspondence matrix is relaxed to take values under sub-bistochastic constraints. In these two cases, another major drawback of the game-theoretic method becomes more obvious: while we already emphasized in a number of experiments that the obtained solutions, although very robust, are also rather sparse, this does not necessarily imply that the obtained correspondence be “uniformly spread” across the shapes. This calls for a third direction of analysis, namely the introduction of some regularization term over the map of correspondences, which has the specific aim of localizing within connected patches corresponding shape regions. With these tools at disposal, we could be able to attack other complex scenarios (such as articulated matching) in a robust manner. A final point we would like to discuss regards the computational and spatial complexity of the framework. Since each match candidate is modeled as a strategy, the size of the payoff matrix between strategies grows with the square of the candidates. This can be a problem in many cases since it is not infrequent that matching candidates are in the order of thousands, leading to data chunks in the order of several millions of values. The problem is further exacerbated when dealing with higher order problems, rendering them practically unfeasible even in the most simple cases. We plan to tackle this drawback by taking benefit from recent technological advances through the use of general purpose GPU programming. While this could be seen as a mere technical issue, a fast implementation of the method would make huge or high order problems practicable, especially in those cases where the payoff matrix is not materialized, but rather its values computed on the fly during the evolutionary process itself.

Bibliography

[17] Agrawal, M. A lie algebraic approach for consistent pose registration for motion estimation. In *Proc. Int. Robotics Symposium* (2006).

[18] Ahn, Y.-K., Park, Y.-C., Choi, K.-S., Park, W.-C., Seo, H.-M., and Jung, K.-M. 3d spatial touch system based on time-of-flight camera. *WSEAS Trans. Info. Sci. and App. 6* (2009), 1433–1442.

[19] Aiger, D., Mitra, N. J., and Cohen-Or, D. 4-points congruent sets for robust surface registration. *ACM Transactions on Graphics 27*, 3 (2008), #85, 1–10.

[20] Akagündüz, E., Eskizara, O., and Ulusoy, I. Scale-space approach for the comparison of hk and sc curvature descriptions as applied to object recognition. In *Proc. of the 16th IEEE International Conference on Image processing* (Piscataway, NJ, USA, 2009), ICIP'09, pp. 413–416.

[21] Albarelli, A., Rota Bulò, S., Torsello, A., and Pelillo, M. Matching as a non-cooperative game. In *ICCV 2009: Proceedings of the 2009 IEEE International Conference on Computer Vision* (2009), IEEE Computer Society.

[22] Aubry, M., Schlickewei, U., and Cremers, D. The wave kernel signature-a quantum mechanical approach to shape analyis. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition* (2011).

[23] Bariya, P., and Nishino, K. Scale-hierarchical 3d object recognition in cluttered scenes. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2010* (2010), pp. 1657–1664.

[24] Batlle, J., Mouaddib, E. M., and Salvi, J. Recent progress in coded structured light as a technique to solve the correspondence problem: a survey. *Pattern Recognition 31*, 7 (1998), 963–982.

[25] Bay, H., Ess, A., Tuytelaars, T., and Gool, L. J. V. Surf: Speeded-up robust features. *Computer Vision and Image Understanding 110*, 3 (2008), 346–359.

[26] Beardsley, P. A., Zisserman, A., and Murray, D. W. Sequential updating of projective and affine structure from motion. *Int. J. Comput. Vision 23*, 3 (1997), 235–259.

[27] Benjemaa, R., and Schmitt, F. Fast global registration of 3d sampled surfaces using a multi-z-buffer technique. In *Proc. Int. Conf. on Recent Advances in 3-D Digital Imaging and Modeling* (1997), pp. 113–120.

[28] BERGEVIN, R., ET AL. Towards a general multi-view registration technique. *IEEE Trans. Patt. Anal. Machine Intell. 18*, 5 (1996), 540–547.

[29] BESL, P. J., AND MCKAY, N. D. A method for registration of 3-d shapes. *IEEE Trans. Pattern Anal. Mach. Intell. 14*, 2 (1992), 239–256.

[30] BORRMANN, D., ELSEBERG, J., LINGEMANN, K., NÜCHTER, A., AND HERTZBERG, J. Globally consistent 3d mapping with scan matching. *Robot. Auton. Syst. 56* (February 2008), 130–142.

[31] BOSCH, A., ZISSERMAN, A., AND MUNOZ, X. Image classification using random forests and ferns. In *Proc. 11th IEEE International Conference on Computer Vision – ICCV '07.* (2007), pp. 1–8.

[32] BRADSKI, G., AND KAEHLER, A. *Learning OpenCV: Computer Vision with the OpenCV Library*, 1st ed. O'Reilly Media, Inc., October 2008.

[33] BRONSTEIN, A. M., BRONSTEIN, M. M., CASTELLANI, U., DUBROVINA, A., GUIBAS, L. J., HORAUD, R. P., KIMMEL, R., KNOSSOW, D., VON LAVANTE, E., MATEUS, D., OVSJANIKOV, M., AND SHARMA, A. SHREC 2010: Robust correspondence benchmark. In *Proc. EUROGRAPHICS Workshop on 3D Object Retrieval* (2010).

[34] BRONSTEIN, A. M., BRONSTEIN, M. M., AND KIMMEL, R. Generalized multidimensional scaling: a framework for isometry-invariant partial surface matching. *Proc. National Academy of Science (PNAS) 103*, 5 (2006), 1168–1172.

[35] BRONSTEIN, M., AND BRONSTEIN, A. Shape recognition with spectral distances. *PAMI*, 99 (2011).

[36] BRONSTEIN, M. M., AND KOKKINOS, I. Scale-invariant heat kernel signatures for non-rigid shape recognition. In *IEEE Conf. on Computer Vision and Pattern Recognition* (2010), pp. 1704–1711.

[37] BROWN, D. C. Close-range camera calibration. *Photogrammetric Engineering 37*, 8 (1971), 855–866.

[38] BROWN, M., AND LOWE, D. G. Unsupervised 3d object recognition and reconstruction in unordered datasets. In *3DIM '05: Proceedings of the Fifth International Conference on 3-D Digital Imaging and Modeling* (Washington, DC, USA, 2005), IEEE Computer Society, pp. 56–63.

[39] BUSS, S. R., AND FILLMORE, J. P. Spherical averages and applications to spherical splines and interpolation. In *ACM Transactions on Graphics* (2001), vol. 20, pp. 95–126.

[40] CARMICHAEL, O. T., HUBER, D. F., AND HEBERT, M. Large data sets and confusing scenes in 3-d surface matching and recognition. In *3DIM* (1999), pp. 358–367.

[41] CHEN, C.-S., HUNG, Y.-P., AND CHENG, J.-B. Ransac-based darces: A new approach to fast automatic registration of partially overlapping range images. *IEEE Trans. Pattern Anal. Mach. Intell. 21*, 11 (1999), 1229–1234.

[42] CHEN, H., AND BHANU, B. 3d free-form object recognition in range images using local surface patches. *Pattern Recognition Letters 28* (July 2007), 1252–1262.

[43] CHEN, W., YANG, H., SU, X., AND TAN, S. Error caused by sampling in fourier transform profilometry. *Optical Engineering 38*, 6 (1999), 1029–1034.

[44] CHEN, Y., AND MEDIONI, G. Object modeling by registration of multiple range images. *Image and Vision Computing* (1992), 145–155.

[45] CHUA, C. S., AND JARVIS, R. Point signatures: A new representation for 3d object recognition. *Intl. J. of Comput. Vis. 25*, 1 (October 1997), 63–85.

[46] CHUM, O., AND MATAS, J. Matching with prosac - progressive sample consensus. In *CVPR 05: Proceedings of the 2005 IEEE Conference on Computer Vision and Pattern Recognition (CVPR05) - Volume 1* (Washington, DC, USA, 2005), IEEE Computer Society, pp. 220–226.

[47] CHUNG, D. H., YUN, I. D., AND LEE, S. U. Registration of multiple-range views using the reverse-calibration technique. *Pattern Recognition 31*, 4 (1998), 457–464.

[48] CLIFFORD, W. *Mathematical Papers*. Macmillan, London, 1882.

[49] COIFMAN, R. R., AND LAFON, S. Diffusion maps. *Applied and Computational Harmonic Analysis 21* (July 2006), 5–30.

[50] CORTELAZZO, G. M., AND ORIO, N. Retrieval of colored 3d models. In *3DPVT* (2006), pp. 986–993.

[51] CURLESS, B., AND LEVOY, M. A volumetric method for building complex models from range images. In *SIGGRAPH '96: Proceedings of the 23rd annual conference on Computer graphics and interactive techniques* (New York, NY, USA, 1996), ACM, pp. 303–312.

[52] DANIILIDIS, K. Hand-eye calibration using dual quaternions. *Int. J. of Robotics Research 18* (1999), 286–298.

[53] DOUXCHAMPS, D., AND CHIHARA, K. High-accuracy and robust localization of large control markers for geometric camera calibration. *IEEE Trans. Pattern Anal. Mach. Intell. 31*, 2 (2009), 376–383.

[54] Duchenne, O., Bach, F., Kweon, I.-S., and Ponce, J. A tensor-based algorithm for high-order graph matching. *IEEE Trans. Pattern Analysis and Machine Intelligence 33* (2011), 2383–2395.

[55] Eggert, D. W., Fitzgibbon, A. W., and B., F. R. Simultaneous registration of multiple range views for use in reverse engineering. Tech. Rep. 804, Dept. of Artificial Intelligence, University of Edinburgh, 1996.

[56] Elad, A., and Kimmel, R. On bending invariant signatures for surfaces. *IEEE Trans. Pattern Analysis and Machine Intelligence* (2003), 1285–1311.

[57] Faugeras, O. D., and Toscani, G. The calibration problem for stereo. In *Proceedings, CVPR '86 (IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Miami Beach, FL, June 22–26, 1986)* (1986), IEEE, pp. 15–20.

[58] Fischler, M. A., and Bolles, R. C. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM 24*, 6 (1981), 381–395.

[59] Fitzgibbon, A. W., and Zisserman, A. Automatic camera recovery for closed or open image sequences. In *ECCV '98: Proceedings of the 5th European Conference on Computer Vision-Volume I* (London, UK, 1998), Springer-Verlag, pp. 311–326.

[60] Frome, A., Huber, D., Kolluri, R., Bülow, T., and Malik, J. Recognizing objects in range data using regional point descriptors. In *ECCV 2004, 8th European Conference on Computer Vision* (2004), pp. 224–237.

[61] Furukawa, Y., and Ponce, J. Accurate, dense, and robust multiview stereopsis. *IEEE Transactions on Pattern Analysis and Machine Intelligence 32* (2010), 1362–1376.

[62] Gelfand, N., Ikemoto, L., Rusinkiewicz, S., and Levoy, M. Geometrically stable sampling for the ICP algorithm. In *Int. Conf. 3-D Digital Imaging and Modeling* (2003).

[63] Gelfand, N., Mitra, N. J., Guibas, L. J., and Pottmann, H. Robust global registration. In *Symposium on Geometry Processing* (2005), pp. 197–206.

[64] Ghosh, D., Amenta, N., and Kazhdan, M. M. Closed-form blending of local symmetries. *Comput. Graph. Forum 29*, 5 (2010), 1681–1688.

[65] Gold, S., and Rangarajan, A. A graduated assignment algorithm for graph matching. *IEEE Trans. Pattern Analysis and Machine Intelligence 18* (1996), 377–388.

[66] Granger, S., Pennec, X., and Roche, A. Rigid point-surface registration using an em variant of icp for computer guided oral implantology. In *MICCAI* (London, UK, 2001), Springer-Verlag, pp. 752–761.

[67] Guan, C., Hassebrook, L., and Lau, D. Composite structured light pattern for three-dimensional video. *Opt. Express 11*, 5 (2003), 406–417.

[68] Guehring, J. Reliable 3d surface acquisition, registration and validation using statistical error models. In *Int. Conf. 3-D Digital Imaging and Modeling* (2001).

[69] Hall, E. L., Tio, J. B. K., McPherson, C. A., and Sadjadi, F. A. Measuring curved surfaces for robot vision. *Computer 15*, 12 (1982), 42–54.

[70] Harris, C., and Stephens, M. A combined corner and edge detector. In *Proc. Fourth Alvey Vision Conference* (1988), pp. 147–151.

[71] Hartley, R. I. In defence of the 8-point algorithm. In *Proceedings of IEEE International Conference on Computer Vision* (1995), IEEE Comput. Soc. Press, pp. 1064–1070.

[72] Heikkilä, J. Geometric camera calibration using circular control points. *IEEE Trans. Pattern Anal. Mach. Intell. 22*, 10 (2000), 1066–1077.

[73] Heyden, A., Berthilsson, R., and Sparr, G. An iterative factorization method for projective structure and motion from image sequences. *Image and Vision Computing 17*, 13 (November 1999), 981–991.

[74] Horn, B. K. P. Closed-form solution of absolute orientation using unit quaternions. *Journal of the Optical Society of America. A 4*, 4 (Apr 1987), 629–642.

[75] Jin, H., Duchamp, T., Hoppe, H., McDonald, J. A., Pulli, K., and Stuetzle, W. Surface reconstruction from misregistered data. In *Proc. SPIE vol. 2573: Vision Geometry IV* (1995), pp. 32–328.

[76] Johnson, A. E., and Hebert, M. Using spin images for efficient object recognition in cluttered 3d scenes. *IEEE Trans. Pattern Anal. Mach. Intell. 21*, 5 (1999), 433–449.

[77] Kanezaki, A., Harada, T., and Kuniyoshi, Y. Partial matching of real textured 3d objects using color cubic higher-order local auto-correlation features. *The Visual Computer 26* (2010), 1269–1281. 10.1007/s00371-010-0521-3.

[78] Kavan, L., Collins, S., J. Žára, J., and O'Sullivan, C. Geometric skinning with approximate dual quaternion blending. *ACM Trans. Graph. 27*, 4 (2008), 105.

[79] KAVAN, L., COLLINS, S., O'SULLIVAN, C., AND ŽÁRA, J. Dual quaternions for rigid transformation blending. Tech. Rep. TCD-CS-2006-46, Trinity College Dublin, 2006.

[80] KAVAN, L., AND ŽÁRA, J. Spherical blend skinning: a real-time deformation of articulated models. In *Proc. 2005 Symp. on Interactive 3D Graph. and Games* (2005), pp. 9–16.

[81] KAZHDAN, M. Reconstruction of solid models from oriented point sets. In *Proceedings of the third Eurographics symposium on Geometry processing* (Aire-la-Ville, Switzerland, Switzerland, 2005), Eurographics Association.

[82] KE, Y., AND SUKTHANKAR, R. PCA-SIFT: a more distinctive representation for local image descriptors. In *Proc. IEEE Comp. Soc. Conf. on Computer Vision and Pattern Recognition – CVPR '04* (2004), vol. 2, pp. 506–513.

[83] KIM, V. G., LIPMAN, Y., AND FUNKHOUSER, T. Blended intrinsic maps. *Trans. on Graphics (Proc. of SIGGRAPH)* (2011).

[84] KRISHNAMURTHY, V., AND LEVOY, M. Fitting smooth surfaces to dense polygon meshes. In *Proceedings of SIGGRAPH 96* (1996), pp. 313–324.

[85] LEORDEANU, M., AND HEBERT, M. A spectral technique for correspondence problems using pairwise constraints. In *Proc. IEEE International Conference on Computer Vision* (2005), vol. 2, IEEE, pp. 1482–1489.

[86] LEVENBERG, K. A method for the solution of certain non-linear problems in least squares. *Quarterly Journal of Applied Mathmatics II*, 2 (1944), 164–168.

[87] LILIENBLUM, E., AND MICHAELIS, B. Optical 3d surface reconstruction by a multi-period phase shift method. *JCP 2*, 2 (2007), 73–83.

[88] LIU, Y. Replicator dynamics in the iterative process for accurate range image matching. *Int. J. Comput. Vision 83*, 1 (2009), 30–56.

[89] LOWE, D. Distinctive image features from scale-invariant keypoints. In *International Journal of Computer Vision* (2003), vol. 20, pp. 91–110.

[90] LOWE, D. G. Object recognition from local scale-invariant features. In *Proc. of the International Conference on Computer Vision ICCV, Corfu* (1999), pp. 1150–1157.

[91] LUCCHESE, L., AND MITRA, S. K. Using saddle points for subpixel feature detection in camera calibration targets. In *APCCAS* (2002), IEEE, pp. 191–195.

[92] MALLON, J., AND WHELAN, P. F. Which pattern? biasing aspects of planar calibration patterns and detection methods. *Pattern Recogn. Lett. 28*, 8 (2007), 921–930.

[93] MARR, D., AND HILDRETH, E. Theory of edge detection. *Royal Soc. of London Proc. Series B 207* (1980), 187–217.

[94] MASUDA, T. Registration and integration of multiple range images by matching signed distance fields for object shape modeling. *Comput. Vis. Image Underst. 87*, 1-3 (2002), 51–65.

[95] MATAS, J., CHUM, O., URBAN, M., AND PAJDLA, T. Robust wide-baseline stereo from maximally stable extremal regions. *Image and Vision Computing 22*, 10 (2004), 761–767. British Machine Vision Computing 2002.

[96] MCCARTHY, J. M. *An Introduction to Theoretical Kinematics*. MIT Press, 1990.

[97] MÉMOLI, F. Spectral Gromov-Wasserstein distances for shape matching. In *Proc. NORDIA* (2009).

[98] MÉMOLI, F., AND SAPIRO, G. A theoretical and computational framework for isometry invariant recognition of point cloud data. *Foundations of Computational Mathematics 5* (2005), 313–346.

[99] MIAN, A. S., BENNAMOUN, M., AND OWENS, R. Three-dimensional model-based object recognition and segmentation in cluttered scenes. *IEEE Trans. Pattern Anal. Mach. Intell. 28* (October 2006), 1584–1601.

[100] MIAN, A. S., BENNAMOUN, M., AND OWENS, R. On the repeatability and quality of keypoints for local feature-based 3d object retrieval from cluttered scenes. *Int. J. Comput. Vision 89* (September 2010), 348–361.

[101] MIKOLAJCZYK, K., AND SCHMID, C. An affine invariant interest point detector. In *ECCV '02: Proceedings of the 7th European Conference on Computer Vision-Part I* (London, UK, 2002), Springer-Verlag, pp. 128–142.

[102] MIKOLAJCZYK, K., AND SCHMID, C. A performance evaluation of local descriptors. *Pattern Analysis and Machine Intelligence, IEEE Transactions on 27*, 10 (2005), 1615–1630.

[103] MOREL, J.-M., AND YU, G. ASIFT: A new framework for fully affine invariant image comparison. *SIAM J. Img. Sci. 2*, 2 (2009), 438–469.

[104] NEWMAN, T. S., AND JAIN, A. K. A system for 3d cad-based inspection using range images. *Pattern Recognition 28*, 10 (1995), 1555 – 1574.

[105] NG, A. Y., JORDAN, M. I., AND WEISS, Y. On spectral clustering: Analysis and an algorithm. In *Advances in Neural Information Processing Systems* (2001), pp. 849–856.

[106] Novatnack, J., and Nishino, K. Scale-dependent/invariant local 3d shape descriptors for fully automatic registration of multiple sets of range images. In *Proc. of the 10th European Conference on Computer Vision* (Berlin, Heidelberg, 2008), Springer-Verlag, pp. 440–453.

[107] Ovsjanikov, M., Huang, Q.-X., and Guibas, L. J. A condition number for non-rigid shape matching. *Comput. Graph. Forum 30*, 5 (2011), 1503–1512.

[108] Pelillo, M., and Torsello, A. Payoff-monotonic game dynamics and the maximum clique problem. *Neural Computing 18* (May 2006), 1215–1258.

[109] Perl, Y., Itai, A., and Avni, H. Interpolation search – a log log N search. *CACM 21*, 7 (1978), 550–553.

[110] Pollefeys, M., Koch, R., Vergauwen, M., and Gool, L. V. Hand-held acquisition of 3d models with a video camera. *3D Digital Imaging and Modeling, International Conference on 0* (1999), 0014.

[111] Pottmann, H., Wallner, J., Huang, Q., and Yang, Y. Integral invariants for robust geometry processing. *Comput. Aided Geom. Des. 26* (2009), 37–60.

[112] Pulli, K. Multiview registration for large data sets. *Int. conf. on 3D Digital Imaging and Modeling* (1999), 160–168.

[113] Remondino, F., and Fraser, C. Digital camera calibration methods: considerations and comparisons. In *International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences,* (Dresden, Germany, 2006), Isprs, Ed., vol. Vol. XXXVI.

[114] Rosten, E., Porter, R., and Drummond, T. Faster and better: a machine learning approach to corner detection. *CoRR abs/0810.2434* (2008).

[115] Rota Bulò, S., and Bomze, I. M. Infection and immunization: A new class of evolutionary game dynamics. *Games and Economic Behavior 71*, 1 (January 2011), 193–211.

[116] Rota Bulò, S., and Pelillo, M. A game-theoretic approach to hypergraph clustering. In *Advances in Neural Information Processing Conference (NIPS2009)* (2009), vol. 22, pp. 1571 – 1579.

[117] Rusinkiewicz, S., and Levoy, M. Efficient variants of the icp algorithm. In *Proc. of the Third Intl. Conf. on 3D Digital Imaging and Modeling* (2001), pp. 145–152.

[118] Salti, S., Tombari, F., and di Stefano, L. A performance evaluation of 3d keypoint detectors. In *International Conference on 3D Imaging, Modeling, Processing, Visualization and Transmission (3DIMPVT)* (2011), IEEE, pp. 236–243.

[119] SALVI, J., ARMANGUE, X., AND BATLLE, J. A comparative review of camera calibrating methods with accuracy evaluation. *Pattern Recognition 35*, 7 (2002), 1617–1635.

[120] SALVI, J., MATABOSCH, C., FOFI, D., AND FOREST, J. A review of recent range image registration methods with accuracy evaluation. *Image Vision Comput. 25*, 5 (2007), 578–596.

[121] SALVI, J., PAGÈS, J., AND BATLLE, J. Pattern codification strategies in structured light systems. *Pattern Recognition 37* (2004), 827–849.

[122] SARFRAZ, M. S., AND HELLWICH, O. Head pose estimation in face recognition across pose scenarios. In *VISAPP (1)* (2008), pp. 235–242.

[123] SCHARSTEIN, D., AND SZELISKI, R. A taxonomy and evaluation of dense two-frame stereo correspondence algorithm. *Int. J. of Comp. Vision 47*, 1 (2002), 7–42.

[124] SEITZ, S. M., CURLESS, B., DIEBEL, J., SCHARSTEIN, D., AND SZELISKI, R. A comparison and evaluation of multi-view stereo reconstruction algorithms. In *CVPR '06* (2006), pp. 519–528.

[125] SHASHUA, A., ZASS, R., AND HAZAN, T. Multi-way clustering using super-symmetric nonnegative tensor factorization. In *European Conference on Computer Vision* (2006), pp. 595–608.

[126] SHI, J., AND TOMASI, C. Good features to track. In *1994 IEEE Conference on Computer Vision and Pattern Recognition (CVPR'94)* (1994), pp. 593 – 600.

[127] SHOEMAKE, K. Animating rotation with quaternion curves. In *Proc. of the 12th annual conference on Computer graphics and interactive techniques* (1985), pp. 245–254.

[128] SMITH, S. M., AND BRADY, J. M. Susan—a new approach to low level image processing. *Int. J. Comput. Vision 23*, 1 (1997), 45–78.

[129] SNAVELY, N., SEITZ, S. M., AND SZELISKI, R. Photo tourism: exploring photo collections in 3d. In *SIGGRAPH '06: ACM SIGGRAPH 2006 Papers* (New York, NY, USA, 2006), ACM, pp. 835–846.

[130] SNAVELY, N., SEITZ, S. M., AND SZELISKI, R. Modeling the world from internet photo collections. *Int. J. Comput. Vision 80*, 2 (2008), 189–210.

[131] SRINIVASAN, V., LIU, H. C., AND HALIOUA, M. Automated phase-measuring profilometry: a phase mapping approach. *Appl. Opt. 24*, 2 (1985), 185–188.

[132] STROBL, K. H., AND HIRZINGER, G. More accurate camera and hand-eye calibrations with unknown grid pattern dimensions. In *ICRA* (2008), pp. 1398–1405.

[133] Sturm, P. F., and Maybank, S. J. On plane-based camera calibration: A general algorithm, singularities, applications. In *CVPR* (1999), pp. 1432–1437.

[134] Sturm, P. F., and Triggs, B. A factorization based algorithm for multi-image projective structure and motion. In *ECCV '96: Proceedings of the 4th European Conference on Computer Vision-Volume II* (London, UK, 1996), Springer-Verlag, pp. 709–720.

[135] Su, X., and Chen, W. Fourier transform profilometry: a review. *Optics and Lasers in Engineering 35* (May 2001), 263–284.

[136] Sun, J., Ovsjanikov, M., and Guibas, L. A concise and provably informative multi-scale signature based on heat diffusion. In *Proc. of the Symposium on Geometry Processing* (2009), Eurographics Association, pp. 1383–1392.

[137] Sun, Y., Paik, J., Koschan, A., and Abidi, M. A. Point fingerprint: A new 3-d object representation scheme. *IEEE Transactions on Systems, Man, and Cybernetics 33* (2003), 712–717.

[138] Surrel, Y. Design of algorithms for phase measurements by the use of phase stepping. *Appl. Opt. 35*, 1 (1996), 51–60.

[139] Takeda, M., and Mutoh, K. Fourier transform profilometry for the automatic measurement of 3-d object shapes. *Appl. Opt. 22*, 24 (1983), 3977–3982.

[140] Tarel, J.-P., Civi, H., and Cooper, D. B. Pose estimation of free-form 3d objects without point matching using algebraic surface models. In *Proceedings of IEEE Workshop Model Based 3D Image Analysis* (Mumbai, India, 1998), pp. 13–21.

[141] Thorstensen, N., and Keriven, R. Non-rigid shape matching using geometry and photometry. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition* (2009).

[142] Tomasi, C., and Kanade, T. Shape and motion from image streams under orthography: a factorization method. *International Journal of Computer Vision 9* (1992), 137–154. 10.1007/BF00129684.

[143] Tombari, F., Salti, S., and di Stefano, L. Unique signatures of histograms for local surface description. In *ECCV 2010 - 11th European Conference on Computer Vision* (2010), pp. 356–369.

[144] Torr, P., and Zisserman, A. Robust computation and parametrization of multiple view relations. In *ICCV '98: Proceedings of the Sixth International Conference on Computer Vision* (Washington, DC, USA, 1998), IEEE Computer Society, p. 727.

[145] TORSELLO, A. Point invariance of the screw tension minimizer. Tech. Rep. DAIS-2011-2, DAIS, Università Ca' Foscari Venezia, 2011.

[146] TORSELLO, A., ROTA BULÒ, S., AND PELILLO, M. Grouping with asymmetric affinities: A game-theoretic perspective. In *CVPR '06* (2006), pp. 292–299.

[147] TRIGGS, B., MCLAUCHLAN, P., HARTLEY, R., AND FITZGIBBON, A. Bundle adjustment – a modern synthesis. In *Vision Algorithms: Theory and Practice* (2000), B. Triggs, A. Zisserman, and R. Szeliski, Eds., vol. 1883 of *Lecture Notes in Computer Science*, Springer-Verlag, pp. 298–372.

[148] TSAI, R. A versatile camera calibration technique for high-accuracy 3d machine vision metrology using off-the-shelf tv cameras and lenses. *Robotics and Automation, IEEE Journal of 3*, 4 (1987), 323–344.

[149] TSAI, Y. R. An efficient and accurate camera calibration technique for 3D machine vision. In *Proc. CVPR* (1986).

[150] TURK, G., AND LEVOY, M. Zippered polygon meshes from range images. In *SIGGRAPH '94: Proc. of the 21st annual conference on Computer graphics and interactive techniques* (New York, NY, USA, 1994), ACM, pp. 311–318.

[151] VEDALDI, A., AND FULKERSON, B. VLFeat: An open and portable library of computer vision algorithms. http://www.vlfeat.org/, 2008.

[152] VERGAUWEN, M., AND VAN GOOL, L. Web-based 3d reconstruction service. *Mach. Vision Appl. 17*, 6 (2006), 411–426.

[153] WANG, C., BRONSTEIN, M. M., BRONSTEIN, A. M., AND PARAGIOS, N. Discrete minimum distortion correspondence problems for non-rigid shape matching. In *Proc. Scale Space and Variational Methods* (2011).

[154] WEIBULL, J. *Evolutionary Game Theory*. MIT Press, 1995.

[155] WEINSHALL, D., AND TOMASI, C. Linear and incremental acquisition of invariant shape models from image sequences. *IEEE Transactions on Pattern Analysis and Machine Intelligence 17* (1995), 512–517.

[156] WENG, J., COHEN, P., AND HERNIOU, M. Camera calibration with distortion models and accuracy evaluation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on 14*, 10 (1992), 965–980.

[157] WILLIAMS, J., AND BENNAMOUN, M. Simultaneous registration of multiple corresponding point sets. *Comput. Vis. Image Underst. 81*, 1 (2001), 117–142.

[158] XIAOBO, C., TONG, X. J., TAO, J., AND YE, J. Research and development of an accurate 3d shape measurement system based on fringe projection: Model analysis and performance evaluation. *Precision Engineering 32*, 3 (2008), 215 – 221.

[159] YUE, H.-M., SU, X.-Y., AND LIU, Y.-Z. Fourier transform profilometry based on composite structured light pattern. *Optics Laser Technology 39* (Sept. 2007), 1170–1175.

[160] ZAHARESCU, A., BOYER, E., VARANASI, K., AND HORAUD, R. P. Surface feature detection and description with applications to mesh matching. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition* (June 2009).

[161] ZHANG, Z. Iterative point matching for registration of free-form curves and surfaces. *Int. J. Comput. Vision 13*, 2 (1994), 119–152.

[162] ZHANG, Z. A flexible new technique for camera calibration. *IEEE Trans. Pattern Anal. Mach. Intell. 22*, 11 (2000), 1330–1334.

[163] ZHANG, Z. Camera calibration with one-dimensional objects. *IEEE Trans. Pattern Anal. Mach. Intell. 26*, 7 (2004), 892–899.

[164] ZHANG, Z., DERICHE, R., FAUGERAS, O., AND LUONG, Q.-T. A robust technique for matching two uncalibrated images through the recovery of the unknown epipolar geometry. *Artif. Intell. 78*, 1-2 (1995), 87–119.

[165] ZHONG, J., AND ZHANG, Y. Absolute phase-measurement technique based on number theory in multifrequency grating projection profilometry. *Appl. Opt. 40*, 4 (2001), 492–500.

[166] ZIEN, J. Y., SCHLAG, M. D. F., AND CHAN, P. K. Multi-level spectral hypergraph partitioning with arbitrary vertex sizes. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems 18* (1999), 1389–1399.