Qualifying 4D Deforming Surfaces by Registered Differential Features

Timothy Campbell Lukins

Doctor of Philosophy
Institute of Perception, Action and Behaviour
School of Informatics
College of Science and Engineering
University of Edinburgh
2008

Abstract

Recent advances in 4D data acquisition systems in the field of Computer Vision have opened up many exciting new possibilities for the interpretation of complex moving surfaces. However, a fundamental problem is that this has also led to a huge increase in the volume of data to be handled. Attempting to make sense of this wealth of information is then a core issue to be addressed if such data can be applied to more complex tasks. Similar problems have been historically encountered in the analysis of 3D static surfaces, leading to the extraction of higher-level features based on analysis of the differential geometry.

Our central hypothesis is that there exists a compact set of similarly useful descriptors for the analysis of dynamic 4D surfaces. The primary advantages in considering localised changes are that they provide a naturally useful set of invariant characteristics. We seek a constrained set of terms - a vocabulary - for describing all types of deformation. By using this, we show how to describe what the surface is doing more effectively; and thereby enable better characterisation, and consequently more effective visualisation and comparison.

This thesis investigates this claim. We adopt a bottom-up approach of the problem, in which we acquire raw data from a newly constructed commercial 4D data capture system developed by our industrial partners. A crucial first step resolves the temporal non-linear registration between instances of the captured surface. We employ a combined optical/range flow to guide a conformation over a sequence. By extending the use of aligned colour information alongside the depth data we improve this estimation in the case of local surface motion ambiguities. By employing a KLT/thin-plate-spline method we also seek to preserve global deformation for regions with no estimate.

We then extend aspects of differential geometry theory for existing static surface analysis to the temporal domain. Our initial formulation considers the possible intrinsic transitions from the set of shapes defined by the variations in the magnitudes of the principal curvatures. This gives rise to a total of 15 basic types of deformation. The change in the combined magnitudes also gives an indication of the extent of change. We then extend this to surface characteristics associated with expanding, rotating and shearing; to derive a full set of differential features.

Our experimental results include qualitative assessment of deformations for short episodic registered sequences of both synthetic and real data. The higher-level distinctions extracted are furthermore a useful first step for parsimonious feature extraction, which we then proceed to demonstrate can be used as a basis for further analysis. We ultimately evaluate this approach by considering shape transition features occurring within the human face, and the applicability for identification and expression analysis tasks.

Acknowledgements

Thanks must first and foremost go to my supervisor Bob Fisher. Without his sage advice, eye for detail, rigourous scientific ethos, and implacable patience - none of this research would have been carried out, let alone written up. Consequently, any faults in this thesis are solely from me failing to understand or listen to his proffered pearls of wisdom.

In addition, I would never have been able to do this without help from the following people:

Colin Urquhart for all his support and encouragement.

Jon Oberlander for his enthusiasm and helpful pointers.

My examiners Eric McKenzie and Dave Marshall for a stimulating viva and the help afterwards.

Matt Szenher, Rowland Sillito and Mark Payne for reading earlier draft chapters.

Toby Breckon, Scott Blunsden, Toby Collins and Ernesto Andrade in the vision lab for making turning up at KB worthwhile.

Craig Robertson for reminding me that Computer Vision is actually meant to be fun.

Dug Green and Ewan Borland at Dimensional Imaging for helping me do some real work.

Toby Bailey for inspiration in matters mathematical.

All the rest of iPub (and former members) for pints and understanding: Paul Crook, Pete Ottery, Jan Wessnitzer, Richard Reeve, Hugo Rossano, Darren Smith, Finlay Stewart, *et al*.

The numerous other excellent folk in Informatics and Edinburgh that I have got to know over the years who have offered me friendship and support.

I would also like to thank Emanuele Trucco and Douglas Armstrong for their patience and understanding whilst working for them at the same time.

The UK government in the form of EPSRC should get a mention for the money they gave me.

And last, but certainly by no means least, my family deserve all my thanks and love for always being there for me.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(*Timothy Campbell Lukins*)

“He put the glass to his lips and drank at one gulp. A cry followed; he reeled, staggered, clutched at the table and held on, staring with injected eyes, gasping with open mouth; and as I looked there came, I thought a change - he seemed to swell - his face became suddenly black and the features seemed to melt and alter - and the next moment, I had sprung to my feet and leaped back against the wall, my arm raised to shield me from that prodigy..."

The Strange Case of Dr Jekyll and Mr Hyde
Robert Louis Stevenson
1886

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

"Look things in the face and know them for what they are."

Marcus Aurelius - Meditations VIII.5

Surfaces are something we literally come in contact with all the time. In our daily lives we touch and feel the textures of a myriad number of different types. Whether it be hard or soft, rough or smooth, flat or curved, complex or simple. They are a fundamental fact of physical reality - explaining why such a large proportion of the vocabulary of human language is given over to their description.

What is also notable is that we have the additional facility to perceive - through sight - the surfaces that surround us. Through this we can, even at a considerable distance, assign many of these properties we associate through touch. Not only this, we can additionally describe other qualities that only occur through the interaction of the surface with light. Properties such as colour, reflectance, and shadowing.

Furthermore, when a surface moves, or alters, over time - it then presents an entirely new set of properties to describe its motion and deformation. Terms such as stretch or fold, bend or flatten, twist or straighten - these immediately conjure up an understanding of how the surface is changing. Fundamentally, these variations can be viewed as relaying the dynamics of the underlying structure and purpose. In the most complex case - such as the human face - these changes signify an important transfer of information, and thus communication.

It is these changing surface properties, and what information they can represent, that form the inspiration for this thesis.

1.1 Motivation and Research Questions

The central idea that motivated this work was one of attempting to describe the subtlety of change in the human face. Many works in the fields of psychology, physiology, philosophy, and art have looked at the face, and how it can relay information. One of the greatest early scientific works in this respect is that of Darwin in his 1872 treaty on the *"Expression of the Emotions in Man and Animals"* [Dar98]. Often eclipsed by his work on evolution, it presents for the first time a structured analysis of the nature and perception of expression.

However, rather than seeking to understand the underlying physical structure, social motivations, evolutionary purpose, or cultural universality of expression - the focus to our work is driven purely by the novel challenges present in interpreting visual information. The ability to go from images, to the recovery of surfaces, and eventually to objects has developed into the field of Computer Vision. Its fundamental aim in understanding the core human ability to perceive and understand the world motivated early researchers such as Marr [Mar82] to propose the tractability of vision to be based primarily on the successful extraction of features and their properties.

From this inspiration we are motivated to investigate the potential for attempting to understand complex visual dynamics, through the reconstruction and analysis of the underlying deformation exhibited by natural surfaces such as the face. The key theoretical questions that then lie at the core of this research are:

---

1. What observations of a naturally deforming surface can be used in order to resolve the correspondences between points on it as they move over time? Could colour or unique feature points, for example, be used to improve robustness?

2. Does there then exist a useful computational vocabulary for describing observations in the geometry of such a changing surface? Can we qualify a local patch on a surface as bending or folding, for example?

3. Would such a vocabulary lend itself to higher level analysis of the information being relayed by the changes? For example, as shown on the face?

From investigating these questions, we aim to answer the following single hypothesis that expresses the central goal of this research:

That there exists a set of differential features governed by intrinsic, extrinsic and metric local changes in the geometry around each point on a surface registered over time, and that these can in turn reveal useful dynamic properties for naturally deforming forms.

1.2 Area of Research

This research forms a timely contribution, tied to the recent advances made in the dense 4D capture of real, moving surfaces. The definition that needs to be clarified here is that 4D refers to the position of an infinitesimal (but ultimately sampled) point in space as it then proceeds to move instantaneously in time. The result of tracking these points is a set of spatial-temporal trajectories that can be replayed and further analysed.

This idea can be most easily related to current existing techniques for motion-capture, where easily distinguished markers can be placed on a surface and then tracked - so long as it is visible from a number of sensors. This has evolved into a robust and useful technology that can provide a complete set of motion trajectories for use. It has opened up the whole area of research concerned with "data driven" animation and modelling of biological dynamics.

However, a number of issues are immediately apparent from the use of motion-capture - representing as it does only a sparse capture of a dynamic surface. Firstly, the isolated points do not contain any direct shape information - and so consequently cannot describe deformation accurately. Secondly, the placement of the markers represents a complete bias in the data towards selective knowledge.

Given that new dense stereo and range finding technology can operate at video rate (and beyond) the opportunity now exists to overcome these issues. It becomes possible for the first time to enable a form of *holistic* motion-capture in which all regions of the surface under observation have the potential to be tracked and analysed in much more interesting ways. In particular, the aim is to ideally perform interpretation automatically with the aid of a computer.

Consequently, this research lies within the provenance of Computer Vision - in its purpose of seeking to understand and model artificial systems capable of perception from sensory data. Much previous work in this field has looked to the static reconstruction and interpretation of surfaces - particularly for traditional object-based recognition based on rigid structure. More recent effort in modelling articulated and deformable structures have mostly attempted the acquisition of non-rigid shapes (invariably using some variation of a "shape-from-X" approach). In general, there has been little by way of analysis of any deformation - except for applications lending themselves towards specialised medical based applications.

The application of this research could therefore potentially enhance traditional motion-capture areas such as animation. It could also offer useful high level interpretation for biometrics and (ultimately) potential security based applications. This is particularly relevant given the recent trends for robust 3D identification of the face, invariant of expression. We are furthermore conscious of this research also addressing some of the engineering problems associated with handling and processing such large amounts of novel data.

1.3 Overview

This thesis describes the results of our enquiry into the measurement and description of deforming natural surfaces: how they move, how they change, and how they can be analysed on the basis of these observations. This has been a linear progression, which could be considered a "bottom up" approach to the problem. In summary, each of the chapters looks at one particular aspect as follows:

- Chapter 2 looks at what prior works tell us about deforming surfaces and their properties. The focus is both on some of the general issues involved, before increasingly describing the state-of-the-art topics of direct relevance to this research. In summary, a final section offers a critique of where the interesting, unexplored, gaps in the field exist.

- The first problem to be tackled is as described in Chapter 3 - that of being able to track the localised movement of points on a surface, such that they correspond over time. One enhanced approach - using dense range flow based methods - is detailed, along with a number of experiments into its effectiveness in improving accuracy.

- Chapter 4 then looks at the complementary problem of how to then describe what the surface is doing in a qualitative manner that best encapsulates the dynamics. Even for a relatively sparse sequence, it is shown through the experiments that these higher level descriptors can effectively integrate the differences occurring on the face over time.

- Taking our experiences from the two previous chapters and combining them within a more rigourous framework is the objective of Chapter 5. By observing the changes over time in the fundamental forms for local tracked surface patches, the potential for recognising additional properties in sequences of change is demonstrated. We seek to apply this to real $4D$ face data, using a combined KLT/TPS based tracker to also accommodate for larger scale surface displacements.

- In Chapter 6, it is finally possible to review the contributions and implications of this research. Given that only one particular approach out of many available ones was adopted, it is useful to present some alternative ideas and analysis. We conclude by highlighting what has been actually been achieved in terms of contributions, and how this research stands amongst other work in the field.

Throughout this thesis, the objective has been to present a concise, accurate and hopefully rendering of this ambitious body of work. The aim is to be brief - but clear - as to the reasons, methods, results, understanding and conclusions of the research. Readers are directed to the comprehensive bibliography at the end for additional details of all the related and earlier works touched on by the research presented here.

1.3.1 Publications

Portions of this thesis were published in the proceedings of the following conferences:

- *"Colour Constrained 4D Flow"*, British Machine Vision Conference (BMVC), Oxford, UK, 2005 [LF05].
- *"Qualitative Characterisation of Deforming Surfaces"*, 3D Presentation, Visualisation and Transmission (3DPVT) conference, Chapel Hill, USA, 2006 [LF06].

Early results for Chapter 5 were also presented at the following symposium:

- *"A Mark-up Language for Describing Dynamic Surfaces"*, British Machine Vision Association Symposium on 3D Video: Analysis, Display and Applications, London, UK, 2008 [LF08]

Chapter 2

Prior Research

"To the pure geometer the radius of curvature is an incidental characteristic - like the grin of the Cheshire cat."

Sir Arthur Stanley Eddington

The study of shape is a topic which has greatly supported and influenced the thinking of many scientific fields. For example, in biology when describing variations between species, or in geology when qualifying the changes occurring to entire landscapes, or even in astrophysics when attempting to model the very structure of the universe. Underlying all these run a common mathematical, geometric, basis: giving us the ability to model and describe form.

Our own field of research concerns the area of computer vision - particularly $3D$ vision, but extended into the temporal domain (which we refer to here as $4D$). This enables us to focus on the question of *dynamic shape*. Our core work consequently builds on the previous work that has looked at tracking the motion of surfaces over time, and also how descriptions of static surfaces can be further extended to best describe these new types of observation. We are ultimately interested in how such a system can be applied to understanding the dynamics of the human face - another huge body of research which we describe here.

In this chapter we then look at what prior work has gone into the capture, modelling, analysis and applications regarding real surfaces. This work also intersects with the related fields of computer graphics, geometric modelling, and aspects of statistical machine learning with inspiration from psychology. We seek to present here an overview of how these interrelate and inspire the resulting research. The intention is to introduce a comprehensive general overview and to highlight in more detail the specifically related work, before finally offering up a critique of what particular aspects we address. Specific mathematical and technical aspects of the most directly related techniques are not presented, but where appropriate they are detailed in depth within sections at the start of each of the subsequent chapters.

To focus our attention, we can generally partition the literature as follows:

- Surface Representation - the description of shape, comparison of form, and modelling of non-rigid change.
- 4D Data Acquisition - the capture of motion, static shape, and both simultaneously.
- Motion Tracking Estimation - resolving the correspondences and registration between forms, particularly with dense flow techniques.
- Face Interpretation - the psychology and machine based analysis of the human face.

This breakdown is shown in the complete overview schematic Figure 2.1. This structure is reflected in the sections and subsections that follow in the rest of this chapter.

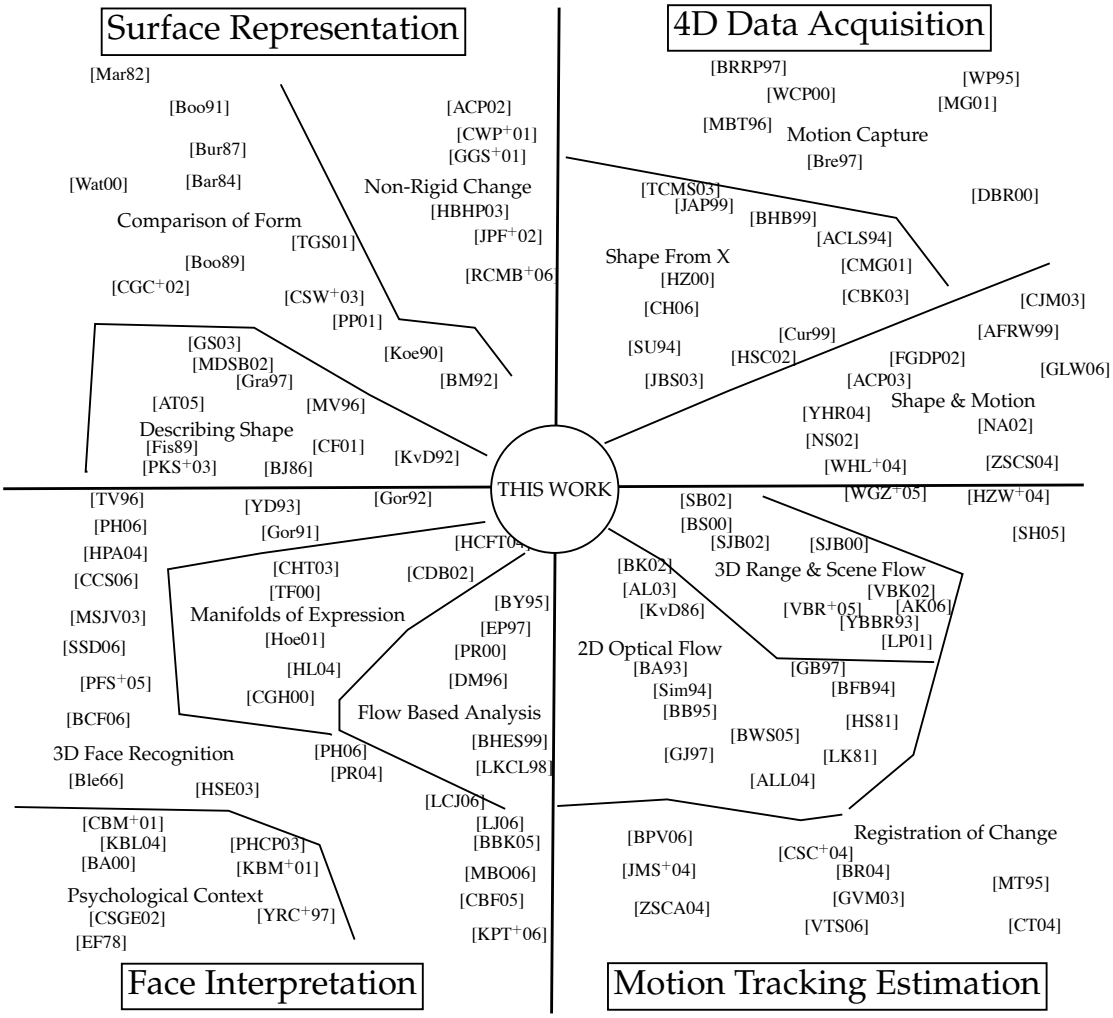


Figure 2.1: Complete overview of the related research contributing to this work.

Illustrating where the work of this thesis lies in the “gap in the research” at the centre of these converging fields. Those related works that contribute, or inspire, the most are indicated by their proximity to the centre.

2.1 Modelling Shape and Change

Shape, in one sense, can be thought of as the perception of a surface. It is the cognitive understanding and model that the human brain can apply to make sense of the physical forms it encounters. It is also the construct in language that can then be used to transmit a description of everyday objects and scenes. It is this desire to describe the specific shape properties and features of a surface that underlies the basis of this work.

In particular, our interest in shape extends to understanding the means by which to analytically compare two forms - in order to establish if they share any similar basis for correspondence. This has given rise to various taxonomies and techniques for revealing such similarities and relationships between diverse, yet related, forms. The natural extension to this is what if the shape to be compared is actually the same object, as it has altered between two instants. Perhaps the difference may be a few milliseconds, or it may be a few aeons. Yet the fundamental question remains as to how to describe changes such as "that bit has moved over there", or "it has flattened over all", or "that area became more elongated". How can such descriptions be reconciled with the underlying geometry, and how can they be automatically applied to analysing complex dynamic surfaces (such as the face)?

In this section we look at the more theoretical and computationally focused work in simply defining what shape actually is, what properties it can have, and how these can alter. This is consequently a topic that has its roots in the very earliest days of geometry established by such giants as Euclid, Pythagoras and Descartes. In particular the field of *differential geometry* as established by Euler, Monge, and Gauss as related to 3D surfaces is of direct relevance to this work. A number of excellent textbooks are available on this, such as Gray's "Differential Geometry of Curves and Surfaces with Mathematica" [Gra97] - but our intention here is to highlight those concepts and definitions that directly influence and impact on this research. We are particularly motivated in this section to stress the practical, computational aspects of working with shape.

2.1.1 Morphology and Geometry

The general scientific study of shape is regarded as the field of *morphology*. This is usually quantified further to determine the particular context, such as *bio-morphology* for living creatures, or *geo-morphology* for landscapes. An excellent overview of challenges associated with this is presented by Bookstein [Boo91]. In this, he sums up how traditionally the field has been the domain of biological taxonomy, broadly divided into biometric and geometric based approaches. The biometric approach seeks to identify specific landmarks and compare the measurements between them, while in contrast, the more modern geometric approach tries to preserve a more complete description of the shape.

The aim in both cases is to provide quantitative assessment of the variability of form in response to other experimental conditions. This must take account of the fact that the description of solid form is itself fundamentally not only a property of the object, but is also based upon the operation (method) or representation (model) used. This is made clear in the excellent book on the topic by Koenderink "Solid Shape" [Koe90] which also covers many of the other topics we introduce in this section.

However, we seek to present this topic in terms that are of direct relevance to our work. Consequently we first look at the issues bound up with representation - which can be said to define the actual surface. This leads us naturally to the properties that can be derived - and thus can be said to define the actual shape.

Representation

Given the concept of a surface, the first question is how to model it. Generally, this can be encoded continuously - or approximated discretely. All approaches are subject to the ultimate resolution to which computation can be performed. Representing surfaces is a problem that also requires consideration of the trade-offs regarding speed, realism, stability (of the representation in response to changes in the data), compactness, complexity, and with regards to supporting the dynamics that can ultimately change it. This last point is the most important for this work - since we are concerned with what representation can best cater for deformation and its analysis.

In the most simple geometric terms, the most intuitive definition of a surface is one of a "solid" 3D surface existing in Cartesian space (that is, it can be expressed in terms of co-ordinates (x, y, z)). A number of other immediately useful additional properties can then also be defined for the surface, or *manifold* it represents embedded in $\mathbb{R}^3$. The surface locally (i.e. the neighbourhood in the immediate vicinity of a point) can be defined within Euclidean $3D$ space, especially as it approaches flatness. This is a desirable property since it clarifies how we can represent and interpret the surface.

The *topology* of a surface represents the spatial properties preserved under bi-continuous deformation (stretching without tearing or gluing). As such, two surfaces may be the same if there exists a homeomorphism that describes the transition of one topological space into the other. The commonly used analogy here is that a coffee cup and a do-nut are equivalent, from a topological view. The definition, and description, of such a transition is what ultimately interests us in this work, in that we want to look at the nature of deformations that retain surface topology, yet alter in the other properties of their shape.

However, these points aside, there are many other concerns to do with the more practical options that are currently employed when modelling and manipulating surfaces within a computer. In effect, this area of work intersects with that of computer graphics. An excellent review of this topic is offered in a number of graphics text books, for example Watt's "$3D$ Computer

Graphics" [Wat00]. With respect to real surfaces - particularly complex ones such as the human face - we summarise below the options and properties that can be accommodated. We stress here that this is of direct relevance to this work since (as will be seen) related higher-level research is fundamentally driven by the way in which the surface is represented, particularly with regards to allowing it to be manipulated or used to derive higher-level characteristics.

The options for the representation of a surface must first and foremost tackle the trade-off between modelling the surface continuously or discretely. Continuously by parametric, implicit, or geometric primitive based approaches have many advantages in terms of the "pure" mathematical definitions, often leading to very compact models. However, such an approach must often be fitted to the data it is to model, which may introduce errors, particularly if the data is to be modelled dynamically. Alternatively, non-parametric approaches - specifically triangulated meshes and volumetric lattices - are often simpler to fit and are more flexible. Yet fundamentally, these schemes are only an approximation (and so lack accuracy), as apparent when viewing the piece-wise linearities along a silhouette edge. Furthermore, their ability to support deformation is inhibited by the increasing complexity required to control the movement of their elements in response to dynamics. Ultimately, modern approaches often employ a hybrid of the two approaches by use of subdivision surfaces to progressively defining a continuous surface towards the limit of an increasing level of detail.

Surface Properties

Given the representation of a surface, there are then many ways in which it can be measured, and other properties that can used to describe it. One of the most fundamental metrics is that of the *geodesic distance* across the surface between two arbitrary points. Depending on the convex or concave complexities of the surface, the straightforward integration along a curve can actually prove difficult to compute. This task is made even more difficult if the path between the two points does not lie along an intersecting plane (instead varying along another trajectory). For mathematically pure objects it can be computed accurately - but with varying degrees of complexity. For other surfaces, the approximation of distance is related to the representation. For example, with a triangulated surface, numerous search based algorithms attempt to find the closest edge path across the surface between vertices

Another interesting local property of a surface is that of *iso-contours* which can describe (and connect) points of related height. The issue here is in determining the concept of height. For relatively flat surfaces it can be ascertained by first determining an orthogonal direction relative to the other variables. In many ways, this is synonymous with the concept of a $2.5D$ surface in computer vision terminology - otherwise known as a depth-map or range-image. Many such representations of a surface can be described as the distances from a particular point or plane, with sensor readings occurring across a regularly sampled grid. For closed $3D$ surfaces it can be

defined by the distance from the centroid, or by *projection* onto spherical or cylindrical surfaces - as in the case of the wide variety of different approaches to mapping the Earth. Similarly, the projection of a depth-map can be performed as a result of a perspective transform, particularly if the data has been resolved relative to a camera. Inevitably, the choice of projection can lead to issues with *distortion* whereby a degree of error can occur in the result. This leads to more complex attempts at preserving relative distances and topology in projections onto harmonic or spherical maps.

This is an idea very much related to the fundamental concept of *gradient* of a surface. More precisely, this is the directional *first derivative* of the surface (greatest change in height with respect to distance). As such, this forms the first fundamental tenet of differential geometry, and facilitates the ability to look at second order properties (i.e. the changes in the gradient itself). This also presents the basic idea of an *intrinsic* property of the surface. To understand this, consider the commonly applied analogy of an ant resting on a bent piece of paper. The intrinsic geometry is that which is immediately visible and measurable to the ant at that point. No other information is required, and - most importantly - such local observation is invariant to the overall global rigid transformations that might have been applied to the entire surface (i.e. where the paper lies in space).

Another intrinsic property is thus the *normal* of the surface at any given point - the vector that is orthogonal to the tangent plane at that point, for a local neighbourhood. Any oriented surface shows consistency in its normals, that is, neighbouring ones all point roughly the same way. Extrema in the local change in normal directions represent surface *discontinuities* which can be used to indicate any number of useful features on the surface, such as edges and corners. This is further related to problems concerning *singularity theory* as encountered on a surface. These are regions on the surface that can be possibly defined by any number of alternative parameterisations.

From these properties, and key to the definition and understanding of shape, is the further second-order concept of *curvature*. In its simplest form this represents the implicit idea of deviation from flatness. More specifically, any given point on the surface can be characterised by imagining its intersection with the tangent plane around the normal at that point. This defines a local co-ordinate frame which can then be considered in order to reveal two extrema, representing the two principal curvatures (usually denoted κ_1 and κ_2) and their respective directions (e_1 and e_2). These are by definition orthogonal to one another, since local variation must occur perpendicularly. By combining these, the other interesting properties of Mean (usually denoted H) and Gaussian (denoted K) curvature can also be defined. Mean curvature represents an extrinsic property of the surface - since it depends entirely on an embedding in another space - whereas Gaussian can be considered intrinsic (in the analogy above, if the ant on the surface were tied to a stretched length r of string round a point it would then integrate the distance of the

circumference relative to r that can be related to K [BJ86]).

These derivations of curvature, and their definition by effectively 6 equations, can completely determine local surface shape and intrinsic geometry. This is ultimately embodied in the *shape operator* formed from the *first* and *second fundamental forms*. The shape operator effectively provides the mapping between parametric (u,v) surface space and local Cartesian (x,y,z) space for any given point (hence it is intrinsic). The first fundamental form matrix I represents first-order intrinsic properties of the surface, whereas the second fundamental form matrix II handles second-order metrics. Together they encapsulate all of the possible changes a surface can undergo, representing a powerful means of analysing surface change.

The use of curvature to describe the local geometry of a surface patch is a fundamental property that can be used for any number of purposes. These are not limited to shape analysis, but provide unique signatures and the ability to segment different regions of the surface to allow object recognition. A number of works in range analysis of depth-maps in computer vision have exploited such classification - for example as shown in Fisher's book "From Surfaces to Objects" [Fis89]. The great advantage to exploiting curvature is that it represents a useful property that is unique to the surface, and invariant to location or any global translation. Invariance represents quantities that are not altered by the underlying curvature of the surface, nor by any rotation or translation.

The necessity of then being able to accurately estimate curvature for any given surface patch becomes a crucial factor (if we are then to rely on it to describe shape precisely). Recent work continues to refine this calculation - particularly in the case of discrete approximations as described in Agam and Tang [AT05]. In that work, for example, an algorithm is proposed to effectively sample and search for a local curve that best resolves the directions of principal curvature. The key idea being that the polynomial curve used has a larger set of coefficients, and thus more degrees of freedom to better match the local geometry. This is viewed as an improvement on those techniques that directly operate on triangular meshes. The classic work of Meyer *et al.* [MDSB02] is an example of this that introduces the topic of operators that estimates curvature directly from the angles and arcs of a triangle fan centred at the vertex in question. Another approach is that of Garimella and Swartz [GS03] where the vertices of an unstructured mesh are used as the basis of a parametric quadric fitting (in turn based on earlier work by McIvor and Valkenburgh [MV96]). One of the acknowledged major problems with reliance on curvature to describe deformation is that, because it is based on the second order derivatives, it is particularly susceptible to noise. It should be noted that most of the approaches operating on discrete surfaces for defining curvature are driven by requirements in computer graphics algorithms in order to smooth or simplify a surface representation. Discrete surface representation (such as triangulated meshes) are the chosen rendering framework of most modern graphics pipelines. In this respect, curvature provides a useful factor for anisotropic shading and effects. Furthermore,

various 3D morphological operators have also been devised to operate over such surfaces - much as 2D operators can be applied to image and 2.5D equivalents).

Combining the principal curvatures into composite features - to better relate surfaces to physical descriptions - has a long legacy for $3D$ object recognition in computer vision. The classic work on this is that of Besl and Jain [BJ86] and their treatment of surfaces to provide a *rich description* for data-driven analysis. In this they compare the positive or negative signs of both the principal curvatures to define six different classes of local shape - peak, ridge, saddle, flat, valley and pit. They furthermore look to the Mean and Gaussian curvatures to define eight more specific classifications (plus one extra which is non-existant in the case of $H = 0$ and $K > 0$) and distinguish between saddle ridges, saddle valleys and minimal surfaces. Again it is important to realise the distinction that the Gaussian curvature is an intrinsic property of the surface and thus represents underlying shape.

Another seminal work along these lines is that of Koenderink and van Doorn [KvD92] in devising a Shape S and Curvedness C index. This was created out of a desire for a much more continuous - rather than discrete - representation of shape. It effectively forms a polar representation of the six different classes of shape, relating them next to their neighbouring forms and to their diametric opposites. The other axis is then free to describe the magnitude of the combined curvatures to describe a continuity from flat to infinitely curved. In this way the value of shape becomes a measurable property that is seen to vary continuously from one form to another, excluding transitions that can not occur between radically different shapes (for example, a peak cannot reform into a pit without first undergoing some intermediate shape).

Comparisons have been made of the various schemes for describing curvature of surfaces; mainly to try and ascertain their robustness to noise in the data, especially with respect to the levels of scale-space over which the surface may be considered. Certainly, some descriptions - such as the Mean curvature - are inherently more robust since they represent averaging (and thus, smoothing). More quantifiable experiments to establish the suitability of particular surface property calculations have been carried out. For example, Cantzer and Fisher [CF01] present an investigation into varying types of recognition between $H + K$ and $S + C$ schemes. This highlights the importance of threshold choice in the matter of defining zero crossing boundaries - that is, level below which a value in the principal curvatures can be considered "flat".

Interesting other viewpoints on the computational nature of shape can be found in the literature. In the work of Page *et al*. [PKS+03] for example a very interesting approach is proposed that unifies the curvature analysis of discrete surfaces with information theory. This is based on an underlying metric of entropy computed by a bin approximation on the Gaussian curvature. In such a way, the effective *complexity* of an entire surface can be quantifiably measured as shape information.

The use of all these properties such as curvature, gradient, contouring, and distance - form

very much the core for static 3D face recognition as surveyed in Section 2.4.3. A common problem is then in recognising the same conceptual terms in order to express equivalence, given that there are many different examples of a particular shape - especially when it deforms. Such formalism is particularly relevant when we seek to not only capture, but to also describe the dynamic nature of the transitions between forms. We review these issues in the following section.

2.1.2 Comparison of Form

Given two representations of surface shape, the question then arises as to how these can be compared to one another. The classic landmark based technique applied in most *morphometrics* (i.e. measurement of shape) to achieve this is the famous Procrustes method - named after the character in Greek mythology who invited passers by to lie on his (adjustable) bed and then promptly cut off their head or feet, or else stretched them, to fit. In principle this underlies the idea behind the mathematical method, in that it seeks to find the optimal set of rotation, translation, reflection and scaling transforms that result in the minimal distances between landmark points. This is usually effectively achieved through standard least-squares solution based on the sum of squared distances centred on the mean center of the shapes in question. A more computer vision orientated topic that relates to this is one of *registration* (which we consider in more detail for deformable surfaces in Section 2.3.1).

However, the classic rigid alignment of two $3D$ surfaces in computer vision is via the famous Iterative Closest Point (ICP) algorithm devised by Besl and McKay [BM92]. This is, in many ways, a similar idea to Procrustes (i.e. minimising the Sum Squared Distances) but it does not rely on specific landmark points, instead using all the data and selecting matching points by their current proximity to one another. Initial alignment is thus crucial to the success of the algorithm - if the surfaces are too far apart then only a small, closest matching region may be used to drive the fitting. Many variations attempt to get round this problem, and to also speed up the process (e.g. using KD-trees for fast neighbour lookup).

ICP furthermore allows for different densities and resolution (i.e. number of points) between the two surfaces. However, there must exist at least a reasonable degree of overlap, or similarity, between the data sets. The worst case can be when matching a 3D deformable surface, such as the face with different expressions. The question of identifying what has not changed on the surface is crucial to the matching process. Attempting to then handle, reverse or correct for these changes can form a particularly difficult challenge.

Fundamentally, ICP is therefore based on the *congruency* of the surfaces involved. That is, the concept that they have equivalent isometry (i.e. distances) and that their shape is the same (i.e. the shape operators are identical). This holds well for solid, rigid objects - but breaks down for surfaces that have deformed. An alternative view of handling the problem is to effectively warp one of the surfaces, once it has been reasonably registered. One early approach to solving

this non-rigid problem is the use of Thin Plate Splines (TPS) and *principal warps* as proposed by Bookstein (furthering his work on morphometrics) [Boo89]. In this he proposes how the traditional use of TPS as a tool for interpolation of a plausible, smooth surface over scattered landmark points, can be extended to explain deformation. The core idea is one of a superposition of such warps, applied at progressively smaller geometric scales, can be used to explain the transition in shape between the two surfaces. Fuller details of this approach - which we later exploit - are described in Section 5.3.2.

Numerous other models for accommodating, and explaining, deformation between surfaces have been proposed. Many of these intersect with the engineering and computer graphics domains and we describe a few below. A common idea is still to have some embedded structure, or fixed landmark points that serve to guide and control the deformations. This may represent the traditional sense of an *armature* (an underlying supporting structure), or it may actually act as a higher-level space in which the surface is embedded. Controlling deformation thus requires a framework for a greater level of motion control by explicitly allowing manipulation of localised regions of the representation (whether polygon or parametrically defined). The problem then lies in guaranteeing the smoothness of such surface alteration, and in successfully being able to apply the desired result consistently.

The basic idea of applying a transformation to specific regions of an object dependant on location was initially proposed by Barr [Bar84]. In this seminal paper he described a simple hierarchy of transformations that can be used to create any arbitrary complex deformation. This in turn gave rise to the generalised concept of *space warping* in which the object is not directly manipulated, but is effectively controlled by the alteration of the space in which it is embedded - an idea more commonly known as *Free Form Deformation* (FFD) as originally proposed by Sederburgh *et al*. [SP86]. While such a solution can offer smooth global transformations, it is often incapable of representing subtle, small scale dynamics - such as facial expression. Furthermore, it is useful to impose constraints over the physical properties of the object, so that for example it does not alter its volume, especially when directly altered.

Consequently, much research has been focused on finer-grained control and accurate extensions to FFD - which has (somewhat inevitably) led to multi-resolution based proposals, such as the comprehensive treatment given by Capell *et al*. [CGC+02]. In this work, they also extend another popular choice for modelling deformation by use of Finite Element Models (FEM). The central concept of FEM is the decomposition of a complex object into simpler interconnected components called finite elements. Incorporating physical constraints with topological flexibility enables a finer degree of control which is able to accommodate very subtle, localised, deformations.

Other attempts at using hierarchical finite element models find their way into the computer vision literature. One outstanding recent example of this is Tsap *et al*. [TGS01]. They seek

to recover the non-rigid motion between sequences of registered range and intensity images, using only a single model driven approach (the FEM) to explain the relationships. To achieve this, they effectively incorporate ideas from "Snakes" [KWT87] into the model to control the tracking by integration of appearance, and by performing a multi-scale approach to recovering correspondences. They actually focus their work on a simplified grid pattern - the control points of which are equivalent to the FEM nodes. Using such a grid pattern on an elastic surface (the skin of the arm) they are able to demonstrate their model accommodating not only translation, but also stretching and strain change over the course of the sequence.

2.1.3 Describing Non-Rigid Deformation

In the previous section we considered how surfaces are treated and modelled from a theoretical and computational viewpoint, with regards to how individually they can be represented, their intrinsic properties discovered, and used to then relate one instance of a surface to another. What happens next is in attempting to further describe these resulting deformations, or to relate them to underlying structures or signals. The problem here is primarily one of complexity, as the modelling of deformation is at heart a non-linear problem, as opposed to the simpler description of how a rigid surface moves (modelled as a simple transformation matrix). A lot of the work in the literature is concerned with structural and material deformation modelling, but here we constrain our review to those works in 4*D computer vision*. In many ways, the problem of modelling (and consequently understanding the nature of) deformation can also be viewed as one of reverse engineering the surface back to a neutral state. This is very much a more active topic in attempting to resolve invariance to expression on the face, as discussed in Section 2.4.3.

One appealing, and useful, theoretical tool to the understanding of deformation is a means of visualising change as a sequence of *Gauss* (or *spherical*) images. These provide a mapping between areas of interest as they occur on an arbitrary surface, and their transfer onto a unit Gaussian sphere. In effect, the unit length normals for all the points of interest on the surface are relocated to a single centroid at the origin, but with their orientations preserved, piercing the sphere to describe the shape. The upshot of this is that it is feasible to watch how a surface can evolve over time from a consistent viewpoint. The excellent survey of this presented by Pae and Ponce [PP01] gives many examples of such analysis.

Earlier, but useful, work in describing the differences and challenges between rigid and non-rigid surfaces from a computer vision perspective is comprehensively surveyed in Aggarwal *et al*. [ACLS94]. In particular they focus on the distinction between two types of non-rigid motion: articulated and elastic. Furthermore, they divide these into model-based approaches and those without any *a priori* model information. The emphasis on accommodating non-rigid motion is highlighted in order to support higher-level applications such as gesture recognition, data compression, biomedical modelling and face recognition.

For example, the recent the work by Allen *et al*. [ACP02] presents an appealing solution to the approach of learning an articulated model of human deformation from example training scans. They manage to recover the initial skeletal pose of each scan by using coloured markers on the subject taken at the same time. From this they then construct a consistent sub-division surface representation of all the scans and represent the actual deformations as displacements from this model. It then follows that for any arbitrary pose, within the pose space of the original examples, they can interpolate and reconstruct a novel and realistically deformed instance.

Somewhat similar ideas are expressed by a statistical technique for analysing the growth trajectory of a deformable surface (i.e. a face as it ages) as proposed by Hutten *et al*. [HBHP03]. In this, the importance of a single generic mesh across the training set is emphasised (see reference to conformation on page 26). They compute this for an entire population of 400 faces scans, using a thin-plate spline to help interpolate the alignment. Following this they then show how eigen-analysis can reveal strong correlation in the changes of face shape with age (particularly up to the age of 18).

Many other works similar to this are focused further within the medical domain, particularly using volume based representations. For example Chung *et al*. [CWP+01] seek to analyse volume change over time in the brain, and to map individual subjects onto a single generic model. To achieve this they propose a voxel based model able to cater to the rate of morphological change, directed by the observed velocity field describing the change. In so doing, they give a rigourous overview of the techniques for describing how a structure evolves over time using a stochastic evolution equation, to estimate this displacement field. By looking at the actual rate of change (by the Jacobian) they are able to highlight and measure the local volume deformations. What is particularly interesting from our point of view is the way they then describe the types of morphological change based on tensor analysis of their resulting displacement gradient matrix. From this they are able to decompose the rates of change into rotational and translational components, and strain - which is further defined by the overall *length* and *shear*. This idea of decomposition of the flow field into more descriptive types is similar to the idea of decomposing optical flow fields (see page 29). It should be noted here that they only consider first-order change, again appreciating that second-order calculations require more data to accommodate variability across spatial and temporal dimensions.

The work of [JPF+02] is somewhat similar in this respect. However, it considers the use of medial representation (i.e. a principal axis, skeleton, or manifold that runs “through the center” of the object, which is then described by displacements from this), used on anatomical objects for segmentation and shape characterisation. They phrase the problem as one of deriving the use of *M-reps* - single, atomic, implicit boundary objects that can be linked together to form the overall shape. Using this, they perform the fitting to image data based on a Bayesian deformable templates method, incorporating prior knowledge of geometry and shape. The variability when

fitting to novel data is accommodated by probabilistic transformations on this template, parameterised by such natural shape operators such as *bending* and *growth*. Analysis of the final segmentation can reveal an invariant set of features based on the relative coarse and fine operators used for the observed bending and growth. This can be used to statistically distinguish shape variability - as they demonstrated for shape of the *corpus callosum*.

Alternative work (but still medically motivated) looks more closely at this task of statistical shape classification based on deformation. Golland *et al*. [GGS+01] use a distance map representation extracted from volumetric data (aligned and clipped to a common size) to produce a feature vector to serve as a shape descriptor. From this they use a Support Vector Machine based technique to learn the optimal inter-class differences within the training set. Having constructed this, they then use it to derive the principal directions of deformation to explain, and visualise, how shape differences in feature space translate back to the original input data. Using this they are able to highlight where, and the extent to which these deformations actually occur.

Some of the most recent work for shape based classification has harked back to the approach of learning symbolic descriptors for the surface. For example, Ruiz-Correa *et al*. [RCMB+06] look at object recognition in range-data, for varying degrees of occlusion, clutter and (importantly) accommodating for shape variability. They set out to achieve this by first breaking up the training surfaces into individual component patches (all input is first aligned using a point correspondence algorithm). These components then take on the numeric values calculated from a spin-image representation, and then related to one another spatially by a symbolic signature which forms the ultimate shape descriptor for the classifier to learn from. Using this representation they are then able to reasonably successfully recognise and classify novel data - even when it varies in overall shape.

2.2 Capturing Deformable Surfaces

For this work, we are interested in the acquisition of non-rigid shape and its deformation. There are a number of challenges associated with this, not least with how to capture both simultaneously, how to pre-process the raw data (to perform hole filling, registration, etc.), and how to the resolve the correspondences over the surface as it moves (how to deal with appearing and disappearing points). This is directly relevant to our work since we seek the most accurate and dense - both *spatially* and *temporally* - model of real deforming surfaces.

In this section, we are interested in previous approaches in computer vision for the acquisition of complex, real, surfaces - particularly those that managed to capture the human form. The problems encountered with this can broadly be described as a compromise between accuracy in either motion or in shape, by disregarding one in order to reliably capture the other. A very good overview of the issues encountered for shape capture (primarily as input for computer graphics)

is presented in Hilton *et al*. [HSC02]. Similarly, the survey presented by Moeslund and Granum [MG01] gives a comprehensive overview of performing human motion capture from a computer vision perspective. What is immediately of note is that there is very little research that directly brings together these two distinct threads into a single, unified model to describe *dynamics* - an issue that lies at the very core of this work. In both cases, the processes are often simplified by constraining the environment e.g. its background, camera set-up, etc.

The classic example of this is the capturing of **motion from active sensors** using the range of approaches commonly known as "Motion Capture" by which a human subject is fitted with magnetic or optical sensors at key positions on the body. This enables the motion they make to be transformed onto an representation of skeletal structure as an articulated, hierarchical rigid-body object. While this approach is not strictly a "pure" computer vision solution it does rely on various signal processing and geometric translations to accurately interpret the data and to fit it to a skeletal model through inverse kinematic optimisation. An overview of this process is presented in Bodenheimer *et al*. [BRRP97]. Its great benefit, which has made it a favourite with the entertainment industry, is in the accuracy and ease with which complex motion can be used to control any arbitrary articulated model. Such useful information can also be employed in biometrics for gait analysis and athletic performance. The disadvantages are foremost in its sparsity when representing the accompanying shape and deformation of the motion, and in the fact that the information captured is only valid for discrete points.

2.2.1 Shape From X

It is actually possible to attempt the derivation of **shape from motion** (also known a *structure-from-motion*) by exploiting the fact that motion itself can yield information about shape. This exists as a fine line between extracting motion for its own sake, or to then use it to reconstruct shape. Traditionally, this seeks to exploit aspects of geometry, including parallax motion and optic flow, to yield a factorisation able to recover form in the classic work Jebara and Pentland [JAP99]. Unfortunately, this only generally holds for rigid shapes. Consequently recent attempts have been made to extend this to the non-rigid (varying over time) case. The seminal work of Bregler *et al*. use such techniques to directly recover $3D$ models of the human face from image streams [BHB99]. However, the results (while impressive for such an unconstrained problem) are rather coarse and only capture a crude approximation.

The **shape from silhouette** approach (also known as *visual hull construction*) on the other hand, provides a relatively accurate means to capture both structure and motion. This is achieved by combining the outlines of a subject taken from a surrounding array of accurately calibrated cameras. The most common way of achieving this is within a strictly contrived environment, in which the cameras have been positioned to take images from orthogonal angles of a subject against a (fairly) easily differentiable background. This can then be used, for example, in a static

case to estimate human body position and pose, as investigated by Cohen *et al*. [CMG01]. This approach has recently been extended further by Cheng *et al*. [CBK03] to offer real-time support in capturing full dynamic estimations of shape *and* motion simultaneously.

Shape from shading attempts to recover 3D information purely on the basis of visual cues offered by the natural surfaces. This is achieved by modelling the properties of (generally one) light source and the surface. Interesting recent work by Castelan [CH06] has improved this estimation for capturing faces by incorporating additional cues based on knowledge of local surface curvature.

Capturing **shape from scanned range data** offers the ultimate in detailed shape capture, but at the cost of the subject remaining completely static for a period of a number of seconds. A survey of the problems of acquisition of 3D models from range data (from a graphics modelling view) is presented by Curless [Cur99]. This highlights the necessary pre-processing and interrelated issues of *registration*, *reconstruction* and *segmentation* that needs to be tackled when successfully handling range data when attempting to further derive relationships to the underlying skeletal structure.

Alternatively, a **shape from multiple view stereo** approach (details of which are well described in the book by Hartley and Zisserman [HZ00]) can be effectively viewed as a similar problem to that of shape from motion - in that points in two images need be identified as being the same on the surface of the object of interest. Such a concern is known as the correspondence problem - and is one that has been often dealt with in vision by various means. One of the simplest solutions is to rely on accurate calibration of the cameras responsible for capturing the images, from which it is then possible to employ *epipolar geometry* to solve for the depth of the points. Other solutions make attempts at exploiting the control of the focal length properties of the cameras.

One of the simplest ways to improve the stereo process is to utilise a *structured light* approach to provide more effective local texture when resolving correspondences, as for example the early work of Siebert and Urquhart [SU94] demonstrates. The importance of initial calibration of the cameras involved, such that their intrinsic and extrinsic properties are recovered, is also crucial to the accuracy of the results gained. The novel enhancement of using a robust scale-space approach to reconstruction lies at the core of this work, by decomposing the input into a pair of spatial band-pass filtered image-pyramids. This is detailed more fully in the later work by Ju *et al*. [JBS03] which also demonstrates the crucial fact that with modern high resolution cameras, there is sufficient texture on most natural non-specular surfaces to permit the recovery of depth data *without* the need of structured lighting. We detail more fully the set-up and workings of such a system, along with the later 4D enhancements, in Chapter 5.

2.2.2 Acquiring both Shape and Motion

A number of the techniques described above either implicitly, or by extension, could be used to capture both motion and shape at the same time. In many ways, this can ultimately be viewed as an issue of sensor resolution, or rather the ability to retain a dense enough amount of data to adequately describe what the observed surface is and how it alters. For example, there are a number of problems with motion capture, as described by Chu *et al*. [CJM03]: expensive equipment, limited space, inconsistent markers, and problems with conversion of marker data to articulated models. However, crucially, in terms of describing shape, a few markers can hardly reveal the form of the subject. Furthermore, the question arises whether the positions of markers are indeed truly representative of the motion to be captured in the first place. In fact, they propose a technique extending from shape-from-silhouette that allows them to extract skeletal points, yet retain some shape information.

Similarly, at the other end of the spectrum for range data, there have been approaches in attempting to interpolate the variations in shape between multiple scans. For rigid data the problems can mean interpreting multiple exemplar data sets to discover the locations of articulations, as for example shown by Ashbrook *et al*. [AFRW99] for rigid articulated structures. More advanced attempts look even to the extent of modelling deformation from a set of scanned instances. The work of Allen et al. [ACP03] represents one of the most comprehensive attempts at this, and has be extended to statistically investigate the shape of entire bodies. However, the problem remains in then relating this to true motion, especially for smaller-scale subtleties of deformation, and in resolving correspondences on the surfaces over such widely varying instances.

Various other approaches have looked to combine techniques to further constrain both motion and shape recovery. Usually this involves incorporating a model within the tracking process to help explain and maintain consistency. Fua *et al*. [FGDP02] use tracking in 3D to fine-tune a deformable model represented as an implicit surface, in which each articulated limb is composed of “metaballs”. They fit the model to sequences of three synchronised and calibrated cameras from which they can recover a $3D$ point cloud of the subject moving. Only some initialisation is required to fit the location of the key joints to the first frame of data, and the benefits in using a geometric surface representation mean it is possible to accurately compute the distances to the observed data. Additional constraints of the final shape of the body can also be computed by comparing the silhouette from the original images to the projected surface outline. A similarly interesting work that also combines a visual hull with stereo for real-time capture within a multi-resolution subdivision surface based model is presented in Neumann *et al*. [NA02].

In recent years the increasing high-resolution and relative costs of video systems have also led to the possibility of video-rate capture of range data by stereo, leading to new commercial and academic systems. These are able to capture increasingly accurate and dense data from synchronised multiple views, although the recovery of each frame usually requires a considerable

amount of off-line processing. An example of a very recent system used in the automotive industry is presented in Godding *et al*. [GLW06] in the form of a number of high-frame rate cameras (up to 1000 f.p.s) and a very accurately engineered beam-splitting stereo scanner (used to negotiate synchronisation issues and cost). They also mention the challenges associated with tracking the deformations on a textured surface using such a system, in particular the need for high contrast in images taken under the low-lighting conditions due to high frame rates.

In looking at capturing more complex surfaces with such systems - such as the human body or face - the work carried out by Ypsilos *et al*. [YHR04] demonstrate the need for an active sensor approach to stereo for greater accuracy. To this end they employ a three pod set-up arranged around the face in which two of the cameras are delegated to recover depth by the addition of infra-red filters to enable them to see a structured light pattern projected onto the subject. The third camera in the pod only responds to visible wavelengths, and so is able to recover unaffected colour texture information of the subject. Once the stereo depth data is recovered, the head model is integrated from all the pods by incorporating both shape and texture into a displacement map (i.e. projected distance from an enclosing ellipsoid). They further incorporate temporal registration of the displacement maps to accommodate rigid head motion by minimising the re-projected distance between each frame using the ICP algorithm. Having done this they are further able to remove a degree of noise by integrating each displacement map over a spatial temporal window with a Gaussian smoothing kernel. In so doing they are able to remove some of the artefacts from the stereo match process in each frame - a useful trick for data cleaning, when able to integrate over time.

Many of these more recent works start to intersect with the concepts of scene and range flow (see Section 2.3.2). The similarities between stereo and temporal matching in order to track and recover deformations, while also recovering the 3D model, are highlighted in the work of Nebel and Sibiryakov [NS02]. In this they directly compute range-flow as the re-projection of 2D optical flow calculated from the images applied to the recovered model. As a side effect of this, they can use the motion estimate to then remove background random surfaces recovered by the stereo. Following this, they apply the range-flow estimation to guide the "skinning" of a virtual model, by first determining the position of articulated joints in the data (performed by initial manual division of the data points into associated limbs). From this, correlation in the motion allows calculation of weights applied to each vertex in the skin mesh, to reflect its deformation as a function of the joint angle.

The seminal work of Zhang *et al*. represents one of the most comprehensive work to date in terms of "space-time stereo" and its application [ZSCS04]. This exploits not only stereo, but the temporal relationship over time to further constrain and refine the output of the motion field. Their set-up is (for helping solve the correspondence problem) based on using filtered infra-red patterns to further resolve the correspondence problem. The particular insight offered

by their approach is to then extend on earlier work to generate more accurate and stable results by generalising the stereo matching into the temporal domain. Thus integrating the estimate over anything from a few closest frames to the entire sequence. They also describe a number of useful post-processing steps for removal of ridging artefacts, followed by effective hole-filling from tracking a base mesh model through the sequence using scene flow. Following this they apply a decomposition to the mesh to reveal the underlying kinematic model that allows for novel, subsequent, data-driven animation.

Another of the other most advanced recent works in this area is that of Wang *et al*. [WHL+04] in which they tackle not only the high-resolution acquisition of 3D facial expressions, but also in learning their underlying subspace such that they can be transferred to animate other face models. They use a unique high-speed system constructed around the projection of three phase shifted fringe patterns onto the subject (and by using a beamsplitter to also capture alternating colour information). Following capture, they then stress the importance of fitting a generic mesh reliably across all sequences and all subjects. The data is first roughly aligned by hand using 30 landmarks and a non-rigid shape registration algorithm based on free form deformations. Based on this, two different resolutions of the base mesh are used to track by splitting the mesh into several deformable regions (with few parameters) and using the low-resolution estimate to initialise the higher one to capture more subtle motion. Ultimately, having these tracked sequences with the same generic mesh allows for particularly interesting analysis of the data. For example, they go on to show the bi-linear separation of expression into style and content. This is achieved by using a local linear embedding to discover the underlying manifold of expression, from which to learn a generative model able to modulate along the dimensions of this manifold to either create novel sequences or transfer those motions to another face (see Section 2.4.2 for related work on analysis of the manifolds of expression).

2.3 Tracking Surface Dynamics

Capturing raw data of a surface moving over time is not immediately useful. As described in the last section, many of the more advanced approaches when dealing with both shape and motion need to also resolve where individual parts or points of the surface actually move to. A first stage to this problem can be to try and determine exactly what regions have actually changed, as opposed to those areas that have remained static. The core task is then that of resolving the *correspondences* between two instances of a surface. That is: where discretely identifiable points on the surface move to. Many problems are associated with this task, not least that deformation of the surface can cause self-occlusion. Conversely, the task of resolving how openings on the surface occur and form is also particularly difficult to handle. The classic example of this is the opening and shutting of the mouth on a face - which must be handled correctly in the case of

then trying to analyse speech and expression.

An important distinction of this section is that this effectual "tracking" of the surface is performed without the use of additional fiducial or other marking of the face at particular points. It is a contention of this work that such markers present an additional bias towards capturing dynamics. Techniques then range from exploiting intrinsic properties of the surface (such as colour and shape), to using a complete model to guide or explain non rigid deformation (and so correcting for it), to ultimately using more advanced techniques to entirely "unwrap" a surface so that it can be registered in 2D as opposed to 3D.

2.3.1 Registration of Change

A great deal of the work in capturing dynamic shape in computer vision is concerned with medical imaging. In particular, for describing the alteration in morphology for various organs of the body over time, such as the brain and heart. Such work is mostly based on volumetric data from MRI or CAT scans that must first be segmented from other parts of anatomy and then - crucially - registered as they change over time. In particular, the alignment of data must attempt to accommodate those regions which have actually deformed, as opposed to those areas that have in effect remained constant but simply moved.

In this respect, the local, intrinsic properties of the surface (rather than its overall form) can be useful in identifying change. For example, using the actual surface normals and curvature values can be used as a basis to refine the fitting between scans. The work of Cash *et al*. [CSC+04] is an example of such an approach in identifying soft-tissue deformation. In this they highlight the problem that global, rigid, alignment algorithms such as ICP have with attempting to minimise the distances across the entire surface (leading to misalignment). At the core of their proposed approach they design a cost-function that instead seeks to identify deformed regions and to thereby eliminate them from the overall registration. This is where they use the differential properties of the surface to identify change, and then try to relate these changes over time using the closest-point distance operator (effectively searching over sub-regions of the surface for the optimal number of points to include to retain a good match). Interestingly, they conclude that the central challenge in incorporating curvature changes involves determining correspondences between the points on the deformed and non-deformed surface (the topic of the next section). However, as it stands, using their approach they are able to gain much better registration than simply using ICP.

Another example of this more holistic approach for registration is to compare two surfaces using the thin plate spline (TPS) representation (common in geometric morphometrics) to estimate the actual non-rigid displacement between them, and so delineate the deformation and isolate areas that have not changed. This use of thin-plate splines has continued in recent work for registration of deformable data. One focus of work aims to handle those inaccuracies re-

sulting from the actual 3D scanning process, and to resolve actual small deformations between scans. For example Brown *et al*. [BR04] combine a hierarchical ICP based alignment to provide guiding correspondences that then control the warping of a TPS. Ultimately, they seek to explain the deformation between two mesh based representations - as extrapolated by the continuous, analytical definition of the TPS, which further imposes reasonable solutions due to its inherent minimal bending energy. Using such an approach they are thus able to accurately handle low-frequency warps present in scanned range data.

Related to this approach, particularly by the means of constraining "realistic" deformations, are numerous other works that seek to physically model the non-linear nature of natural tissue for medical applications. For example, the early work of McInerney and Terzopoulos [MT95] applied a FEM based anatomical models to the task of tracking cardiac change. For this they exploit a generic, dynamic "balloon" model as a triangular mesh that is capable of being fitted to data by mediating between local forces and observed motion.

Similarly, more recent work of Vadakkumpadan *et al*. [VTS06] treats registration as an extension of parametric 2D image manipulation. They formulate this as a task to optimise the parameterisation that describes the matching of one mesh onto another. They furthermore constrain the validity of the solution with an elastic model that is made up of a linear combination of eigenvectors that map the deformation to a $2D$ problem. This is the body of work that then begins to intersect with the larger field of model based tracking. Thus, in many ways, this approach is also similar to the use of "Snakes" or "Active Appearance Models" (AAM) [CT04] to fit and direct the dynamics of a pre-designed model to incoming data (even though it deforms). The work of Goldenstein *et al*. [GVM03] for example represents a comprehensive recent attempt at performing image based tracking of a parametric deformable model. They crucially seek to speed up computation by use of Directed Acylic Graphs to effectively combine local displacements into single "cues" that then affect the parameters in the model.

Continuing work along this line has also recently been spurred on by applications that seek to perform 3D face matching invariant to the expression on the face (as described more fully in Section 2.4.3). Other work for ensuring accurate comparative analysis of scans relates to landmark oriented schemes, coupled to physical elasticity modelling as a natural extension of biometric morphometrics. In the work of Basso *et al*. [BPV06] the specific problem of registration of $3D$ facial models for the training of deformable models is advanced. Usually, this process focuses on alignment between a neutral expression and the training instance. However, they propose a technique for feature correspondence based on an energy regularisation algorithm and a $3D$ Morphable Model, from initial correspondences proposed by optical flow.

Similarly, the technique of *conformation* to face data by fitting a generic model to instances of scanned data can be viewed as equivalent to registration. This is performed by alignment crucially at uniquely identifiable landmarks, used to then guide an elastic model to minimise the

rest of the fitting. The key idea is that the same generic mesh is used to fit to all instances of the data, and thus resolves good correspondences (since the landmarks serve to "pin down" the rest of the fitting). The recent work of Mao *et al*. [ZSCA04] is an example of this where they directly seek to tackle the problem of outliers caused by local deformation that can radically affect this process. Their approach is to first perform a global mapping - which unfortunately relies on a degree of manually defined landmarks - following which they locally deform the mesh to fit. At this stage they do not simply fit to the closest vertices in either mesh, but instead establish a similarity measurement for the local area to determine which vertices should "pull together". Here again the intrinsic properties of the surface, specifically the surface normals and principal curvatures are used to identify more likely matching candidates. They use this for comparing scans of the face over a number of years, and assess its accuracy (using only 5 landmark features) to have 95% of the generic mesh triangles within $1mm$ of original model.

This was followed up by the later work of Ju *et al*. [JMS+04] in pointing out the applied benefits of conformation for hole-filling, segmentation of articulated body parts, and, crucially, for describing the variation in shape using a Point Distribution Model. From this they were able to establish (through Principal Components Analysis) the principal eigenvectors along which changes occur from a mean shape, yielding interesting analysis. In many ways this work is thus similar to that of Allen *et al*. [ACP02] presented on page 17.

The use of manual landmarks is a hindrance to the true applicability of registration. More recent work has certainly re-focused on robust, automatic ways of constraining surface fitting. Furthermore, the idea to factor $3D$ registration such that it becomes more a task equivalent to $2D$ registration (and hence simpler) is a very promising research trend. This is however very much dependant on how to preserve the relative distances involved (i.e. geodesics). Fundamentally, when performing registration, the emphasis is very much focused on resolving the correspondences between multiple scans of an object, in order to then guide and resolve any deformation. The main problem encountered here is to handle those situations where points effectively appear or disappear from view (e.g. the opening/closing of the mouth).

A good example of a theoretical framework for resolving correspondences in a sequence of $4D$ data is the work of Huang *et al*. [HZW+04]. In this they use a multi-resolution face model, coupled with hierarchical tracking of global and local deformation. What this effectively means is that they impose very strong smoothness constraints at both the global and local level, resulting in smooth, continuous and dense one-to-one correspondences. In reality they must still perform initial registration (using ICP) and divide the face into a number of regions with controlling parameters. They furthermore also still allow for user defined landmarks, but these then serve as hard constraints to further guide and correct the tracking as necessary. The overall results, particularly for tracking subtle expressions involving points around the corner of the mouth, are very good.

The extension to this work by Wang *et al*. [WGZ+05] provides one of the most advanced solutions to the problem of dense temporal correspondence by using harmonic maps. A harmonic map is a diffeomorphism between two instances of disk topology, which can be viewed as the embedding of each manifold into a planar graph while crucially seeking to minimise the bending energy to retain relative distances (given that the edge of the disk acts as a fixed boundary constraint). The analogy here is of a rubber disk that is covering the object is question, which can also be additionally clamped at interior points (acting as additional feature constraints). This sheet can then be pulled away to reveal a $2D$ representation of the surface for matching, and so is much more tractable. Using this, they then build on their previous work to initially align the data, and then use the harmonic map to track motion between adjacent frames. This is achieved by using a sparse set of features detected using corner detection and other standard techniques on texture and curvature mapped disk images. Using the correspondences between these, it is then possible to reverse the mapping to discover where the points actually move in $3D$.

Similar work should be noted for the use of spherical maps, as performed by Starck and Hilton [SH05]. This instead uses a mapping in the spherical domain that can be applied to matching two temporal instances of data (in this case a entire human body captured from a multi-view system). Again, this manages to reduce the problem to matching non-rigid surfaces in $2D$ instead. They also apply an iterative course-to-fine approach to fine tuning the match in terms of regularisation framework that takes into account the surface disparity in terms of colour and surface normal, along with similarity in the actual deformation. Impressive tracking results of entire people are shown.

2.3.2 Dense Flow Methods

The problem of resolving the correspondence between two sampled points spatially - as in the case of performing dense stereo capture - is exactly the same issue as resolving where that point moves over time (i.e. between frames). This is the crucial realisation when considering 4D tracking - that the correspondence problem occurs within the complete spatial-temporal domain. Whereas "tracking" performed spatially can be used to construct *dense* correspondences between surfaces from multiple views - the exact same techniques can be applied to tracking *dense* flow over time.

2D Optical Flow

In this work we are particularly interested in the nature and formulation of instantaneous surface tracking, the topic of computer vision most often associated with *optical-flow*. A very comprehensive survey of this is presented by Beuchemin and Barron [BB95].

Indeed, it is sometimes forgotten that the modern seminal work that defined the whole field of reconstructing dense flow - that of Lucas and Kande in 1981 [LK81] - was originally intended

as a method for stereo recovery. Yet, this work clearly motivates and defines the challenges associated with the recovery of an accurate dense flow field. Such a field - as they define it - is simply the 2D vectors that describe the translation in the image plane (traditionally in a (u,v) co-ordinated system) between every pixel and its location in the subsequent image. In this way, the algorithm can be effectively compared with an image-alignment problem and the many ways in which this can be resolved (or optimised).

In most image based optical flow - only the intensity (grayscale) changes are considered. This can potentially disregard other potentially useful information that could be used to resolve dense flow more accurately. The use of texture has for example been considered by Arrendondo [ALL04], but it is the use of colour in particular that has been proposed as more generically useful, as looked at by Andrews *et al*. [AL03]. This includes consideration as to what colour-space is more useful - a topic that is very much concerned with the capabilities of the actual sensor and storage format of images (since issues with quantisation and encoding have a large impact in accuracy).

The justification for this is that information contained within additional colour channels is more representative of the actual surface properties of the objects in question, as pointed out by Golland *et al*. [GB97]. When considering that intensity depends on more global assumptions about illumination and reflectance, then surface invariant colour characteristics have obvious benefits in resolving certain types of localised motion ambiguity. This is particularly the case in observing non-rigid deformation where attempting to track motion on the surface (even with accurate stereo correspondence) becomes an issue.

Recent quantitative assessment of standard optical flow techniques modified to include colour have shown some improvement over simply adding gray-scale intensity. Again, the rigourous comparison and assessment of the benefits was carried out by Barron *et al*. [BK02] . That work also illustrated the issues in deciding which channels (intensity, saturation, hue) to select and how to adjust the weightings between them, such that they can contribute to the overall accuracy of the flow field. This a particularly appealing idea, in that it seeks to directly resolve ambiguity in the motion by assessing the predictive results of each channel's contribution.

In image based techniques it is common to use optic flow, with the possibility to decompose the resulting vector fields into component elements - such as shear, divergence and curl - that reveal something of the type of motion, and perhaps the deformation. This was originally proposed by Koenderink [KvD86] in his accessible, but early, overview of what optical flow represents from a perceptual point of view. His real insight is that analysing optical flow fields for these "differential invariants" can reveal something of the shape of the object in question, not only from a rigid transform, but also (in the case of shear and divergence) changes in the actual surface. These observations have recently found their way back into the "structure from motion" problem for monocular systems.

3D Scene and Range Flow

An important realisation when calculating 2D image flow, is that the observed motion is a projection of the actual 3D motion present in the scene. That is to say, it would ideally be more accurate to describe the motion intrinsically as it occurs in $\mathbb{R}^3$. The calculation is of course dependent on accurate 3D data captured by a suitable sensor that is also - most importantly - capable of sufficient high frequency over time.

This has its basis in a number of interesting experimental psychophysical works in the 1980's into the perception of *motion flow* and the limits in the human visual system - very much related to later work (as for example that of van Doorn and Koenderink [vDK84]. This included work investigating the seemingly separate, but much connected, pathway for integrating colour information - starting from early work by Ramachandran and Gregory [RG78]. This is a continuing interesting field up to today (e.g. Monnier and Shevell [MS04]).

The computation of a dense instantaneous velocity field that describes the displacements between two or more instances of a 3D surface can be described as the *scene* or *range-flow* (i.e. dense velocity flow fields describing point motion) . The general distinction between the two terms is that the former is usually derived within a dynamic synchronised multi-view camera framework, while the latter usually from discrete moments of scanned data from laser or fixed stereo systems. However, fundamentally these both extend the application of optical-flow techniques with the aim to accurately reflect the motion and deformation of surfaces as 4D flow (i.e. across 3D space and time). Estimates of observed optical-flow can be used to recover depth changes, particularly when using multiple cameras and feature correspondences. In one respect, this is where optical flow intersects with the problem of also deriving shape from motion, since the two are interrelated. As optical-flow is the projection of 3D surface motion onto an image plane, it can be used to recover the scene-flow as defined within the context of a framework that must take account of multiple camera intrinsic parameters and lighting conditions from various angles.

The seminal work of Vedula [VBK02] [VBR$^+$05] is largely responsible for first of all coining the phrase "scene flow" and defining it within this context. The earliest work introduces the main issues concerned with capture, particularly in dividing the process into three main scenarios. These represent the various cases where either complete knowledge of the structure in view is known, whether only partial information as to stereo correspondences on the surface is understood, or (in the worst case) no prior knowledge of the surface is modelled. This in turn led to issues concerned with interpolation of both the shape and the motion across time.

Alternatively, an approach can be adopted for raw range data - whether from a stereo or a triangulated sensor - to derive *range-flow*. In theory any optical flow technique can be applied for estimation, but the most investigated approach (which we directly build on) uses differential estimation techniques for measuring the temporal-spatial gradient as performed by Spies and

Barron [SJB02] [SB02]). One advantage is that range-flow can be computed directly on the original surface by using localised estimates calculated directly from the sample grid as the basis for creating a regularised, smooth flow field. Range-flow can in fact be computed from raw 2.5*D* stereo depth-maps as if they were intensity images in order to track the apparent depth change. However, the resulting 2D flow estimates must also be re-projected into 3D space, potentially leading to error over a large depth-of-field.

Both scene and range-flow are fundamentally similar in that they seek to capture the 4D motion and deformation of a surface. An advantage in using the regularly sampled depth information (e.g. from laser-stripe scanners and dense stereo capture) is that it is often accompanied by aligned intensity or colour information, which may not have been directly utilised in the initial 3D capture. For example Barron *et al*. [BS00] demonstrate this idea of fusion, in which the extra intensity information can further constrain and thereby improve the accuracy, reduce aperture ambiguities, and increase density in the flow estimates. This is further illustrated in the case of incorporating intensity with range data by Spies [SJB00].

Range flow can then be applied in a number of situations to observe the changes in deformation and to then attempt to explain them. For example, it can be used to infer the elastic constraints of an object when manipulated by a robot, as shown by Lang and Pai [LP01]. Increasingly, such tasks then rely on the computation of range-flow in real-time. An early example of this is that of Yamamoto *et al*. [YBBR93] which is based on an underlying deformable "net" model that is manipulated by a linear combination of local motion. Importantly, they also noticed how to improve the estimation by the addition of the image data, as well as the range data, to eliminate ambiguities. Similar work by Aoki [AK06] also takes input from a real-time stereo system. This also further bears out the use of combining local shape "height" and colour into a histogram to allow more robust point-to-point correspondences to be resolved.

The application of 2D optical flow, 3D range flow and combinations of the two estimation methods to the analysis of the face and its dynamics (i.e. expression) is a common trend in computer vision. This is primarily in response to the complex, yet familiar, nature of the data and in the fundamental question it poses in the use of motion for recognition. The issues associated with this interpretation are presented in the following section.

2.4 Recognising Faces and Expression

Images of faces have been the focus of considerable research, particularly with the objective of establishing the identity of the person in question. It is considered a particularly appealing problem, not only for its non-intrusive aspects, but also as it involves investigating how we ourselves can perform so well at visual identification of people. The topic of face recognition has a considerable history within computer vision, for example Bledsoe in 1966 summed up the

challenges as:

> "This recognition problem is made difficult by the great variability in head rotation and tilt, lighting intensity and angle, facial expression, ageing, etc. Some other attempts at facial recognition by machine have allowed for little or no variability in these quantities. Yet the method of correlation (or pattern matching) of unprocessed optical data, which is often used by some researchers, is certain to fail in cases where the variability is great. In particular, the correlation is very low between two pictures of the same person with two different head rotations." [Ble66]

This section presents the more applied works that are related to the interpretation of the face, and particularly expression in $3D/4D$. This invariably has a considerable basis in the psychological literature. However, the core idea for our work is that the face is a surface - just as any other. It is certainly one of the more complex and intriguing real surfaces, and so offers considerable challenges in the topics we have reviewed so for acquisition, tracking and representation. This further explains why it has been such a favourite topic within computer vision, and here we highlight those works that have directly looked at the face.

2.4.1 A Psychological Context

The face can be viewed as the ultimate medium of expression. As such, it has attracted a huge amount of research within psychology, much of which also touches on a number of graphics and vision techniques to enable experimental manipulation. This section gives a brief overview of the importance of motion and shape in face perception, especially as underlying the understanding of face data. An excellent recent review of the wider cultural and scientific understanding of the face is presented in the book by Kemp *et al*. [KBL04].

The considerable investigation conducted by Paul Ekman over the last 40 years specifically into *facial expression* [EF78] has resulted in the key establishment of the *Facial Action Coding System* (FACS) for identification of expression. Initially, Ekman proposed that there were effectively six basic forms existing at various physical positions corresponding to the basic extremes of emotions (sadness, anger, joy, fear, disgust, surprise). However, it was soon realised that this approach was unable to handle the diversity and ambiguity inherent in the variation in face shape and individual performances requiring a more detailed approach. The FACS system is still applied very much in recent works, particularly in marking-up and analysing data within experimental psychology. For example, a particularly interesting work by Cohen *et al*. [CSGE02] attempts to rationalise the claim how expression can affect identification of the individual.

Continuing work by Pollick *et al*. has further applied motion analysis to the study of the face [PHCP03]. In this they captured a number of expressions as marker data, which they further manipulated to exaggerate both spatially and temporally. From this they showed that (relative to a baseline neutral expression) exaggerated spatial expressions were perceived with greater

intensity, while temporally enhanced expression showed less of a trend for perceived intensity (excepting a small factor for slower expressions such as a frown).

Further experimentation into the effects of dynamics in the surface of the face on the role of modulating the intensity of the expression is the topic of research performed by Kamachi *et al*. [KBM+01]. In this they play back morphed sequences of images from neutral to peak expression. Surprisingly, the speed of play back influences the recognition of the expression. For example sadness is more identifiable when played slower, whereas happiness (and to some extent surprise) are more recognisable when played at a faster rate.

Attempts to quantify the degree to which the perception of facial expressions can be differentiated have raised some particularly interesting results. The work of Young *et al*. [YRC+97] construct experiments utilising techniques in computer graphics to "Megamix" faces in order to assess how much percentage of an expression is required to consistently recognise it. They effectively discovered that about 70% of an expression was required to differentiate it from the "mixed" one. Further work also revealed the role of intensity to caricature and thereby enhance identification; and also how configural coding - the relationship between facial features - is also important to the recognition of expression. This result is backed up by later work that justified how expression is separately processed from the use of features for identity, as seen seen by PCA based analysis performed by Calder *et al*. [CBM+01].

2.4.2 2D Interpretation of Dynamic Faces

Similar to the experimental manipulation of face data to see how it is perceived, in this section we look at the inverse problem of how to interpret face data to understand expression. The majority of work involved in the dynamics of expression has been driven by $2D$ video based capture, primarily because real-time $4D$ acquisition of the face has only recently become achievable (as seen in Section 2.2.2). None-the-less, a great deal of important work in computer vision has attempted to automatically analyse the face which is highlighted below. We are not however interested here in the task of image based $2D$ recognition of neutral faces. We are only particularly interested in dynamic flow-based methods, especially ones that attempt to describe underlying shape change. One of the main applications touted by this desire to recognise expression is for Human Computer Interaction (HCI), such that software is able to react to the emotional state of the user.

Perhaps the biggest challenges in this area of research is (as hinted at the end of the last section) the fact that people often display a "mix of emotions" on their face at any one time. In many cases the systems described below are only really capable of handling the extreme prototypes offered. Related to this is the fact that the traditional six classifications of expression are perhaps too limited to capture the entire range of human emotions, and that it can be hard (particularly depending on cultural aspects) to sometimes categorise one type or another.

Flow Based Analysis

While the issues related to static recognition of identity also affect dynamic sequences (e.g. detecting the presence of a face in the scene, tracking and aligning the rigid motion of the entire head) the main differences between works for expression recognition is in using either static images (i.e. the final state of a smile or frown), or to use the complete dynamics of the sequence (i.e. raw video or tracked data for every frame of a face moving). The excellent but now slightly out-of-date survey by [PR00] highlights this distinction and presents an excellent overall framework. However, it is the latter case of actual dynamics that interest us in this work.

Most of these approaches can be directly contrasted with the early work that was more directly engineered towards seeking to reconcile direct comparison to the Facial Action Coding System. One such work (Bartlett *et al*. including Paul Ekman himself) attempts to automatically relate video of expression to Facial Action Units [BHES99]. In this they employ a number of different techniques in order to create a hybrid system that includes spatial analysis, a specific "wrinkle detector", and optical flow motion. The spatial analysis is based on a statistical treatment of the difference images between neutral and expressive faces before performing PCA (in a form of "eigen-expressions") using a neural network for classification. This is combined with a measurement of linear structures (i.e. wrinkles) in particular regions of the face (i.e. brow, eyes), along with template based matching of the optical flow fields calculated using a Horn and Schunck based approach. The information from all these classifiers was then fed into a single neural network to achieve results nearly comparable with a human expert at labelling FACS action units.

In the work of Lien *et al*. [LKCL98] they are also motivated to try and directly estimate FACS using a multi module approach combining feature point tracking, spatial "furrow" detection and optical flow analysis. Their approach to flow analysis is interesting in that it directly seeks to derive the respective principal components - or eigenflows - indicating the main directions and regions of motion. Final classification is performed using a Hidden Markov Model to explain the transitions and relationships between different regions of the face. Furthermore, they also attempt to gauge the intensity of the expression, by using the correlation property of the PCA to find the best motion example from the training data set that has already been labelled as having a particular level of intensity.

Additionally, the research can then be broadly divided into those approaches that treat dynamics with either a template based or raw flow based manner (or hybrid between the two). Possibly one of the earliest, and most comprehensive works that combine these two methods was that of Essa and Pentland [EP97] who directly attempted to relate a derived optical flow field with an underlying geometric (and physical muscle) model of the faces structure. This produced a reliable parametric representation of which muscle groups were active, which could be directly related to the facial motion. Recognition of the expression could then be performed either by

directly comparing the muscle activation, or to use the model to then generate spatial-temporal motion templates of the expression to then match to. This then allows a data-driven approach (not based directly on FACS) instead deriving underlying probabilistic motion and muscle activation. Importantly, the underlying physical model is also used to help guide the optical flow estimation (i.e. to constrain it to plausible motion).

This theme of integrating flow with a model was continued in work by DeCarlo and Metaxas [DM96]. In this the motion acts in combination with the physical constraints encoded within a mesh to create a single FEM based model. This includes an interesting way in which deformations are directly handled as localised rigid and bending operations. These are specified using a small number of parameters, and can be combined to construct the entire face model. Special concern is given to the tracking of edges, for example as occur around the mouth. The high-point of their approach is then in being able to extract both the shape and motion from image sequences in order to track motion and deformation.

Another particularly interesting early paper (from our point of view) is that of Black and Yacoob [BY95]. In this they explore the middle ground between pure template and flow based methods by exploiting piece-wise parametric models to explain localised expression. Specifically they separate the face into rigid and non-rigid image regions, and then use a set of operators such as curl, divergence, and deformation to relate the flow field to the underlying structural change on the surface. These are more robust since they do not require explicit modelling of facial structure and geometry, plus they form more useful abstractions for describing the expression. This they apply once a sequence has been successfully tracked, by quantifying the resulting parameters into various mid-level and high-level predicates that describe the start and end of an expression. For example, the beginning of “surprise” is related to “raising brows and vertical expansion of the mouth”, in turn related to upward translation of the brow region and elongated deformation of the mouth area above certain thresholds. By relying on such local descriptors, this approach is able to implicitly handle head “looming” and additional motion.

Manifolds of Facial Expression

From these earlier works which have focused on feature, template, and model based flow interpretation, the trend has shifted towards more statistical and purely image based analysis of expression dynamics. These have initially found their basis in the transitional modelling to try to capture the interplay between regions of the face, especially in the presence of other signals (such as speech). Such techniques are particularly relevant for sequences of longer duration where any number of different expressions and motion can be interwoven together. Another interesting realisation is the potential to de-couple the “style” motion from the underlying structure, in a similar way as has been applied to motion-capture data to perform “motion editing”. The overarching theme to creating statistical models of the variations in the face is in many ways seeking

to understand the manifolds that best describe the sub-space of expression. Many such works apply statistical techniques to complete sequences of images simply in order to discover intrinsic dimensionality, with respect to then attempting to perform some form of classification. What it is noticeable from this work is that expressions are indeed separable, but only on the basis of accurate tracking and in employing suitable dimensionality reduction and embedding functions to discover a suitable manifold.

Earlier work has focused on the interesting idea of using a Hidden Markov Model (HMM) to label temporal cues during a sequence of expression. This was initially explored in the work by Cohen *et al*. [CGH00] which exploit a multilevel HMM to automatically segment and recognise expression based on increasing levels of separation by using a Maximum Likelihood classifier at the top level. This attempts to increase the discriminatory ability of the six basic expressions (represented by lower level HMM's) by trying to find the probability that a sequence not only shows one class of expression, but also that it does not show any other others. Training of these individual models is performed by pre-labeled sequences and the Viterbi algorithm.

From this has stemmed even more complex versions of unsupervised systems that automatically learn the variations of expression. For example, Hoey *et al*. have constructed a hierarchical dynamic Bayesian network based on a mixture of Hidden Markov Model Chains [Hoe01]. With this they were also able to perform supervised learning of five different types of expression, gaining an average reported recognition rate of 98%. Further work by them extended this to even more complex POMDP models for unsupervised clustering of sequences and associations to the given context (for example, when playing a simple card game) [HL04].

Some of the more interesting recent work on the face continues this theme of data-driven learning and separation. One approach by Chuang *et al*. [CDB02] de-couples a model of expressive features from underlying content. This then allows them to retarget a sequence in which the speaker, for example, speaks with a happy expression to one in which the same sequence is played but the speaker might look instead angry or neutral. This is achieved by a bi-linear approach based on [TF00] that inter-links the shape and texture of observed faces. This is used to create a factorisation model encoding on the one hand the expression alongside the "visieme" (visual phoneme) for the content of speech. From this it is then possible to generalise to novel configurations of expression applied to new sequences of content, or to interpret novel sequences for a known expression.

More recent research has directly focused on the underlying manifold of facial expressions. The key observation made by Chang *et al*. [CHT03] is that images of a subject's facial expressions define a smooth manifold in the high dimensional image space. The ideal here is to try to create a global analytical framework for the space of all possible expressions. To achieve this, they note that the choice of an embedding algorithm is crucial (e.g. both Lipschitz and LLE were investigated). To handle invariance to scale and illumination in the images they use Active

Wavelet Networks (a method of replacing the PCA based texture model in Active Appearance Models with a wavelet representation). This results in a number of reliably tracked features that act as input to embedding. Further work by them [HCFT04] use an Isomap embedding, and they extend the Active Shape Model further to probabilistically select the appropriate model for reliable tracking within a more robust framework.

2.4.3 3D Analysis of the Face

Another important aspect is the way that an image of the face is actually encoded, such that correspondences can be made between different human subjects - as pointed out by Troje and Vetter [TV96]. These can be grouped into two sets of general problems associated with **alignment** and **normalisation**. Normalisation is primarily an issue of correcting appearance, to primarily reveal *invariance to lighting conditions*. It can also refer to the removal of texture, colour, or even elements of disguise. An extreme example of this is to further compensate for the actual expression of the face. Alignment is most often considered a problem of registration by performing *head pose recovery* and possibly further *warping* in order to register the face in a neutral pose and size. Some of these factors they note (particularly for normalisation) can be handled by using a different modality - such as infra-red - in order to try and increase robustness. However, the important realisation is that many of these problems are as a result of the fact that the face is an inherently $3D$ structure and that the task is made hard because there are so many degrees-of-freedom and environmental factors that alter its appearance when projected onto $2D$.

Consequently, it is claimed that $3D$ recognition of faces has much better potential than simple image based recognition, since all the effects of lighting, orientation, texture, etc. are eliminated. For the most up-to-date review of $3D$ (and combined $2D$) face recognition the survey conducted by Bowyer *et al*. [BCF06] offers a very comprehensive review and assessment of the performance of various approaches. One interesting point made by them is attempting to dispel the "myth of illumination invariance" by making the valid point that this only necessarily holds true for sensors or techniques that are unaffected by changes in lighting conditions.

It should be stressed here that in general the current basis of research in using $3D$ face scans to date is very much geared towards static capture, unlike the dynamic aspects enabled by high frame-rate video. One way to approach the problem is simply to extend the concept of "eigenfaces" but instead performed on the depth-map of a face, as for example in Hesher *et al*. [HSE03]. However, recent renewed interest in the topic has focused on more "pure" $3D$ interpretation, fuelled by the creation of the Face Recognition Grand Challenge data-set (see the overview presented by Phillips *et al*. [PFS+05]). This has been used to finally push for reliable recognition, mainly in response to demands for security and biometric applications. In the remainder of this section we highlight some of the more interesting research that directly influences our own work.

Recognition by Alignment

Most systems for 3*D* face recognition are based on alignment of query face data to a previously stored model. The decision is to be made by effectively measuring some degree of similarity between the two, in order to decide if it is the same person. This is then a question of *registration*, for which the most widely applied solution is the Iterative Closest Point algorithm. Various enhancements to this approach have been presented in the literature (particularly with respect to improving speed), but in essence it remains a problem of minimising the distances between two sets of points - usually expressed as the Root Mean Square (RMS) error (see Section 2.3.1 for more details of ICP). In this work we are interested in how such approaches are then extended to looking at the face.

A typical example of this approach is seen in Lu *et al*. [LCJ06] which has the entire pipeline of acquisition, registration of multiple scans, and cleaning of the data. What is particularly novel about their approach is the course-to-fine alignment they adopt by utilising a shape index value for every point to first discover the location of key features (nose, eyes, etc.) onto which they can project the probe depth-map, before they then fine tune with a hybrid ICP based on the point-to-plane distance metric. They also compare texture information by using appearance matching based on a constrained Linear Discrimant analysis model to synthesise new textures that complement the 3*D* registration. The combined system integrates the registration and appearance based approaches to improve overall accuracy.

A combined multi-modal approach using “4D” has been investigated [PR04] that complements the 3*D* information with texture. Note that the concept is at odds with the concept of 4*D* expressed in this work (i.e. 3*D* over time). The distance between the model and candidate data is measured by the Euclidean distance between 3D and texture similarity. The contribution of the texture can be weighted by distinctiveness, which is useful in the case of poorly illuminated expression data where texture cannot help resolve the alignment. Their primary result is to show that combining texture information in the registration process can, in the case of frontal views, help improve the performance of recognition.

Pears and Heseltine [PH06] use an alternative approach to perform registration, relying instead on isoradius contours. These are defined relative to a single point of reference, for which they use the tip of the nose (discovered as the point closest the sensor). From this, the entire face can be decomposed into a set of such contours of different radii, along which different surface properties (including curvature) can be recorded. Alignment can then be implemented as as simple 1*D* correlation along these contours. They point out how this method could furthermore compensate for expression by selecting to align using those contours that give the strongest correlation, discarding the ones that are most affected by changes in the surface, and thus have greater variance.

Curvature Based Features

One of the main problems often associated with ICP are its requirements for a good initial pose between the data - i.e. alignment. Otherwise the attraction of local minima may not be overcome in order to discover the optimal solution. When matching the face, an initial stage of feature extraction can be performed in order to perform a reasonable initial alignment (and so avoid "falling" into any minima). Many such schemes exploit aspects of curvature of the face - a topic of direct relevance to this work.

For example, the early work of Gordon [Gor91] considers the depth-map in such a way, but utilising other higher-level descriptors based on curvature, such as watershed and contour operators. In this she directly performs calculation of the fundamental forms from raw depth data, following which Gaussian (K) and Mean (H) curvature can be computed. The importance of smoothing when handling real data is highlighted by this work, particularly since second derivatives are particularly sensitive to noise. An adaptive technique can work best in this case, in order to preserve subtleties in local regions that might be obliterated with a global smoothing operation. She also mentions that simple volumetric comparisons will be affected by expression, particularly around the mouth and cheeks. Interestingly, she also proposes in future work the need for more qualitative feature based descriptors to describe particular regions of the face.

One of the earliest works that directly tried to analyse range data of the face is that of [YD93] who showed techniques for directly labelling the components of the face from range data, by exploiting a diffusion process to find convex and concave points, followed by a threshold to label particular regions of the face. More advanced work by Gordon [Gor92] actually presents a recognition system which demonstrates the usefulness of the high level facial feature extraction methods. In this she also stresses the importance of view-point invariance - since such curvature based features are embedded on the actual surface. Using curvature enables extraction of features such as the nose ridge, and the corners of the eyes, and the width of the head - following which additional scalar values are computed between the discovered features, to construct a feature vector that can be used for classification. Even with such a simple feature set recognition was reported at 70%.

In the work of Moreno *et al*. [MSJV03] they continue this trend for using curvature to derive a set of 3D surface feature descriptors to improve recognition to 78%. Their input is a database of 420 faces captured using a laser scanner, which they then segment on the basis of HK curvature into a number of distinct regions. From each of these regions they then extract the descriptors based on quantifiable properties such as area, distance between centres of mass, mean H curvature, etc. Each of the 86 features is furthermore assigned a discriminating power, computed using the Fisher coefficient. They then follow the traditional path of feature selection and classification (based on experimentation in the selection of the optimal feature vector).

This question of what are the best features forms the focus of other work. In Heseltine *et al*.

[HPA04] a combination of features are used, based on a "Fishersurfaces" sub-space commonly used for classification in 2D. They perform a comprehensive set of tests to establish the optimal set of discriminatory features to use (settling on 184 dimensions extracted from 16 surface representations). They are able to achieve this as a result of first acquiring a unique dataset of 1770 different face models of 280 people.

Using such contour surface features also persists, for example Samir *et al*. [SSD06] focus on the use of "facial curves". The extraction of a level curve function, operating over a depth-map, requires a number of pre-processing stages to fill and smooth the data, before comparisons can be made based on the geodesic distance between curves. Using 6 such curves can reveal enough discriminatory information to enable classification. They claim that such an approach is more robust than curvature based approaches to the presence of noise. However, this technique must work only on already constrained or aligned captures that are looking straight ahead. This is similar to work by Beumier and Archeroy [BA00] in showing that the central facial profile from a 3D scan can be used, and is better than using a profile from the side of the face.

Curvature of faces has also been used for the slightly different task of face detection. In Colombo *et al*. [CCS06] the emphasis is then in trying to find faces, even in the presence of multiple subjects, or occlusions. Again the emphasis is in trying to discover the location of salient face features such as the nose and eye sockets. These are found by exploiting a thresholding scheme to find areas of high curvature and classification based on HK to consider only positive convex and elliptical concave regions. Once discovered, the triangular region between the two eyes and the tip of the nose is extracted and used in an eigenfaces based scheme based on the depth-map image. This approach offers a certain robustness to expression since this area of the face is relatively static.

Invariance to Expression

The other main issue concerning alignment of 3D face data for recognition is how to handle the significant changes that can occur between "gallery" and "probe" data depending on expression. In many regards 3D face recognition is more sensitive to expression than 2D recognition - since the changes of the surface are actually much more considerable than they might appear. This is currently a much investigated topic since it is perceived as the greatest obstacle to genuinely useful recognition systems. From our own point of view we are interested in this work for the issues it raises in what portions of a surface actually change, especially in terms of the geometry. It should be noted however that only static scans of final expression are considered (i.e. there are no 4D systems that directly exploit the deformations themselves as a means of recognition).

The work of Lu and Jain [LJ06] has focused on the problem in matching when an expression is present. They approach this by first learning a set of template deformations from a control set of subjects, which can then be applied (by using a Thin Plate Spline mapping) to generate a set of

synthesised models for the query to then match against. They learn these templates by marking up a set of fiducial landmarks on the example faces, and then maintain the geodesic distances between neutral and expressive face scans. The actual recognition is then performed by a two stage optimisation, using ICP to first establish any rigid transformation, followed by adjustment of the weighted contribution of the available deformation templates to try and replicate the expression. Overall, they show a useful improvement to overall performance in identification (from 87.6% without to 92.1% when accommodating deformation).

One of the most comprehensive approaches to invariance of recognition under expression was carried out by Bronstein *et al*. [BBK05]. In this they aim instead to match a canonical form representation, based on the core observation that the expressions on face can be modelled as isometrics (that there exists an isometry in Riemannian terms between a surface and a diffeomorphism) - such that the intrinsic geometric properties of the facial surface are expression-invariant. This takes the form of an embedding of the face structure in a lower 2D dimensional manifold by measuring the geodesic distances on the surface and performing classical scaling. They assume all faces to have the mouth closed so not to introduce violations in the isometry (i.e. holes). An important precursor to this work was in proving that the variations in geodesic distances between expressions is insignificant (by tracking 133 markers on the face). The important fact is that there is no distinction between intrinsic and extrinsic geometry in the embedding space represented by the canonical form. They include in their pre-processing stage a feature detector based on HK classification to locate fiducial markers to then allow generation of the geodesic mask (from which the embedding can be performed). Ultimately they perform surface matching by high-order moments (as opposed to ICP). They achieve considerable accuracy even in the presence of quite extreme expressions, and (more surprisingly) for simple neutral recognition.

Kakadiaris *et al*. [KPT+06] used an actual deformable model for the alignment process. They call this an Annotated Face Model that is constructed from anatomically correct data using standard graphics techniques, marked up into particular regions of the face. Initial alignment is using a standard spin-model and ICP, following which the deformable model is fitted based on a FEM like scheme to minimise implausible deformations. For further performance and accuracy this iterates over multiple scales by a subdivision surface model. For classification the important realisation is that the 3D deformable model's topology can be represented by simpler 2D geometry. They also make the matching process scale invariant by encoding the final face models in wavelet-based Harr and pyramid coefficients. Actual matching is then performed by computing distances in the wavelet domain.

Chang *et al*. [CBF05] return to the idea of using portions of the face that change relatively little. They propose the use of a technique they term Adaptive Rigid Multi-region Selection (ARMS), which they compare against a PCA-ICP recognition baseline. Their extraction of these regions is guided by initial extraction of skin coloured points on the surface, enabling them to

proceed to find and extract small local patches based on $H + K$ curvature (their calculations are derived from a least-squares fit to a quadratic patch). In this way they automatically find the nose, bridge, and eye sockets. They note that even these areas can deform a little - so they finally advocate selection of multiple local surfaces around the nose region. These multiple regions are then all compared to the probe using ICP and a decision fusion process to then perform recognition. This approach is then shown to yield better results when compared to their baseline, which after all, considers the entire face as a single rigid surface.

Finally, Mian *et al*. [MBO06] present a complete, automatic face recognition system based on a number of approaches. They first detect faces in the data by discovery of the nose tip, and then crop the face by the intersection of a sphere centered at this point. They also correct for the pose using the Hotelling transform (which is also applied to the 2D colour image so that it remains registered with the 3D data). The most important contribution of this work is however the use of an efficient 3D spherical face representation to quickly classify candidate faces for matching. These are then confirmed using a novel expression correction approach that makes use of the fact that the face contains regions that are not altered. In particular, they note from their training data that the variance of the forehead and regions around the eyes and nose is most unaffected by expression, whereas the mouth and cheeks are most affected. From this knowledge they are able to build masks to guide a final ICP based match to verify a person.

2.5 Summary and Critique

In conclusion, we have outlined in this chapter how an emerging gap in the research exists in the theory of $3D$ surface change over time. This is uniquely driven by recent advances in capturing dynamic surfaces, and offers fascinating avenues describing these observations, leading to possibilities for further analysis of facial identity and expression.

To achieve this, we note the requirement for a **dense, accurate and colour retaining data-capture** approach. Many novel systems have been proposed for this, many of which follow the approach of spatial-temporal stereo context. To this end, we directly build on the 3D stereo over time approach based on Siebert *et al*. [SU94] and their later more conformation focused work [ZSCA04]. It should be noted that this aspect of the research is greatly enabled by our industrial collaborators.

In many ways, this chapter has served to describe the options and research behind the various stages in a hypothetical "pipeline" of creating a system that is able to capture, track, describe and classify the variations of a dynamic surface (such as the face). Such a pipeline has recently been made a reality by a number of the key works described above (e.g. [ZSCS04] [WGZ+05] [LJ06]). However, excellent and inspirational though these works are, we clearly differentiate the work described in this thesis along the following lines:

- The key outstanding problem is one of **establishing correspondences**. While the trend of more recent work looks towards the discovery of an embedding surface for registration and resolving this, we feel that the possibilities of treating the problem within $\mathbb{R}^3$ have not been fully exhausted nor explored. We therefore return to the concepts of range and scene for defining accurate displacements - particularly building on the work of Spies *et al*. [SJB02]. The issues of then improving the estimation lead us to furthermore look at the idea of employing additional constraints, particularly the use of colour (as proposed by Barron *et al*. [BK02] [BS00]).

- Fundamentally, we consider that the **analysis of dynamic curvature** and other intrinsic properties as a crucial point - not only for the purposes of establishing correspondences. Much work has derived from using curvature to find unique static features (i.e. the classic works of Besl, Jain [BJ86] and Koenderink [KvD92]). However, while recent advances allow for capture of actual dynamic surfaces, there has not been much work to date that seeks to analyse, or describe, the observed deformations. Indeed, we note in general that there is an absence of common terminology, or any complete geometric definition. This motivates our focus on producing a simplified and more symbolic technique for presenting the dynamics of the data in a meaningful and accessible way. To this end we seek inspiration from those works that have at least attempted to describe types of deformation - such as Chung *et al*. [CWP+01].

- Curvature has been applied to looking at faces before - namely in the classic work by Gordon [Gor92], and more recently by the likes of Moreno *et al*. [MSJV03]. In general however, there appears a lack of research focusing on how the dynamics of shape can be applied. Our main contention is that current motion-capture studies are inherently biased towards a sparse sampling of the motion as it appears in the face. Furthermore, the majority of the work that currently aims to accommodate expression in 3D scans is geared towards **invariance for recognition**. No work appears to actually recognise the expression directly, or to use it to perhaps identify the person. This is a particularly interesting topic - especially from the point of view of extending the work of Chuang *et al*. [CDB02] and Chang *et al*. [CHT03].

In the next chapters we go onwards to investigate these gaps in the research. We also present some additional very recent research in the final chapter (Section 6.3.1) within the context of future directions.

Chapter 3

Tracking Dense Non-Rigid Correspondences

"Errors, like straws, upon the surface flow,
Who would search for pearls must dive below."

John Dryden. 'All for Love' (1678)

In this chapter we look at the problem of resolving correspondences. Both **spatially** to reconstruct range-data from stereo colour images, and **temporally** to track deformable motion in sequences of depth images. Our objective is to thus go from a sequence of paired images, to an accurate $4D$ vector flow field that describes the individual motion of dense localised points on the surface. We are fortunate to be able to build on the vast body of research and available implementations for stereo recovery, and similarly to refer to the research carried out on the topic of *flow*.

Consequently, we first present a brief overview of the techniques and technology we actually use to acquire data from high resolution stereo. We then give an equally brief exposition on the derivation and computation of optic and range flow. The main body of this chapter then presents our proposal for improved colour constrained flow estimation. This includes a review of the nature of separate channels and how independent contributions can be weighted together. We then demonstrate the resulting approach on some synthetic and real face data.

To place this chapter in context, it sets out to answer our first research question: *"What observations of a naturally deforming surface can be used in order to resolve the correspondences between points on it as they move over time? Could colour or unique feature points, for example, be used to improve robustness?"*

3.1 Capturing High-resolution Dense 4D Data

The principles of stereo capture are founded on the established science and techniques of *photogrammetry*, in turn based on knowledge of optics and projective geometry to recover measurable structure from images. Many computer vision textbooks provide excellent reviews of this process (e.g. Forsyth and Ponce's "Computer Vision: A modern Approach" [FP03]). In this section we present an overview of the process we rely on in this work and to introduce the reader to the nature of the data we base our work on. Understanding the scope and limitations of the capture system are crucial to the work presented in later Chapters.

Stated simply: the task is one of first establishing the internal properties and external relationship between a pair of cameras located in space and looking at the same scene. This enables the definition of the *epipolar geometry* between them and to, crucially, enable the search for similar features (i.e. *correspondences*) in both images. Having established the location of such featues then enables the straight forward trigonometric recovery of the *depth* of the feature from the co-ordinate frame of one of the cameras. An overview of this process is shown in Figure 3.1.

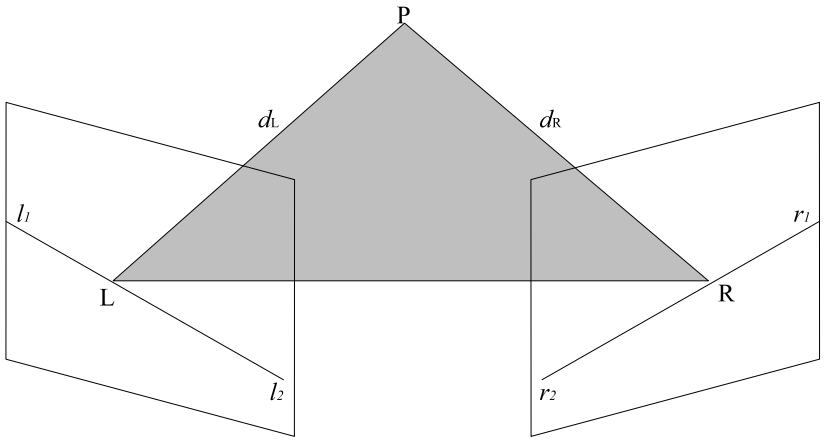


Figure 3.1: Stereo camera epipolar geometry.

The relationship between a point P as observed by two perspective cameras located at *principal-points* L and R defines the plane PLR. The intersection of this plane with the respective image planes of the cameras defines the *epipolar lines* $l_1 \rightarrow l_2$ and $r_1 \rightarrow r_2$ used in turn to *rectify* the images and to guide the search for neighbouring correspondences, leading to the recovery of depth as the distance of the point from either camera's image plane (d_L and d_R).

The concept of "depth" in a stereo context is essentially that of a ray passing from the principal point of the camera and terminating at a point on the subject surface. The length - or range - of this ray can then be recorded at the point where it intersects the image plane of the sensor. Consequently, this also leads to the notion of "dense" data, since every such intersection

down to the level of the sensor grid records a value. This gives rise to the concept of a *depth-map* (or *range-image*) representing a measurement at every single pixel for the distances to the captured surface. Throughout this work we work predominately on such data as our input. It gives a naturally compact and implicit $2.5D$ representation of the surface, and can furthermore be re-projected, using the original camera calibration, to reconstruct the surface accurately in $3D$.

Other important aspects of working with dense stereo data involve the fact that the features used to establish the correspondences between the images may only be visible from one of the cameras, in which case *discontinuities* arise in the recovery. Usually this effect can be negated by choosing a suitably short base-line between the cameras, so that the perceived *disparity* between images of the object is not too great. Another common effect noticable in stereo capture is that only the portion of the scene that is well focused (i.e. lies within the *depth-of-field*) can give good quality results. As a result, portions of the surface in the far distance or near foreground are overcome by noise as the recovery process mismatches to progressively more ambiguous correspondence (a fact exploited by shape-from-defocus and confocal microscopy). Fundamentally, accurate results require good photography - that the images are taken with appropriate levels of exposure, and the subject is well framed and in focus.

The initial work in this thesis, as presented in this Chapter and the next, is based on the use of a commercial Dimensional Imaging $3D$ stereo rig, as shown in Figure 3.2. This uses commercially available cameras that are capable, thanks to modern CCD technology, of capturing extremely high resolution images. There is first a necessity to establish the camera model as the *extrinsic* relationship to the world co-ordinate frame, and also the projective *intrinsic* properties. An excellent treatment of perspective camera models can be found in the book by Hartley and Zisserman [HZ00]. In summary, a 4×4 transformation matrix defines the extrinsic properties of the camera - its location and orientation in the world co-ordinate frame. Similarly, a projection matrix and associated lens distortion parameters define those intrinsic properties that define the internal formation of the imaging process.

The establishment of the camera model is performed by first calibrating the cameras. To achive this a number of images of the "dot array" calibration target are taken, of the target under varying degrees of rotation and slight translations to maintain the target at a comparable distance of approximately $70cm$. From these images, the intersecting points on the grid can be automatically extracted, and used (on the assumption that they lie on a relatively accurate planar surface) to fit an equivalent model of the target. Sub-pixel accuracy can be attained by using such a model, and by further considering the region around each intersection. Non-linear optimisation of this model over all the images leads to convergence on the reprojection of the current estimation of the grid back through the camera, for a Root Mean Square objective error metric. Most successful calibrations for this camera and set-up fall within the final range of

Figure 3.2: High resolution stereo camera capture rig.

Constructed of two Canon EOS 300D digital SLR cameras mounted side by side on a robust and stable pedestal base. Each one is set to capture images at a resolution of $2035 \times 3070\ pixels$ ($6.8\ Megapixels$). The cameras are positioned with an approximate $30cm$ baseline and are connected by a joint electronic shutter release. Images are taken using either the on-board flash or halogen bulb lamp, and are then downloaded to the attached computer for processing. The checker-board calibration target is also shown at the standard capture distance of $\approx 70cm$.

$0.0005m < RMS < 0.001m$.

Following calibration, the stereo rig is used to capture and reconstruct models of a subject placed at the equivalent distance (i.e. in focus and within the spatial volume limits of the depth-of-field). The basis for reconstruction [SU94, JBS03] is a multi-resolution correlation based image matching approach. This effectively involves a two-stage process, which consists of first filtering to generate a scale-space of increasing spatial frequency, followed by a matching stage which interpolates current estimates up the image pyamid and so incrementally improves the estimation of correspondences. Crucially, there is no need to project any form of "structured lighting" in order to help matching as the cameras provides enough resolution to use the actual texture of human skin (i.e. each pixel represents a scale of around $0.1mm$). The upshot of this is that it allows instantaneous capture, with already perfectly registered colour information. From this the $2.5D$ and $3D$ reconstructions are then generated, as shown in Figure 3.3.

There are still limitations in the accuracy of the captured model. The lack of texture over the object in question can result in poor reconstruction, particular for man-made objects (such as plastics). Specular highlights can also result in local ambiguity leading to surface artefacts.

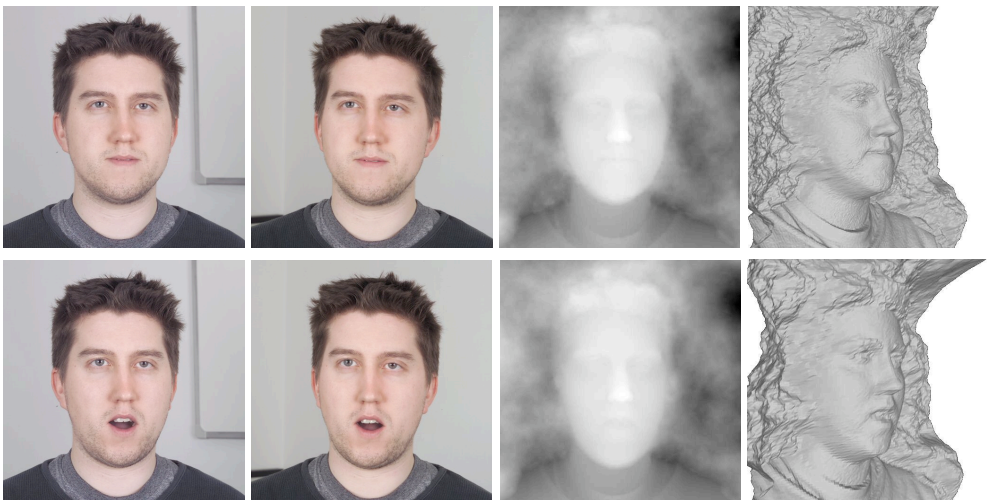


Figure 3.3: The stereo process: from images to depth-maps to $3D$.

The recovery process for two frames (shown on separate rows) for a neutral and open-mouthed expression. To the left is the original stereo pair of images, from which the dense $2.5D$ depth map and resulting $3D$ triangulated re-projections are generated, as shown on the right. (Images and data while on placement with Dimensional Imaging Ltd.)

Ambiguity is also induced by those portions of the scene that are out of focus, and consequently blurred. Furthermore, more complex distortion - as for example caused by the subjects wearing corrective eye lenses - cannot be accommodated by the camera model (all subject in this work are asked to first remove their glasses).

One particular artefact of the multi-scale approach, and how it attempts to resolve ambiguity, can be prevalent in the final reconstruction. So called fine "orange peel" undulations can be generated for progressively worse scales. One means of countering these is to reintroduce a degree of structured lighting, combined with an infra-red filter, so not to make the pattern visible in the colour image. Another approach is to simply apply some Gaussian smoothing to the $2.5D$ depth map, another approach is to integrate a number of frames. The unique burst capture feature of the cameras is leveraged in order to acquire *spatially dense* and *temporally sparse* $4D$ data - as shown in Figure 3.4. Other pre-processing steps include masking the background by performing "chromakey" colour-based background removal (for example by placing a blue screen behind the subject).

3.2 Review of Calculating Optic and Range Flow

The same problem of calculating dense correspondences on a surface also applies to the *temporal correspondences* of a surface when viewed over time. This is the fundamental search that

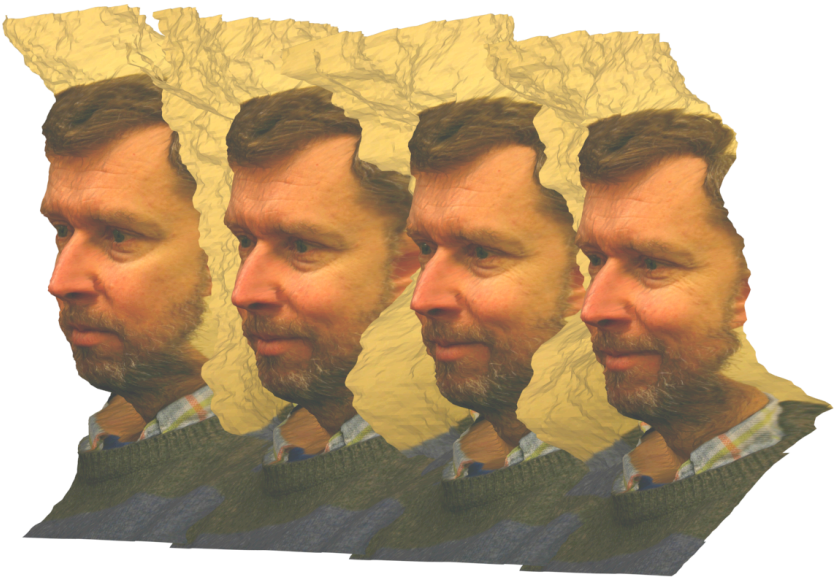


Figure 3.4: Example four frame sequence for a “smile”.

The $4D$ result of the stereo capture performed over time for a sequence of data captured at $2.5 f.p.s$ (4 frames over a period of approximately 1.6 seconds) using the camera burst mode functionality. Notice the reprojection of noisy (out of focus) data to the periperal. This can be removed by sucessfully *masking* on the basis of the background colour.

defines *optic flow*, i.e.: given a set of points in one image, find the same set of points in another image. Expressed simply for an image intensity $I(x,y,t)$, and assuming that such intensities do not change due to other causes (such as lighting), the **optical flow constraint equation** can then be phrased as:

$$I_x u + I_y v + I_t = 0 \tag{3.1}$$

where the partial derivatives of image I are denoted with subscripts along the x, y and t dimensions. This can in turn be re-written for desired *flow field* $\vec{f} = [u,v]$ and image change $\nabla I = [I_x, I_y]$ as:

$$\nabla I \cdot \vec{f} = -I_t. \tag{3.2}$$

From this, it becomes apparent that in order to accurately calculate the flow from observations, it is necessary to further constrain the search. One such way is by the assumption that a suitably small local image patch will all move together at the same time. This leads us to the most famous difficulty in resolving optic flow: the notorious *aperture problem*. Put simply, the derivation of the flow can only attempt to resolve correspondences through a restricted local window, or “aperture”. This means that certain types of structure may result in ambiguous motion when observed in such a way. For example, consider a linear feature as observed at

two moments. While it will be apparent that all the points of this line have moved, what is not determinable is whether any motion along the line actually occurred. This is the essence of the aperture problem - that there may not simply be enough local information to identify, and consequently resolve, motion within a given region. The only feature that is effectively guaranteed to be sucessfully resolved is a *corner*, explaining why consequently they are often relied on as the best local descriptor.

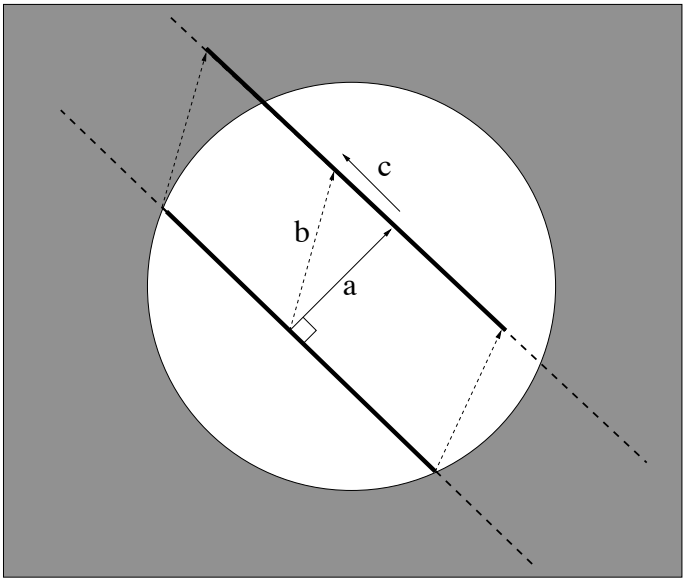


Figure 3.5: The aperature problem for flow calculation.

Given a line (intensity gradient) moving past a small circular window - or *aperature* - there is an implicit problem with resovling the motion as percieved in the direction of the normal a. The tangent component c of the true motion b is lost. From the observer's point of view the motion appears identical and is ambigous, the normal velocity can be recovered - but not the tangential velocity.

The classic solution to this, as proposed by Lucas and Kanade [LK81] [1] is to assume a neighbourhood of N pixels so it is then possible to phrase the integration over a suitably sized region thus:

$$\underbrace{\begin{bmatrix} I_{x1} & I_{y1} \\ \vdots & \vdots \\ I_{xn} & I_{yN} \end{bmatrix}}_{A} \underbrace{\begin{bmatrix} u \\ v \end{bmatrix}}_{\vec{f}} = \underbrace{\begin{bmatrix} -I_{t1} \\ \vdots \\ -I_{tN} \end{bmatrix}}_{b} \tag{3.3}$$

or simply $A\vec{f} = b$. Such an overdertermined system of equations of two unknowns can then

[1] Interestingly, it should be noted that original L+K was originally phrased as a solution to the stereo spatial correspondence problem - by seeking to register two images first.

be traditionally solved by a least squares solution. This leads to the classic Lucas and Kanade solution to optical flow as:

$$\overrightarrow{f} = (A^\top A)^{-1} A^\top b. \tag{3.4}$$

However, there are of course further complications. Due to either viewpoint motion, or objects in the scene undergoing either rigid or non-rigid transformation - it may simply not be possible to resolve an exact match between pixels. Effectively entire portions of the scene may appear or disappear between images. This will, as with spatial estimations, result in *discontinuities* occurring in the output as regions in the flow field where it is impossible to calculate any estimate. Furthermore this is related to the issue of aperture size for local estimation - since large motions can cause structures to pass completely out of view between observations. The issue of sub-pixel accuracy (in the case of an image array) is also important, where most schemes attempt to resolve this by considering a parameterisation of the local neighbourhood as an energy surface. A number of attempts have been made to better quantify the performance of different approaches - as in Barron's paper on the topic [BFB94] (including the standardisation of test data such as the now famous "Yosemite sequence" in order to compare the benefits).

It should be noted that various other algorithms have attempted to tackle these issues by alternative means - primarily by the introduction of some degree of *global* decomposition or regularisation. The intention is then to effectively introduce a smoothness constraint over the entire flow field - thus allowing neighbouring regions to propagate those areas where an estimate is impossible (i.e. due to discontinuities). The greatest danger lies in over-regularisation, by which smaller and more detailed motions are completely smoothed away. The acknowledged earliest attempt at this approach is the work of Horn and Schunk [HS81] and its smoothness assumption (usually represented as a first order membrane model). This can be directly contrasted with the work of Lucas and Kanade - which relies entirely on the local constraints imposed by a single neighbourhood of data values - and indeed, other works have sought to unify these two views (e.g. Bruhn *et al*. [BWS05]). This represents the state of the art in optical flow computation - producing the best result - as seen in the recent evaluation work performed by Baker *et al*. [BSL+07].

In terms of actual calculation of dense flow fields for image data most approaches follow a minimisation (such as a least-squares approach based on derivative calculation). Thus, good derivative estimation is crucial. In the case of a sequence of 2D images, it is often more appropriate to consider these as a 3D stack of data with the additional axis of time. This then allows for more advanced calculations that consider change both spatially and temporally, as in Giaccone *et al*. [GJ97]. In terms of then performing the calculation, most implementations apply convolutions with varying types of filter. Such filters can be fairly simple derivative filters (such as a Sobel) or can be more advanced multi-tap designs with inbuilt noise reduction characteris-

tics. One of the most successfully employed versions for dense flow calculation is the design by Simoncelli [Sim94].

The natural 3D extension to this work is the computation of an instantaneous velocity field that describes the displacements between two instances of a surface, known as *range-flow*. This is effectively the application of optical flow techniques to 3D data. The idea has inspired some researchers to investigate the use of optical flow itself within a stereo context, and to look at the problem as one of projecting the observed planar motion to recover the additional depth changes. Such an approach is often considered in terms of the derivation of scene-flow proposed by Vedula *et al*. [VBR+99], particularly when using multiple cameras and sparse image correspondences. It is also encountered when operating on raw depth-maps as if they were intensity images, to track the apparent depth/intensity change, as for example by Nebel and Sibiryakov [NS02].

However, a more intrinsic approach operates directly on the 3D data itself, by considering the actual depth disparities as they occur in $Z(x,y,t)$. These focus not only on tracking rigid motion, but on the deformable shape which can in turn describe and act as input for direct physical modelling of observed objects (e.g. Yamamoto *et al*. [YBBR93]). Computation can be performed directly on the data in its original sample grid (i.e. a depth-map) using a combination of localised estimates as the basis for creating a regularised, smooth flow field. This has advantages in speed of calculation of derivatives within the sample grid via convolutions. However, the aperature problem still applies (in the form of planar and ridge/depth edge features).

In our formulation we extend the original work of [SJB00] which considers the varying depth of an orthographically captured surface $D(x,y,z,t)$ as a function of time t, locally constraining a *range flow* field in 3D $\overrightarrow{f} = [u,v,w]^\top$ by:

$$D_x u + D_y v + D_z w + D_t = 0 \tag{3.5}$$

where subscripts indicate partial derivatives for the simultaneous change in depth D with respect to local position in (x,y,t).

3.3 Adding Additional Colour Constraints

Given our dense stereo reconstructed data with aligned colour information, and using the formulation for calculating optical and range flow described above, we now describe our proposed framework for generating more robust $4D$ vector flow fields. This is primarliy an extension to the work of Spies and Barron [SJB00] in which only $2D$ gray-scale intensity information is considered and added (where appropriate) to help the range-flow estimation.

The key assumptions in building on this concept are related:

1. That intensity represents an implicit loss of information about the motion of an object.

2. That retaining colour can help resolve any ambiguity in the motion of that object.

The first assumption is naturally intuitive, but seldom actually quantified by research. The representation of colour as *information* can be considered in terms of the classic sense introduced by Shannon [Sha48] in communication theory. In this case, the definition of the degree of additional useful information relayed by colour can be based on the respective entropy, and mutual information shared with the equivalent intensity image. Related recent work has exploited this, for example Bart and Ullman introduce the use of mutual information for image normalisation by using it to decide which PCA basis vectors contribute most effectively to the make up of the image [BU04]. Similarly, Finlayson *et al*. [FDL04] look to establish the intrinsic image by minimising entropy - resulting in the useful result of effectively being able to remove shadows. Other work backs up this assertion of maximising information, for example Tsagaris *et al*. [TGA05] directly look at measuring the content of colour images by use of Kullback-Leibler (KL) divergence as a measure of information gain.

The second assumption follows from these ideas, and is most simply illustrated as shown in Figure 3.6. This is the key (if simple) concept that retaining the perception of colour can reveal motion, whereas intensity information alone cannot be used. This is particularly the case when the intensity may also be affected by illumination and shadows - introducing a large number of other ambiguities that are unrelated to the actual motion of the object in question.

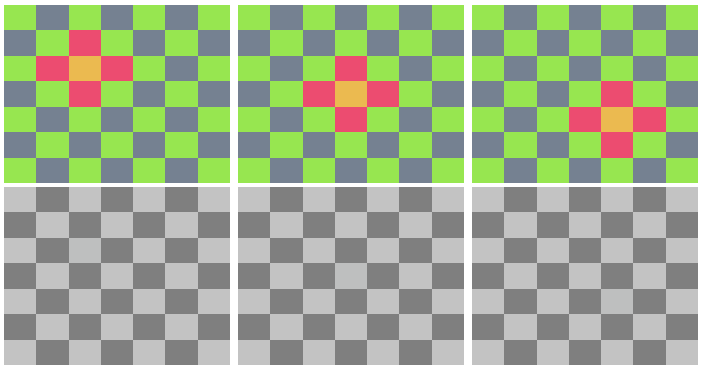


Figure 3.6: The ambiguity problem resolved with colour.

In the top row, the apparent motion from left to right of the orange/red region to the viewer (if not colour-blind) should be easily apparent. The same sequence, but using only the intensity of the colours, is shown on the bottom row. The perception of motion is lost, along with the information the colour channels relay.
(The effect is also lost if this figure is printed on a gray-scale printer).

This effect and the benefits of adding further colour constraints in the context of the calculation of optic flow, are summed up in the key paper of Barron and Klette that looks directly at the use of colour information in improving optic flow estimation [BK02]. This adopts a qualitative

approach to assessing the impact of additional colour information. More formally, they reimplement Golland and Bruckstein's [GB97] initial approach of combining multiple colour channels. In addition they can modulate the contribution made by each by weighing the contributions. They also control the combination by using the smallest eigenvalue or condition number of $A \top A$ to assess the numerical stability for the calculation based on a certain channel. If it is below a certain threshold, the flow is considered undefined at that point - and the estimate is not included in the final result. They conclude that colour optic flow estimates calculated via the full three colour channels (with weights 1.0) are slightly better than simple grey or staturation value based optic flow.

Another advantage to using the regularly sampled depth information as usually provided by laser-stripe scanners and dense stereo capture, is that it is often accompanied by perfectly aligned colour information. Such information can act to further constrain and thereby increase the accuracy of the flow estimates, as shown by initial experiments for improving optical flow estimates by incorporating texture by Arrendondo *et al*. [ALL04] and colour by Andrews and Lovell [AL03]. The use of colour, in particular, is proposed under the key assumption made by Golland that the use of additional channels from RGB, HSV, CMY, UCS or other colour spaces are **more representative of the actual surface properties of the objects in question** [GB97]. When considering that intensity depends on more global assumptions about the illumination and reflectance, then surface invariant colour characteristics have obvious benefit in resolving certain types of localised motion ambiguity.

In adapting this to range-flow we assume that while the underlying surface may deform and change shape in any number of ways, any apparent planar motion observed is consistent when reprojected back onto the surface. Key to achieving this is the concept of preserving a meaningful distinction of colour, and using this to help us impose additional constraints on the range flow estimation. However, crucial to this is determining what suitable colour spaces work best for the retention of information - a question we deal with in the next section.

3.3.1 Colour Representation

The idea of percieving the same colour under varying degrees of illumination is often phrased as the question of *colour constancy*. The key objective of which is to try and mimic the same level of competency exhibited by biological systems in distinguishing the same colours, irrespective of environmental factors. An up-to-date treatment of this topic can be found in a number of books, for example the recent one by Ebner [Ebn07] discusses this topic from its perceptual psychophysics basis. Core to all work in this are is the task of then defining the varying *colour spaces* that can be used in order to actually measure and represent the property and response to colour, by removing us from the underlying physics between light and surfaces - focusing purely instead on perception.

For simplicity, we do not consider more device-dependent, hardware oriented colourspaces such as CMYK, YIQ, and YUV. Despite being often faster to compute, they are specifically tuned to printing and display requirements.

It is important as a pre-cursor to our formulation, to realise what in real terms colour represents, and how can in turn be *quantitized* and manipulated. In terms of dealing with real sensors this is often a level of precison dictated by the *dynamic range* of the sensor and its encoding (e.g. 24 bits per pixel). Many cameras in fact really use only 8 bits for all colours, with various coding schemes. The desire is often to accommodate the broadest *gammut* of colours possible that can actually be sensed. Most cameras output information in the standard three component tristimulus values **RGB** Red-Green-Blue scheme by simply returning the value of each separate *channel* independently. This system is primarily adopted for its simplicity and intuitive nature for transferring digital information between sensor and display. In our work we consider the encoding of these values in floating point form:

$$\begin{bmatrix} R:\{0...1\} \\ G:\{0...1\} \\ B:\{0...1\} \end{bmatrix}. \tag{3.6}$$

RGB's greatest failing is that it does not directly separate the concept of colour from intensity, or lightness. The lack of defining chromaticity by a singular value is countered by an alternative and straightforward mapping from the RGB space to a polar-coordinate representation - giving rise to the **HSV** Hue-Saturation-Value colourspace. In this, the actual colour component is defined solely by the hue term, modulated by the degree of lightness (saturation) and strength (value). In this way the colour can be considered independent of these other factors:

$$\begin{bmatrix} H = \begin{cases} \text{undefined} & \text{if } max(R,G,B) = min(R,G,B) \\ 60^\circ \times \frac{G-B}{max(R,G,B)-min(R,G,B)} + 0^\circ, & \text{if } max(R,G,B) = R \text{ and } G \geq B \\ 60^\circ \times \frac{G-B}{max(R,G,B)-min(R,G,B)} + 360^\circ, & \text{if } max(R,G,B) = R \text{ and } G < B \\ 60^\circ \times \frac{B-R}{max(R,G,B)-min(R,G,B)} + 120^\circ, & \text{if } max = G \\ 60^\circ \times \frac{R-G}{max(R,G,B)-min(R,G,B)} + 240^\circ, & \text{if } max(R,G,B) = B \end{cases} \\ S = \begin{cases} 0, & \text{if } max = 0 \\ \frac{max(R,G,B)-min(R,G,B)}{max(R,G,B)} = 1 - \frac{min}{max}, & \text{otherwise} \end{cases} \\ V = max(R,G,B) \end{bmatrix}. \tag{3.7}$$

It should be noted that an alternative approach to invariance in the degree of brightness can be also performed directly within the RGB colour-space by removing the effects via *normalised* **NRGB** values (which sum to 1) as:

$$\begin{bmatrix} NR = \frac{R}{R+G+B} \\ NG = \frac{G}{R+G+B} \\ NB = \frac{B}{R+G+B} \end{bmatrix}. \tag{3.8}$$

However, the question of colour can also be directed by considering the response of the human eye itself, as the paridigm for a sucessful biological solution. The effort to standardise the range of colours possible to perceive was carried out under the auspices of the Commission Internationale de l'Eclairage (CIE). This lead to the definition of the base CIE-**XYZ** (1931) colourspace representation (relative to a standard observer). The three tristimulus values roughly correspond to red, green and blue components, but serve to define the entire *gamut* of percievable colours. It is furthermore controlled by a range of standard luminants - as definined by the corresponding *whitepoint*. This acts to define the response to the perception of colour under different wavelengths of light. Crucially, it then acts to govern the conversion, as for example using the CIE 'E' whitepoint ($[X_r, Y_r, Z_r] = [\frac{1}{3}, \frac{1}{3}, \frac{1}{3}]$) in equation 3.9 below.

$$\begin{bmatrix} X \\ Y \\ Z \end{bmatrix} = \begin{bmatrix} 0.488718 & 0.176204 & 0.000000 \\ 0.310680 & 0.812985 & 0.0102048 \\ 0.200602 & 0.0108109 & 0.989795 \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix}. \tag{3.9}$$

A modified alternative is the CIE **LAB** which is derived from XYZ, but formulated in such a way to maintain perceptual linearity (i.e. a change in one component leads to a similar degree of visual affect). This is designed to model the actual human visual system, and the way it retains the distinct tone of a colour under different intensities, and in the way it can smoothly transition from one colour to another. It represents the *lightness* L, with A as the measure along red$(+)$ to green$(-)$ axis, and B as the level of yellow$(+)$ to blue$(-)$.

$$\begin{bmatrix} L = 116 \times f(Y/Y_r) - 16 \\ a = 500 \times (f(X/X_r) - f(Y/Y_r)) \\ b = 200 \times (f(Y/Y_r) - f(Z/Z_r)) \end{bmatrix} \text{where} \begin{cases} f(t) = t^{1/3} & \text{if } t > 0.008856 \\ f(t) = 7.787 \times t + 16/116 & \text{otherwise} \end{cases} \tag{3.10}$$

With all these colourspaces we are interested in two aspects related to :

1. Isolating the *luminance* component (i.e. brightness, value, intensity) in order to remove it and thus identify the chromatic components.

2. Preserving the *linearity* of chromatic change in order that even if a colour is subtly different it will not skew subsequent calculations.

3. Avoiding the *w*rap-around issues when using numerical polar co-ordinate schemes (such as Hue jumping from zero to 2π.

These desirably properties - as governed by the **shape** of the colourspace - can be illustrated by considering Figure 3.7 below.

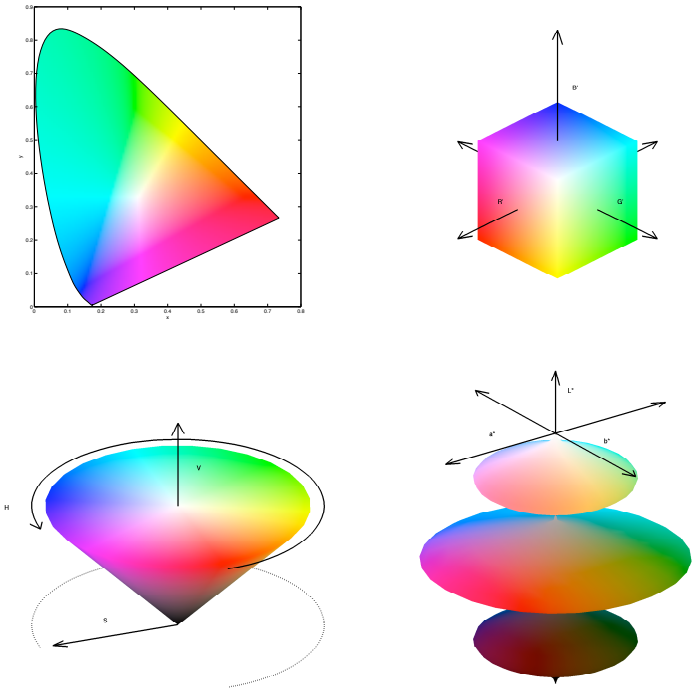


Figure 3.7: Colourspaces: XYZ *(top left)*, RGB *(top right)*, HSV *(bottom left)*, LAB *(bottom right)*.

XYZ forms a single planar representation (based on whitepoint) capturing colour within the range of human colour perception. The RGB representation contains all possible combinations with equal weighting. HSV decouples luminance from colour, but sacrifices realistic distance between perceptually similar colours. The LAB colourspace corrects this by modifying the distances between colours depending on the separate luminance (shown here as 3 separate slices through this space).

3.3.2 Combining Channel Constraints

We generalise the optical/range flow equation constraint provided by any arbitrary *channel* C for observed 2D/3D flow partial derivatives C_X, C_Y, C_T (and optionally C_Z) as:

$$C_X u + C_Y v + C_Z w + C_T = 0 \tag{3.11}$$

where $C_Z = 0$ for colour (e.g. R,G,B,NR,NG,NB,H,S,A,B) and intensity channels (e.g. V,L), and $C_Z = 1$ for range channels (e.g. Z). The Z channel case for depth-map data is provided by the orthographic projection of C_X and C_Y. This value of 1 then acts to effectively enable the calculation of depth motion (w) on the basis of unit displacement.

By adding these channel constraints together, we aim to resolve, or reduce, the aperture problem. This can occur for example in the 2D case of a line, and in the 3D case of a fold-edge, for which there is not enough localised information to resolve motion along the direction of that line or edge. Only in the case of a singular point feature can the full direction of motion be estimated. In the case of a region containing entirely flat range data samples, where there are no localised surface features, then the core idea is that another channel can be used instead (as for example when there is no 3D features, but plenty of 2D colour information). In combining multiple estimates from separate unrelated channels (e.g. range and colour) we thus anticipate fewer aperture problems due to the combined constraints serving to cancel out such ambiguities.

As with optical/range flow algorithms, we wish to locally estimate the motion by solving for $\overrightarrow{f} = [u, v, w]^{\top}$. To achieve this further assumptions must be made. One is to assume nearby velocities are equivalent in relative direction and magnitude, and thus use a *Least Squares* minimisation to integrate the velocities into full flow estimates. Here we assume orthographic equally spaced data, but the technique can be adopted for other cases - for example when the data lies within a mesh topology. So long as the concept of topological distance between neighbouring values is maintained.

Fundamental to our approach is the way in which the individual channels must be combined via a scaling value β. Spies *et al*. [SJB00] calculate a single weighting on the basis of the averages in intensity and depth gradient magnitudes (using a representative training data set, but often only setting the value uniformly to 1.0). They also suggest the possibility of adding multiple channels constraints by simply summing all contributions. We instead desire a more robust and invariant estimation of the contribution by the different channels to the combined estimate.

It is furthermore important to realise that range and intensity values have different units. The weighting is consequently performed on the basis that we cannot assume that any two channels are in any way related to the same underlying process or sensor capabilities, and so cannot be compared directly in terms of their scale or distributions. Thus we scale all channels to have the same mean and variance as the depth data. This enables us to overcome the initial discrepancies in sensor ranges and readings and to directly compare the reliability of each channel.

We also wish to give precedence to channels that are more reliable over those that are worse affected by noise and by aperture ambiguity. A Lucas and Kanade based Least Squares approach can incorporate weighting when constructing a system of linear equations to solve $\overrightarrow{f}$ for a local neighbourhood of size N pixels, over M channels as shown below. This equation is derived as a re-factoring of the channel motion constraint Equation 3.11 above as a Least Squares solution (i.e. solving for $[C_X, C_Y, C_Z][u, v, w]^{\mathsf{T}} = C_T$).

$$\underbrace{\begin{bmatrix} \sum_{i=1}^{M} \beta_i C_{X_1} & \sum_{c=1}^{M} \beta_i C_{Y_1} & 1 \\ \vdots & \vdots & \vdots \\ \sum_{i=1}^{M} \beta_i C_{X_N} & \sum_{c=1}^{M} \beta_i C_{Y_N} & 1 \end{bmatrix}}_{A} \overrightarrow{f} = \underbrace{\begin{bmatrix} -\sum_{i=1}^{M} \beta_i C_{T_1} \\ \vdots \\ -\sum_{i=1}^{M} \beta_i C_{T_N} \end{bmatrix}}_{B}. \tag{3.12}$$

Note that we always assume the contribution of just one depth channel in this formulation, hence the singular 1 values in the 3^{rd} column for C_z. In this way we are simply seeking to weight and combine the information provided by all the channels within the same framework, as dictated by their β values.

In order to actually calcuate the β values, each **individual** channel's coefficient matrix for the local 2D neighbourhood (equivalent to a single channel $A^\top A$):

$$D = \begin{bmatrix} (\sum_{j=1}^{N} C_{X_i}^2) & (\sum_{j=1}^{N} C_{X_i} C_{Y_i}) \\ (\sum_{j=1}^{N} C_{X_i} C_{Y_i}) & (\sum_{j=1}^{N} C_{Y_i}^2) \end{bmatrix} \tag{3.13}$$

can be constructed and assessed for its reliability ρ as the *reciprocal of the condition number* derived from the ratio of its maximum to minimum eigenvalues:

$$\rho = \frac{1}{\lambda_{min}^{D} / \lambda_{max}^{D}}. \tag{3.14}$$

This represents the *numerical stability* of that channel's contribution to the Least Squares calculation. In other words: the matrix is more ill-conditioned as this ratio rises (as either eigenvalue approaches zero), but we take the reciprocal so that its contribution is increased if the inverse is true. The beta value then expresses this contribution as a weighting for each channel in the neighbourhood over the summation of all other channel reliabilities:

$$\beta_C = \frac{\rho_C}{\sum \rho}. \tag{3.15}$$

If the summation $\sum \rho$ is not greater than a threshold level θ (i.e. no amount of combined channel estimate provides any contribution), we can reject outright the estimation for this neighbourhood as unreliable. It should be noted that this rejection is a common technique employed in the optic flow literature, but we employ it uniquely here to combine it with range flow (Spies and Barron [SJB02] also employ a β term, but select it as an arbitrary constant to weight only combined intensity and range estimates).

The question of selecting a suitable flow threshold value θ is naturally of critical importance to the "sensitivity" of the algorithm. It acts to effectively regulate the acceptable degree of error, above which we do not consider the sum contribution of any of the estimates to be reliable. Since the algorithm is designed to overcome ambiguity by deciding that at least one channel is reliable, then the choice of threshold may be biased to this - such that only a small contribution by another

channel is able to confirm estimation. Given the varied nature of even simple data (as seen in the experiments below) it is often only possible to set the final value empirically, by finding a suitable level for a sufficient density of estimate versus the overall error.

3.3.3 Deriving Final Flow

To numerically generate a flow estimate for each uniformly spaced pixel of the aligned input channels we construct the respective matrices A, B as defined above in Equation 3.12. We perform this separately for each channel to derive C_X, C_Y and C_T. First it is necessary to arrange the original data into an aligned stack, in order to consider the surrounding neighbourhood of $5 \times 5 \times 5$ *pixel* values along each direction, as illustrated in Figure 3.8. In the case of fewer frames, it may be necessary to interpolate or duplicate this data in the T direction.

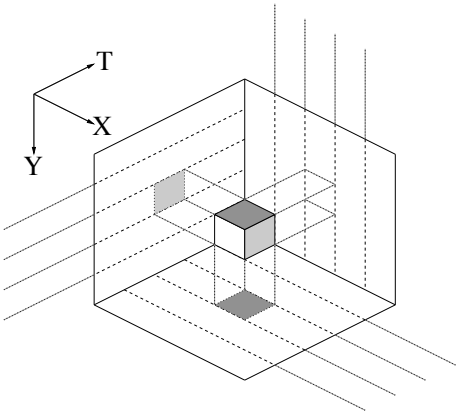


Figure 3.8: Derivative calculation window around a value.

We then use balanced Simoncelli filters [Sim94] for derivative estimation. These have been specifically constructed to provide robust results by employing two stage low pass (noise reduction) and high pass (differential) 5-tap filters. The calculation of the derivatives are then performed as convolutions across the appropriate dimensions. For example, to calculate C_X we first convolve across the T dimension using the smoothing kernel:

0.036	0.249	0.431	0.249	0.036

followed by similar smoothing across the Y dimension, before finally convolving the differentiation kernel:

-0.108	-0.283	0.0	0.283	0.108

across the X dimension. For Least Squares, the flow estimate can then be computed from these combined derivatives by the pseudo inverse of Equation 3.16:

$$\overrightarrow{f} = (A^\top W^2 A)^{-1} A^\top W B \tag{3.16}$$

where the W is an additional component expressing a diagonal weighting matrix of size $N \times N$ over the neighbourhood drawn from a zero mean 2D Gaussian with standard deviation 1.5. This crucially matches the same ordering of the point values in the matrix A and B - so that values closer to the central point are thus able to contribute proportionally more to the final estimate.

3.4 Experiments: Tracking with Extra Channel Constraints

Having outlined our extension to the combined estimation of range flow, we now seek to measure any level of improvement gained. To establish this we design and perform a number of experiments based on both synthetic and captured real surface data-sets. The synthetic data represents fairly straightforward **rigid** translations, and can be directly contrasted with the work by Spies and Barron [SJB02]. The real data looks to the more complex issue of **non-rigid** deformation, using the sequence captured from the actual stereo system. In all cases we are particularly interested in which (if any) colourspace acts to provide the best contribution towards resolving ambiguity.

In all these experiments we use the same underlying code (written in Matlab) to perform the estimation. Note again that all the channel values are first converted to floating-point representation in the range 0..1. As described in the previous section, there is a flow threshold used to discard any combined estimate that is sufficiently unstable. Here we use a value of $\theta = 0.5$ to reject estimation in those regions. This value was determined empirically in order to highlight the subtle differences between the estimates - by selecting a value that initially generated an estimate for each point (i.e. 100% density) for a displacement of 1 *pixel*).

3.4.1 Synthetic Data

To quantitatively test the benefits of additional constraint channels we assess their accuracy in predicting the known rigid displacement of a surface. The two orthogonal synthetic data-sets we rely on are a sloped plane ("slope") and a sinusoidal plaid pattern ("splaid"), each of a sampling size 100×100 *pixels* as shown in figure 3.9.

Altogether we seek to compare 6 methods of combining additional channel information: depth by itself (Z), depth plus raw colour (ZRGB), depth and CIE LAB (ZLAB), depth plus only luminance/intensity (ZI), depth plus three-channel brightness invariant colour (ZNRGB), and depth plus one-channel hue (ZH).

The splaid surface is generated over the functional space of $2\pi \times 2\pi$ for a frequency of 3 (that is, repeated every 32 pixels). Of critical importance to these experiments is that the colour and depth values are generated over the same functional space, such that the observed displacement

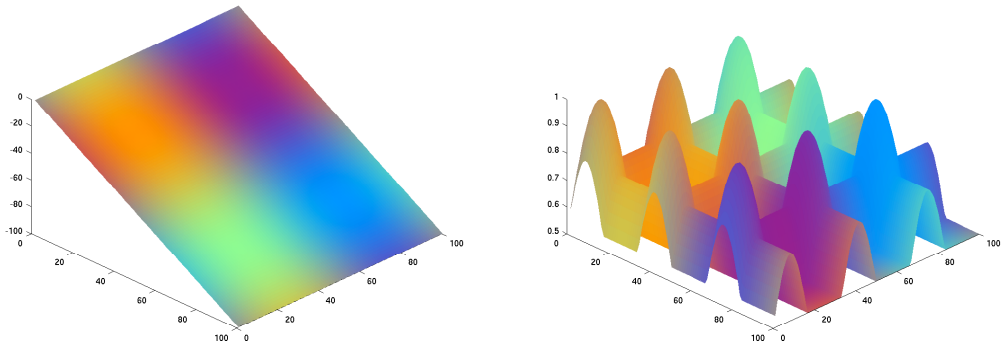


Figure 3.9: Synthetic slope and splaid data-sets with depth and RGB colour.

The slope represents the simplest type of surface in which rangeflow is subject to unconstrained estimation in planar motion, but for which the colour information should correct. Conversely, the sinusoidal plaid ("splaid") data represents a surface in which range-flow is unambiguous, but for which different colour channels may in fact confuse the estimation if they provide a stronger bias.

in depth must match exactly the motion of colour values on the surface. We also seek to create a varied pattern that contains different aperture ambiguities when represented in various colour channels, generated again by an interwoven combination of sine waves - as seen in figure 3.10. Noise can also then be added by a random amount drawn from a Gaussian distribution of standard deviation σ separately to each input channel.

Figure 3.10: Synthetic colour data in RGB, Intensity, Normalised RGB, LAB, and Hue.

The generation of colours by orthogonal functions in the $[u,v]$ plane is designed to create smooth, but varied, neighbourhoods around every pixel.

To assess the level of accuracy we use the standard metrics described in Barron [BFB94], to calculate the relative *magnitude error* (E_m) as a percentage (of the true motion), thus:

$$E_m = \frac{|\|f_{est}\| - \|f_{cor}\||}{\|f_{cor}\|} \times 100. \tag{3.17}$$

Similarly, the *angular error* (E_a) is measured in degrees as:

$$E_a = arccos(\frac{f_{cor}}{\|f_{cor}\|} \cdot \frac{f_{est}}{\|f_{est}\|}). \tag{3.18}$$

Both are derived by comparing the flow between estimate f_{est} and known correct displacement f_{cor} used to generate the sequence.

Using these, we take the mean error over all the estimated flow vectors, and observe the effects of 1D translation $(1 \leq T_X \leq 20)$ of the slope dataset as shown in figure 3.11 where $1 pixel \approx 1 unit$. Similarly, the effects of a more complex uniform translation in 3D $(1 \leq T_X, T_Y, T_Z \leq 20)$ for the surface of the splaid dataset are shown in figure 3.12. The addition of varying levels of noise to the slope dataset produce consistently inferior results for NRGB and RGB as shown in figure 3.13 for the magnitude error of the slope dataset translating up to a smaller amount on the $X-Y$ plane $(1 \leq T_X, T_Y \leq 10)$.

3.4.2 Real Data

As a more qualitative assessment of the accuracy of the colour enhancement, and how it can resolve more complex motion, we look at using $4D$ scene flow to resolve an expression occurring on the human face. The stereo capture rig was calibrated and used in burst mode to capture a sequence of a subject making a "surprised" expression at 2.5 frames per second (each pixel resolved at $\approx 2mm^2$). Dense stereo data was recovered from each of the two simultaneous images via stereo photogrammetry, and transformed back to the coordinate frame of the original left image for aligned colour data - as shown in figure 3.14. For computation, and to avoid aliasing, we reduce the final resolution to $80 \times 120\, pixels$ (i.e. one pixel represents about $\approx 0.5cm^2$ on the surface). A higher threshold of $\theta = 0.8$ was determined empirically to highlight the differences in density estimation.

Using our least-squares approach we derive a range flow estimate between the first and last frames of data, incorporating additional channels as shown in figure 3.15. Overall, the quality of the results for such a sparse temporal sequence would appear to be very crude. Aliasing (i.e. the motion is so great and out with the aperture that complete false matches are generated) is very much in evidence and would seem to affect the amount of estimation considerably. This is primarily due to the amount of rapid displacement for such an expression.

From this LAB works best due to its better gammut preserving aspects. This is confirmed by comparing the Sum Squared Differences (SSD) between the actual surface displacement in the two original frames of depth data, and the predicted displacement of the range-flow (i.e. warping the previous frame) shown in table 3.1. Additional qualitative analysis of the actual vector flow fields on an additional subject are furthermore shown in Figure 3.16 which also point to the accuracy of LAB based channels.

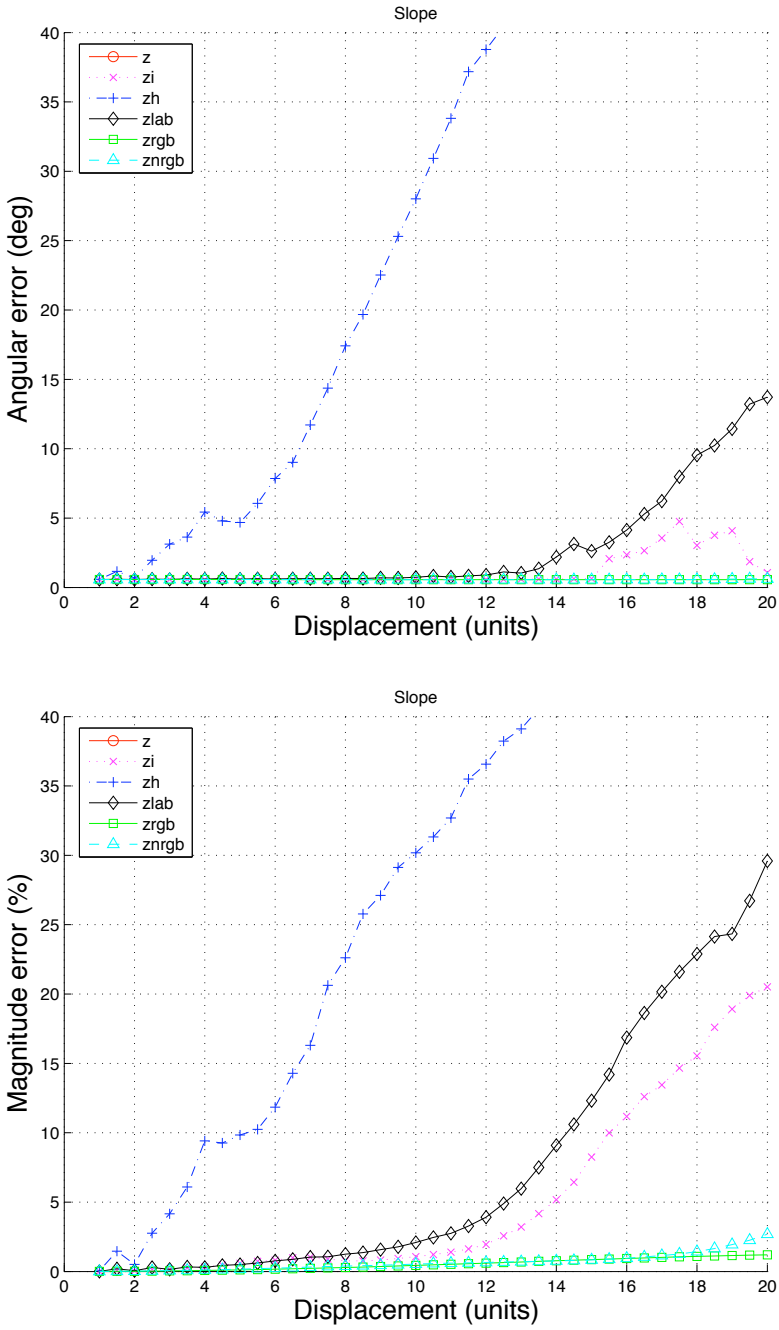


Figure 3.11: E_a and E_m for translation $(1 \leq T_X \leq 20)$ of the synthetic slope dataset.

These show a complete failure to derive an estimate based on Z alone due to the aperture ambiguity of the slope creating an ill-conditioned solution that is rejected. However, combining additional information from the other channels allows this to be resolved. These estimates progressively worsen as the plane translates further across X, yet the ZRGB and ZNRGB solutions consistently provide the overall best results. The addition of LAB also provides a relatively robust solution, while the addition of H erratically affects the estimation, seemingly due to its polar representation of colour which can result in sudden colourspace transitions in spatial gradients.

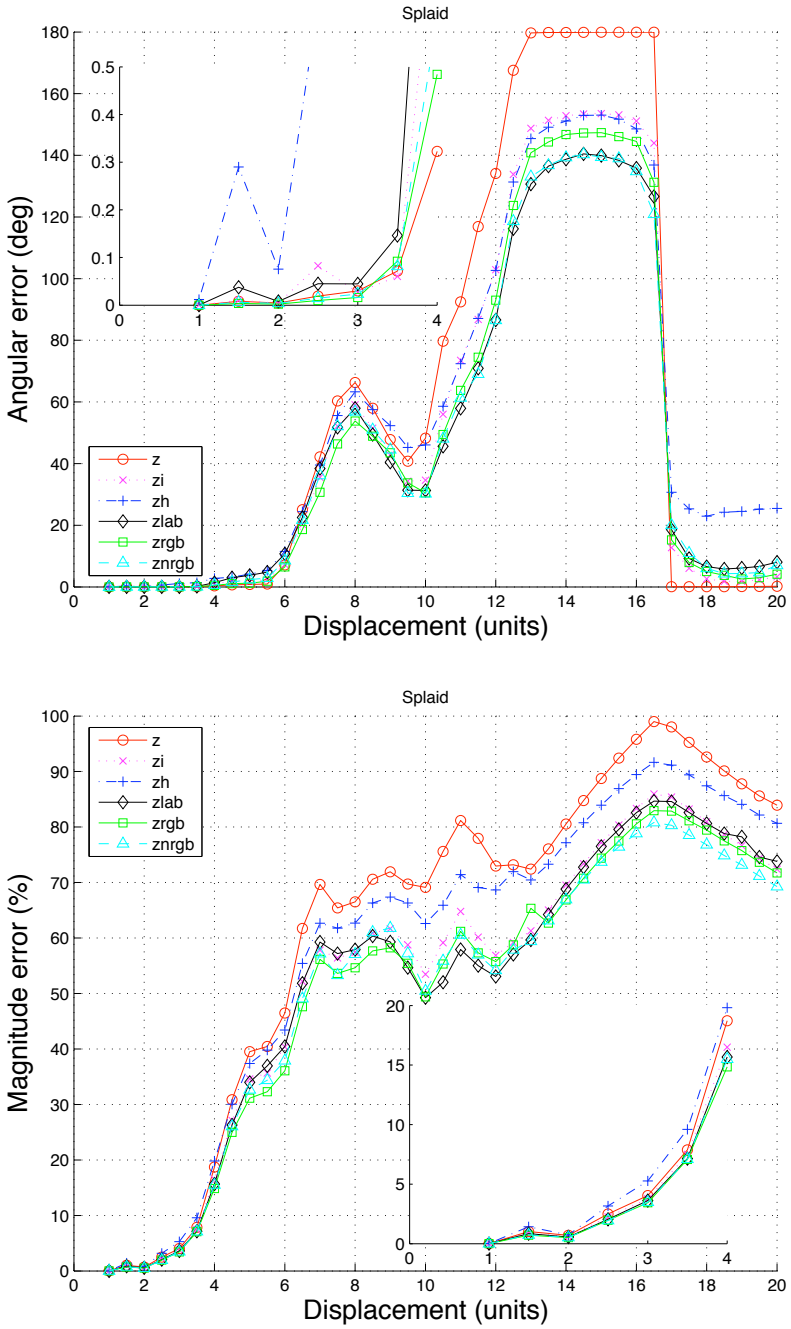


Figure 3.12: E_a and E_m for translation $(1 \leq T_X, T_Y, T_Z \leq 20)$ of the synthetic splaid dataset.

Here the Z channel has enough local information from the surface to make a good estimate and perform well for estimation of smaller translations (especially in angle). Interestingly, the combined use of RGB and NRGB has little improvement as they have an overall lower reliability than the Z estimate. However, the addition of LAB colour provides the most robust correction for larger translations due to the A and B channels having a relatively higher stability, as the represent colour by smoother, more linear transitions.

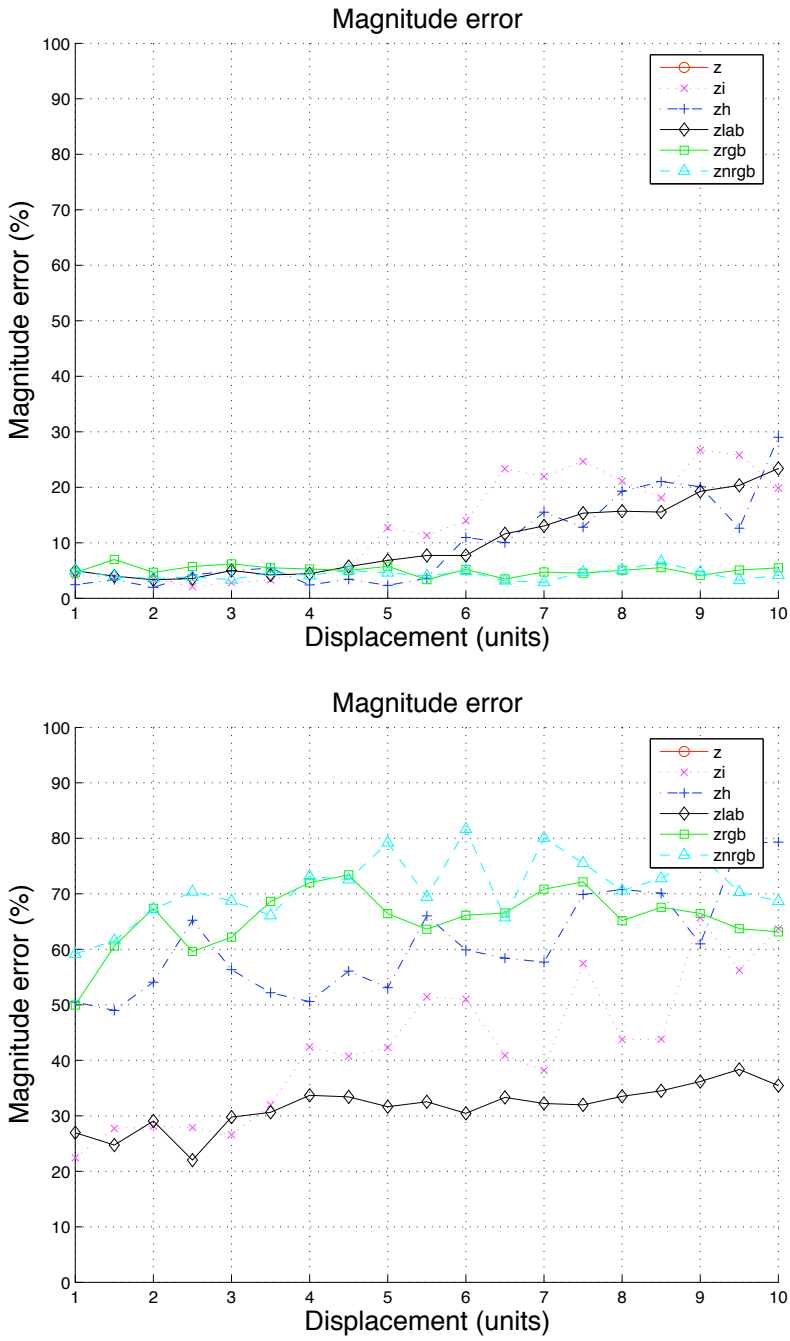


Figure 3.13: E_m for noise $\sigma = 0.001$ (top) and $\sigma = 0.01$ (bottom) in the translation $(1 \leq T_X T_Y \leq 10)$ of the synthetic slope dataset.

The combined Intensity and LAB channels act better for a greater level of noise in this experiment, which can be attributed to the effectual smoothing that occurs when combining separate channel RGB input data (although Hue still suffers from its polar based representation). If the intensity channel were from a true intensity sensor, the noise level would be higher and thus results would be worse for intensity. For smaller noise levels the separate RGB channels do however preserve better estimates.

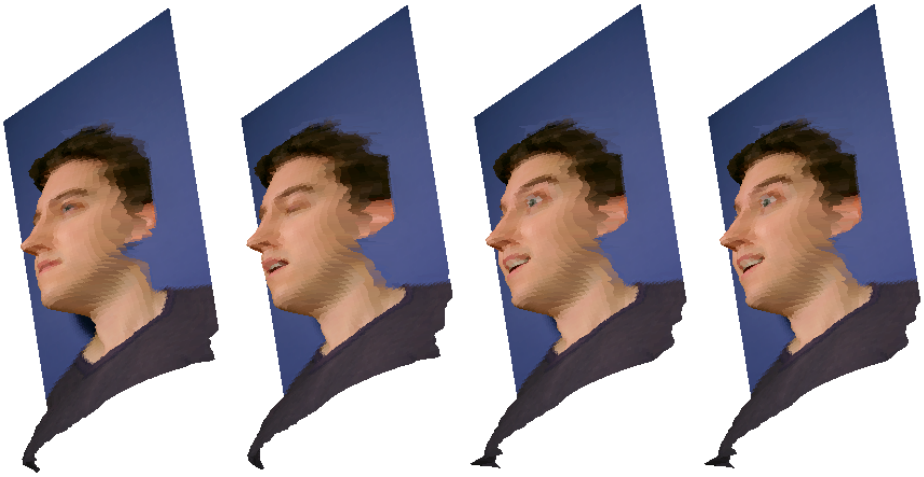


Figure 3.14: Real data “surprise” sequence frames.

Notice that four frames were captured, but, due to the timing of the expression, the fourth is effectively the same as the third. For computation we then copy the first frame to provide a total of five. We also perform a “chromakey” detection on the blue background in order to mask it and (as shown here) remove the noisy surrounding depth data.

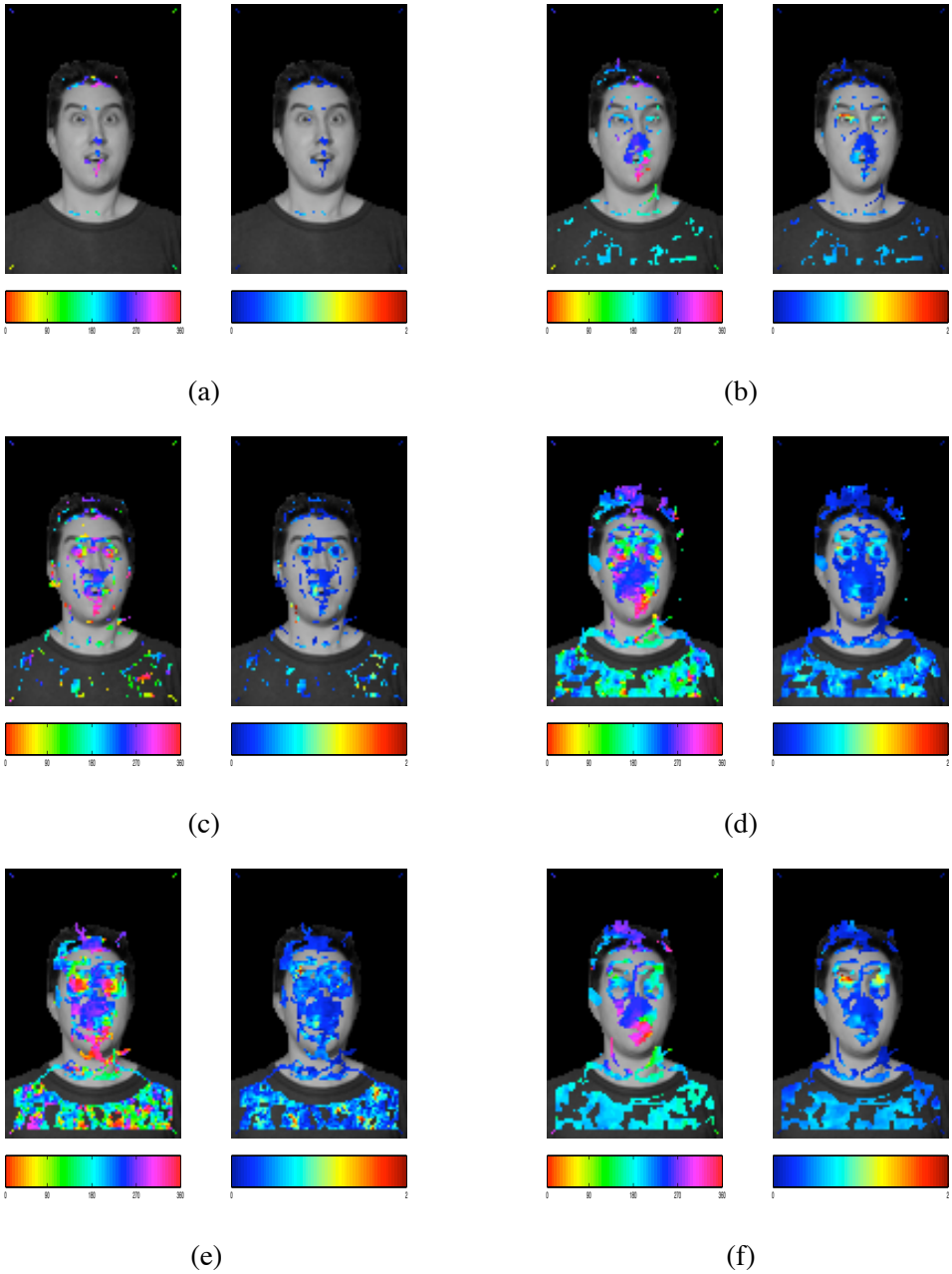
Channels:	Z	ZI	ZH	ZLAB	ZNRGB	ZRGB
SSD	0.110679	0.111088	0.111171	0.112795	0.109312	0.117954

Table 3.1: Sum Square Difference between warped and actual surfaces.

3.5 Discussion

The addition of colour (or intensity) constraints within the range-flow calculation has been shown to improve the estimation and resulting flow, as opposed to using depth alone. It effectively acts to reduce aperture errors and thereby increase the contribution by separate - potentially independent - modalities. In particular, the combined use of aligned raw RGB channels and other colour-spaces can generally reduce error further than simply combining with Intensity. Furthermore, the use of more compact colour representations do in general provide additional help in resolving ambiguities, and appear to provide robustness in the presence of greater levels of noise.

The biggest advantage is certainly that more channels lead to a greater density of estimation than using depth data alone. However, the fact that these additional estimates only apply for planar motion (i.e. parallel to the imaging plane) can unfortunately fail to accommodate for the true motion in depth. This may result in inaccuracies - particularly in the case of non-rigid surfaces that behave in a variety of complex ways (i.e. the face).



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

Figure 3.15: Reprojection of range flow estimates for real data “suprise” sequence with additional channel contributions

Each pair of images shows the angle estimate (left) in polar representation with $0^o 360^o = down = red$, and the magnitude estimate (right) ranging from 0 to 2 pixels of displacement. These calculations are performed for Z *(a)*, ZI *(b)*, ZH *(c)*, ZLAB *(d)*, ZNRGB *(e)*, and ZRGB *(f)*. As can be seen, the additional channels immediately lead to an increased density estimation compared to the Z channel alone. However, this can vary considerably due to the large amounts of movement, and can in consequence predict too much planar displacement relative to the range flow. Overall, the ZLAB solution appears to capture best the motion of the eyes and jaw.

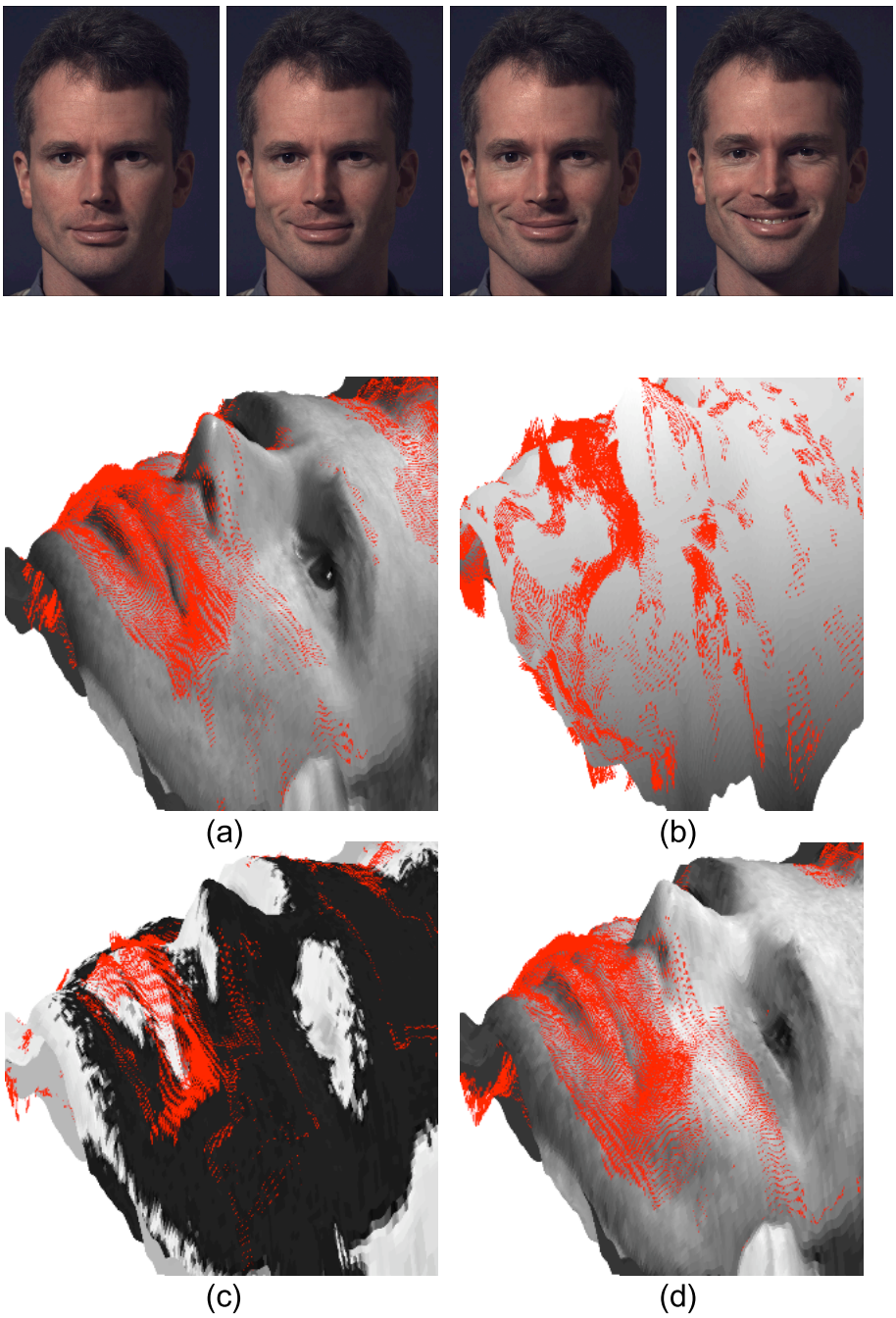


Figure 3.16: Vector flow field in example real data smile sequence *(top)* based on channels: intensity *(a)*, depth *(b)*, hue *(c)*, and red-green from LAB *(d)*.

The estimates formed by intensity *(a)* and red-green *(d)* form very similar and qualitatively reasonable motion around the mouth. The red-green estimate in particular seems more consistent, whereas the intensity produces a greater density (in particular, capturing the motion of the eyebrow). The hue estimate *(c)* appears to miss and propose additional motion (e.g on the side of the brow) and the depth estimate *(b)* appears very erratic - possibly due to the local smoothness of this real example surface.

Perhaps the most surprising result is the effect of the choice of colour-space on the estimation. Firstly, it would appear that exploiting the linear smoothness LAB for representing chromatic change leads to more regularised and stable computation in the case of more complex motion. This also applies in general to using the raw RGB and NRGB representations, which of course preserve the original data as much as possible. Conversely it is those representations that attempt to condense the colour or intensity down to a single value (H and I) that perform the worst.

Fundamental to the generation of these results is the consideration of *aliasing*. As we only consider an integration window of $5 \times 5 \times 5$ *pixels* any displacement greater that 3 units between frames will lead to increasing errors in the estimation. Further enhancements may be possible by exploiting additional texture information and by integration over a longer temporal window. Using an intrinsic image representation, further improvement to colour consistency may well improve the estimation.

A final outstanding topic here is the fact that the stereo spatial reconstruction of the surface from the real data is itself based on the same RGB data that is consequently used to derive the flow. This two stage process, and dependency on the same sensor, leaves open the question of whether independent data sources would perform even better. Furthermore, it suggests that perhaps a single framework would eliminate errors further and simplify the process (as for example performed by [ZSCS04]).

3.6 Summary

This chapter set out to introduce the problems and methods required to resolve spatial and temporal correspondences. Our focus is on how to go from a sequence of paired stereo images, to a vector flow field that attempts to describe at every instance a set of points on the surface, and where they move to in the next moment.

Our key results from this work can be summarised as follows:

- That the use of colour as additional constraints can help resolve localised ambiguity in planar surface motion and improve the density of estimation.
- That the choice of colour space, and its degree of compression down to a singular value can determine the accuracy of the resulting flow.
- That temporal aliasing (too much movement between frames) and ambigious planar motion may result in incorrect calculation for complex moving surfaces.

In conclusion, colour does indeed offer many possibilities for more accurately resolving the motion and deformation of a surface. We proceed to try and exploit this with our dynamic shape descriptors in Chapter 5 in order to unify the description of the surface change. Other additional

improvements and potential investigations to the flow calculation are further detailed in Chapter 6 under future work.

Chapter 4

Describing Dynamic Deformation by Curvature Change

> "Thus we are led to a remarkable theorem: If a curved surface is developed upon any other surface whatever, the measure of curvature in each point remains unchanged."
>
> *C. F. Gauss, 'Theorem Egregium' (1827)*

In this chapter we look to extend the idea of classification schemes from static surface curvature into the temporal domain. We seek to identify regions in sequences of 2.5D depth data that exhibit variations in shape change, and to characterise the deformations. From observing the change in principle curvatures we show how it is possible to de-couple the type of change into one of fifteen classes, and also reveal the extent of alteration. In so doing, we maintain that this produces a compact and parsimonious representation of dynamics for qualitative characterisation of deforming surfaces.

We show how the calculations for this can be performed by applying techniques from differential geometry to 4D data sequences. In order to achieve this we first present the background theory to curvature calculation, particularly in how it can be performed accurately by quadric surface fitting. From these foundations we introduce our proposed scheme and test it with experiments on synthetic and real data. We conclude with some discussion into its effectiveness - particularly the question of accurate comparison between different instances of a surface, as for example when analysing different people and expressions.

To set this chapter in context, we return to our second original research question: *"Does there then exist a useful computational vocabulary for describing observations in the geometry of such a changing surface? Can we qualify a local patch on a surface as bending or folding, for example?"*

4.1 Review of Static Curvature Descriptors

Exploiting the *curvature* of a surface for analysis has a long history of use in Computer Vision. Its main advantage is that it can describe the localised properties that are invariant to any particular viewpoint, illumination or texture (although this can in fact depend on the technique used to capture the surface accurately). The properties calculated for each local *patch* can furthermore be clustered to group similar neighbouring regions, and so arrive at a very useful segmentation and set of features as a means of object recognition.

One approach to this is to consider the "raw" values of the principal curvatures (denoted κ_1 and κ_2). These are by themselves not overly intuitive, so instead it is natural to consider more descriptive, composite forms. The most widely known scheme in Computer Vision in this regard was introduced by Besl in 1986 [BJ86], based on the calculation of the *mean* curvature:

$$H = \frac{1}{2}(\kappa_1 + \kappa_2) \tag{4.1}$$

and the *Gaussian* curvature:

$$K = \kappa_1 * \kappa_2. \tag{4.2}$$

By considering the positive and negative variations for these values, we arrive at 8 possible - and 1 impossible - shape classifications as shown in Table 4.1.

	$K < 0$	$K = 0$	$K > 0$
$H < 0$	Saddle Valley	Concave Cylinder	Concave Ellipsoid
$H = 0$	Minimal	Plane	Impossible
$H > 0$	Saddle Ridge	Convex Cylinder	Convex Ellipsoid

Table 4.1: Mean (H) and Gaussian (K) surface classifications.

While this discretization to a subset of shapes is useful, such a scheme loses the means to describe the *degree* to which the surface is actually curved, and the *progression* in shape. In response to this Koenderink [KvD92] proposed an alternative polar based representation which directly decouples *shape*:

$$S = \frac{2}{\pi} arctan\left((\kappa_1 + \kappa_2)/(\kappa_1 - \kappa_2)\right) \tag{4.3}$$

from an independent value of *curvedness*:

$$C = \sqrt{(\kappa_1^2 + \kappa_2^2)/2}. \tag{4.4}$$

This results - as shown in Figure 4.1 - in a continuous gradation between concave ($-1 < S < -1/2$), hyperboloid ($-1/2 < S < 1/2$) and convex ($1/2 < S < 1$) shapes - with special cases ($S = 1/2, S = -1/2, S = 0$). Each of these shapes can in turn vary their value of curvedness from zero (indicating flatness) towards infinity for ever-increasing, highly curved regions around a point.

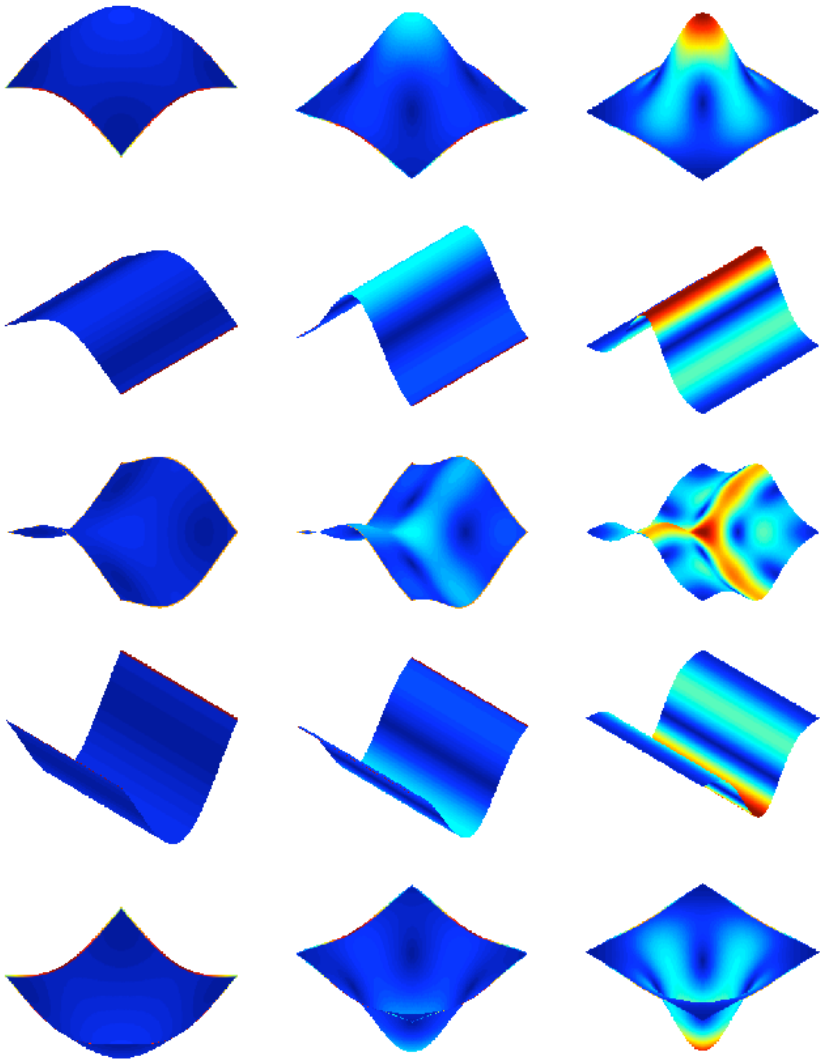


Figure 4.1: Shape and Curvedness.

The ultimate value of zero for curvedness indicates completely flat surfaces (not shown here, but would occur on the far left for all instances shape). Notice that the transitions between the shapes form a naturally descriptive path in alternative directions: from saddle to ridge to peak, or from saddle to valley to pit.

One very important aspect in both these schemes, especially in the case of handling real data, is the necessary step of applying *zero boundary thresholds* - particularly in order to accommodate noise [CF01]. These thresholds effectively allow the classification of either of the principal

curvatures as approximately "flat" at some level of scale, and so enable those classifications that rely on zero curvature in one (or both) directions. Finding the most effective value that can provide this, at the cost of mis-clasifying regions, must often be determined empirically. Despite this, both the $H+K$ and $S+C$ techniques continue to have many applications in analysing static range/depth data from scenes, and inspire us to consider how they might be extended to apply over time.

4.2 Formulation for Changes Over Time

We propose a natural extension to the classification of surface deformation by observing the variation of the principal curvature values over time. As illustrated in the previous section, various classification schemes are available for describing a static surface. In particular, the primary 6 classes of *prototypical shape* which are possible via comparison of the relative signs for the principal curvatures κ_1 and κ_2 are shown in Table 4.2. For these classifications to work, it is often necessary to define the zero boundary regions with a threshold θ for each of the principal directions ($\kappa = 0 \iff -\theta_{shape} < \kappa < \theta_{shape}$). As noted by [CF01], the importance in selecting this threshold is based primarily on a compromise around how locally "flat" a curved surface can be regarded. In the case of real, noisy data it is often useful to select a higher threshold, so as to reveal the form of the true, underlying, surface. This then, in our approach, becomes a trade-off in terms of the accuracy - and scale - with which we seek to describe changes. In all cases throughout this section we attempt to empirically determine the suitable level of scale in conjunction with the threshold value that yields the qualitatively best result.

	$\kappa_1 < 0$	$\kappa_1 = 0$	$\kappa_1 > 0$
$\kappa_2 < 0$	concave ellipsoid ("pit")	concave cylinder ("valley")	hyperboloid ("saddle")
$\kappa_2 = 0$	concave cylinder ("valley")	plane ("flat")	convex cylinder ("ridge")
$\kappa_2 > 0$	hyperboloid ("saddle")	convex cylinder ("ridge")	convex ellipsoid ("peak")

Table 4.2: Principal shape classes based on κ_1 and κ_2 (as proposed by Koenderink).

If one then considers the alteration of κ_1 and κ_2 over time, then the transitions that can occur from one prototype to another, and the dynamic relationships between the classes, can be

visualised as the graph shown in Figure 4.2.

It is furthermore evident from Figure 4.2 that it is impossible for some shape classes to deform into others without first transitioning through an intermediate form (e.g. for a peak to turn to a valley, it must first become flat, or else move from a ridge to a saddle). In reality, when considering a sequence of data, this is entirely dependent on the sampling rate over time. It is then entirely possible for the region in question to transition between any two shape classes, if such a change occurs at a rate that the intermediary stages are not observed. In this work we assume that the data is captured at sufficient speed to preserve these distinctions.

On this basis, it can be realised that there are only a limited number of different types of transition which we enumerate in more detail here:

1. Those that typify, and exemplify further, the formation of a prototype (e.g. "**protrude**" a peak, "**subside**" a pit, "**fold**" a valley, "**bend**" a ridge,"**warp**" a saddle).

2. Those that move the opposite way from the prototype towards flat (e.g."**flatten**"). In combination with the first transition, this serves to capture the equivalent variation in curvedness expressed by Koenderink.

3. Those bi-directional transitions that are also applicable only between neighbouring non-flat prototypes (e.g. to "**squeeze**" a pit to form a valley, to "**collapse**" a valley to form a pit, to "**dimple**" a valley to form a saddle, to "**crumple**" a saddle to form a valley, to "**crease**" a saddle to form a ridge, to "**dent**" a ridge to form a saddle, to "**bulge**" a ridge to form a peak, and "**stretch**" a peak to form a ridge).

4. Those shapes that do not make any transition - as they have no observable change in curvature (e.g. they are "**constant**").

This results in a total of **15 different deformation classes**, which we formalise as the *type* of *deformation*:

$$T \in [1, ..., 15] \tag{4.5}$$

that can occur over any given duration as defined by *the relative change in the principal curvatures* $\Delta\kappa_1$ *and* $\Delta\kappa_2$ *for the* **same** *local surface region*. These are shown formally in the extended Table 4.3 - indicating the transitions that can occur based on each initial shape class and the detected changes (c.f. Table 4.2).

As with the initial shape classes, in order to define the zero boundary region it is necessary to employ a threshold term ($\Delta\kappa = 0 \iff -\theta_{change} < \Delta\kappa < \theta_{change}$). Otherwise, it would be impossible - particularly in the case of real, noisy data - to capture those transitions in which at least one of the principal curvatures remains constant. Given the nature of the data, the use of

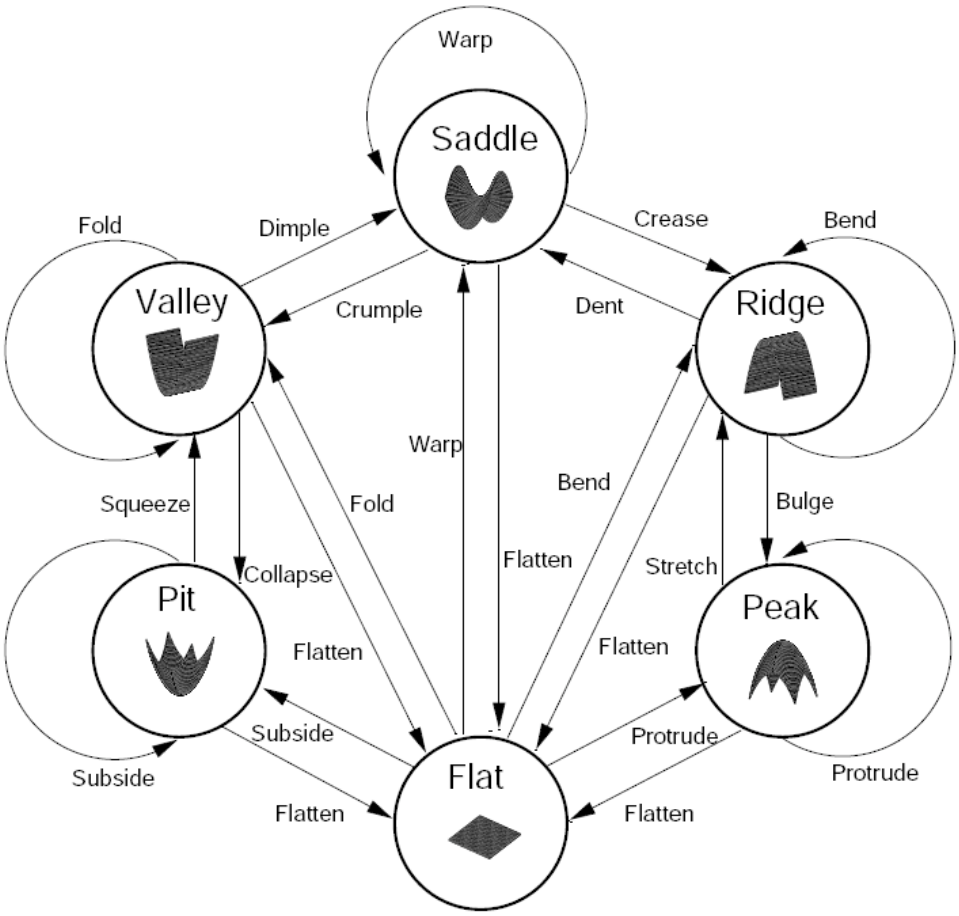


Figure 4.2: The 15 transitions between between principal shapes.

This important figure summarises at a glance one of the primary contributions of this thesis. The simple, but crucial realisation that there exists a limited subset of terms (15 in number - 14 shown in the figure, plus the term “constant” for no discernible change) that can be used to describe the transitions within the space of all principal shapes. Notice that this is effectively based on capturing the dynamics of Koenderink’s $S+C$ scheme, in that every principal shape can increase or decrease in effectual curvedness (C) in moving away or towards flat. Similarly there are transitions between different shapes which correspond to an increase or decrease in shape (S) value. In codifying the terms for the transitions we have attempted to naturally define the processes of deformation. However, notice that certain rotational changes associated with stretching, twisting or skewing the surface (in terms of diffeomorphisms that modify the extrinsic co-ordinate frame) are not directly catered for in this mechanism (see Chapter 5 for discussion and further extensions).

such thresholds are best investigated empirically - particularly in light of the accuracy in second-order derivative terms when calculating curvature for an arbitrary sized local surface patch.

We furthermore can define the *extent of change* (E) to also measure the degree to which this deformation occurs over the duration. This effectively formalises the dynamics of the Koenderink C "curvedness" value:

$$E = \sqrt{\frac{\Delta\kappa_1^2 + \Delta\kappa_2^2}{2}}. \tag{4.6}$$

At this point, we reiterate that our objective here is to propose this scheme as a means of performing higher-level *feature extraction*. As opposed to looking at an arbitrary range of points in $2.5D$ or $3D$ as they move over time, we instead hope to summarise the discrete set of changes that occur in the observations of a surface. However, these can only be effective as determined by the *accuracy* by which they are derived. Ensuring this is the topic of the following section.

4.3 Calculation on Sequences of 2.5D Data

In the following sections, we illustrate how $\Delta\kappa_1$ and $\Delta\kappa_2$ can be calculated robustly. We assume here that the incoming data forms a sequence of $2.5D$ depth-maps $Z(x,y,t)$ - such that there is a neighbouring "patch" of associated data that can easily be accessed around any given point. Many other techniques rely on an underlying topology or triangulation to determine this neighbourhood - and can use the same underlying curvature estimation based on those representations (e.g. [GS03]). However, we instead directly exploit the array structure of the raw data depth-map. In Chapter 5 we look towards more explicit ways of handling re-projected $3D$ data, especially with regard to how it must be registered (tracked) using our experiences from Chapter 3.

Our approach here follows a straightforward pipeline in which we first must register our depth-maps on top of one another (and so accommodate for any rigid translations). We then proceed to extract a surface patch around every point of data, and to fit a quadric to this. From the quadric co-efficients it is then possible to derive the principal curvatures, and so calculate the delta changes between frames. We then show in the next section how these calculations are applied to a number of synthetic and real objects.

4.3.1 Temporal Registration

Registration is the term in $3D$ Computer Vision that describes the process of aligning two surfaces. It can be broadly divided into *rigid* and *non-rigid* cases. Even the simpler case of rigid registration between instances of the same surface can be difficult to achieve in the absence of any mapping between unique landmark features. If these are however defined, then it is possible to find the affine transformation that minimises the distance between them (i.e. Procrustes

Initial Shape:		$\kappa_1 < 0$			$\kappa_1 = 0$			$\kappa_1 > 0$		
Dynamic Change:		$\Delta\kappa_1 < 0$	$\Delta\kappa_1 = 0$	$\Delta\kappa_1 > 0$	$\Delta\kappa_1 < 0$	$\Delta\kappa_1 = 0$	$\Delta\kappa_1 > 0$	$\Delta\kappa_1 < 0$	$\Delta\kappa_1 = 0$	$\Delta\kappa_1 > 0$
	$\Delta\kappa_2 < 0$	Subside	Squeeze	Squeeze	Collapse	Fold	Dimple	Crumple	Crumple	Warp
$\kappa_2 < 0$	$\Delta\kappa_2 = 0$	Squeeze	Constant	Squeeze	Collapse	Constant	Dimple	Crumple	Constant	Crease
	$\Delta\kappa_2 > 0$	Squeeze	Squeeze	Flatten	Collapse	Flatten	Dimple	Flatten	Crease	Crease
	$\Delta\kappa_2 < 0$	Collapse	Collapse	Collapse	Subside	Fold	Warp	Dent	Dent	Dent
$\kappa_2 = 0$	$\Delta\kappa_2 = 0$	Fold	Constant	Flatten	Fold	Constant	Bend	Flatten	Constant	Bend
	$\Delta\kappa_2 > 0$	Dimple	Dimple	Dimple	Warp	Bend	Protrude	Bulge	Bulge	Bulge
	$\Delta\kappa_2 < 0$	Crumple	Crumple	Flatten	Dent	Flatten	Bulge	Flatten	Stretch	Stretch
$\kappa_2 > 0$	$\Delta\kappa_2 = 0$	Crumple	Constant	Crease	Dent	Constant	Bulge	Stretch	Constant	Stretch
	$\Delta\kappa_2 > 0$	Warp	Crease	Crease	Dent	Bend	Bulge	Stretch	Stretch	Protrude

Table 4.3: The 15 types of deformation based on change in principal curvatures from initial shape.

method). Alternatively, if we have a representation of the surfaces as a unique set of points, then each one of these can be considered a landmark in its own right - giving rise to stochastic search methods that seek to minimise the distance between all the points (i.e. the Iterative Closest Point method). If there exist variations in the subjects then alignment may not be overly accurate. This is illustrated in Figure 4.3, where a composite of 20 male neutral and smiling $3D$ captures are registered only by the tip of the nose. Consequently this portion of the composite face model is reasonably sharp, but is progressively worse due to the variations out with of the central area.

Non-rigid registration can accommodate for these variations. It can likewise be viewed as a task of finding a matching non-linear/affine warp that can serve to uniquely define the mapping between a set of points. However, this is much harder to achieve, and gives rise to many approaches that seek to constrain the problem so that it is tractable with additional knowledge of the surface (e.g. by using geodesics). In reality many surfaces have both rigid and non-rigid components which often results in a compromise in the techniques used. The face is a classic example of this where the regions around the eyes and mouth can vary a great deal, but the brow, nose and chin do not.

In this instance, we are concerned with what can be termed *temporal* registration - that is, determing the **same region** between instances of a surface over time. In essence, we wish to register on the basis of those rigid portions of the data, so as to highlight the deformations. This would then mean that we can compare how the surface changes along the lines of our types and extent of deformation. We do not however employ any non-rigid warping at this stage, particularly as we only consider a limited number of frames of data, and we only focus on explicit regions of change. Consequently, we first only compensate for global motion in intra-sequence frames. We assume here that any such local motion is minimal, given that the data has been captured at a suitably fast frame rate, or has already been externally aligned.

We use a modified Iterative Closest Point algorithm [BM92] to minimise the error distance between the two frames represented by the data $Z(x,y,t)$ and $Z(x,y,t+1)$ projected into $\mathbb{R}^3$. Since the data is already roughly aligned, this process should produce the result we desire of registering the non-rigid portions and accentuating the changes on the surface. In order to give precedence to this emphasis between rigid and non-rigid elements we employ the Tukey M-estimator as robust means of totally rejecting outlying points (i.e. those points that have deformed) in a similar way as presented in [ME03]. We seek to only allow the best possible fitting for a rigid subset to the data, and so remove the 50% greatest distances outright in order to reject outliers caused by deformation. Having a fitting, we then project the final registered point back via the first camera frame for the resulting depth data alignment of frame $Z(x,y,t+1)$. In this way, by registering the entire head, the regions around each point in the resulting aligned depth-maps are directly comparable.

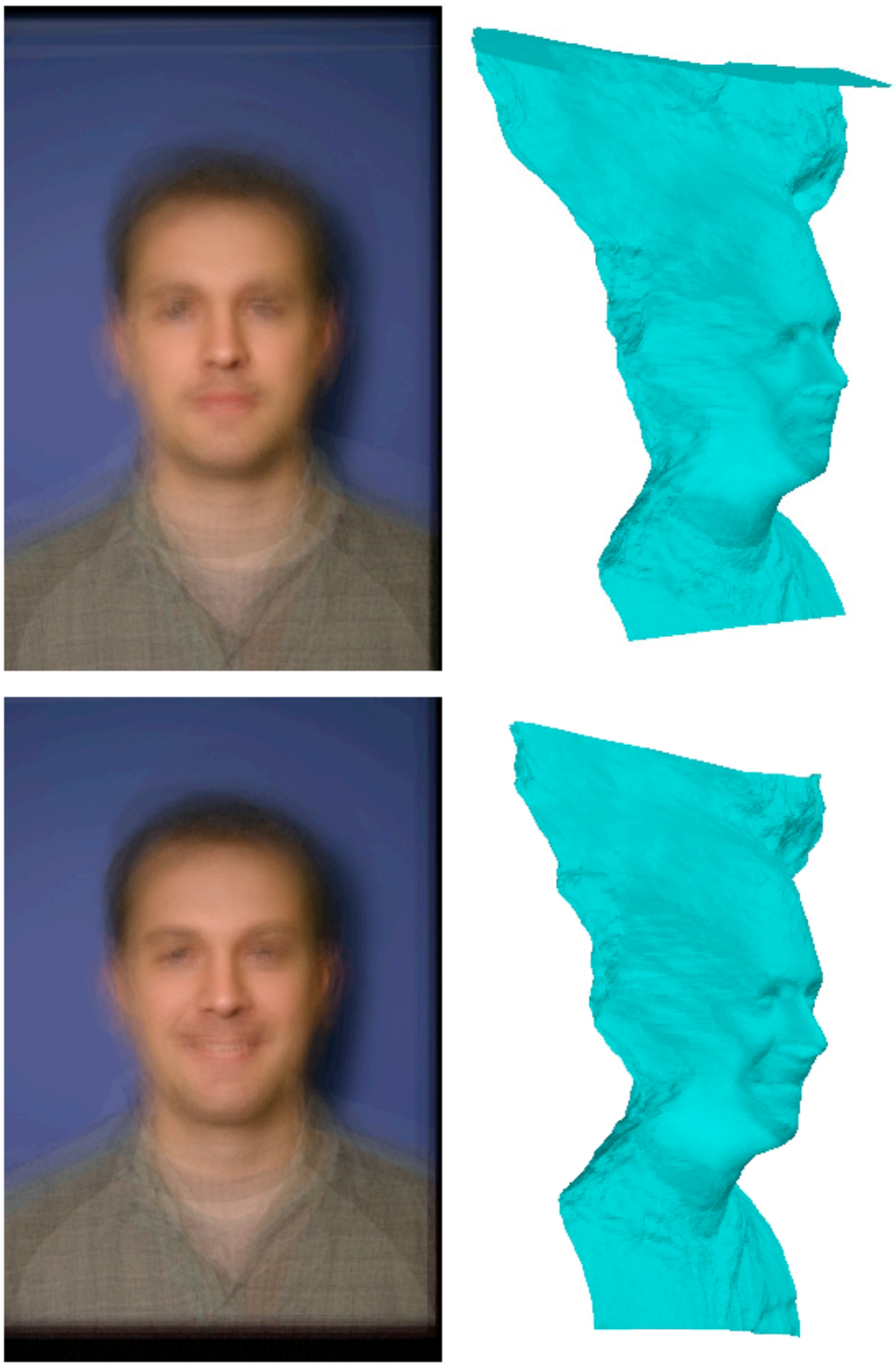


Figure 4.3: Image and $3D$ of an average neutral expression and smile.

This rigid registration was created from 20 male subjects by exploiting the peak on the nose to align the depth-maps (and consequently the images with the same transform). The depth data is then re-projected in $3D$ to create the full model. Notice how, since no transform is applied based on other landmark features, the data is "sharpest" around the nose - but smooths outwards and is indistinct around the eyes. However, since all the heads are roughly the same size, the resulting average still manages to preserve a good degree of detail.

4.3.2 Quadric Fitting

Having performed the registration of the data, our next step is to calculate an analytic model of the surface for each local neighbourhood surface patch. This will allow us to accurately calculate the principal curvatures at the point in question. For this purpose we fit a second order polynomial - specifically a *quadric* - to the neighbourhood of each point. A further advantage in using quadrics is that they can implicitly smooth through their minimisation to underlying data while fitting. In this way, any noise in the data is accommodated down to the level of scale for the size of patch to be fitted to by the quadric. In effect, the quadric can be used to accommodate small variations and undulations in the data by finding the minimal underlying surface that satisfies the best fit. This technique can be further employed to effectively down-sample particularly dense data, as dictated by the distance between sample points, and the size of the surrounding data patch.

This simplistic, yet powerful, surface model has been used in the interpretation of range data (for example see [RFWA00], [MLM01] and [FF01]). However, as summarised by these works - there is a range of different options (focusing on the relative merits of algebraic, Taubin and Euclidean minimisation during the fit to the data). Since we are concerned with only small, localised patches - we simplify the process by seeking to align the quadric by first moving the data to the principal frame, centred at the origin with the normal aligned on the z-axis. Furthermore, we do not (in this case) re-project the data into $3D$. This allows us to focus on fitting only an "extended" quadric of 5 coefficients (as opposed to the full 10 coefficient generalised form):

$$z = ax^2 + bxy + cy^2 + dx + ey. \tag{4.7}$$

We start by iteratively refining the fitting following the approach proposed by McIvor and Valkenburg [MV96]. This proceeds as follows:

1. Take a patch of data around a point: $[X,Y,Z]$.
2. Translate the patch's center point to the origin.
3. Fit a plane to the patch to derive initial normal.
4. Align the patch normal to the z-axis.
5. Perform a Least Squares fitting of the quadric: $\left[X^2,XY,Y^2,X,Y\right][a,b,c,d,e]^\top = Z$.
6. Derive a new normal: $[-d,e,1]^\top/(1+d^2+e^2)$.
7. Repeat from step 4 until normal aligned: $d = e = 0$.

Having performed our initial alignment we then focus further on fitting the data to the "principal" quadric (since the co-efficients that determine the normal have been resolved):

$$z = ax^2 + bxy + cy^2. \tag{4.8}$$

Here we seek to minimise the Euclidean Sum Squared Distance (SSD) between the patch data and the quadric. As a relatively hard, but linear problem, with the potential of some local minima, we use the *Simplex* optimisation algorithm [NM65], where our initial estimate is provided by the extended quadric fitting performed above (which serves to minimise the *algebraic* distances between the points and the quadric surface). We can then operate upon the following objective function that iteratively considers the *euclidean* distance (*dist*) between the surface Z and the points x (see [FF01]):

$$argmin(a,b,c) = \Sigma dist(x_p, Z(a,b,c))^2 \tag{4.9}$$

An example of the results of this process for fitting a quadric to real data is illustrated in Figure 4.4 below.

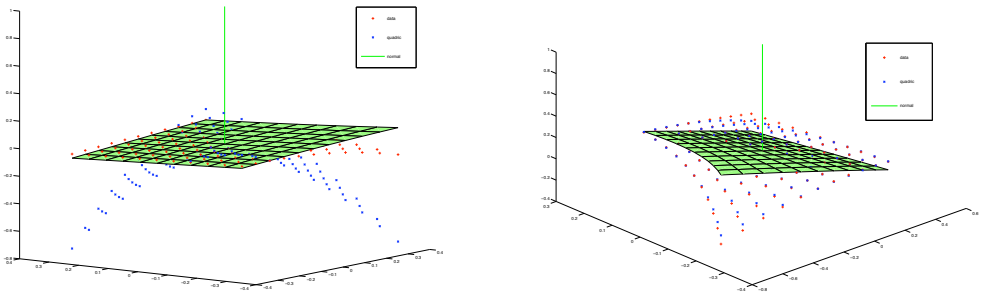


Figure 4.4: Example quadric fitting to a patch of data.

The first phase of establishing the normal by fitting a plane (green) through the data is shown on the left, with the resulting normal (green line) being aligned to the z-axis. The initial extended quadric is indicated (red) - but varies considerably from the data points (blue). On the right however, the results of 100 iterations of the Simplex optimisation shows the resulting quadric points virtually indistinguishable from the target data points. The final error in this case is $1.0E^{-9}$ indicating a very good fit.

In many instances the final fitting represents an accurate, descriptive, compact model of the local surface patch. However, in certain cases the fitting cannot accommodate for the effectual "complexity" of the surface around the point in question. In these cases, the quadric can at best only approximate the minimal shape (i.e. plane) that transects the data. Instances of such surfaces include "corner valleys" and "Y ridges". To a large extent, the occurrence of these complexities is a factor of *scale* of feature the quadric must accommodate. The larger the patch, in proportion to the subject, the more this is evident. This is best illustrated in Figure 4.5 below,

and is furthermore reflected in the final error value of the each of the fittings (this forms a metric we exploit later on in smoothing).

4.3.3 Curvature Change Calculation

The calculation of curvature for the quadric at a point in frame $Z(x,y,t)$ can be derived from the 3 coefficients of the fitted principal quadric (Eq. 4.8) at that point, such that the first principal curvature is:

$$\kappa_1 = (a+c) + ((a-c)^2 + b^2)^{1/2} \tag{4.10}$$

and the second principal curvature is:

$$\kappa_2 = (a+c) - ((a-c)^2 + b^2)^{1/2}. \tag{4.11}$$

We then desire to accurately calculate the relevant delta changes in curvature, in order to then apply our scheme for describing type and extent of deformation. Fundamental to our approach here is to robustly accommodate for local variations due to noise and the error in quadric fitting. We therefore employ both spatial and temporal smoothing over a local neighbourhood of curvature values (as stored in the original format of the depth-map array for every point to which a quadric has been fitted).

Thus, given a sequence of temporally registered frames at each time-step t for $\kappa_1(x,y,t)$ and $\kappa_2(x,y,t)$ then the temporal variation can be calculated for a window of duration d. This is computed by a convolution (denoted with $\star$) with a $1D$ Derivative of Gaussian (D_t) filter in the t dimension, then by spatial integration via convolution with Gaussian filters ($G_{x,y}$) in the x and y dimensions as follows:

$$\Delta\kappa_1(x,y,t) = G_x \star G_y \star D_t \star \kappa_1(x,y,t \pm d/2) \tag{4.12}$$

and similarly:

$$\Delta\kappa_2(x,y,t) = G_x \star G_y \star D_t \star \kappa_2(x,y,t \pm d/2). \tag{4.13}$$

For both spatial and temporal convolutions we employ a Gaussian window with standard-deviation equal to 1. The window width d is the same both spatially and temporally over the period of integration (it is the same size so not to bias the calculation). This is then able to accommodate larger motion which will potentially travel further and extend beyond the limit of the local registration. Ultimately, relying on convolution will lead to aliasing failure for motion that extends over longer intervals - and so requiring more advanced tracking (as we detail in Chapter 5). However, for our purposes we assume that the gross level of change for any local region is not too great.

Figure 4.5: Reconstruction of face data by individual quadric patches.

Here is shown how the majority of this face depth-map is accurately reconstructed, with each patch translated after fitting back to the global co-ordinate frame. Each patch is around $120 \times 120\, pixels$ - which at this scale represents around $1cm^2$. However certain regions are not accurately fitted - as indicated by the closely banded areas which are exaggerated to indicate the gradient. These errors are introduced in those patches where the surface data cannot be modelled accurately by the principal quadric.

4.4 Experiments: Qualifying Variation in Deformation

Having described our approach for the calculation of the changes in curvature for every point (in this case, arranged by neighbourhoods of values in the original depth-map format) - we now seek to validate our claim that we can accurately, and meaningfully show the qualitative advantages of our scheme for describing deformations. It should be stressed that we only here seek to consider very short, episodic sequences - of only 4-5 frames. This allows us to focus on the most significant and relevant changes without (yet) handling any of the aspects related to timing or reversals.

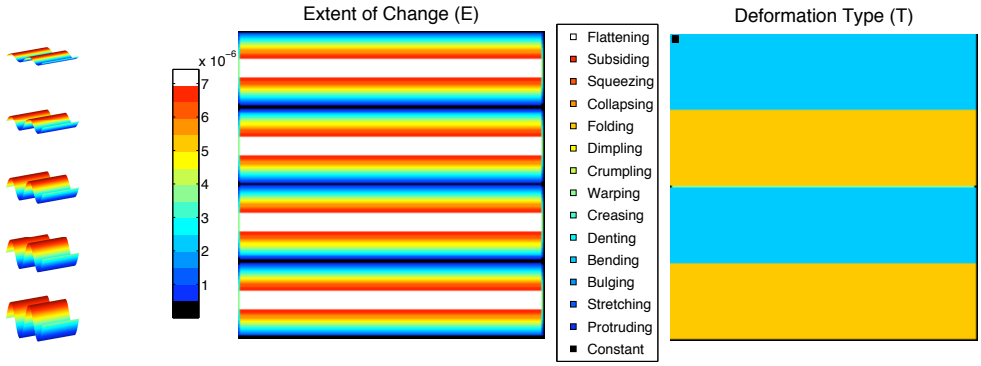
We first seek to validate our approach on a set of synthetically generated objects with known morphology. We then include a number of examples of real test objects - card and putty based - that are manipulated to present some further interesting cases, and to highlight the issues with real data. From this we then progress to looking at the complexity of actual facial expressions over a number of different subjects.

In all the experiments below it should furthermore be noted the considerable degree of processing required - as individual quadrics must be fitted for every point, using a patch of a given size. In this event - all the processing was performed "off-line". We must also set the θ_{shape} and θ_{change} thresholds experimentally according to the scale of the data, and in order to empirically best capture the range of deformation.

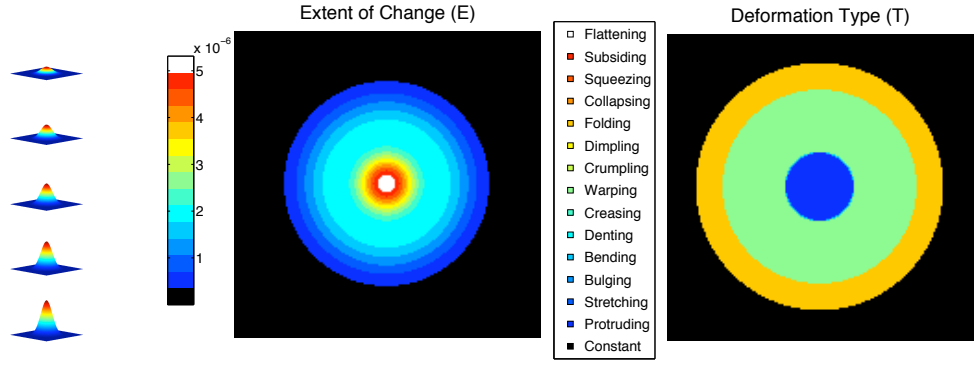
4.4.1 Synthetic Test Cases

To numerically verify and illustrate our formulation, we first apply the calculation to synthetically generated data - as shown in Figure 4.6 below. Both sequences are generated over 5 frames at a resolution of $500 \times 500\, pixels$, with no additional global motion (so registration is effectively perfect already and need not be performed). We then down-sample and fit to this data at every 3^{rd} point with a 7×7 quadric, producing a final result array of $166 \times 166\, pixels$ for each κ_1 and κ_2 value which is more efficient to process. The 5 frames are integrated together, by derivative of Gaussian convolution, in order to calculate the cumulative change over the entire sequence.

In sequence *(a)* the data is generated from a sinusoidal wave with the amplitude increasing by 0.1 every frame. We set the thresholds $\theta_{shape} = 1.0$ and $\theta_{change} = 0.00001$ to effectively define all initial shapes as locally flat, and only verify the effects of altering the principal curvatures. In sequence *(b)* we consider a $2D$ Gaussian with standard deviation $\sigma = 0.05$. This is scaled in z over the sequence by a factor of 10% between frames, using the same thresholds as above. In both cases we still integrate the values spatially and temporally with the convolutions applied to a $5 \times 5 \times 5\, pixel$ volume. This can, as with all such operations, introduce artifacts to the edges of the data where missing data must be duplicated.



(a)



(b)

Figure 4.6: Synthetic data for expanding sinusoidal wave *(a)* and Gaussian peak *(b)*.

In *(a)* the extent of change shows the simultaneous increase in shape change toward the maximal bends in the surface. The types of deformation are also shown correctly to be primarily of bending and folding, with slight boundaries of buckling where the transition is made (only one pixel wide in the plot above).

In *(b)* the extent of change here nicely shows the differential increase expected, which is particularly large towards the apex of the data. For the types of deformation, the flat region round the peak is labelled as protruding, with folding then occurring as the relative angle between peak and base increases. The sides of the peak then suffer warping/bending as the peak protrudes further and increases the elongation of the surrounding curvatures.

4.4.2 Verifying Simple Real Objects

Here we conduct experiments using our dense stereo rig to capture a sequence of $3D$ surface models of interesting deformable, but relatively simple, test objects. We apply various forces to these objects using additional apparatus - allowing us to perform a simplified "stop-motion" capture at progressively more advanced moments of change (as shown in Figure 4.7 below). This results in 5 frames of data from which we select a 500×500 $pixels$ region. As the object is held rigidly, we do not need to perform any additional surface alignment. We do however down-sample to and fit each localised data patch to a quadric in order to calculate the shape and change in curvature. Even for such simple objects this still represents a large volume of data requiring considerable processing.

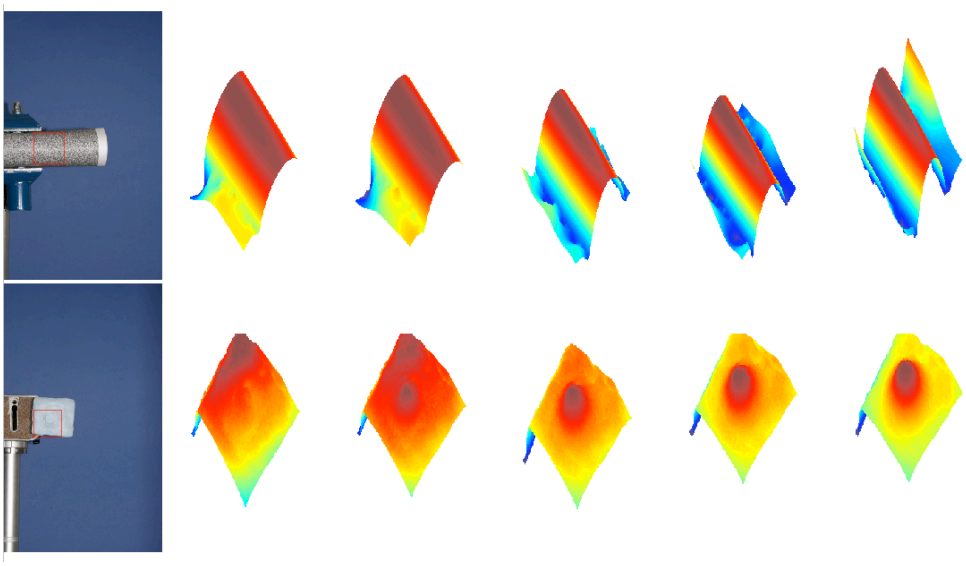


Figure 4.7: Real test deforming object sequence: paper fold *(top)* and peak protrusion *(bottom)*.

Notice the use of random texture noise to enable accurate reconstruction in the case of the folded paper. Surfaces constructed from "Blue-Tac" also exhibit enough stochastic texture and is suitably non-specular to also permit effective reconstruction. The 500×500 $pixels$ regions in question cut from the entire reconstruction is shown in the photographs by the red rectangles in the images. The colour for the $3D$ reconstruction is purely to highlight the changes in the depth data - showing the formation of the shape changes.

The results of these experiments are presented in Figure 4.8 and Figure 4.9 below - showing purely qualitative analysis of the extent and type of deformation. Immediately apparent here (when using real data) are the effects of setting the zero crossing thresholds in order to highlight significant patterns. Here we have empirically set $\theta_{shape} = 0.000001$ and $\theta_{change} = 0.0001$ to counteract the reality of handling noisy real data that is nowhere locally flat. Similarly, we also

highlight here the importance of selecting the relative *scale* at which to perform sub-sampling and patch fitting, and the consequent results of selecting different patch sizes for quadric fitting.

4.4.3 Looking at Sparse Face Sequences

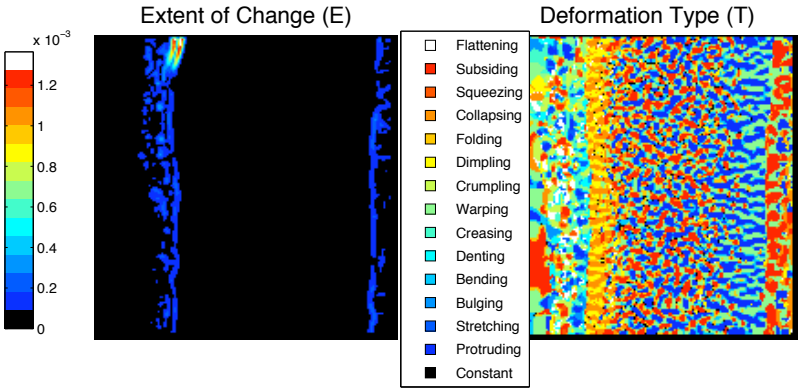
Our dense stereo, high-resolution, capture rig constructed from two 6 $Megapixel$ cameras was calibrated and used this time in "burst" mode (i.e. able to rapidly capture images) to create a sequence from three pairs of subjects making different expressions of happiness, sadness and surprise - a total of 6 sequences (in 3 pairs of two). The intention was to illicit as natural and yet as prototypical an expression as possible - although these expressions are known to be relatively similar given they all involve considerable movement of the mouth and eyes. The sequences were sampled at 2.5 frames per second for a duration of 1.6 seconds - capturing 4 pairs of images, as shown in Figure 4.10. Since we require 5 frames of data, we simply copy the first set of images at the start (this is a limitation of the stereo system when used in burst mode).

Following this, dense stereo recovery was performed on each pair of simultaneous images to create a single high resolution 2048×3072 $pixels$ frame of depth data. Additional masking of the data was achieved by removing any blue background by chroma-keying. We also apply here the crucial pre-processing step of non-rigid registration via ICP with robust outlier removal. This is particularly necessary in the case of the happiness expressions which elicit the natural muscular response of pulling back the head when smiling. In total, the registration stage took up to ≈ 100 iterations and resulting in a final SSD error of the order of $0.1cm$.

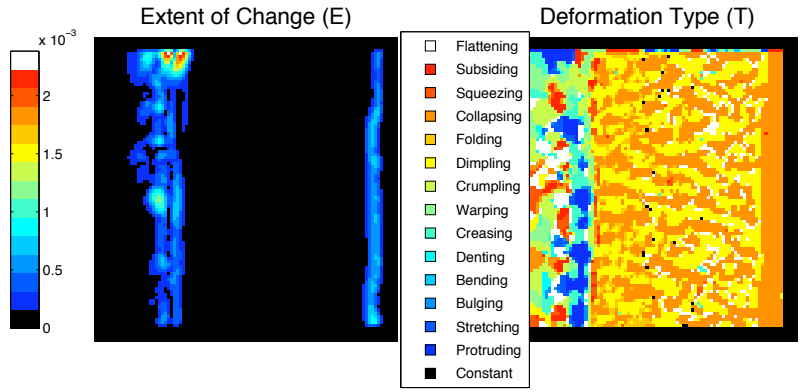
This was then followed by down-sampling to every 15^{th} point with a 121×121 $pixels$ neighbouring patch ($\approx 1cm^2$ in the original data) fitted to a quadric. The final κ_1 and κ_2 data arrays are then 136×204 $pixels$. After some initial experimentation, we set the thresholds $\theta_{shape} = 1.0 \times 10^{-9}$ and $\theta_{change} = 1.0 \times 10^{-5}$ to define the zero boundaries. These are selected to preserve a fair degree of the surface as constant as possible - and so highlight those regions that actually change. However, this has a noticeably greater effect on the surprise sequences, which can in part be attributed to less non-rigid motion.

As with our synthetic data-set, we again consider a spatial-temporal integration window of $5 \times 5 \times 5$ $pixels$ when convolving the data. The results for extent and type as shown in Figures 4.11, 4.12 and 4.13 reveal deformations that exemplify the areas of the face that alter for particular expressions. In particular, the regions around the eyes and mouth which are typical for all expressions - but in subtly different ways that can be immediately visualised by our approach.

To emphasise these distinctions further, we also perform a simple count of each of the 15 types and plot the results as a un-normalised histogram for each expression. Given the relative size differences of the subjects in question (a factor of their different sized heads and position), and the differences in the extent of change, the histograms still appear to maintain a considerable degree of similarities. They also enable us to qualify the distinctions between the expressions.



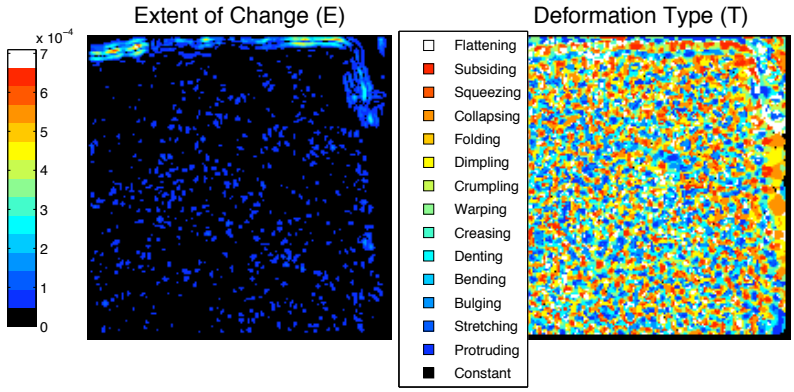
(a)



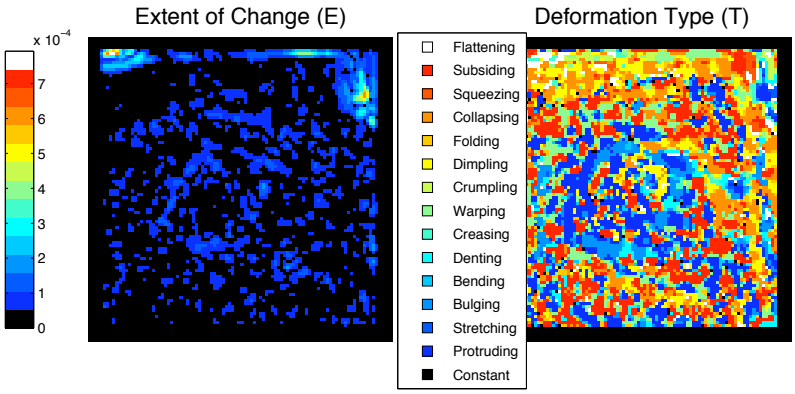
(b)

Figure 4.8: Test object extent and type for paper fold.

In *(a)* the data is sub-sampled at every 3^{rd} point with a quadric fitting to a patch $17 \times 17\, pixels$ - resulting in a $166 \times 166\, pixels$ array. In *(b)* the data is sub-sampled at every 5^{th} point with a quadric fitting to a patch $51 \times 51\, pixels$ - resulting in a $100 \times 100\, pixels$ array. Apparent from the extent of change is the tendency of large values - caused here by the "edges" of the fold moving inwards - to dominate the more subtle changes, irrespective of the size of quadric patch used. Similarly, the detail in the type of change can be hard to discern at lower level of scale, which accentuates the subtle localised variations in the surface. This effect can be somewhat mitigated by considering larger patches - seen in *(b)* - to reveal that the expected types of "folding" and "dimpling" are occurring.



(a)



(b)

Figure 4.9: Test object extent and type for peak protrusion.

In *(a)* the data is sub-sampled at every 3^{rd} point with a quadric fitting to a patch $17 \times 17\ pixels$ - resulting in a $166 \times 166\ pixels$ array. In *(b)* the data is sub-sampled at every 5^{th} point with a quadric fitting to a patch $51 \times 51\ pixels$ - resulting in a $100 \times 100\ pixels$ array. Here again the extent of change is dominated by larger changes occurring on the edges of the data. Similarly, the types of change are again dominated by the complex nature of the surface and the subtleties at finer levels of detail. Only by reverting to larger patch areas commensurate with the scale of the deformation do the more global variations become apparent - seen in *(b)* - moving ever more towards the types expressed by the similar synthetic case (Figure 4.6).

Figure 4.10: Left images (from the original stereo pair) showing 4 frame sequences of happiness (*top*), sadness (*middle*) and surprise (*bottom*).

For example, the histograms shown in Figure 4.12 for *surprise* indicate a good degree more "bending" than exhibited by *happiness* in the histograms in Figure 4.11.

For further statistical analysis of these results, we consider the Bhattacharyya Distance Measure[Bha43] between histograms in order to establish feature consistency among the two subjects for a particular emotion and inconsistency when comparing one emotion with another. This effectively treats each histogram as a unit vector, and returns the cosine of the angle between them. We first take the histograms generated by our technique and *normalise* to length 1.0. Following this, in the discrete case of two normalised histograms a and b (representing distributions over the same sample domain X) it is expressed as:

$$BDM(a,b) = (1 - \sum_{x \in X} \sqrt{a(x)b(x)})^{\frac{1}{2}}. \tag{4.14}$$

This metric can serve to qualify the respective similarity between the different expressions,

where 0.0 = similar and 1.0 = dissimilar. As can be seen in Table 4.4 below, it is the corresponding intra scores between the similar expressions that are indeed most similar - particularly in the case of Happy. Conversely, the inter scores reflect very little variation. This provides evidence that we are measuring a certain degree of consistent patterns of shape change for the two expressions.

4.5 Discussion

The results presented here show an encouraging degree of distinguishable variation, given the relative simplicity of the idea (i.e. measuring and categorising changes in principal curvature). Our initial experiments using synthetic - and perfectly aligned - data show the type of results achievable under ideal conditions. However, while the results for real data are promising, and reveal at a glance interesting regions in surface deformation, they could certainly be improved.

Fundamentally, there is the crucial aspect of point-to-point surface correspondence (temporal registration) between frames, especially over greater intervals of duration. Our use of only 5 frames was mitigated by the relatively small changes which we wished to highlight, whilst removing any rigid component between frames by using the robust statistic version of ICP. For longer sequences this is no longer truly feasible, although it is still possible to choose a suitable "window" for integration of change, but the question of reliably aligning the values in question, and also of the duration of the window to be considered, then becomes crucial.

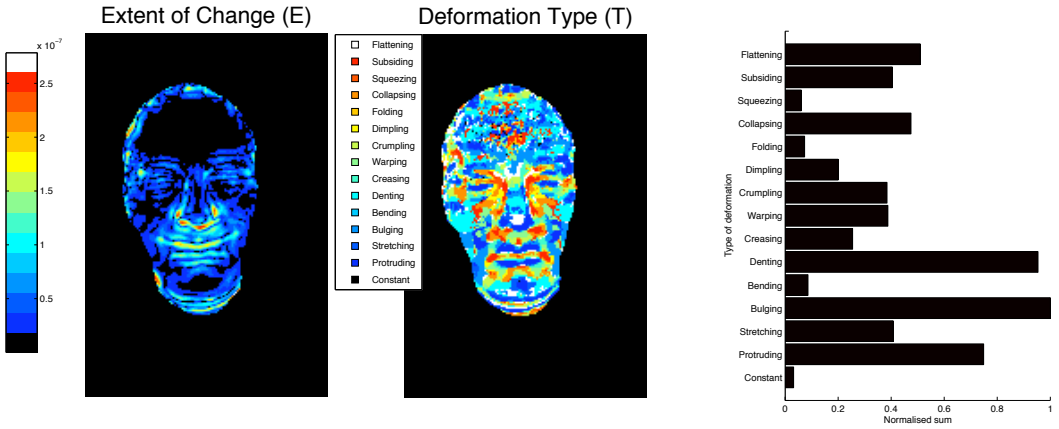
The first issue is then in reliably tracking the motion of the surface. As we have already seen in Chapter 3 - this can be a particularly difficult issue to resolve, especially for large displacements, and for surfaces that change their underlying structure (by opening or closing holes). Combining surface characteristics (flow, colour, texture, etc.) within a global constraining topology - could enable more robust tracking and local registration of points as they move and deform.

The second issue - the episodic nature of real surfaces changes - relies entirely on the accuracy in tracking, but then raises the interesting question of how to handle the fact that a reliably registered surface that bends, and then folds back again would be described as constant in our current scheme. Maintaining a simple integration of the extent of change for the duration of the sequence could be performed (as a simple "bending energy" approach). However, mapping the transitions from one shape class to another would require a more advanced approach. The possibility for explaining observations as most likely paths could be described by a Markovian framework. This could certainly be used to describe the even longer transitions that occur between entirely different expressions, which in turn suggests a hierarchical approach.

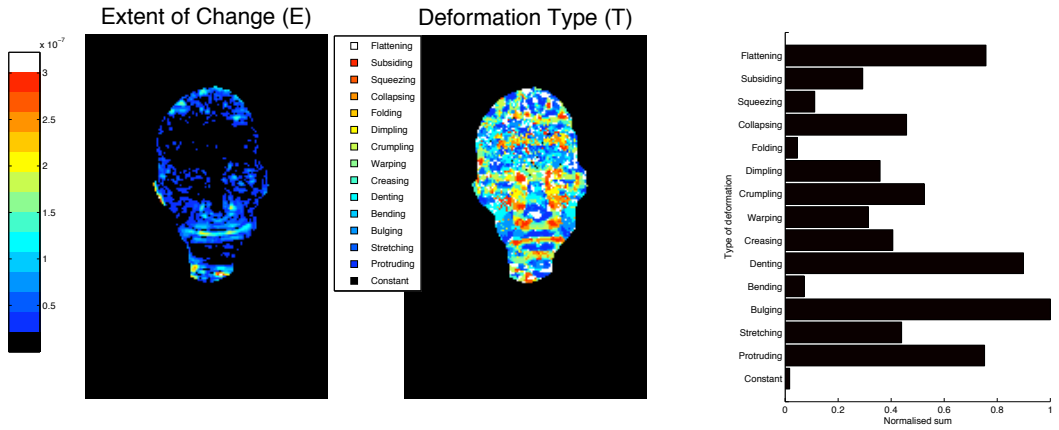
A further important point with the scheme proposed here - especially in response to real data - is the accuracy in calculating the principal curvatures. This mainly concerns the difficulties in resolving noisy and arbitrary complex surfaces, in order to resolve their true underlying

	Happy-A	Happy-B	Surprise-A	Surprise-B	Sadness-A	Sadness-B
Happy-A		**0.0875**	0.4230	0.4627	0.3949	0.3438
Happy-B			0.4155	0.4852	0.4633	0.4112
Surprise-A				**0.2123**	0.5632	0.4476
Surprise-B					0.4391	0.5178
Sadness-A						**0.2619**
Sadness-B						

Table 4.4: Bhattacharyya distances between histograms of 15 expression types.



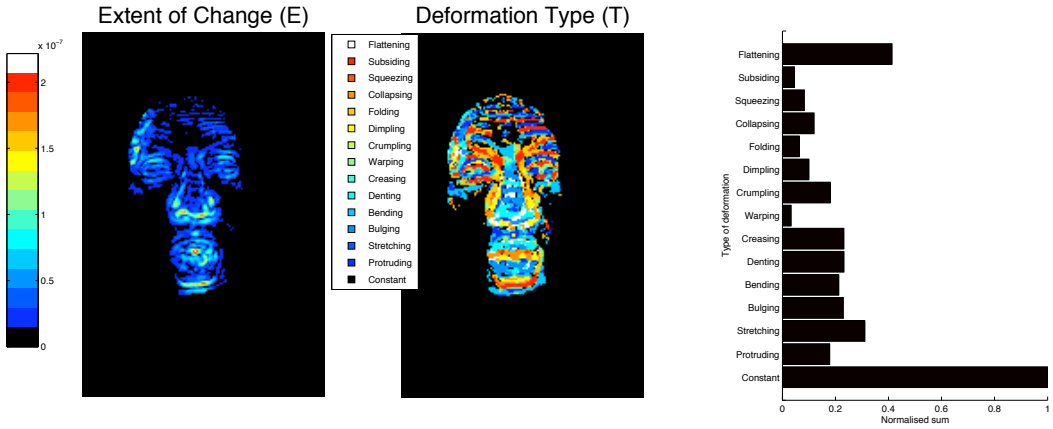
Happy-A



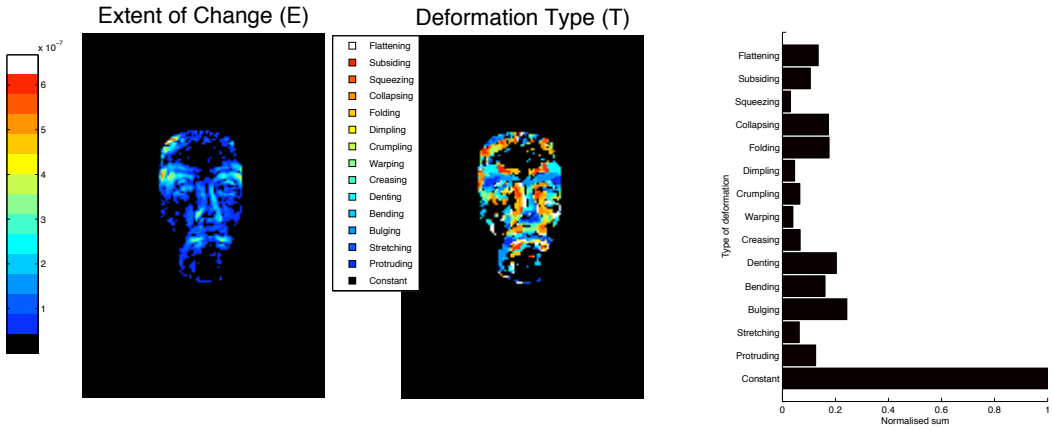
Happy-B

Figure 4.11: Real data extent and type of deformation for *happiness*.

In these examples of *happiness* shown for two different subjects (Happy-A) and (Happy-B), it is the regions round the mouth, nose and eyes that vary the most. This is most prominent in the relative extent of change, which clearly shows a smile forming. The distributions of types of deformation show a very similar pattern between subjects - as further revealed by comparing the histograms.



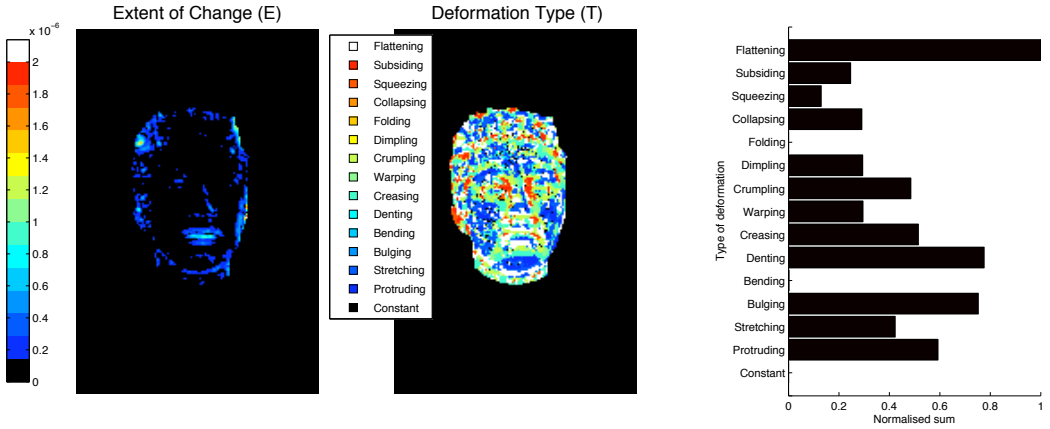
Surprise-A



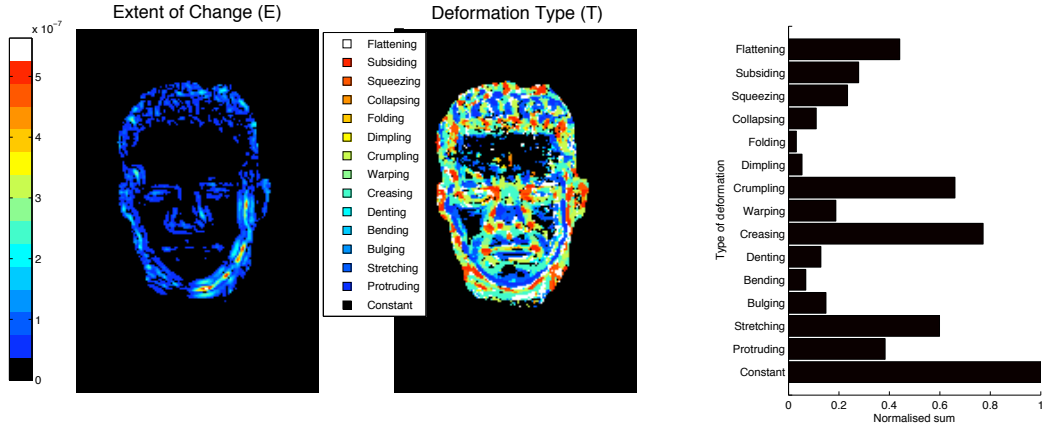
Surprise-B

Figure 4.12: Real data extent and type of deformation for *surprise*.

In these examples of *surprise* as shown for two different subjects (Surprise-A) and (Surprise-B), it is the brows which are seen to move in the characteristic upward direction, with ridges and valleys forming. Large portions of the face are also more constant, particularly around the cheeks. The extent of change again shows this most clearly, while the distribution of types indicates the occurrence of more bending and folding - especially on the brow where it typifies this expression.



Sadness-A



Sadness-B

Figure 4.13: Real data extent and type of deformation for *sadness*.

In these examples of *sadness* as shown for two different subjects (Sadness-A) and (Sadness-B), it is the mouth region that seems to alter most, with a fair degree of bending and creasing occurring. The eyes also appear to vary (in a similar way to happiness). The extent of change appears to be more subtle - as governed by the particular threshold values. The distribution of types points towards a greater degree of flattening than the other expressions, although in the case of Sadness-B the distribution is quite similar to that of happiness.

curvatures. Of critical importance here is the relative scale at which the computation around a local surface patch is carried out. This is, in effect, governed by the size of feature and thus the deformation that can occur. The inherent multi-resolution nature of complex surfaces should ultimately be addressed, in which successive levels of detail in the surface could define the variations occurring at different scales (suggesting some form of hierarchical quadric fitting).

Related to this is the issue of discovering the best threshold levels which enable the most revealing description of the surface changes. Finding these in the experiments described here involved a large degree of empirical validation. Ultimately, this should also be governed by the scale of the surface and the variation in deformation expected, and similarly by the sampling rate over time. Furthermore, varying of levels of noise could be introduced to the synthetic data sequences, thereby directly comparing the effect with the data gathered from the equivalent simple test sequences (for a more meaningful comparison of the overall approach).

However, in general this promises to be a useful technique that can be employed when analysing data captured from novel video-rate $3D$ systems at higher-frame rates (with a suitable lag for integration). One aspect is that it immediately suggests the potential use as a feature-extraction stage prior to classification or identification. This is related to the idea that when considering the face it would be useful to de-couple extent (from the type) and so more effectively represent a measurement of the intensity of expression. This approach then provides a possible means to further investigate the automatic analysis of faces along the variational axes of both form and intensity, with the possible identification of particular expressions (or person) using a statistical approach.

Additionally, other types of rotational deformation could be determined from the changes in the principal directions. This could be calculated within a more rigorous mathematical framework that takes advantage of the changes occurring in the first and second fundamental forms. Other extrinsic properties such as stretching and skewing of the surface would reveal further types of change.

4.6 Summary

In this chapter we have introduced a novel technique for the description of surface change in terms of the changes in differential geometry. This is proposed as a temporal extension of static curvature analysis, which uses initial and variations in the principle curvature values to characterise the type and extent of deformation ($T+E$). The control offered by the threshold values allows us to specify the bearing that the initial shape and amount of change has on these categories. We have shown the importance in non-rigid alignment in the process, and how we can calculate - via quadrics - and integrate the resulting values spatially and temporally for more accurate results.

Our key results from this work can be summarised as follows:

- That there exists a set of 15 *types* of deformation between the principal shapes. These can act as a visually useful basis for analysis of surface change, especially as a first step of feature extraction from the wealth of data.
- Similarly, we can measure the *extent* of deformation by the magnitude of change in the principal curvatures. This too can reveal interesting variations in the data.
- Furthermore, we can use this scheme to qualitatively assess the variation in change on real surfaces. In particular, the expressions of the face show encouraging similarities and differences along the intuitive distinctions that occur for actual deformations.

In conclusion, this approach offers a potentially useful insight into the nature of $4D$ video data. However, we have yet to fully apply it to true, non-episodic data over multiples of 30 to 100 frames. More fundamentally, this scheme does not accommodate those extrinsic changes that can affect a surface - such as rotation and stretching. In Chapter 5 we combine our understanding from Chapter 3 with the results of this Chapter in order to look at more complex and varied surface changes. Further insights and thoughts on potential further enhancements are reserved for Chapter 6.

Chapter 5

Analysing Expression by Changes in Form

"Frons, oculi, vultus persaepe mentiuntur; oratio vero saepissime."
(The face, eyes, and expression frequently lie; but the tongue lies the most.)
Cisero - Epistulae ad Quintum Fratrem

In this chapter we finally seek to bring together the two aspects of both **tracking** a deformable surface, along with seeking to **describe** what it is doing. Central to our approach is the idea that a richer set of features lends itself better to the analysis of the dynamics of complex surfaces. To this end we propose an enhanced framework for expressing the nature of deformation based on analysis of the fundamental forms. In so doing, we devise a further set of *qualifiers* that express additional changes on the surface: rotation, skew and expansion. We then focus on attempting to determine the variations in expression over a larger data-set of different subjects' facial expressions.

To achieve this we first present an overview of the $4D$ capture framework we use to acquire our data. In particular we show how to unify the comparison over different subjects by conforming a generic mesh. Tracking larger displacements are addressed by using a global thin-plate-spline and feature based tracker. We review the derivations of the fundamental forms, before expanding on how to calculate the feature set we require. We then show how these can be applied to the tracking and analysis of facial expression in 4D data sequences - highlighting the problems still encountered that hinder this approach in terms of final classification.

To set this chapter in context, we follow on from the previous chapters in resolving correspondences and in establishing a vocabulary of change to then ask: *Would such a vocabulary lend itself to higher level analysis of the information being relayed by the changes? For example, as shown on the face?*

5.1 Capturing Lower Resolution Video Rate 4D data

To enable us to further investigate the nature of complex surface deformation we use an enhancement to our existing stereo system - in the form of the commercial system as shown in Figure 5.1 below. This allows us for the first time to truly engage with dense temporal data - at the cost of slightly less high-resolution spatial reconstruction. The system works on exactly on the same principal as the basic "static" $3D$ approach introduced in Chapter 3 where each frame is independently produced on the basis of stereo correspondences. The main obstacles are primarily hardware driven in order to ensure adequate synchronisation and buffering during the capture, governed by the speed with which data can be stored to disk.

Figure 5.1: $4D$ Dimensional Imaging^{TM}Capture System.

The system is constructed of 3 Pulnix $1.4\,Megapixel$ cameras - two of which are black and white (mounted top and bottom) with the third central one capturing colour. All the cameras are synchronised via their respective frame-grabbers, which also allow high-speed buffering of images up to a maximum of approximately 300 frames. The camera can be selectively programmed to focus on smaller regions of interest (ROI) to gain faster speed - but at the cost of less detail in the reconstruction. We consequently use the cameras as full resolution ($1040 \times 1392\,pixels$) for each expression at $25 f.p.s$ - resulting in around $30-40$ frames of data for a duration of $1.5-2$ seconds.

The system is reliant on using an external studio lamp system in order to ensure constant good exposure (we cannot employ flash lighting as this would require rapid recharging). This is done in balance with the smallest aperture then possible, in order to increase the depth of field, resulting in an effectual volume of approximately $10 \times 10 \times 10cm$ in which accurate reconstruction can be performed. The subject can be placed at around $70cm$ away in this configuration, following calibration using the same "dot" pattern target, in which the calibration falls within the range of $\approx 0.05cm$ RMS error.

One additional unique and noteworthy enhancement to the system is that when processing the reconstruction it can exploit the temporal coherency across multiple frames, and so reduce the effects of noise and the systematic "orange peel" artefacts. This is simply achieved by using a $4D$ Gaussian window to smooth the data including the temporal (as well as spatial) domain - as shown by the example in Figure 5.2.

Furthermore, the software that accompanies the system also provides a manual interface for the user to define a set of landmark points between two different data-sets. This then provides the initial registration transform for an implementation of the conformation process similar to that used in Mao *et al*. [ZSCA04]. Applying this to the registration of a generic mesh model onto captured data-sets allows us to exploit this tool so that all sequences share the same generic mesh (fitted to the first frame and tracked across all frames in the sequence).

5.2 Deformable Qualifiers

In Chapter 4, we introduced the concept of deriving a useful set of features to classify the underlying observed changes in curvature. This resulted in 15 descriptive classes for localised *types* of surface change, along with a single metric to capture the total *extent* of change. As relatively simple and straightforward as these can be derived, there are however effectively two major enhancements that could conceivably capture much more of the dynamics.

Firstly, we seek to enhance our feature descriptors, so that we can also describe how the local geometry of a surface can change over time relative to an external co-ordinate frame. In so doing, these would then quantify the *extrinsic* properties of the local surface area in question as embedded within an ambient space. The most useful such measurement would effectively summarise how much a surface can locally *rotate* in the tangent plane between subsequent time-steps. Furthermore, other measurements describe how the metric properties of the surface $z(u,v)$ itself can change, particularly how it can *expand* or *shear* around the local region in question. A summary of these three measurements is illustrated in Figure 5.3. These *qualifiers* (as we term them) can occur at the same time as the transitions that define type and extent, but - crucially - they do not alter the definition of type, which is a wholly *intrinsic* property defined by the alteration of curvature.

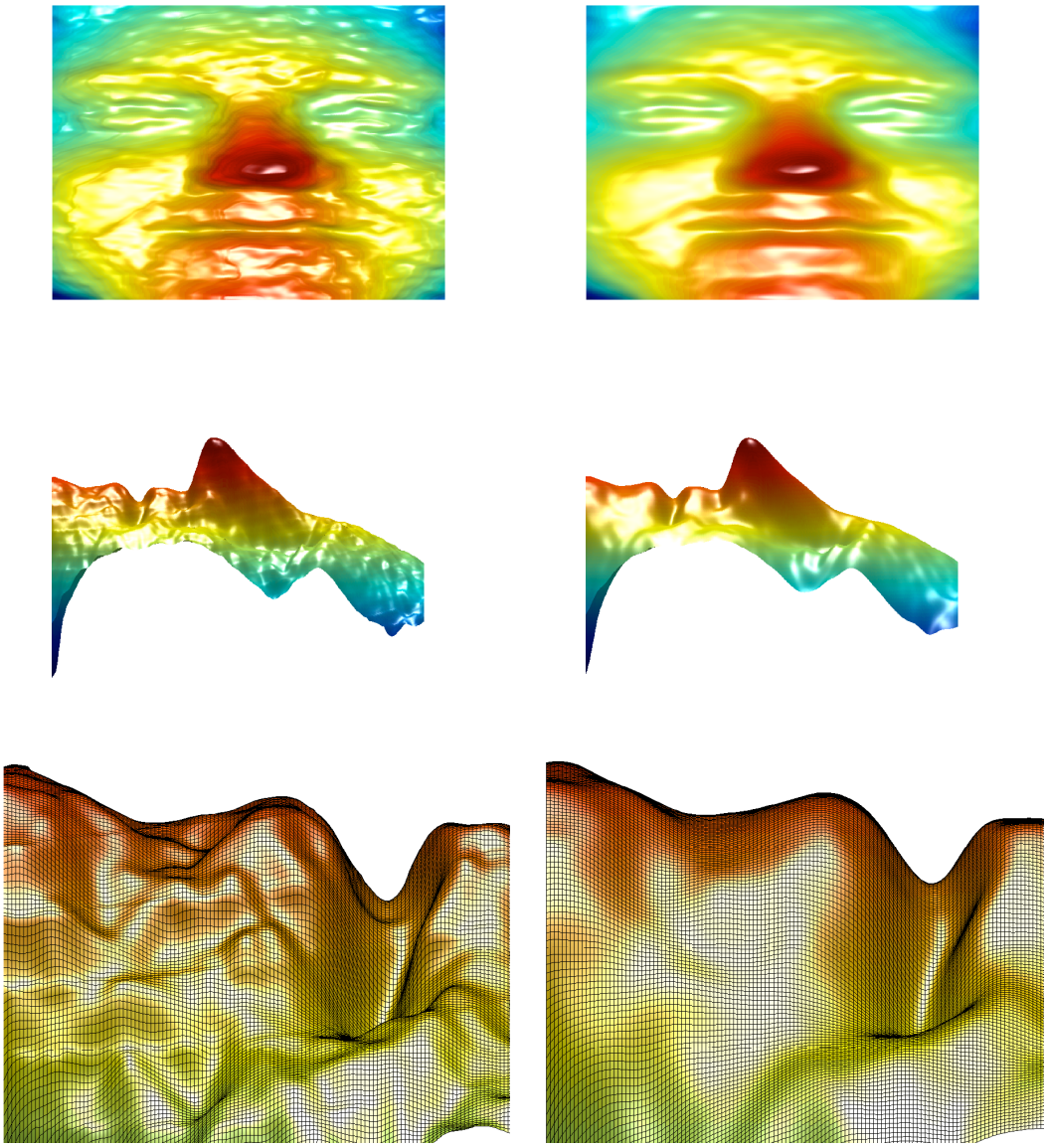


Figure 5.2: Reduction in surface artefacts with $4D$ smoothing.

Before smoothing *(left)* and after *(right)*. Due to the lower spatial resolution of the incoming data (compared to using higher resolution cameras), artefacts on the surface of the reconstruction can systematically appear - particularly in regions of lower quality texture. However, by applying a spatial-temporal Gaussian smoothing function these artefacts are effectively removed, without destroying too much of the true underlying surface detail. Here a $5 \times 5\, pixels$ Gaussian with $\sigma = 2.5$ in the temporal dimension, followed by a $15 \times 15\, pixels$ Gaussian across the spatial dimensions. The colour in the figure only indicates relative height for clarity.

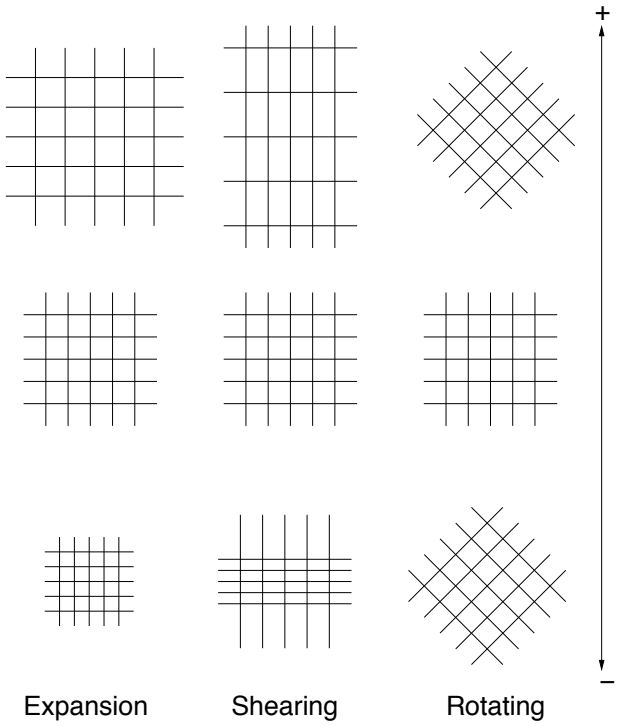


Figure 5.3: The three types of extrinsic surface qualifiers: expansion, shearing and rotation

Each of these is based on changes to the underlying orthogonal co-ordinate system for the surface, which may cause it to alter in both positive and negative extents for expansion and rotation. Notice however that the rotation can be ambiguous, unless based on a nominated direction perpendicular to the surface normal at the point of rotation. The case of shearing forms a specialisation of expansion where the ratio of spacing along one aspect only is increased or decreased.

Secondly, we note that up to now we have only considered changes from one instantaneous frame to the next. However, now that we come to consider sequences of longer duration, it is conceivable that any number of different changes may occur around the localised surface patch - particularly during the cumulative effects of an entire sequence. This could be represented by a summation of the total change, or as the representation of the change as a trajectory in the space of deformation, or alternatively as the integration of the various different types. Ultimately, the calculation of such values could also accommodate for the speed in which the transitions are made.

In the remainder of this section, we propose how these properties (along with the existing extent and type) can be calculated within a single mathematical procedure based on decomposition of the first and second fundamental forms.

5.2.1 Review of the Fundamental Forms

Just as a 2D curve is determined locally by its curvature (first order) and torsion (second order), so in the 3D case, the shape of a surface can be completely determined by its first and second fundamental forms. For a given parametric surface patch $z(u,v)$, the *first fundamental form* is expressed as a matrix:

$$\mathrm{I} = \begin{bmatrix} E & F \\ F & G \end{bmatrix} \tag{5.1}$$

where:

$$\begin{aligned} E &= \vec{z_u} \cdot \vec{z_u} \\ F &= \vec{z_u} \cdot \vec{z_v} \\ G &= \vec{z_v} \cdot \vec{z_v} \end{aligned} \tag{5.2}$$

with $\vec{z_u}$ and $\vec{z_v}$ equal to the gradient vectors at the nominated point p on the surface $z(u,v)$ such that $\vec{z_u} = \frac{\delta x}{\delta u}\vec{i}$ and $\vec{z_v} = \frac{\delta x}{\delta v}\vec{j}$ as computed at p.

From this it follows that the *second fundamental form* expressed as a matrix is:

$$\mathrm{II} = \begin{bmatrix} L & M \\ M & N \end{bmatrix} \tag{5.3}$$

where similarly:

$$\begin{aligned} L &= \vec{z_{uu}} \cdot \vec{n} \\ M &= \vec{z_{uv}} \cdot \vec{n} \\ N &= \vec{z_{vv}} \cdot \vec{n} \end{aligned} \tag{5.4}$$

with $\vec{z_{uu}}$, $\vec{z_{uv}}$ and $\vec{z_{vv}}$ equal to vectors describing the second derivatives of the surface gradient at the point p on the surface $z(u,v)$. The unit normal n at that point is derived from:

$$\vec{n} = \frac{\vec{z_u} \times \vec{z_v}}{\|\vec{z_u} \times \vec{z_v}\|}. \tag{5.5}$$

Together, the fundamental forms serve to capture both *intrinsic* and *extrinsic* properties of the surface. There exists a third fundamental form (III), but it provides no additional information, since it is expressed solely in terms of the first and second forms. However, it should also be noted that one of the most powerful surface relationships can be determined by combining the first and second fundamental forms as the *shape operator* $S = \mathrm{II}^{-1}\mathrm{I}$. This represents the effectual mapping from each point on the surface to the equivalent point on the Gaussian sphere.

5.2.2 Decomposition of Change

The shape operator itself cannot however reveal directly any temporal variations. For our purposes we are interested in how the localised changes to the fundamental forms over time express and relate to particular types of deformation. Suppose we have a small deformation of the surface parameterized as:

$$z(u,v) = z'(u,v) + \delta z(u,v) \tag{5.6}$$

representing the time derivative change in the surface over period δ. Then, in matrix form this can be represented as perturbations to the fundamental forms:

$$\begin{aligned} \mathrm{I} &= \mathrm{I}' + A \\ \mathrm{II} &= \mathrm{II}' + B \end{aligned} . \tag{5.7}$$

For a given instant, we can derive the principal curvatures κ_1 and κ_2 as the relative eigenvalues of II with corresponding eigenvectors e_1 and e_2.

From these, let P be the 2×2 matrix with the principal direction vectors e_1 and e_2 as columns. Similarly, let D be the diagonal 2×2 matrix with the principal curvatures κ_1 and κ_2 as its elements. These can be extracted directly as the two eigenvectors and eigenvalues from II.

Given, the previous instant's P and D matrices, and the changes to the fundamental forms A and B we then construct a matrix representing how the entire surface varies over time: [1]

$$Q = D' + P^{-1}BP'A. \tag{5.8}$$

The decomposition of the matrix Q onto a set of orthogonal basis matrices can then be performed (nominally through a Least Squares solution) as follows:

$$Q = \varepsilon \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} + \beta \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} + \gamma \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} + \rho \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} \tag{5.9}$$

to reveal as follows:

- ε- The uniform expansion of the surface in all directions.
- σ- The shearing of the surface, where $\sigma = |\beta + i\gamma|$.
- ρ- The rotation of the principal axes on the surface.

The type T and extent E can also be calculated in this manner from the eigen-decomposition of B to return the difference in principal curvature.

[1] This more elegant solution was suggested to us by Toby Bailey, School of Mathematics, University of Edinburgh. Alternatively, the metric expansion and shear of the surface can be calculated from the simple relative change in the underlying surface along the basis of the principal directions. The differential change in rotation can also be calculated directly from successive analysis of the principal directions.

5.2.3 Application to a Sequence

To apply this scheme to describing the changes occurring at the point p on the surface, we seek a total description of change, normalised by the duration of the sequence. The simplest metric is the displacement δ_p of the point over time, as it travels from its location on the surface at the start ($t = 1$), to the end of the sequence ($t = \ell$). However, this must accommodate the rigid transform of all the points as the entire surface moves, by correction of the mean displacement $\mu_\delta = \frac{1}{|S|}\sum_{p\in S}\delta_p$ of the set S of all surface points:

$$\hat{\delta_p} = \delta_p - \mu_\delta. \tag{5.10}$$

We can also perform summation of the **absolute values** of the three surface descriptors defined above to record the total expansion, shear and rotation of the surface at each point.

$$\begin{aligned} \varepsilon_p &= \sum_{t=1}^{\ell} |\varepsilon_t| \\ \sigma_p &= \sum_{t=1}^{\ell} |\sigma_t|. \\ \rho_p &= \sum_{t=1}^{\ell} |\rho_t| \end{aligned} \tag{5.11}$$

Notice that in the case of expansion, taking the norm does not accommodate for a negative value representing *retraction* - i.e. the surface undergoes a diminishing collapse towards a singular point. Such an ambiguity can also occur with regard to shearing (as an ever increasing ratio). Due to the omnidirectional calculation of rotation, this is always expressed as a positive value.

Calculation of the qualifier values for various synthetic test data sequences (of 10 frames duration) are shown below in Figure 5.7 in the case of a rotating cylinder (Figure 5.4), an expanding sphere (Figure 5.5) and a stretched ellipsoid (Figure 5.6). In each of these we also show how each of the surface regions in question can be fitted to a quadric in order to calculate the necessary terms, and how this can robustly (in the case of pure synthetic data) lead to calculation of the principal curvatures and directions. These sequences are textured to also provide example test input for the following experimental section.

More complex descriptions of the trajectory of any point as it moves through this deformation space are also possible, but require additional representation (e.g., splines). In this initial approach we look at recording the most representative change by simply building a feature vector at each local neighbourhood of a point that represents the cumulative change in each one of the three qualifiers.

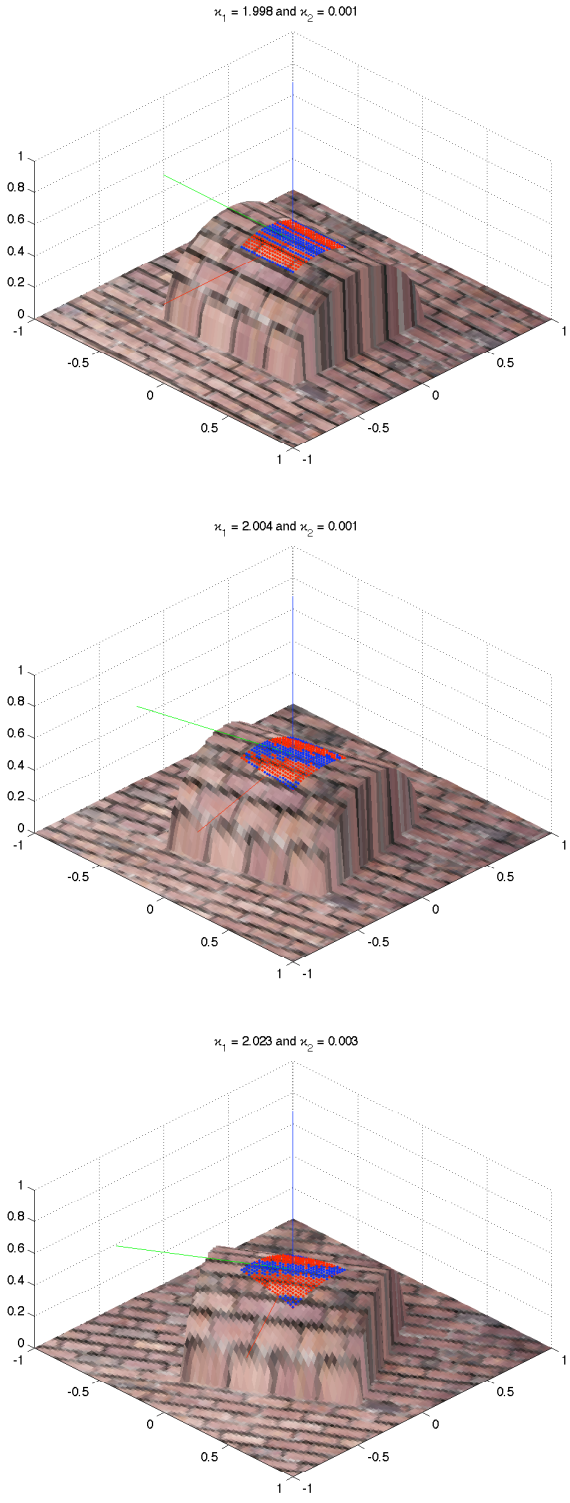


Figure 5.4: Synthetic data sequences for *rotating cylinder*.

Original surface 10×10 *pixels* patch *(red)* fitted to quadric *(blue)* with average error 0.003 MSE. Rotation between individual frames of $\pi/60$ (approximately 3 degrees). Note principal directions (red/green axis located at centre of patch) rotating in accordance with motion. Principal curvature of approximately 2.0 is consistent for cylinder of radius 0.5.

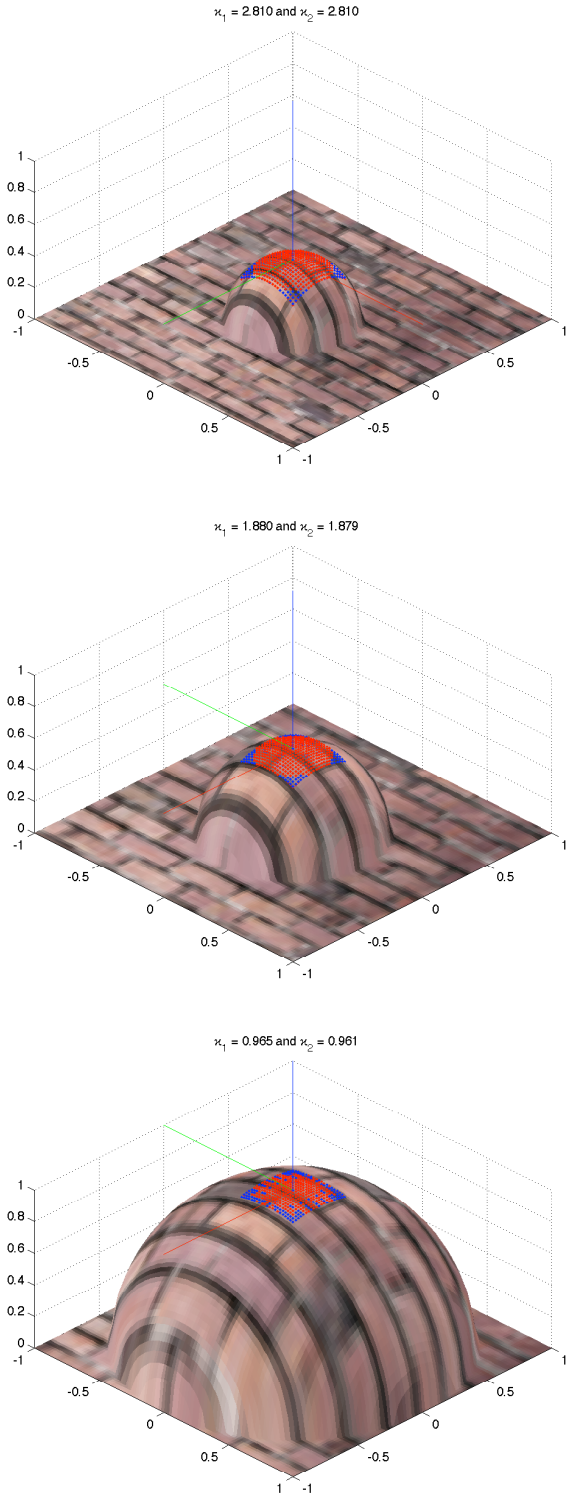


Figure 5.5: Synthetic data sequences for *expanding sphere*.

Original surface 10×10 *pixels* patch *(red)* fitted to quadric *(blue)* with average error 0.0043 MSE. Displacement between vertices in successive frames of 0.07. Note principal directions are undefined at the umbilic point and thus have a tendency to "flip" as the sphere expands - due to minimal difference between their respective principal curvature values.

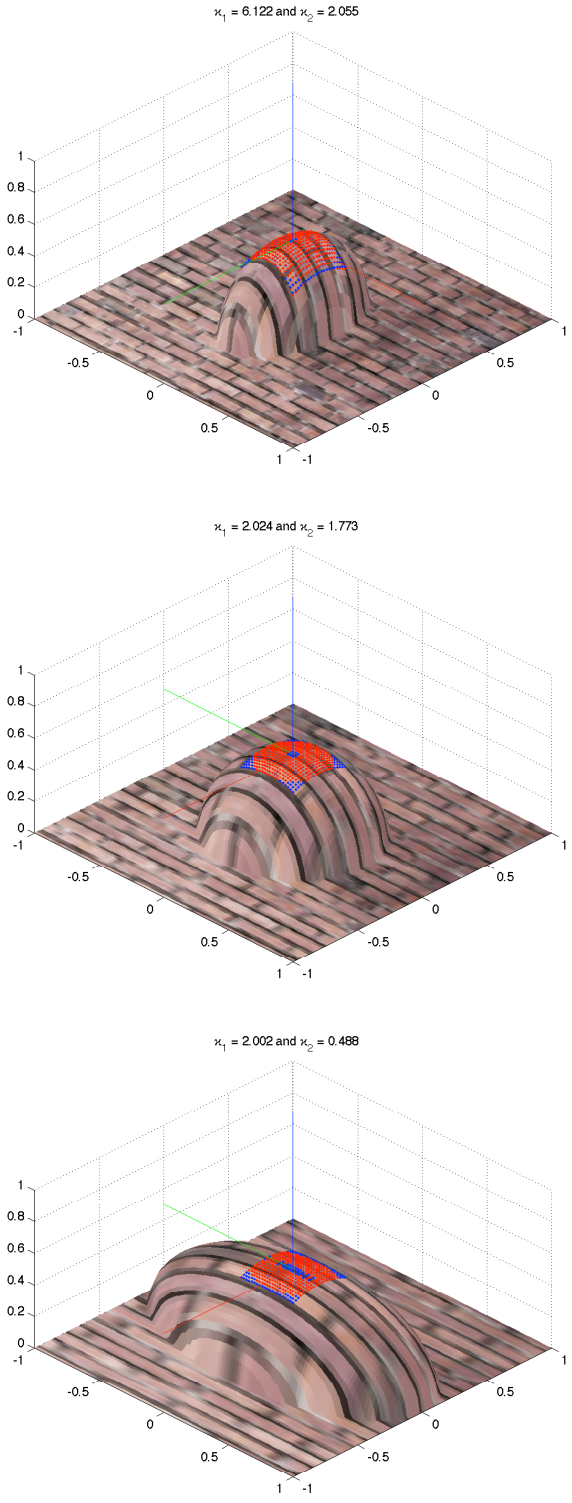
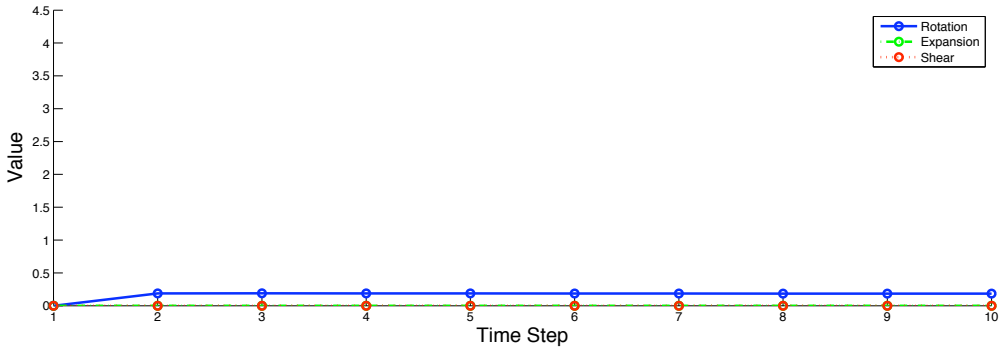
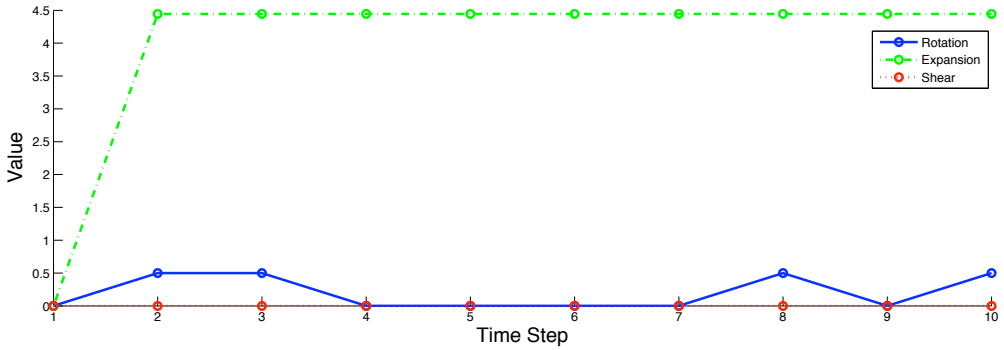


Figure 5.6: Synthetic data sequences for *stretching ellipsoid*.

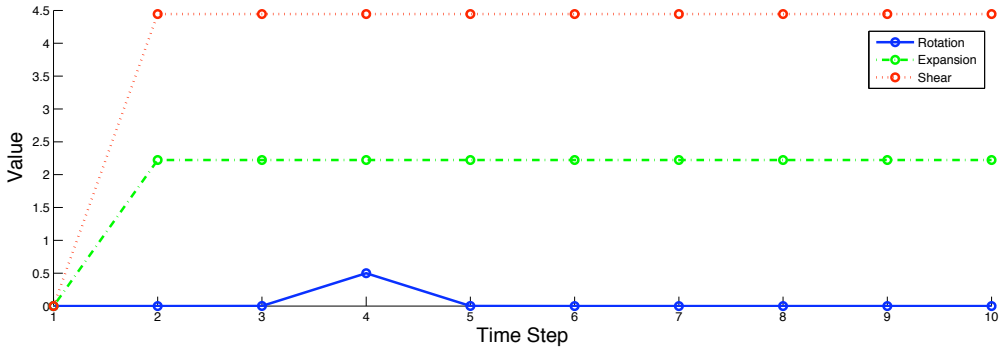
Original surface 10×10 *pixels* patch *(red)* fitted to quadric *(blue)* with average error 0.007 MSE. Displacement between vertices in the u direction between successive frames of 0.07. Notice in this case the principal directions do change direction mid-way through the sequence (at frame 4) in response to the new dominant principal curvature.



(a)



(b)



(c)

Figure 5.7: Changes in qualifier values over time for pure synthetic sequences: rotating cylinder (a), expanding sphere (b), and stretched ellipsoid (c).

Calculated over 10 frame sequence as shown in previous Figures 5.4, 5.5 and 5.6. Cylinder correctly represents only constant rotation, with zero shear and expansion. Sphere undergoes only constant expansion, with occasional orthogonal “flips” in terms of the principal directions and zero shearing. Ellipsoid has zero rotation (apart from one flip corresponding to change in dominant direction) but constant shearing alongside proportional expansion which is ambiguous along one principal direction.

5.3 Improving Tracking with Global Warping

One of the most crucial issues involving the use of flow guided temporal registration is in accommodating for extremely large and rapid displacements. This is effectively controlled by the window, or aperture, considered around each surface point in question. However, as we have seen in Chapter 3, this *local* treatment of surface motion can have particular problems - especially in the absence of enough information to resolve the local ambiguity (in the form of the aperture problem). Irrespective of modality - intensity, colour, depth - only point features can provide enough information to correctly resolve the correct motion estimation. In general, the flow based techniques we have so far considered are accurate for such features, but are relatively *sparse* and are limited to providing estimates only for the local neighbourhood they consider.

An important distinction with other approaches to optical flow (such as Horn and Schunk [HS81]) is to sacrifice some degree of accuracy in order to provide up to 100% density of estimation, by employing some form of energy preserving *regularisation* term or assumption. The survey paper of Barron *et al*. [BFB94] directly compares these approaches. Indeed, some of the most recent work in optical flow - such as that of Brox *et al*. [BBPW04] has made important improvements to the overall accuracy by using a combination of these terms and assumptions (both intensity and gradient constancy along with spatial-temporal regularisation, within a coarse-to-fine warping strategy).

5.3.1 Deriving KLT Features

This leads us to the idea of instead focusing on the question of selecting a robust collection of features in the data that are reliably tracked, even over large scale motion. If such a selective collection of points is relatively sparse, these features can in turn be used to define a *global* warp over the entire rest of the surface. The *KLT tracker* first introduced by Lucas and Kanade [LK81] and since refined by Tomasi and Kanade [TK91] (hence the name: KanadeLucasTomasi tracker) is one such means of reliably detecting good features between subsequent frames. This was originally framed within the context of an image registration method, to discover the suitable set of parametrised warps (displacements plus some noise) that lead to minimisation of the error residue between two regions.

An extension of this idea is not just to perform registration, but to extend it to find good features to track over longer periods of time. A good feature is defined (as with localised optical flow) as a region with high texture variation in both x and y directions - such as a corner point - which can be tracked well in a precise mathematical sense. This can be formulated by first determining a window size to define the size of features to consider, within which the pixel partial derivatives of the intensity function $I(x,y)$ are considered as defined by the Hessian matrix:

$$H(I) = \begin{bmatrix} \left(\frac{\partial I}{\partial x}\right)^2 & \frac{\partial I^2}{\partial x \partial y} \\ \frac{\partial I^2}{\partial x \partial y} & \left(\frac{\partial I}{\partial y}\right)^2 \end{bmatrix}. \tag{5.12}$$

The criteria for a good feature are that the matrix H is well conditioned, and that the signal does not encompass too much noise. The eigenvalues of the matrix (for an acceptable feature window) must satisfy $min(\lambda_1, \lambda_2) > \lambda_{min}$, where the threshold λ_{min} can usually be determined halfway between the upper and lower bounds determined by sampling from uniform and extremely textured surfaces in the image. To quote [TK91]: *"In practice, when the smaller eigenvalue is sufficiently large to meet the noise criterion, the matrix ... is usually also well conditioned. This is due to the fact that the intensity variations in a window are bounded by the maximum allowable pixel value, so that the greater eigenvalue cannot be arbitrarily large."*

Tracking using these selected windows can then proceed, by considering how the windows may change between subsequent frames (i.e. not just simple rigid translation if the surface is increasingly slanted, or warped). This is in essence handled by *residue monitoring* - if a matching window differs, such that the discrepancy error cannot be accommodated by pure displacement (and some noise), then the feature is dropped. More detail of this approach is clearly presented in [TK91], and an example of it working is shown in Figure 5.8.

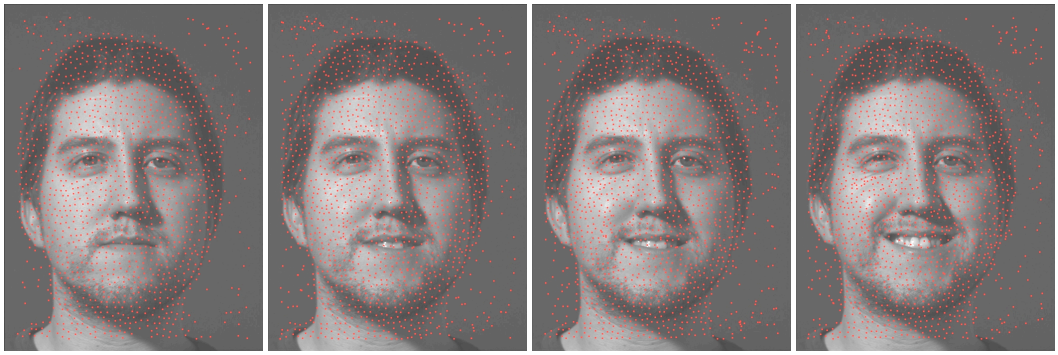


Figure 5.8: Example KLT detected and tracked points on a face sequence.

Between $1420 - 1764$ points were tracked in this sequence, with approximately 100 points replaced each frame to accommodate those lost via occlusion or changing surface appearance (i.e. specularities, shadowing). However, in each frame, a new set of features can be generated to replace those lost. In particular, the generation of new points around the region of the mouth is thus able to accommodate for motion when smiling. Using the resulting displacements between successive frames enables us to the guide the warp interpolant between the instances of the surface represented as a Thin Plate Spline.

5.3.2 Warping via a Thin Plate Spline

We then employ the resulting tracked features to guide and generate a *dense warp field* describing the continuous motion of every other point. The idea of using these features to guide a warping function between the two frames is not necessarily new, and indeed has recently been applied to other work on face tracking. For example, [MTL+06] combined a KLT tracker with a modified Active Shape Model in order to perform 3*D* face tracking.

However, in our work we do not aim for any explicit model (at this stage) for explaining/tracking the face. We instead simply wish to generalise a smooth and dense deformation field for the entire surface - in order to complement and enhance additional estimates of surface motion. To this end we consider the use of a *Thin Plate Spline* (TPS) to analytically define the entire surface in question. This technique has the powerful capability of expressing a realistic property of continuous surfaces as the minimal bending energy as guided by a set of control points. As a physical analogy, consider the TPS as a sheet of metal that is pushed down orthogonally over an armature of control points arranged at different heights above the plane. More formally, the TPS is defined by a set of radial basis functions:

$$U(r) = r^2 \ln r^2 \tag{5.13}$$

where r is radial distance $\sqrt{x^2+y^2}$ from the control point (x,y). The solution seeks to minimise the "bending energy" E (as the integral of the squares of the second derivatives) for each point $f(x,y) \in \mathbb{R}^2$ on the surface:

$$E[f(x,y)] = \int\int_{\mathbb{R}^2} \left(f_{xx}^2 + 2f_{xy}^2 + f_{yy}^2\right) dxdy. \tag{5.14}$$

This is achieved in the context of a regularisation term σ, specifying the exactness with which the interpolant need pass through the control points. The seminal work of Bookstein [Boo89] introduced the concept of defining the warps between two instances of a TPS. This can be performed by fitting to the *displacements d* between matched sets of points on two instances of the surface. This in turn defines the warp as the parameter a (defining the central co-ordinate frame of the TPS) and weights w associated with each control point, which is estimated via a least-squares solution:

$$\begin{bmatrix} K & P \\ P^{\mathsf{T}} & O \end{bmatrix} \begin{bmatrix} w \\ a \end{bmatrix} = \begin{bmatrix} d \\ o \end{bmatrix} \tag{5.15}$$

where K is the square matrix of distances between all p control points, modulated by the regularisation term σ as $K_{ij} = U(\|(cx_i, cy_i) - (cx_j, cy_j)\|) + I\sigma$, P is the $p \times 3$ matrix of $[1, cx_i, cy_i]$ control points, O is the square 3×3 zero matrix, and similarly o is $[0,0,0]^{\mathsf{T}}$.

Having fitted our TPS to the displacements defined by the KLT tracked points, it is then possible to define the new position of every point $\hat{z}(u,v)$ on the surface via the warp interpolant

based on the cumulative distances from the original p control points (cx_i, cy_i) to the point in question:

$$\hat{z}(u,v) = a_1 + a_2 u + a_3 v + \sum_{i=1}^{p} w_i U(\|(cx_i, cy_i) - (u,v)\|). \tag{5.16}$$

In summary: our global, dense deformation framework combines the estimate of motion derived from the KLT tracker to create a smooth continuous TPS warp that can be applied to all points in a frame of data. Examples of this combined KLT/TPS working for the synthetic test data sequences presented earlier are shown in Figures 5.14 and 5.15 below.

5.4 Experiments: Describing Face Surface Motion

We now seek to apply our enhanced set of descriptors, along with our enhanced surface tracking approach, to the task of analysing, and potentially classifying, the complex dynamics and surface motion of the face. In considering multiple subjects we first conform the same generic mesh to all of the incoming data. The captured surfaces must then be registered/tracked over time, such that point-to-point correspondences at each vertex of the mesh are resolved. We seek to verify first if indeed useful tracking of the data is possible using the KLT/TPS approach, alongside that of our earlier simple flow based method. Next, accurate computation of the fundamental forms must be performed around the region of each vertex. From these, our final set of features are extracted and compiled into unique feature vectors that can then form the basis for analysis.

5.4.1 4D Data Acquisition and Conformation

We first captured our data using the $4D$ stereo system detailed earlier in Section 5.1. This consisted of 3 male subjects making 3 different expressions: *happy*, *surprise*, and *disgust*, over 3 separate instances. This then formed a total of 27 sequences of $30-40$ frames (approximately one and half seconds duration). We furthermore verified the expression relayed by each sequence by playing the images back as a movie (at $25 f.p.s.$) and asking 2 independent people to label it as one of the classifications. The output is calculated for the corresponding frames of depth-maps, constructed to 50% full size at a resolution at $520 \times 696\, pixels$, with an effective resolution of $1 pixel \approx 0.05 cm$. The data is also 4D smoothed with a Gaussian kernel (as described in Section 5.1). The original colour image data is also reduced to 50% to provide matching 1:1 pixel resolution.

In order to apply our set of features to a set of very different real surfaces, we required a single geometric representation that could accommodate variation and allow us to describe the motion of the surface over time. For this purpose we use a triangulated mesh, in which each vertex is moved in relation to the discrete frames of data. Our generic mesh is a modified, front only, version of the generic head model provided by [Pig99]. Of notable importance is the topology

of this mesh - it dictates the effective resolution at which we ultimately sample and track the surface. This allows us to capture a larger degree of motion in these areas and to accommodate greater change. The final generic mesh has a total 3579 vertexes and 7008 faces - as shown in Figure 5.9.

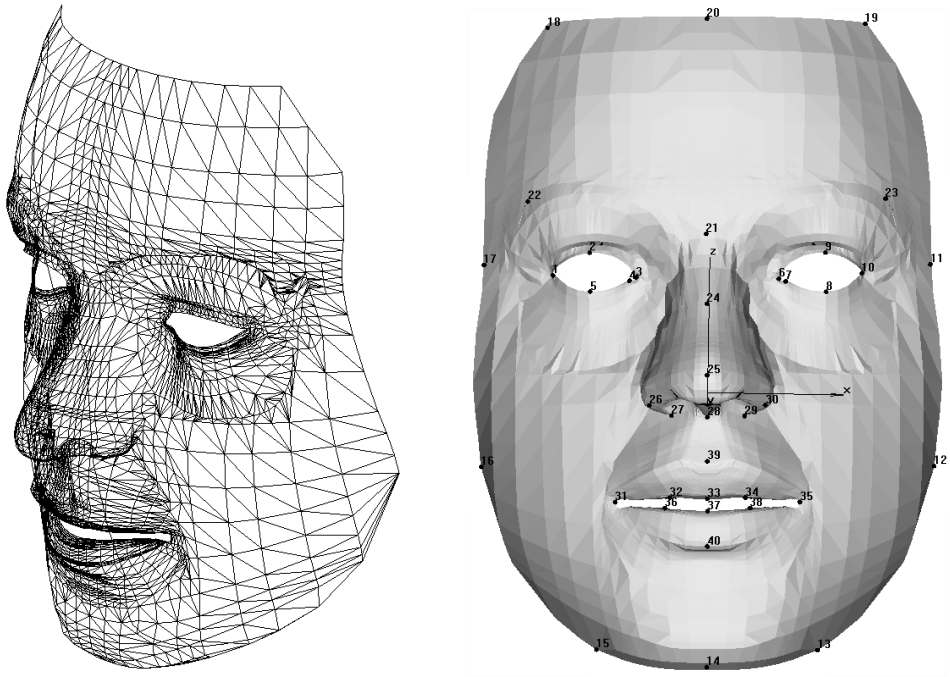


Figure 5.9: Generic mesh *(left)* and landmark points *(right)* used for conformation.

This mesh provides a useful topology in that it is dense in the regions where more motion is defined during expression. In particular the areas around the eyes, nose and mouth are almost 10 times denser in terms of vertices. Ideally, this allows a great deal of flexibility in tracking the motion we anticipate to be greater around these areas. The disadvantages are however in accommodating for unexpected motion outwith these regions, and in having to very carefully select suitable landmarks for the initial conformation. The 40 landmark points used for conformation are shown on the right. They are selected to provide the maximum level of constraint around the eye and mouth regions.

For initialisation, we require that the generic mesh is conformed to the first frame in the sequence. This process involves human interaction to define 40 landmark points (also shown in Figure 5.9) to guide a least squares minimisation to the scanned data, similar to the approach used in [JMS+04]. The scanned data is represented as dense reprojected mesh, and displayed alongside the generic mesh to allow this initial fitting to be calculated. Conformation then proceeds with results as shown in figure 5.10.

The results of this conformation for a single subject are shown in Figure 5.11. In general the process is reasonably robust on a qualitative level, but on closer assessment a number of

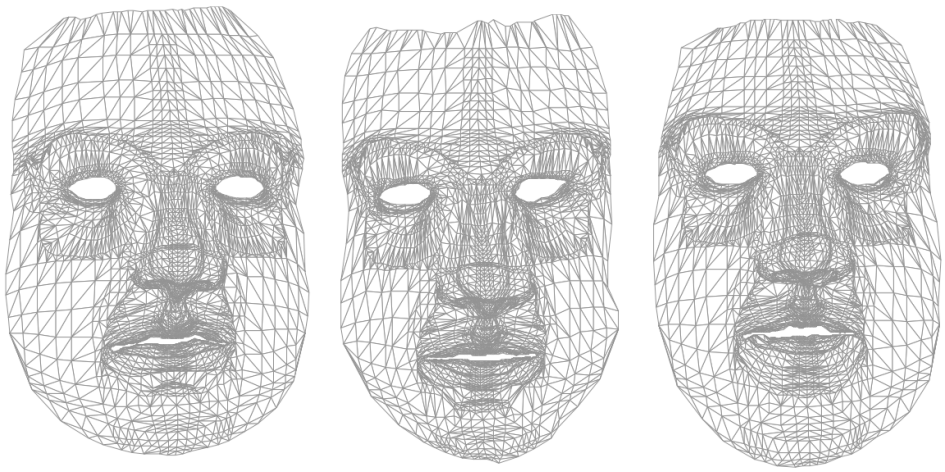


Figure 5.10: Generic mesh conformed to 3 different subjects.

In each case it is crucial to initialise the mesh so that the same vertices are successfully located in equivalent regions of the face. While this is successfully achieved for the larger, less dense mesh areas it can result in a degree of mesh compression and overlap in the denser regions around the eyes and mouth (depending on the relative scaling of data).

vertices are subtly misplaced - particularly in the denser regions around the eyes and mouth, which are regions that have especial significance for later tracking. This can in part be attributed to differences in the reconstruction (and the fact that the subject is not exactly in the same position each time). The biggest factor is however in the influence of the landmark points, which even on the same subject can not be absolutely guaranteed at the same location each time. This has implications for later comparison even between the expressions made by the same subject.

5.4.2 Calculation of Features

Given a mesh we then wish to calculate the cumulative shape change features at each vertex, in order to then compare the dynamics of each sequence. For each local neighbourhood around a vertex v on the mesh, we seek to create a parametric surface representation $x(u,v)$. This is done by projecting the vertex back onto the original depth-map data (to which it was conformed). A surface patch of size $n \times n$ is then selected and orientated such that the vertex point lies at the origin with the surface normal transformed to align on the z-axis $[0,0,1]$.

We then fit an extended quadric to the data, in a method again described in [MV96]. We are particularly concerned here again with robustness to noise - especially in the case of the necessary second order calculations presented below. As introduced in Chapter 4 the extended quadric, which can be alternatively represented in a parameterised form:

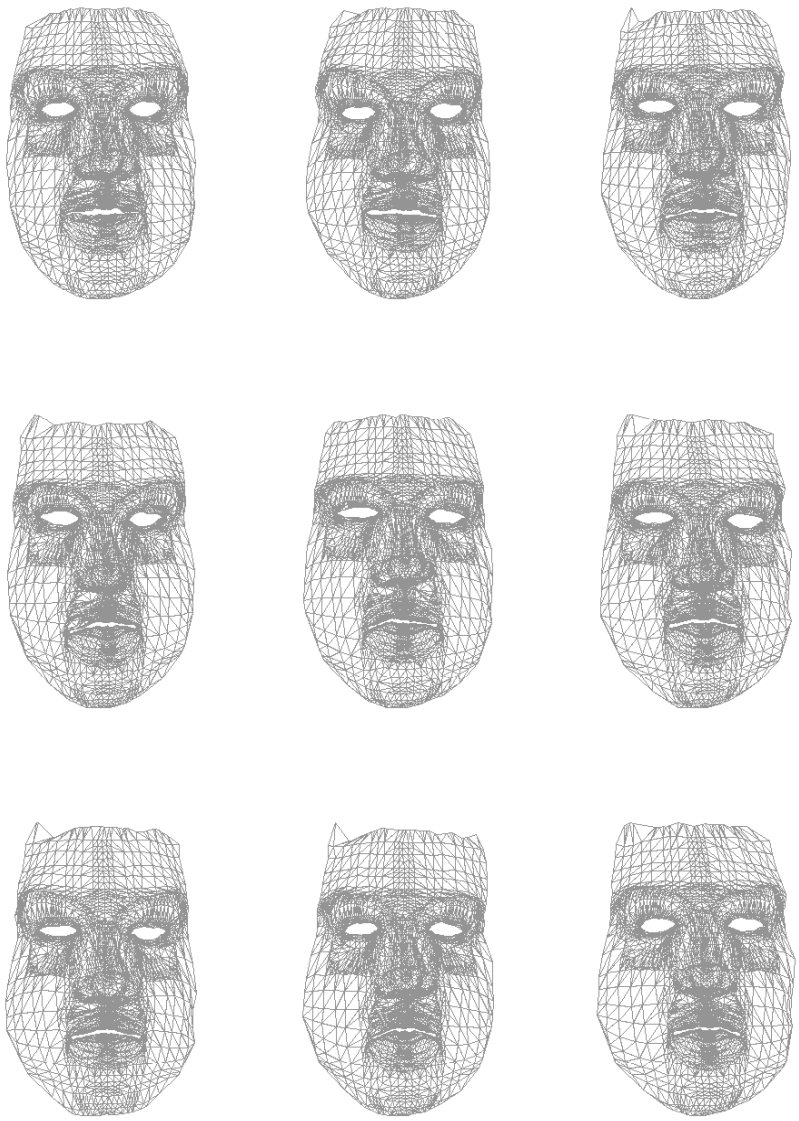


Figure 5.11: Generic mesh conformed to same subject 3 initial neutral expressions for 3 different sequences: happy *(top)*, surprise *(middle)* and disgust *(bottom)*.

Data representing the same face, held for the same neutral expression - yet subtle changes in the optimised mesh fittings are evident. Discounting the sometimes extreme displacements occurring on the periphery of the mesh (where the underlying stereo data starts to break down) - it is in the outlines of the mouth and eyes that show variation in the optimised placement of vertices to best conform to the given set of landmarks.

$$z(u,v) = au^2 + buv + cv^2 + du + ev. \tag{5.17}$$

We gain an initial estimate of the vertex normal coefficients d and e by calculating the average normal of the neighbouring 6 faces in the mesh. Following this, the principal quadric co-efficients (a,b,c) can be used to directly calculate the first and second order derivatives of the parametric surface:

$$\begin{aligned} \vec{z_u} &= [du, 0, 2au + bv]^\mathsf{T} \\ \vec{z_v} &= [0, dv, bu + 2cv]^\mathsf{T} \end{aligned} \tag{5.18}$$

and similarly:

$$\begin{aligned} \vec{z_{uu}} &= [0,0,2a]^\mathsf{T} \\ \vec{z_{uv}} &= [0,0,b]^\mathsf{T} \\ \vec{z_{vv}} &= [0,0,2c]^\mathsf{T} \end{aligned} \quad . \tag{5.19}$$

Notice that the underlying metric changes in u and v (terms du and dv above in Equation 5.18) are calculated by considering the mean motion of the 6 neighbouring vertices and their displacements across the surface (if any) projected onto the patch as it lies in the original 2D array. This may seem at odds with the vector representations of z_u and z_v which would expect unit directions that define the tangent plane (i.e. $[1,0,2au+bv]^\mathsf{T}$ and $[0,1,bu+2cv]^\mathsf{T}$). However, by introducing here the underlying resampling of the tangent plane, we can effectively incorporate these components here. The alternative is simply to use the positions of the neighbouring 6 vertices to guide the actual size of a regularly sampled patch (as the bounding box of the vertices projected back onto the original depth-map data).

In our data sequences, we empirically use a value of $n = 11$ to select a reasonable $n \times n$ patch size for the scale of our data. Constructing the second fundamental form II (as introduced in Section 5.2.1) it is then possible to derive the changes in surface features. A qualitative assessment of this process for each vertex is to then derive the underlying principal directions and principal curvature values located at each vertex - as shown in Figure 5.12.

From these quadric fittings, and the total displacement, we can then calculate the feature vector $[\sigma_p, \varepsilon_p, \rho_p, \delta_p]$ to represent the cumulative degree of shear, expansion, rotation and displacement at every mesh vertex p. However, it should be noted that the estimated principal directions even for static regions can change radically in response to the motion of the mesh as it varies over time. This "jitter" can be attributed to the initial estimate in surface normal varying in response to neighbouring vertex motion, and by the very subtle KLT/TPS drifting of the vertex point to neighbouring regions that have locally different underlying surface structure (either from true variation, or from remaining artefacts not entirely removed by smoothing). The unfortunate side effects of this "jittering" is that the calculation of rotation ρ is very much susceptible to miscalculation, as this is the qualifier most affected by the local surface geometry and

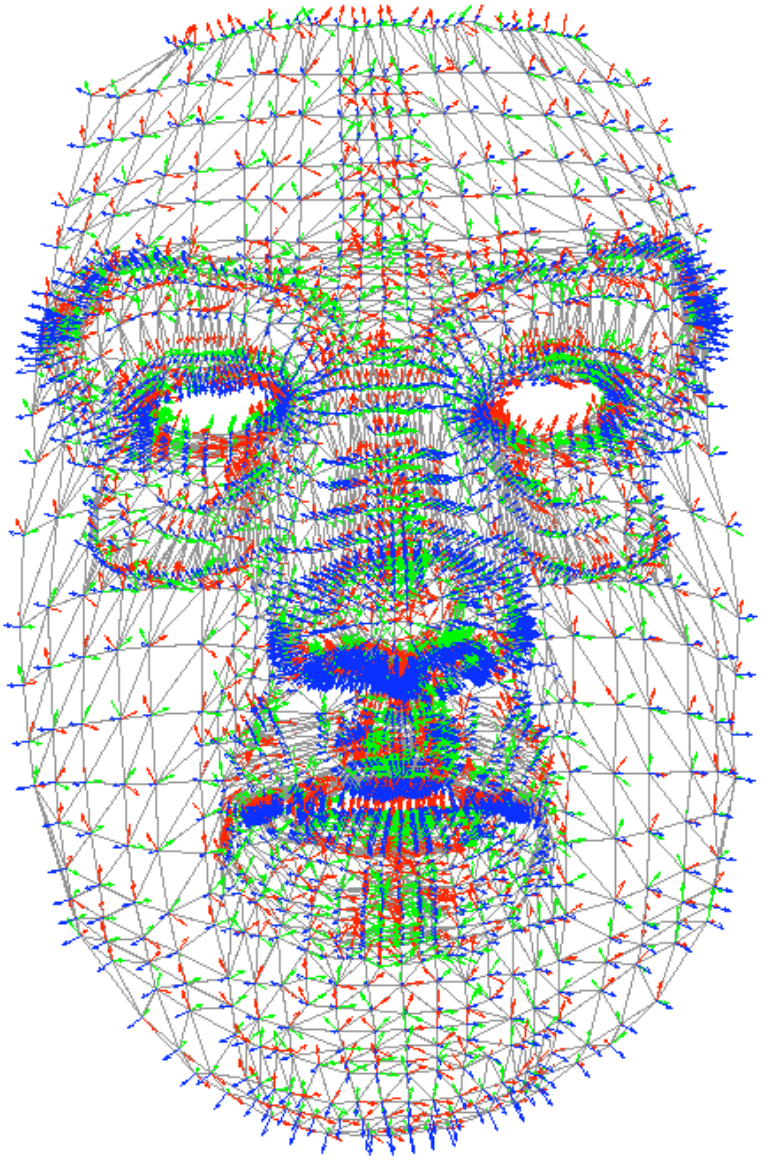


Figure 5.12: Quadric normal and principal directions estimated for each vertex

The normal directions *(blue)* at each vertex are initially estimated from the neighbouring mesh faces and then refined to the underlying surface patch. The principal directions shown as *(red)* and *(green)* are then calculated on the resulting quadric fitting. In this example the local variations across the fixed patch size (of $11 \times 11\, pixels$) do not account for the spacing in the topology of the mesh - but are only concerned with the immediate surface geometry.

ambiguity. Fundamentally, true rotation is also the feature most restricted to larger scale motion when considering the dynamics of the face.

5.4.3 Confirming Limitations with Synthetic Data

The core problem in helping to offset potential "jitter" and drift in the sequences is in guaranteeing that the computation of the feature vector *accurately* records the changes to the *same* underlying surface region. Before tracking the captured face data, we first seek to assess the accuracy with which the KLT/TPS tracker is capable of handling even the relatively simple synthetic data sequences (introduced in Section 5.2.3). In this experiment we consider the same 10 frame sequences for the rotating cylinder, expanding sphere and stretched ellipsoid - but are instead entirely reliant on the tracking of the mesh change based on the KLT tracking of the overlaid textures. This therefore differs from the purely synthetic approach used to generate the sequences initially, where the underlying surface $z(u,v)$ was re-sampled on the same grid at each time-step and (crucially) the centre point in the data did not alter.

The resulting tracking of the textures overlaid on the sequences as shown in Figure 5.13 indicate a good initial set of feature points. These are used to build the TPS warp then used to adjust the respective meshes shown in Figure 5.14 as the sequence progresses. Qualitatively these warps perform reasonable extrapolation of the point motion applied to the entire mesh, which is particularly impressive given the relative sparsity of the KLT points used.

However, the amount of drift in the central mesh vertex evident in Figure 5.14. Even for these simple sequences, over a relatively short time-frame, the distances exhibit quite significant motion. This has further implications for the accuracy with which we can safely say that the same local region around each vertex (and respective surface change) is the same point initially conformed to.

This then impacts the accuracy with which certain of the qualifiers around each vertex are then calculated. The values for the qualifier calculations around the central point highlighted in the test sequences are shown in Figure 5.15 (compare directly with 5.7). These values still mirror overall trends in rotation, expansion and shearing - since the local surface patches fitted are not actually affected so much by the shifting position of the vertex (for example, each point on the expanding sphere surface appears the same). However, it is the changes to the underlying mesh topology that serves to perturb the metric measurement of shear and expansion. The fact that the 6 neighbouring vertices (by which the metric changes in du and dv are calculated) form a skewed arrangement may further bias these calculations along alternating diagonals.

In conclusion, while this synthetic data appears relatively robust to tracking, it only represents an underlying surface exhibiting only one homogeneous type of deformation. Real surfaces (depending on the level of scale) can be expected to alter radically across neighbouring surface regions. In such circumstances the degree of vertex drift becomes increasingly influential on the

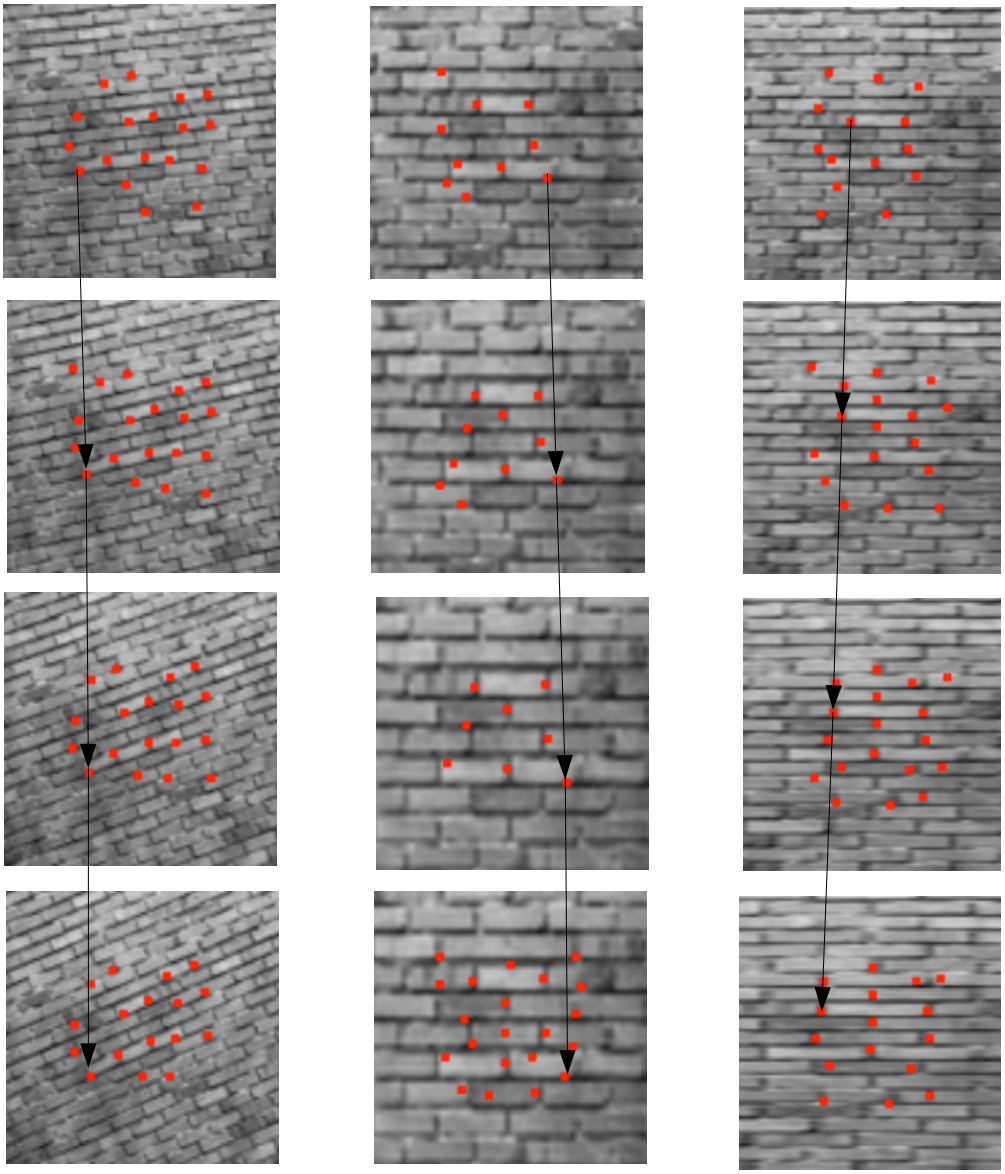


Figure 5.13: KLT tracking for synthetic cylinder *(left)*, sphere *(middle)* and ellipsoid *(right)*.

Shown are the textures generated by the synthetic sequences from Figures 5.4 to 5.6 with the tracked KLT features overlaid. Notice the retention of the central features, which are then replaced as necessary with more reliable tracker features if the displacement is too great (particularly evident in the case of rotation where features further out from the centre undergo larger displacement). The arrows indicate an example point retained throughout the sequence.

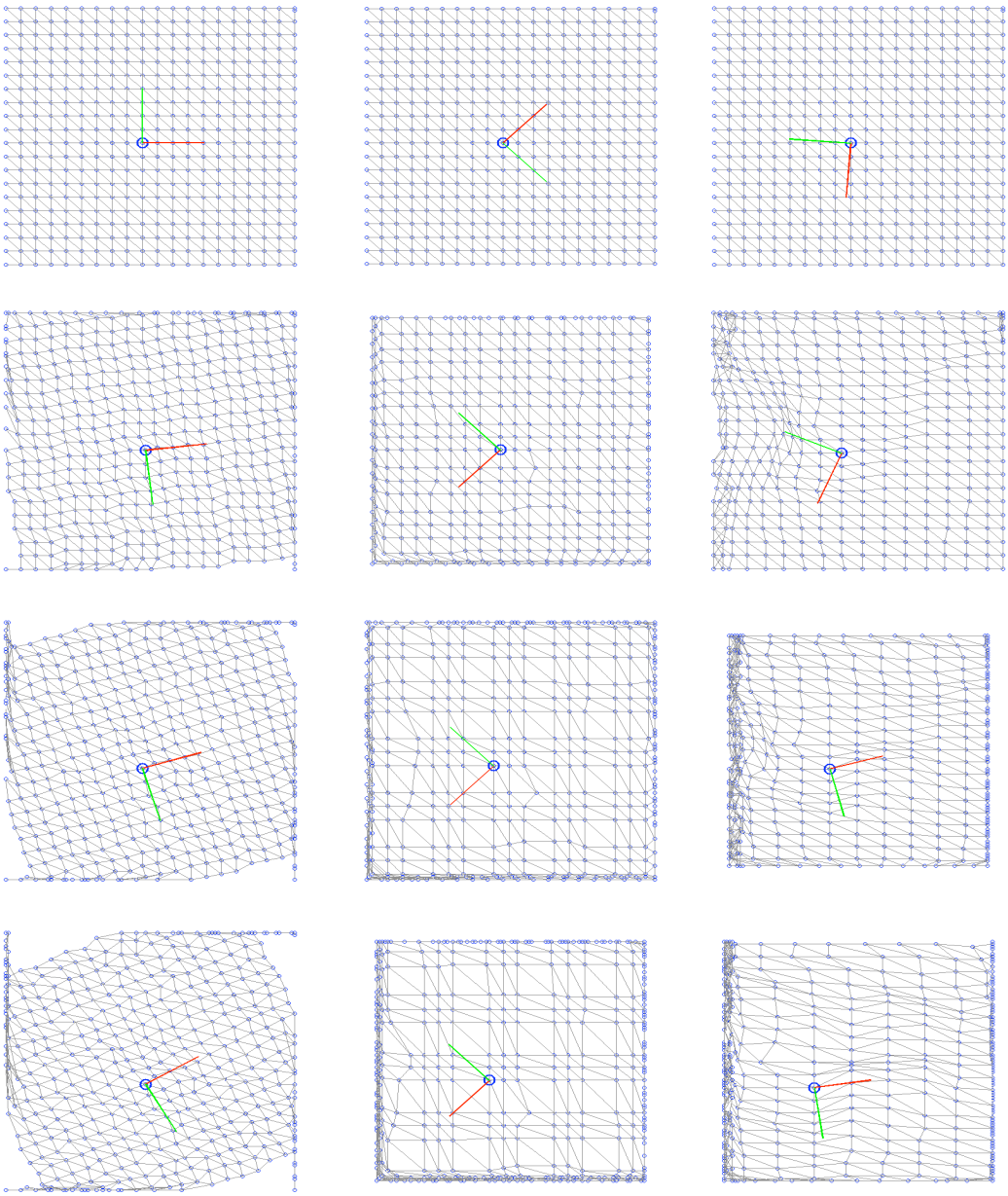


Figure 5.14: TPS warping for synthetic cylinder *(left)*, sphere *(middle)* and ellipsoid *(right)* .

The underlying mesh topology - shown here in top down view - of the data sequences is warped according to the TPS interpolant for all vertices, as driven by the KLT features shown in Figure 5.13. The initial central vertex is highlighted in all cases - along with the estimated principal directions. Of major importance here is the degree of drift shown by this central vertex as the sequence progresses. Notice the "dragging" of the mesh in the case of rotation, the compression of the mesh to the bottom left corner in the case of expansion, and similarly the compression of vertices to opposing sides in the case of stretching.

accuracy of the final result.

5.4.4 Real Data KLT/TPS Tracking

We now employ our KLT/TPS tracker to locate robust features between subsequent frames of data in the real face sequences. With the larger images, on average ≈ 1500 KLT features were maintained between frames (with ≈ 100 being replaced) - as shown earlier in Figure 5.8. The resulting TPS warp is then used to provide a $x(u,v)$ vector flow for every vertex in the generic mesh. The new position is then re-projected to ensure that the vertex continues to always lie on the surface of the original data. Vertices that have not already been moved by the flow estimation, are then moved by the KLT/TPS amount.

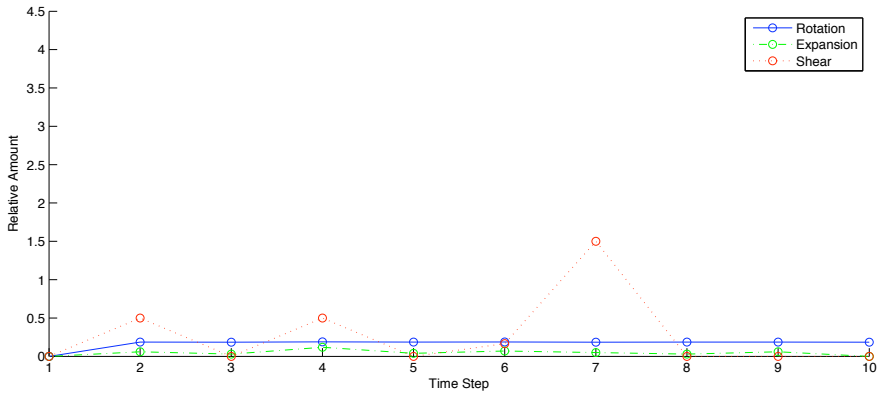
The immediate observation for these longer and more complex sequences is that they can rapidly suffer from vertex drift, as shown in Figure 5.16. This result, comparable with all other sequences in our data-set, effectively leads us to reject the usefulness of the KLT/TPS tracker for use with real data. Instead, we return to considering the use of an image based optical flow approach where the displacement vectors in (u,v) are used to determine vertex motion. Each new position is again re-projected back into $3D$ to guarantee that the vertex continues to lie on the surface of the original dense stereo data.

Example tracking using this flow based approach is shown in figure 5.17. This indicates an overall reasonably stable estimate of vertex motion, but which unsurprisingly breaks down around the region of the mouth. Although it may look correct, the actual vertices do not move in accordance with the true underlying motion (instead simply remaining fixed and "filling" the mouth as it opens). The flow based technique also fails to accommodate wherever more rapid motion occurs for certain sequences within the data-set.

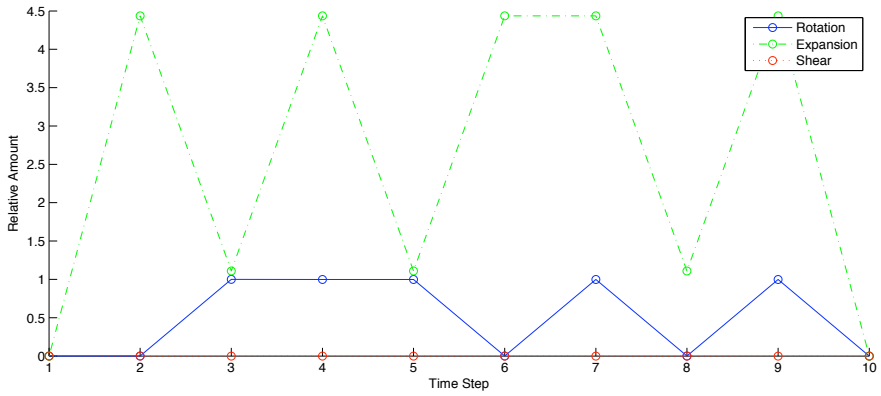
Consequently, the assessment of these real data sequences can only remain at best qualitative, especially given the complexity of motion, and reliability in the original fitting (i.e. initialisation) of the conformed mesh. The main issues can still occur around too large a displacement or occlusions, leading to vertex drift - as further discussed below in Section 5.5. However, by using this approach we are at least able to visualise the extraction of features around each vertex.

5.4.5 Analysing Faces and Expression

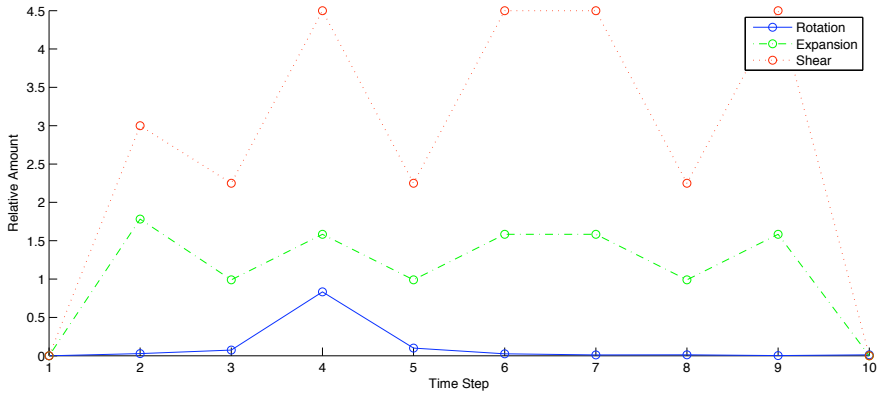
Having detailed the initial conformation, the calculation of the respective qualifiers, and the tracking of each vertex, we next proceed to visualising and analysing the extracted values over the entire sequence. In particular, we seek to establish if these qualifiers express any immediately *useful* information. By this, we mean that they can be applied to the task of revealing underlying surface dynamics, which in turn may enable classification of the patterns of variation on the surface of the face.



(a)



(b)



(c)

Figure 5.15: Changes in qualifier values over time using KLT/TPS warped sequences for rotating cylinder (a), expanding sphere (b), and stretched ellipsoid (c).

The same overall pattern expected for these sequences is generated. The cylinder indicates relatively constant rotation which can be attributed to the local similarity of all surface patches. Similarly, during the shear sequence, the one “flip” occurring at the transition of principal directions is correctly indicated. However, it is with regards to expansion and shearing qualifiers that the influence of the underlying mesh topology and the varying influence of the warp leads these values to oscillate.

Figure 5.16: KLT/TPS tracking drift over long sequence of real data.

This shows the 1^{st}, 20^{th} and last 40^{th} frames of original image data with the conformed mesh vertices overlaid in red. The actual displacement around the mouth is quite accurately resolved initially and well fixed, but the regions around the eyes rapidly move upward. Notice the overall degradation in the mesh quality on the last frame where it is difficult to make out the original topology as so many vertices have shifted relative to one another.

One of the simplest and most visible ways of immediately assessing the result of the optical flow tracking is to view motion between first and final vertex positions as cumulative displacement. Since we are tracking on the surface, this will effectively remove any global, rigid component of the motion, and highlight the actual deformation. This is further used (as described in 5.2.3) to compensate and normalise for changes in the other features. Examples, as shown in Figure 5.18, show the deformation of subject expressions following successful tracking using optical flow, projected back onto the initial conformed mesh. Each of these sequences were of 30 frame duration.

Despite these issues with the trackers, select sequences in the data-set can still reveal clear distinctions in the dynamics of the surface as shown in Figures 5.19 5.20 and 5.21. In these examples both displacement and expansion produce reasonable results - indicating the maximal degree of change occurring around the mouth region in the case of happy, the brows for surprise, and the side of the nose for disgust. Shear (in all cases) appears to mirror the inverse values of expansion, which can be explained as those regions where the differences in the underlying surface is close to zero in one direction. As previously discussed, the rapidly changing orientation of the surface normal at each vertex leads to rotation producing erratic and inconsistent values. Notice as well that these examples were drawn from sequences with the *same* subject - indicating again the variations in the conformed meshes (here shown at the final frame in each sequence).

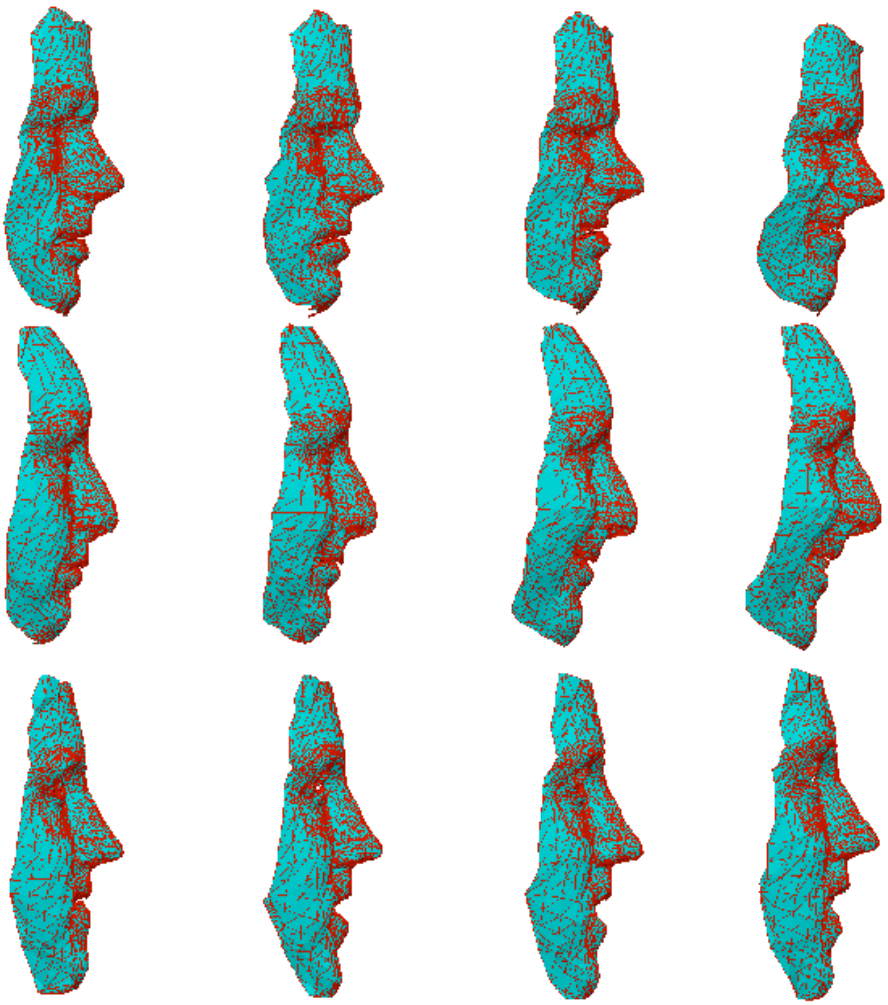


Figure 5.17: Tracked generic mesh for “disgust” *(top)*, “happy” *(middle)* and “surprise” *(bottom)*.

These are side on views for 3 different subjects making one of the prescribed expressions. Notice that in the case of large mouth motion, the flow estimate can lag, and discrepancies may occur where depth discontinues on the edge of the raw data can suddenly displace vertices. However, in these instances, the mesh is able to track and move in accordance with the motion that exemplifies the expression - but may still cumulatively decay in quality towards the end of the sequence.

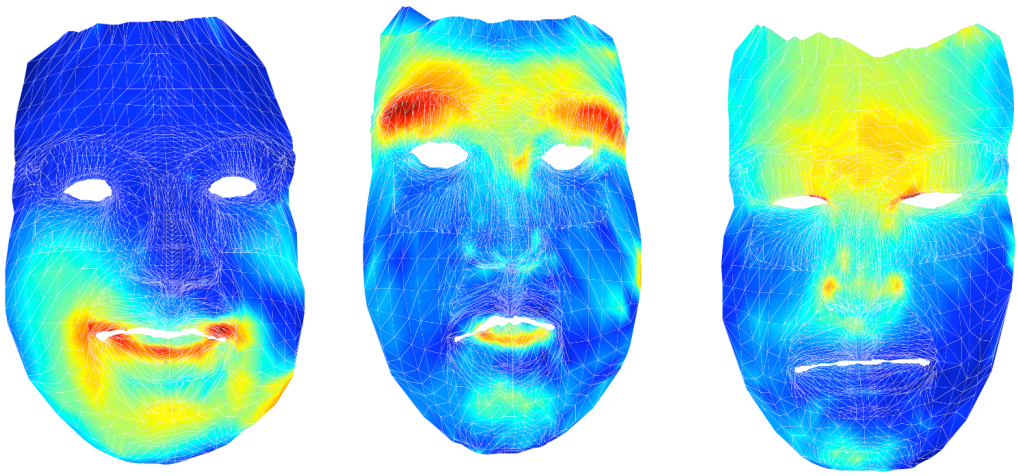


Figure 5.18: Total displacements for “happy” *(left)*, “surprise” *(middle)* and “disgust” *(right)*.

The coloration indicates the magnitude of cumulative displacement. Each of the three expressions show prototypical motion that intuitively matches expected human perception. For happiness the mouth and cheeks move in a characteristically strong response that dominates the face. For surprise the eyebrows travel upwards together and the mouth opens. For disgust the central brow region knits and the corners of the nose expand as the nostrils flare. Of additional interest is the significant degree of subtleties seemingly captured by the expressions - in particular for happiness which shows the asymmetrical motion generally associated with genuine expressions.

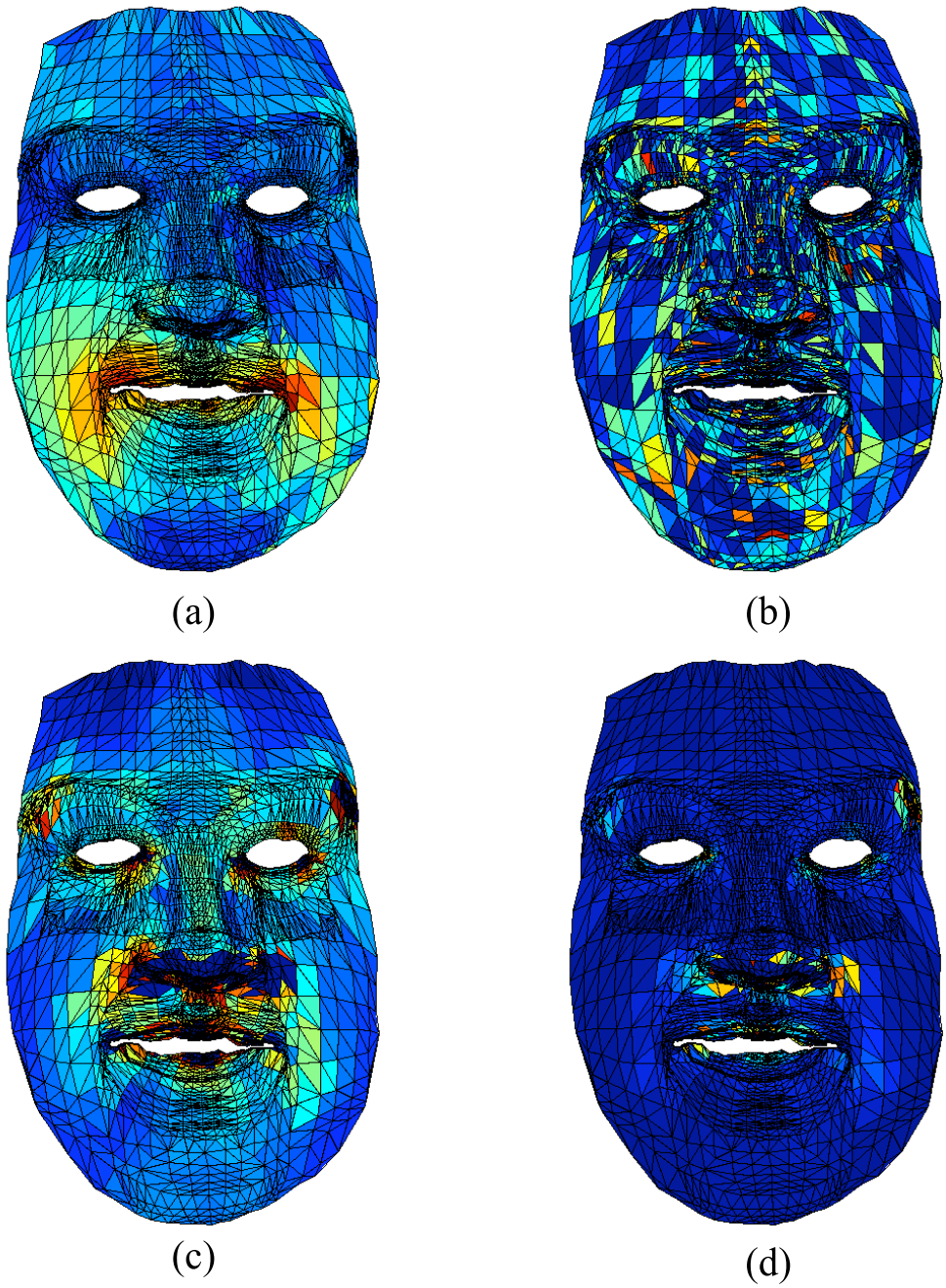


Figure 5.19: Qualifiers displacement *(a)*, rotation *(b)*, expansion *(c)*, and shear *(d)* extracted for a “smile” sequence.

Data is a 40 frame sequence with mesh vertices moved via optical flow guided technique. Relative values indicated by colour and projected back onto conformed mesh at final frame.

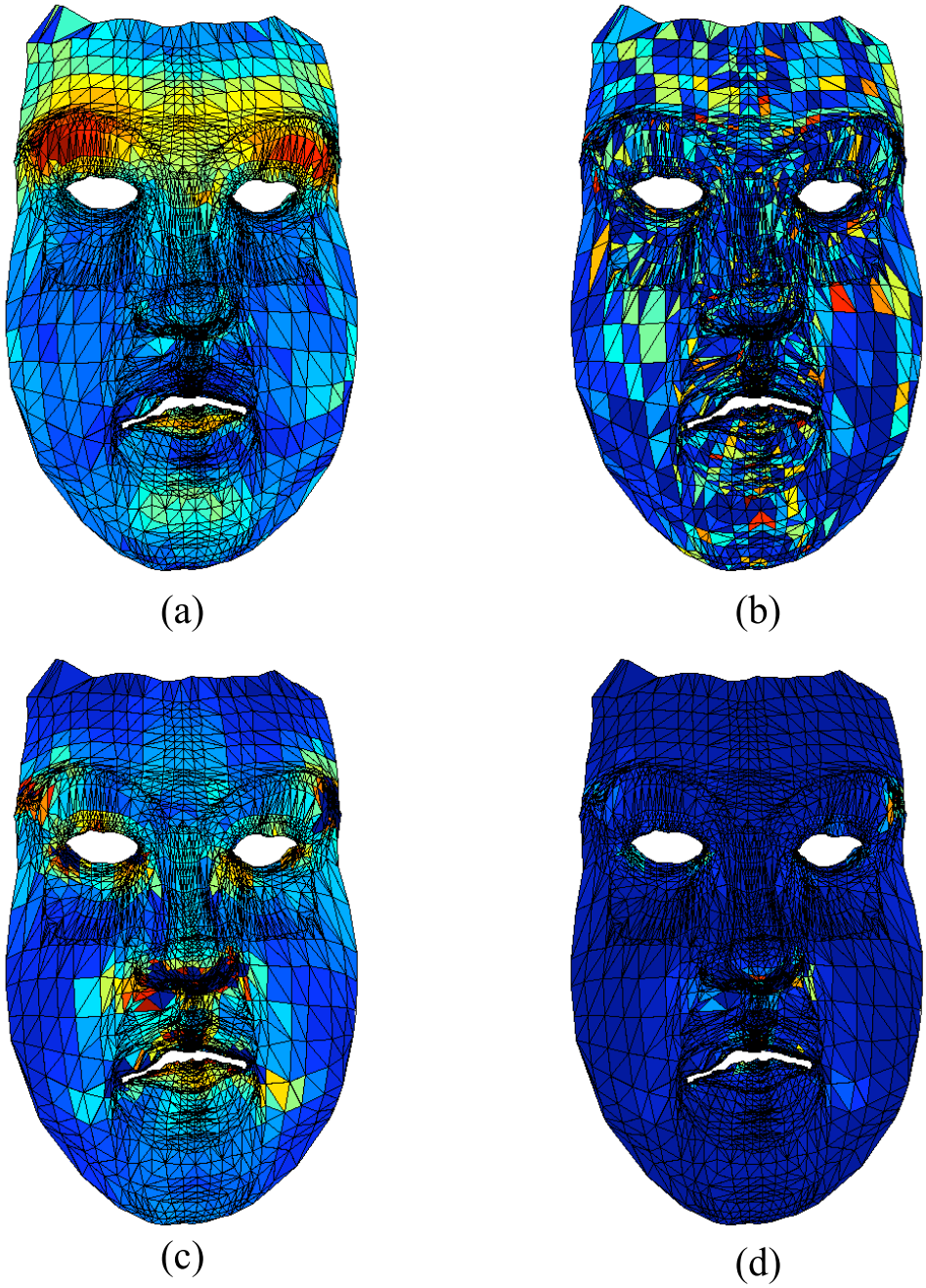


Figure 5.20: Qualifiers displacement *(a)*, rotation *(b)*, expansion *(c)*, and shear *(d)* extracted for a “surprise” sequence.

Data is a 37 frame sequence with mesh vertices moved via optical flow guided technique. Relative values indicated by colour and projected back onto conformed mesh at final frame.

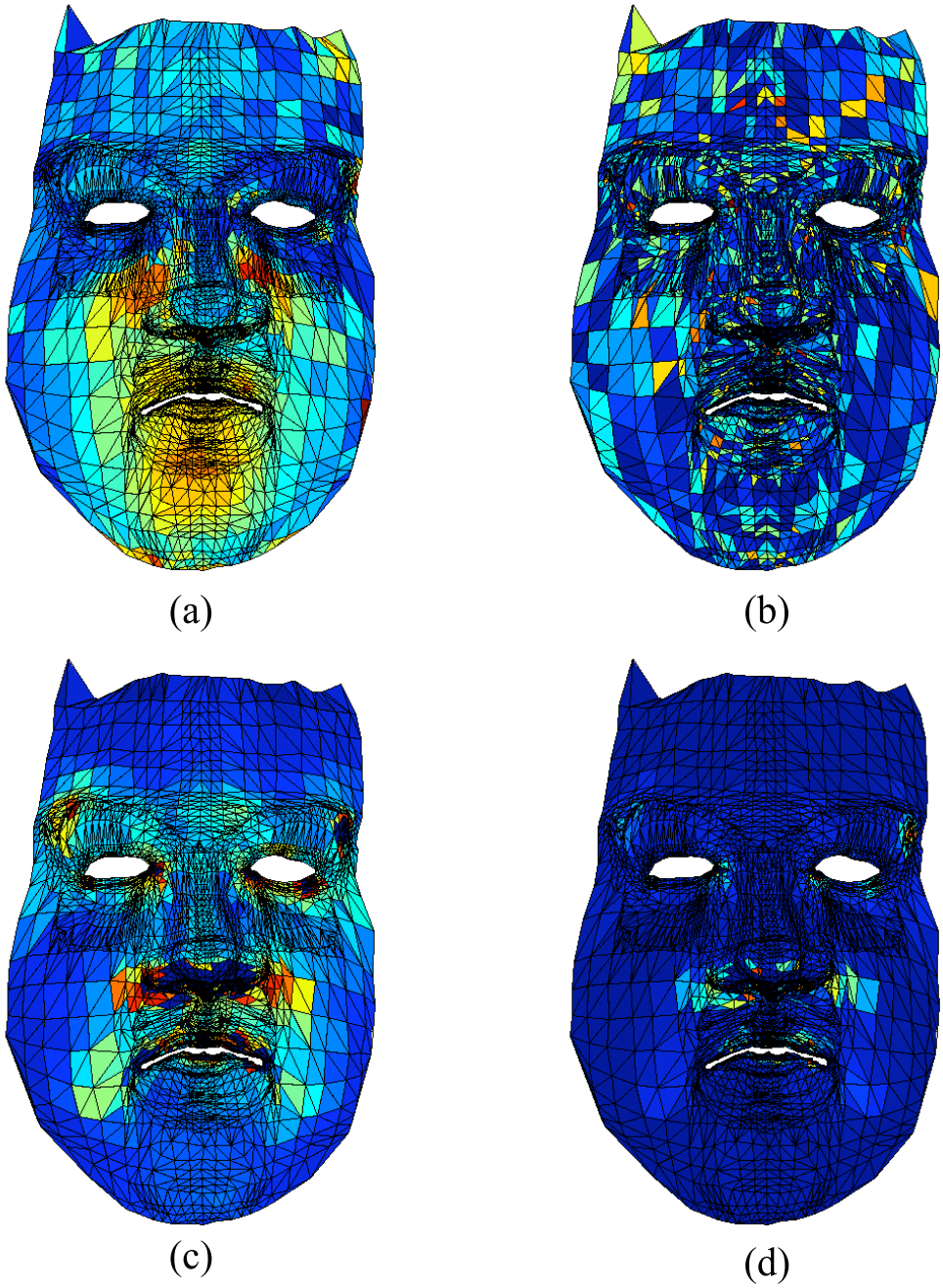


Figure 5.21: Qualifiers displacement *(a)*, rotation *(b)*, expansion *(c)*, and shear *(d)* extracted for a “disgust” sequence.

Data is a 32 frame sequence with mesh vertices moved via optical flow guided technique. Relative values indicated by colour and projected back onto conformed mesh at final frame.

5.5 Discussion

In this approach the outcome was very much dependent on the quality and flexibility of the initialisation of a conformed generic mesh. This was introduced in order that each of the surfaces could be mapped to a common representation for effective comparison between subjects. This would only be meaningful if the same vertices of the generic mesh model actually resolved to similar regions in each instance of surface data. In deciding on the mesh used, we selected a relatively small model - in terms of the number of vertices compared to those used in commercial animation. Crucially, it also provided us with a varying density in its topology around regions of the face with greater flexibility and potential deformation. We anticipated this to be a useful property - as ideally it would capture very subtle changes, such as wrinkles occurring around the eyes. However, it would appear that this had instead a negative effect in both the complexity in the initial conformation, and throughout the tracking process as we detail below.

The manual process of defining 40 landmark feature points was an intensive and difficult operation to perform consistently. In each subject there was a degree of variability in safely determining the same key points over all initial poses. It was empirically determined that such a high number of landmarks was required in order to guide the conformation - the eyes and mouth regions alone required half of these points in order to preserve the contours around openings correctly. The optimisation of the mesh was unable to differentiate between such boundaries in the actual $3D$ data. It was furthermore unable to smoothly handle in certain cases the collapsing distances between vertices for denser regions - leading to artefacts such as overlapping folds.

While the overall process is robust enough to give very reasonable fittings when visualised - there is as yet no complete guarantee by how much each vertex has individually "drifted" across the surface in accommodating for minor surface features, and the global minimum of the cost function for the entire fitting. It could be improved by imposing additional constraints on the extent to which the generic mesh could be altered (perhaps following the elastic/FEM and dynamic model paradigms e.g. [ACLS94]). In addition, it might be possible to also consider the actual stereo image data as well - in order to detect eye and mouth regions automatically as simple enclosed edges. Perhaps even a simpler solution would be to introduce markers on the subjects faces that would then allow us to robustly constrain both conformation and tracking. Ultimately the question remains as to whether the task could be entirely automated within a more geometric framework that removes the need for landmarks altogether and would prove more consistent, as seen in more recent works exploiting harmonic maps (e.g. [WGZ$^+$05]).

With regards to the tracking, the first important realisation is again that both tracking and conformation *are effectively the same task*. In this sense tracking can be considered as conformation of a differently shaped generic mesh to *every* single frame of data. This returns to the recurring concept that spatial and temporal registration are effectively the same problem. The core idea we follow here is that if the initial registration (conformation) is correct, then the temporal regis-

tration can proceed by tracking around each region (i.e. vertex) in turn from one moment to the next. The problem occurs if these regions are significantly changed beyond recognition - either because they have moved too far, or have deformed too much.

Our enhancement introduced here is the combination of a KLT/TPS based tracker in order to best accommodate these more dramatic displacements - along with suddenly appearing or disappearing regions (such as the mouth opening). This was furthermore envisaged as a useful extension in order to handle larger regions of sparse or ambiguous motion by introducing a *global* context, so that no vertex is left isolated from the combined motion of the entire surface (except for right at the boundary of the data - as observed in the synthetic rotating cylinder). The idea of then combining this with an existing *local* context as supplied by multi-channel flow-based estimates was to improve the overall tracking ability. As we showed (in Figure 5.17) the flow based approach can still yield reasonable results, except for the motion around the eyes and mouth (exactly the issues reserved for the KLT/TPS approach).

However, perhaps the most obvious recurring problem is again based in robustly handling the motion for such a large number of vertices and, consequently, the topology of our chosen generic mesh. The most apparent problem here is that some vertices, particularly around the mouth and eyes, may not move in conjunction with their neighbours in a believable or realistic manner. This can be mainly attributed to the breakdown of local estimates for newly occurring surface points (as the mouth opens) or simply failure to resolve ambiguity. The motion is then wholly dependent on the KLT/TPS global warp, which unfortunately does not have the fine-grained level of control for moving the individual vertices in these denser regions.

Furthermore, the mesh does not constrain the motion of the individual vertices in any plausible, pre-determined way. The result is that vertices are inevitably moved in an inconsistent manner, and that the topology of the mesh alters and fragments in a chaotic way. This is increasingly evident by the cumulative degradation in the quality of the mesh as processing proceeds through each frame - thus, longer sequences are particularly affected. Critically, this alteration of the mesh topology in turn affects the initial estimate of the surface normal used by the quadric fitting approach. A simple solution would perhaps rely on shorter sequences.

The problem then lies in part with the overall complexity of the generic mesh, particularly since the denser vertex regions match areas that exhibit the most deformation and sudden appearance/disappearance of surface features. These are unfortunately the areas where more errors occur - as well as being the most sampled. Using a simpler mesh might alleviate some of these issues, but at the cost of failing to capture the more interesting and revealing subtle surface changes. Indeed, many of the subtleties - particularly higher frequency components such as wrinkles - may already have been effectively removed by the $4D$ smoothing performed directly on the captured data. An alternative appealing idea would be to adaptively modify the mesh topology in response to the level of detail required to successfully capture all the expressions.

This would be performed by altering the mesh in response to where the largest errors occur over *all* the sequences - revealing the optimal mesh for all tracking.

The results of the tracking are however reasonably consistent in the overall scale of gross vertex motion that is successfully captured. This is shown by the displacement calculations in Figure 5.18 - which highlights that some of the sequences can still be tracked to relay a good degree of realistic deformation. Taken at a large scale, this is an encouraging result, but ultimately it comes at the cost of losing the precision of tracking individual vertices - especially if they are then to be compared as the same point on each face via the same generic mesh. This approach to tracking is fundamentally undone by the sheer complexity of the task for the level of detail desired. The best solution is perhaps offered by more effectively combining both the local flow with the global estimate, in a similar way to the more recent best performing optical flow techniques (i.e. [BBPW04]).

Ultimately, the accumulation of error in conformation and tracking is unfortunately mirrored by the secondary problem of robust surface fitting. This is a particular problem for the reliable estimation of the fundamental forms - based on the second order derivatives - from which the differential features are derived. As already mentioned, the generic mesh is used to guide the initial estimate of the surface normal. However subtle differences in this direction, as each vertex moves relative to its neighbours, results in a large degree of "jitter". This has the most noticeable effect on the calculation of the rotational qualifiers (one solution would be to perhaps zero this value if both principal curvatures are approximately the same: $\kappa_1 \approx \kappa_2$). When considering the face, this is perhaps not so surprising, since most of the surface is locally smooth and not dramatically non spherical (depending on scale). This does however affect the accuracy of the expansion and shear calculations which are calculated on the basis by which the local neighbourhood of the mesh is stretched and skewed in response to the relative motion of neighbouring vertices.

These problems can again be tied to the question of mesh topology, since ideally, the localised region around a vertex could be selected by projection of the nearest n neighbours on the mesh back onto the surface. Thus for dense regions, correspondingly smaller areas would be fitted, whereas sparse mesh regions (such as the cheeks and brow) would be catered for by larger surface areas. Taking this further, it would be better to consider the actual geodesic distances involved for selection of a surface region for fitting - and to factor these, rather than the projected planar distances, into the qualifier calculations. We have at least shown for the synthetic data sequences how our derivation of features from the changing fundamental forms can be carried out. From these results we again see that even with near perfect data, the accuracy of the underlying quadric fitting to a region of data is often a question of scale and careful selection of patch size.

A final major point must be directed at the quality of the initial data. In particular it should be noted that even the speed at which the stereo system operates, it can fail to capture the rapid and

instantaneous motion of genuine expressions. Even with the addition of the combined KLT/TPS tracker, some of these displacements break down. Furthermore, there are certain regions of the data where self-occlusion has serious implications for the quality of the reconstruction. In these instances, we could first of all leverage the occlusion mask - derived from what is only visible from either camera - to identify potential areas of poor quality tracking.

Overall, while these results are at least encouraging, there is still considerable scope for improvement, and in a reduction of the respective errors introduced at each stage. This must be performed robustly at all stages if there is to be any possibility of reasonably performing meaningful analysis and classification. For the moment we are unable to go beyond the evidence presented here to conclusively say that our set of descriptors can indeed meaningfully contribute to understanding such complex surface dynamics.

We acknowledge here that currently these 27 samples represent a rather small data-set to be performing effective statistical analysis - but we believe it contains enough variation to make it sufficiently interesting. This was primarily driven by the effort involved in capturing, post processing, building the 3D models, manually land-marking each initial frame, processing each sequence (as distributed over Beowulf clusters) and handling the combined 2Gb of final data. It is consequently a relevant point to highlight here the sheer throughput, storage, and computational requirements of our approach, versus the immediacy, and straightforward nature of simple 2D image data.

Even more fundamental is perhaps the simple fact that we do not really all smile, frown or express surprise in the same way. In which case image based methods (as opposed to the surface based interpretation we propose) may be more robust - since they focus instead on higher level interpretation of quite large scale features, such as the entire mouth, eye-brows, etc. and also manage to retain higher frequency features such as wrinkling. Furthermore, even with such cutting-edge $4D$ technology there are certain facets of micro-momentry expression (such as the speed of blinking) which cannot yet be captured.

5.6 Summary

In this penultimate chapter, we have brought together the separate strands of our previous chapters to finally look at the goal of going from sequences of $4D$ to registered motion of individual points and vertices, and to describing how the local neighbourhoods around those points are changing over time. We have exploited for the first time true $4D$ data of the face as it moves, and have been able to compare and contrast motion from a number of different subjects by conforming a generic mesh. We have proposed an extended methodology for describing the local changes on a surface over time by decomposing variations in the fundamental form around a point.

Our focus was on the potential to develop a "vocabulary" for further describing dynamic surfaces, by bringing together and enhancing the elements proposed in previous chapters. For example, we would like to be able to describe a region of points on the surface as a "shearing fold", or "rotating dimple", or "expanding protrusion".

We have shown that the resulting *rich* feature descriptors can be used in this manner, and as the basis for analysis of complex surfaces undergoing change (such as human faces). However, the accuracy of these is unfortunately dependent on the robustness of the intermediary and interlinked stages in their calculation. This work forms part of continuing research into the novel prospects offered by new $4D$ systems and the insights they offer into the dynamics of data from real surfaces, with the benefits for then interpreting such data.

Our key results from this can be summarised as follows:

- By using a global combined KLT/TPS tracker we can capture more density and accuracy in large scale surface motion.
- Successfully conforming a generic base mesh (and resulting tracking) must be well posed - or else cumulative errors can severely hamper and disrupt later calculations.
- We can derive a set of three further extrinsic qualifiers (expansion, rotation and shear) to act as surface feature descriptors, estimating all features by analysis of the first and second fundamental forms.
- Using these features to classify expression has some potential, but requires further investigation. The significance would only be wholly gained from a larger data-set of subjects

We present a final summary and analysis of this ultimate approach in the following concluding chapter, where we also consider the more fundamental problems raised and to be addressed (along with potential future work) in the context of other more recent research.

Chapter 6

Conclusion

> "The past is but the beginning of a beginning, and all that is or has been is but the twilight of the dawn"
>
> *H.G. Wells, 'The Discovery of the Future' (1901)*

As stated in the introduction, this work set out to investigate the original hypothesis that there existed a useful set of higher-level registered features that can serve to better analyse and classify the wealth of data generated by a $3D$ surface which is moving and deforming over time. The application of such a set of techniques would extend the field of Computer Vision by complementing the new and recent developments in $4D$ sensors and capture systems. Such a solution could be useful in addressing tasks concerned with making the data *more meaningful* - especially in certain problems which are limited by their tractability, hindered by the overwhelming volume of data.

So have we achieved this grand objective? In the course of this research we have pursued a bottom-up approach that has touched on many of the problems associated with the acquisition, pre-processing, registration, representation, description, analysis, and classification of real-world dynamic surfaces. If there is one main criticism, it is that this is too large a topic to cover all at once - in particular the problem of deformable registration. However, we have certainly achieved one possible way of building a complete pipeline that goes from sequences of image data to classification from surface change. Yet, this has inevitably given rise to more questions than answers.

In this chapter, we draw together the various strands of this work and offer some verdicts on the approach adopted. We then compare this against the relationship we have to other research in converging fields - particularly with regard to what ongoing, alternative approaches can offer. This in turn leads us to assess what could be done to investigate the open questions that remain and alternative techniques that may offer further interesting results and improvements.

6.1 Review of Achievements

In this section we aim first to highlight just what exactly has been achieved by this work - particularly if we look back to our original set of research questions (Section 1.1). We also present a brief summary of the work undertaken, and reiterate the importance of the key works that we extend.

6.1.1 Contributions

As a consequence of the investigations following the approach described above (and summarised below in Section 6.1.2 below) we claim to have made the following contributions to the analysis of time varying 3D data:

I An improved algorithm for estimating range flow by integrating colour.

II Guiding global motion through KLT feature guided TPS warping.

III Describing intrinsic shape change by variation in curvature change.

IV Further describing metric and extrinsic change by analysis of fundamental forms.

The first two areas of research (I and II) were carried out with the objective of establishing our first question: the extent to which additional surfaces properties and invariant features can be leveraged to help establish true **surface motion**.

The next two areas of research (III and IV) formed the outcome from considering our second question: the construction of a computational vocabulary for describing **surface change**.

Both aspects of research where then deployed in attempting to answer our ultimate question: that accurate surface motion combined with robust derivation of a subset of features, would then lend itself to higher-level analysis of **surface meaning**.

6.1.2 Summary of Work

In summarising the work presented in this thesis, it can be broken down into a number of stages of a conceptual "pipeline" which flows from the incoming data, through registration and conformation, surface tracking, to feature extraction and classification. In order to look at the core problem surrounding our main hypothesis (i.e. focusing on *describing shape change*) it proved necessary to develop algorithms to address a number of intermediary steps before we could get into the ultimate position to approach this in a meaningful way.

At the outset of the research we focused on the nature of data from our initial acquisition system. This system was based on a set of high resolution (6.8 *Megapixel*) cameras. Even

though the core stereo recovery algorithm was provided (by our industrial sponsors) it still required considerable understanding and experimentation to capture good results. Fundamentally, the algorithm works by capturing sufficiently detailed spatial texture to resolve the correspondence problem. The reconstruction worked best for natural surfaces, such as skin, provided the images were well exposed in focus and naturally illuminated. Synthetic, man-made objects with featureless texture were not reconstructed with accuracy, nor were objects with any transparent or specular surfaces.

The quality, speed and resolution of this data naturally forms the foundation of many of the applied results presented in this thesis. The occurrence of systematic artefacts (i.e. "orange peel") on the surface can seriously disrupt the calculation of second order surface characteristics. However, a sufficient degree of temporal smoothing can act to reduce this effect, as can the size of surface patch considered - although this has implications as to the scale of feature considered. Dealing with the real, complex and topological changes occurring on surfaces such as the face still presents considerable challenges. The success of tracking a $4D$ sequence can often depend on considerable empirical work to determine an optimal combination of parameters. In particular, we discovered this to be especially difficult around the area of the mouth, which exhibits rapid and radical deformations.

Extending the use of the high-resolution system by using the cameras burst mode functionality was the first (novel) step we took at capturing dynamic surfaces that changed over time. Similarly, performing "stop-motion" capture with this system allowed us to move real objects within the volume defined by the cameras' calibrated depth-of-field. Considerable care had to be taken with this regard, as the degree of error increased rapidly towards the front and rear of this capture frustum. This also applied to the high-speed capture system, as evidenced by the motion of subject's heads where effort had to be made to restrict the range of movement.

Having acquired our data, we turned to the problem of temporal registration. This proved to be an entire field of research by itself. Our approach here was to attempt to estimate both the local dense motion (via range and optical flow) that could be used to define the deformation warp from one instance of the surface to the next. The main enhancements focused on the combination of colour channel information to help resolve local ambiguities and aperture problems, and weighting these by their relative errors. Similarly, our globally guided approach attempted to regularise the warp governed by reliably tracked points, with constraints to minimise the total bending energy. While the results were encouraging, and indicated some degree of improvement and success at "tracking" the surfaces - the ultimate problems with surface discontinuities introduced by the appearance/disappearance of holes (such as the mouth) resulted in a degree of break-down in those regions. At all stages we have evaluated our approach using both synthetic, real test surfaces, and actual dynamic face data.

Our next stage did not immediately apply the results of this tracking - but instead looked

for the first time at ways to describe changes on a surface. Sequences of five frames were considered. These were, if necessary, registered using a robust (Tukey M-estimator) version of ICP in order to highlight those regions of change by aligning only to rigid regions. Our insight here was to consider all possible transitions from one class of codified shape to another, in terms of the initial and changes in principal curvature. From this we established a total of 15 types of change, which, when coupled with the extent (as defined by the magnitude of change), produced a compact representation which qualitatively illustrated the results. This was validated by testing with synthetic data, while using the approach to look at expressions (of only 5 frames) indicated the potential for a qualitative and useful basis for analysis.

We then turned our attention to combining our first techniques - for improved tracking of a deforming surface - with an extended system for describing deformation on the face. This was done in conjunction with the purchase and arrival of the new 4D system that allowed us for the first time to capture sequences of longer and denser temporal data ($30-40$ frames). Using this system to capture and post-process a number of subjects making three instances of three different types of expression required a large amount of processing and resulted in a large and complex data-set. This formed one of our first realisations that the sheer volume of data to be processed required time and some forethought in leveraging computational resources.

Our first step in order to conform each data-set was to fit a generic face mesh to the first frame of data. This mesh was chosen with a suitably flexible topology to allow it to accommodate surface motion (denser around the mouth and eye regions). Fitting the mesh to the data was performed on the basis of 40 land-marked feature points that were manually selected. This laborious process was the only stage that required direct user intervention in our pipeline (and, unfortunately, a degree of error that had repercussions for later comparison).

We then applied an alternative version of our tracking algorithm using higher-level KLT based features to guide TPS based warping. This was introduced to seek to address the failings in the local flow based approach when handling larger, more rapid displacements. Following synthetic validation with encouraging results, the results of applying this to the real data proved disappointing. We instead reverted back to a simpler flow based system, in order to at least capture some of the variation in the sequences.

Our theoretical work for devising further qualifiers - expansion, rotation, and skew - was then applied to describing the changes undergone. Estimating these accurately still formed an important step - but was helped in this instance by initialising the orientation and size of each patch based on each vertex in the tracked mesh (except in cases where the tracking went awry). Certain aspects of the qualifiers worked reasonably well at capturing some of the dynamics, but meaningful comparison between sequences proved impossible due to miss-matches in the initial conformation and tracking.

6.1.3 Relationship to Other Research

As previously stated, our work comes at a timely period. Research into "real-time stereo" is possible in part due to recent advances in camera sensor technology that can provide frame-rate capture, at sufficiently high-resolution to facilitate dense stereo recovery. This in turn is fuelling a number of novel applications in the area, and allowing us to visualise and measure dynamics to a level of accuracy hitherto impossible. As seen in Chapter 2, a number of prior works feed and inspire this research. However only a selection have a more direct influence. Here we draw attention to the core influences on our work, and briefly re-iterate how we address some of the open questions left by these.

Work by such luminaries as Besl, Jain, Koenderink and van Doorn provide the groundwork for so many lucid observations regarding surface curvature and its application to $3D$ analysis [BJ86, KvD92]. Of particular importance is the formalisation by Koenderink of a continuous shape and "curved-ness" metric for surface analysis and segmentation as presented in his seminal book [Koe90]. This directly led us to consider how this surface classification scheme could be extended, resulting in our derivation of extent and types of change.

Koenderink was also at the fore of relating the dynamic changes in optical flow [KvD86] with his definitions of curl, divergence, and shear in the flow field. These also formed the inspiration for our attempts to qualify extrinsic surface changes by rotation, shear and expansion. If anything, it is an attempt to unify the body of research into flow-based surface motion with the differential geometry underlying surface change that lies at the very center of our work.

Addressing the core problem of how to best "track" such surface motion led us to consider the $3D$ notion of *range flow* as a similar framework. This in turn led us to consider the work of Spies *et al*. [SJB02], particularly with regard to their fusion with image intensity data [BS00]. This also dove-tailed nicely with other independent work conducted by Barron and Klette into the benefits of adding colour information [BK02]. We also found inspiration for our KLT/TPS based tracker from earlier work in deformable registration - such as [BR04].

Curvature analysis of static surfaces, as with range-image interpretation in general, seems out of favour with current trends in Computer Vision (which we discuss below). The exception to this however is in analysing $3D$ scans of faces as a means of recognition. One of the earliest uses of this was seen in the work of Gordon [Gor92] which described registering two instances of a face for identification by using a modified version of ICP - a technique we also use initially to accommodate rigid motion while discarding surface changes with robust statistics.

More recent work has really focused on the challenges associated with invariance to expression in the face, especially by alignment to regions which have not changed, as looked at by Chang *et al*. [CBF05]. In this work, as with others, only two instances of scans of data are considered. To date, the actual dynamics of the expression have not been fully utilised.

6.2 Criticism and Outstanding Issues

Were we then able to successfully validate our original direction for this research? Our main claim was that there exists a useful set of properties that can serve to reveal the dynamic characteristic of a deforming surface. This can be broken down into two interrelated claims: firstly that such properties do indeed exist (and - crucially - can be accurately calculated), and that secondly they are in some sense *meaningful* and *useful* (i.e. can be applied to enable higher-level analysis). However, our approach adopted is not without its flaws - especially in light of the scope and scale of the problems encountered. In this section we look at what are the main outstanding issues and criticisms that can be generally levelled at our work.

We note here that perhaps the most significant general omission in our work is the lack of overall experimentation in conclusively determining the limits, exceptions, and performance of our developed techniques. These omissions apply to varying extents at all stages of work:

1 In data acquisition. (What is the actual accuracy of surface reconstruction? To what extent does $4D$ smoothing remove important features? Does the addition texture markers help landmark conformation?)

2 In tracking and registration. (What is the true performance of both flow and KLT/TPS methods? How do they perform in varying levels of noise? How good is the Tukey ICP method in rejecting outliers?).

3 In describing surface change. (Just how good is the underlying fitting of surface patches to different complexities of surface? How robust are the differential features to the presence of controlled noise? Are the patterns of surface change seen for real face data consistent over a larger data-set?)

4 Finally, in ascertaining the usefulness of the descriptors to classification and other tasks. (With a sufficient training set, do we actually see clusters and underlying subspaces that group similar expressions and identities together? Are there distinct dimensions along which expressions vary?)

We acknowledge that all such questions are valid, but in our defence, we highlight the ambitious scale of the task to accomplish when working with such real complex data. This was particularly true in the case of our objective to quickly move onwards, and so progress to the ultimate research objectives. With hindsight it would have perhaps been more effective to focus first on correctly ascertaining that each stage on the pipeline was working effectively on a single, well chosen collection of data.

Beyond this set of deficiencies, there are also a number of outstanding issues which cloud the existing methodology as it stands. These can be broadly subdivided into theoretical and numerical issues.

6.2.1 Theoretical Issues

As we have shown, our formulation for registering and observing the differential changes derived on the basis of an underlying surface patch, results in a method for describing the extent, type and qualification of deformation that a surface is undergoing. However, at a deeper level of understanding - have we managed to arrive at an actual meaningful set of descriptors for dynamics surfaces? A summary of the issues that are still left open are listed below:

- **Descriptors as useful features.** Perhaps the most important omission from this research is the final stage of classification. This would then form the ultimate validation of our descriptors as useful (particularly if they could be compared to other purely image based approaches). However, due to the inaccuracies in conformation, tracking and calculation, the final set of feature vectors (consisting of the 27 sequences) did not produce any significant distinctions. Furthermore, we do not address which (if any) of the descriptors are more meaningful?

- **The range and equivalence of descriptors.** One point we have not addressed is to consider if all possible descriptors have indeed been defined. Could they in fact be simplified further, or do they already represent a constrained set? Related to this issue is the unresolved equivalence between descriptors - particularly when considered over different scales. For example, a flattening peak (in which the changes in principal curvatures approach zero) can also be considered very similar to qualifying as expansion when only viewing a small local window. We do not satisfactorily denote where exactly these boundaries may lie.

- **Transitions over time.** Our approach at handling the transitions between types of form over time is not really considered. Indeed, in only recording the sum total of change for qualifiers we do not address any concerns with the rate of change. This therefore fails to capture any of the diversity of the dynamics with regards to transitions.

6.2.2 Numerical Issues

The application of our approach is wholly dependant on a degree of accuracy and robustness to variations in the data. Fundamentally, when analysing real surfaces, the cumulative error and numerical instabilities introduced at each stage is extremely hard to mitigate. We highlight these challenges below:

- **Accurate differential surface calculation.** Curvature is a second order property of the surface, reliably describing changes to it involves successfully integrating the underlying analytic form. To this end we focus on using a quadric that is fitted to the local patch

around any point in question. This still requires explicit handling to resolve the direction of the surface normal, before then applying optimisation techniques in order to model the surface at that point. This is inevitably hindered by the relative complexity of the surface - as governed by the the degree of noise and (more indirectly) by the chosen size of patch. Fundamentally, this again links back to the underlying issue concerning the scale of feature to be considered.

- **Accurate motion estimation.** The effective tracking of complex surfaces is incredibly difficult when performed for large displacements, and for longer, non-episodic sequences. The main problem with very dynamic surfaces, such as the face, is successfully registering the changes around the mouth and eyes. These changes in particular can be fast, varied, and complex. Furthermore, directly contrasting different sequences requires a solution to subject conformation as well - by which the same model over all samples can be used when tracking. In this sense tracking and conformation are the same thing. The obvious omission to our work is in failing to successfully unify the local flow based estimation with the global KLT/TPS warping. Further investigation into the limitations and break-downs of each technique is also required.

- **Automatic scaling and thresholding**. Considerable effort in our approach is spent on the task of empirically determining suitable threshold values in the case of determining the type of change, and when considering the values for window sizes when tracking. These are issues that again touch on the notion of the resolution of the data, and the scale of feature to be resolved. We do not investigate here the question of automatically selecting the appropriate selection of values based on the nature of the data.

- **Data quality**. Everything in our approach is ultimately dependant on the accuracy of the data captured from the true surface. Throughout this work we have encountered a number of issues with the detail and reconstruction. Tackling and resolving these more thoroughly would have immediate benefits for improved calculations.

6.3 Directions for Future Work

We have, in the course of this work only scratched the surface of the set of problems that then become apparent. The approach we have adopted has been a pragmatic and engineered solution, driven as it was from a desire to look at the more interesting final problem of describing surfaces from their shape change. We do, however, rely to a large extent on the theoretical underpinnings of differential geometry to provide us with a framework to achieve this. This is fundamentally where the more interesting aspects can be addressed, although inevitably by relying on ever more powerful and abstract techniques. However, this should not replace our core guiding objective

of a principled and parsimonious solution to extracting meaning from shape change. We start our analysis of alternative and novel approaches by considering the very most recent work that aligns with our own (and as a direct comparison to the basis to our work presented in Section 6.1.3 above). Specific enhancements to the three main areas we particularly focus on in this work are then presented in the following subsections.

6.3.1 Current Trends in the Field

Our approach is only one of several possibilities for investigating such an interesting problem, and with such a rich and varied source of data. Whereas we consider our work to stem from more traditional analysis of range-data, other researchers are more directly involved in investigating the issues surrounding systems for capturing novel $4D$ data. At one end of the spectrum there is the ongoing work in improving the construction of visual hulls using shape-from-silhouette techniques [CBK03]. Often, this work provides a solution to generating sequences for transmission and display within fully interactive CAVE-like environments, alongside recent developments in omnidirectional treadmills for fully immersive environments [KG06]. Here the focus is distinctly on the real time acquisition, reconstruction and encoding of such data to produce realistic and useful (i.e. transmittable) results. Recent work has in particular looked at combinations of stereo, volumetric and feature based techniques to improve the overall detail and speed of capture for realistic models [SMH06].

Furthermore, the ability to capture $4D$ is closely linked with the recent push towards enabling novel display technologies. An important issue here is creating sensors that are compact, do not require constant calibration, and are as simple to operate as existing image/video capture. The ideal is to have a solution which provides a point of depth for every pixel. Perhaps one of the most exciting developments for such capture are newly prototyped *multi-aperture* imaging sensors [FGW06]. In these, every group (or even individual pixel) acts as a separate camera across the entire CCD array - meaning they can be calibrated during manufacture. Further work in depth from defocus could also benefit based on plenoptic camera systems [NLB+05].

Another area of current active investigation focuses on the recurring problem of effective *registration* of surface data. Despite decades of research this is still very much an open question - especially in the case of data which deforms or warps. This a core problem looked at by our work in determining *where* a surface is moving, in order to then describe *how* it is moving, and so attempt to utilise this information to *recognise* larger patterns in the surface dynamics. Recent trends are rejecting stochastic techniques - such as ICP (although still investigated [WPWW06] and [YB07]) - in favour of more elegant solutions that reduce the complexity of the task by finding lower dimensional embedding that are more tractable. Examples of this are the spherical and harmonic map techniques [WGZ+05]. Such work has recently been extended to using conformal maps as an alternative approach to preserving surface properties. Additionally, work

also continues to look at the fundamental calculation of underlying surface properties - such as the principal curvatures in response to noise [PWY+06], or as directly based on a triangulated surface [MSR07].

This contrasts with ongoing flow based tracking approaches to resolving surface motion (as a form of temporal registration). Recent elegant solutions for combining local and global motion models in order to calculate more robust surface tracking have been proposed [BF07]. Further improvements to accuracy often include such coarse to fine estimates - and have also directly factored in a number of assumptions within a single framework that accommodates discontinuities based directly on spatial-temporal smoothness in warping energy [BBPW04]. This in turn links with the other recurring problem of handling occlusions, even with these more advanced techniques currently employed. Other image based work focuses particularly on explaining and modeling instances of cloth, paper, and other extremely mobile types of surface where self-occlusion is prevalent. For example [GBBS07a, GBBS07b] consider finding the optimal warping function that also explains self-occlusions which are detected by a suitable degree of shrinkage as pixels vanish. Whereas much of this work continues to be 2*D* image-based, it is interesting to note the increasing reliance on deformable models to explain the motion.

In terms of shape analysis, our work is distinct because it focuses primarily on the issue of creating explicit dynamic shape descriptors to describe the curvature dynamics. Other alternative approaches still only consider static data - but must be able to accommodate variations in order to recognise the same surface feature, even if it has changed. These include multi-resolution Spin Images [DK06] and 3*D* Shape Contexts [FHK+04]. The former work is an interesting case of determining feature descriptors that are invariant to different scales of data, by calculating the appropriate bin size from a coarse-to-fine scale-space. The latter work introduces a new approach to developing region shape descriptors again using a "binning" approach but one more suitable to point-cloud data. Both approaches effectively generate a feature vector that can be compared to other feature vectors - especially by using a spherical harmonic transform to create a rotationally invariant descriptor. This idea of striving for invariance has also led to recent attempts at adopting leading edge image based descriptors for surfaces - for example N-SIFT [CH07] which uses a volumetric framework to discover scale and rotation invariant 3*D* features to guide medical dataset registration. Other work carried out very recently by Miam *et al*. [MBO08] rely on SIFT to propose a "keypoint" feature detection on texture, around which 3*D* shape based descriptors can be constructed.

Most applied work in 4*D* still remains focused on analysis of the face. Ongoing work often continues to lie at the intersection of computer vision, graphics/animation and psychology. For example recent work by Wallraven *et al*. [WBCB08] employ their novel 4*D* scanner system to record accurate motion from human faces. They then use this data to modulate and control an artificial avatar to investigate the perception of realistic expressions. Directly learning from dy-

namic $3D$ data instances of Facial Action Units has also been recently investigated [BKC+07]. A guiding active appearance model is trained on this data, from which principal control parameters based on peak intensities within the action units are identified. New facial expressions can then be generated for novel realistic animation using these discovered parameters.

This problem is related to further work looking at the actual manifold of expression - as of yet this is mainly reliant on image based techniques (such as active appearance models) - for example the work of Chang *et al*. [HCFT04] and Chung *et al*. [CDB02]. These studies served as inspiration for techniques for determining the effective underlying meaning of various expressions from a rich set of features.

Ultimately continuing to look to $2D+3D$ and beyond to $4D$ faces for classification of identity continues to form a particularly active area of research. The recent survey by Abate *et al*. [ANRS07] looks at this across the broad range of techniques, whereas the survey by Kittler *et al*. [KHHI05] focuses more on the direct $3D$ issues. There occur still novel uses of depth and curvature information as the basis for recognition, as for example Lee *et al*. [LKSM06] look to extracting nose based depth features for PCA based analysis. However, additional geometric invariants and their usage also form the focus of work such as [RD07]. Actually applying these techniques and others to dynamic data for identity recognition still remains mostly an open area for investigation.

In summary, much work continues along the trends of acquisition, registration, deformation tracking, shape analysis and facial dynamics. Perhaps the overarching theme that unites these continues to be inclusion of temporal information to constrain, improve and distinguish with greater accuracy the nature of these dynamics. Fundamentally, making sense of such a wealth of data - such that all the complexities can be effectively modelled - remains the core problem to be addressed.

6.3.2 Acquisition and Temporal Registration

- Our approach is based on processing already reconstructed dense stereo data in which the individual frames are not related to one another temporally. Our own understanding of the problem now leads us to believe it is best approached with a single framework that fundamentally computes both spatial and temporal correspondence *together*. In so doing, the issues surrounding noise, continuity and accuracy can all be improved upon. The computation of the spatial-temporal *warp* (for error calculation) is itself a useful first step towards deriving the surface properties. It also raises the interesting idea of using such properties to resolve the correspondences in the first place (similar to using ICP on curvature for example). This would effectively introduce a form of feedback to arrive at a sequence in which the data is not only tracked, but also computes and best highlights the changes.

- The issues in handling occlusion and the opening/closing of the surface still need to accommodate such complex changes - especially when such changes can be so fast and fleeting. The most elegant way to phrase this problem is as a diffeomorphism, expressed as a one-to-one bijective mapping f from one smooth $3D$ differentiable manifold M_t to the next instance of it M_{t+1}, such that $f : M_t \rightarrow M_{t+1}$. The mapping must represent the continuous, localised warping between the two instances, and correctly determine the motion of every point on the surface. First mapping to a reduced subspace, such as a harmonic map, can greatly simplify this process and can preserve the equivalent distances. Further improvement could be gained by including other geometric properties of the surface, such as the geodesics. Ultimately, employing conformal maps could be an elegant and preserving framework to use for such an extension.

- Furthermore, a more informed *information based* approach to colour and its usefulness could present an interesting avenue to integrate further surface properties. Similarly, considering the use of *texture* (on which some recent optical flow algorithms are based) could provide an additional input to the process [ALL04]. Indeed, localised patches of texture, and their variations, have been used to create estimates of normal directions (normal maps) and even surface shape. These could all act as additional input to the surface recovery and simultaneous tracking - since the angle of the surface directly relates to motion and occlusion contours.

6.3.3 Deformation Description

- A more advanced solution to deformable feature extraction may require further aspects of differential geometry and related branches of mathematics. The use of *Riemannian geometry* (a favourite of cosmologists) could prove the ultimate and most powerful basis for expressing the alterations in shape. Through this, there may exist a further simplified set of descriptors. The potential for expressing these changes in a volumetric framework may allow greater numerical robustness and simplicity. Again, a single framework based on a conformal map could be a way of unifying the extraction of features from the underlying diffeomorphisms.

- One of the fundamental problems we do not address in this work is the hierarchical nature of surface change. For example, certain types of interesting changes occur at quite a large scale - for example the furrowing/folding of someone's brow when they frown, and others may occur at a considerably smaller level - such as the wrinkling/bending around the eyes when smiling. Both involve similar changes - but the problem is how can they both be accommodated without prior knowledge of the surface dynamics. This tracking (as we investigated) can also be unified by the use of a single mesh - representing a single topology

across sequences. Another idea would be to adapt the mesh to reflect the complexity of change - by iteratively seeking forwards and backwards through the sequence in order to discover fitting errors and so adapt the model to best fit the data (yet without over-fitting).

- Overall, the greatest current omission is in not accounting for the transitions between episodic changes and how these follow one another. For more complex surfaces a number of changes may happen that are interrelated and indeed de-mark the points where important global meaning is made (i.e. for the face to frown and then break into a smile). This could be addressed by using a higher-level dynamics model (such as a HMM) that would resolve the transitions within the sequence. Other ideas from motion editing could act to segment the motion into salient segments of recurring dynamics. The ultimate validation for this would be to test its ability to relate these to speech, in order to recognise what words someone is saying as they are scanned in $4D$ (continuing the linkage between "visemes" [TF00] and "phonemes" - but instead as "formenes" linking form to speech).

6.3.4 Classification and Other Applications

- Could the foundations laid by this approach actually be used to reliably recognise both people, and expressions? This raises the additional interesting question as to whether people themselves could be recognised by their own unique expression signatures. Our preliminary results showed some interesting variations along these lines, but more empirical work over larger data-sets would be required to assess the viability for bio-metric identification. A related idea would be to establish what is the effectual information content of expression (e.g. entropy of a smile as opposed to a frown?). Indeed, it is still an open question if humans actually interpret the changing shape of the face (i.e. through shape from shading) and use this medial representation instead of simply 2D imagery.

- An important question when using dense surface motion is how do we know in advance what regions best exemplify and describe the deformation? This is part of our justification for not relying on a sparse input - such as motion capture - which is invariably biased towards the placement of tracked markers. However, an appealing idea that stems from this association with motion capture is the process of expression transfer - by which dynamics can be extracted and transferred to entirely new sequences (this is already being investigated - e.g. [CRRM07]). Central to this idea is the separation of invariant stylistic components from an underlying shared content. In this way the generic dynamics of a smile in 4D could also be modulated by such features.

- Another important realisation stemming from this work is that the temporal duration does not necessarily need to be of the order of a few seconds. It can just as easily be used to analyse shape change over a matter of hours, day or even years. This could offer interesting

visualisation for medical and geological data-sets, where the categorisation of the shape change could potentially reveal interesting trends (for example, a volcano, captured over time with aerial side scan radar, could reveal a characteristic bulging expansion associated with magma build-up as a precursor to eruption). Other uses in the medical domain could compare surface motion in the face for stroke patients to assess recovery.

- Finally, might it ultimately be possible to quantifiably measure dynamic 4*D* change truly in real-time? Given the improvements in sensor and processing technology on the horizon, the effective time from images to surfaces to descriptors could be much reduced. This could have major implications for using our work as the basis for tasks where quickly analysing dynamic surfaces would yield immediate benefits.

6.4 Final Word

Casting our vision even further ahead - what does the future hold for 4*D* data and its role for aiding our understanding of the face and other complex deformable surfaces? Certainly such data will become ever more pervasive, especially with the media embracing it as new display and interactive technologies mature. The immediate problems with storage and transmission will most likely be solved first with progressively larger and cheaper memory and faster communications. In terms of analysis and interpretation, there are then many possibilities for applications and systems which can apply and build on the concepts described in this work. These would in turn hopefully enable more fundamental scientific questions to be addressed. From visualisation through to classification, but in also providing a crucial experimental tool to validate new theories in psychology, perception and AI. Ultimately enabling us to understand and make sense of the 4*D* world of surfaces.

Bibliography

[ACLS94] J. K. Aggarwal, Q. Cai, W. Liao, and B. Sabata. Articulated and elastic non-rigid motion: A review. In *Proceedings of the IEEE Computer Society Workshop on Motion of Non-Rigid and Articulated Objects*, pages 16–22, 1994. 2, 2.1.3, 5.5

[ACP02] B. Allen, B. Curless, and Z. Popović. Articulated body deformation from range scan data. *ACM Transactions on Graphics (SIGGRAPH 2002)*, 21(3):612–619, 2002. 2, 2.1.3, 2.3.1

[ACP03] B. Allen, B. Curless, and Z. Popović. The space of human body shapes: Reconstruction and parameterization from range scans. *ACM Transactions on Graphics (SIGGRAPH 2003)*, 22(3):587–594, 2003. 2, 2.2.2

[AFRW99] A. P. Ashbrook, R. B. Fisher, C. Robertson, and N. Werghi. Construction of articulated models from range data. In *Proc. British Machine Vision Conference BMVC99*, pages 183–192, 1999. 2, 2.2.2

[AK06] K. Aoki and H. Koshimizu. Detection of 3D-flow by characteristic of convex-concave and color. In *Proceedings of the 18th International Conference on Pattern Recognition*, volume 1, pages 75–78, 2006. 2, 2.3.2

[AL03] R.J. Andrews and B.C. Lovell. Color optical flow. In B.C. Lovell and A.J. Maeder, editors, *Proceedings Workshop on Digital Image Computing*, pages 135–139, 2003. 2, 2.3.2, 3.3

[ALL04] M.A. Arredondo, K. Lebart, and D. Lane. Optical flow using textures. *Pattern Recognition Letters*, 25(4):449–457, 2004. 2, 2.3.2, 3.3, 6.3.2

[ANRS07] A.F. Abate, M. Nappi, D. Riccio, and G. Sabatino. 2D and 3D face recognition: A survey. *Pattern Recognition Letters*, 28(14):1885–1906, 2007. 6.3.1

[AT05] G. Agam and X. Tang. Accurate principal directions estimation in discrete surfaces. In *Proceedings of the Fifth International Conference on 3-D Digital Imaging and Modeling (3DIM)*, pages 293–300, 2005. 2, 2.1.1

[BA93] M.J. Black and P. Anandan. A framework for the robust estimation of optical flow. In *Fourth International Conf. on Computer Vision (ICCV-93)*, pages 231–236, 1993. 2

[BA00] C. Beumier and M. Acheroy. Automatic 3D face authentication. *Image and Vision Computing*, 18:315–328, 2000. 2, 2.4.3

[Bar84] A.H. Barr. Global and local deformations of solid primatives. *Computer Graphics*, 18(3):21–30, 1984. 2, 2.1.2

[BB95] S.S. Beauchemin and J.L. Barron. The computation of optical flow. *ACM Computing Surveys*, 27(3):433–466, 1995. 2, 2.3.2

[BBK05] A. M. Bronstein, M. M. Bronstein, and R. Kimmel. Three-dimensional face recognition. *International Journal of Computer Vision (IJCV)*, 64(1):5–30, August 2005. 2, 2.4.3

[BBPW04] T. Brox, A. Bruhn, N. Papenberg, and J. Weickert. High accuracy optical flow estimation based on a theory for warping. In T. Pajdla and J. Matas, editors, *Proc. 8th European Conference on Computer Vision*, volume 3024 of Lecture Notes in Computer Science, pages 25–36, 2004. 5.3, 5.5, 6.3.1

[BCF06] K.W. Bowyer, K. Chang, and P. Flynn. A survey of approaches and challenges in 3D and multi-modal 3D + 2D face recognition. *Computer Vision and Image Understanding*, 101(1):1–15, 2006. 2, 2.4.3

[BF07] A. Buchanan and A. Fitzgibbon. Combining local and global motion models for feature point tracking. In *Proc. Computer Vision and Pattern Recognition*, pages 1–8, 2007. 6.3.1

[BFB94] J.L. Barron, D.J. Fleet, and S.S. Beauchemin. Performance of optical flow techniques. *International Journal of Computer Vision*, 12(1):43–77, 1994. 2, 3.2, 3.4.1, 5.3

[Bha43] A. Bhattacharya. On a measure of divergence between two statistical populations defined by their probability distributions. *Bulletin of Calcutta Maths Society*, 35:99–110, 1943. 4.4.3

[BHB99] C. Bregler, A. Hertzmann, and H. Biermann. Recovering non-rigid 3D shape from image streams. In *IEEE Conf. Computer Vision and Pattern Recognition 2000*, volume 2, pages 690–696, 1999. 2, 2.2.1

[BHES99] M. Bartlett, J. Hager, P. Ekman, and T. Sejnowski. Measuring facial expressions by computer image analysis. *Psychophysiology*, 36:253–263, 1999. 2, 2.4.2

[BJ86] P. J. Besl and R. C. Jain. Invariant surface characteristics for 3D object recognition in range images. *Computer Vision, Graphics, and Image Processing*, 33(1):33–80, 1986. 2, 2.1.1, 2.5, 4.1, 6.1.3

[BK02] J. Barron and R. Klette. Quantitative color optical flow. In *ICPR*, volume 4, pages 251–255, 2002. 2, 2.3.2, 2.5, 3.3, 6.1.3

[BKC+07] L. Benedikt, E. Krumhuber, A. Calvert, D. Cosker, P. Rosin, and D. Marshall. Using dynamic 3D facial data to create 3D appearance models of facial action units. In *ACM SIGGRAPH/Eurographics Symposium on Computer Animation (SCA)*, 2007. 6.3.1

[Ble66] W.W. Bledsoe. Man-machine facial recognition: Report on a large-scale experiment. Technical Report 22, Panoramic Research, Inc., Palo Alto, California, 1966. 2, 2.4

[BM92] P. Besl and N. McKay. A method for registration of 3D shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(2):239–256, 1992. 2, 2.1.2, 4.3.1

[Boo89] F. L. Bookstein. Principal warps: Thin-plate splines and the decompostion of deformations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2(6):567–585, 1989. 2, 2.1.2, 5.3.2

[Boo91] F. L. Bookstein. *Morphometric Tools for Landmark Data: Geometry and Biology*. Cambridge University Press, 1991. 2, 2.1.1

[BPV06] C. Basso, P. Paysan, and T. Vetter. Registration of expressions data using a 3D morphable model. In *7th International Conference on Automatic Face and Gesture Recognition*, pages 205–210, 2006. 2, 2.3.1

[BR04] B.J. Brown and S. Rusinkiewicz. Non-rigid range-scan alignment using thin-plate splines. In *Proceedings. 2nd International Symposium on 3D Data Processing, Visualization and Transmission*, pages 759–765, 2004. 2, 2.3.1, 6.1.3

[Bre97] C. Bregler. Learning and recognizing human dynamics in video sequences. In *IEEE Conf. Computer Vision and Pattern Recognition*, pages 568–570, 1997. 2

[BRRP97] B. Bodenheimer, C. Rose, S. Rosenthal, and J. Pella. The process of motion capture: Dealing with the data. In D. Thalmann and M. van de Panne, editors, *Computer Animation and Simulation '97, Eurographics Animation Workshop*, pages 3–18. Springer-Verlag, 1997. 2, 2.2

[BS00] J.L. Barron and H. Spies. The fusion of image and range flow. In *Proceedings of the 10th International Workshop on Theoretical Foundations of Computer Vision: Multi-Image Analysis*, pages 171–189, 2000. 2, 2.3.2, 2.5, 6.1.3

[BSL+07] S. Baker, D. Scharstein, J.P. Lewis, S. Roth, M. Black, and R. Szeliski. A database and evaluation methodology for optical flow. In *Proc. International Conference on Computer Vision (ICCV)*, pages 1–8, 2007. 3.2

[BU04] E. Bart and S. Ullman. Image normalization by mutual information. In *British Machine Vision Conference*, pages 327–336, 2004. 3.3

[Bur87] D.S. Burnett. *Finite Element Analysis*. Addison-Wesley, 1987. 2

[BWS05] A. Bruhn, J. Weickert, and C. Schnörr. Lucas/kanade meets horn/schunck: Combining local and global optic flow methods. *International Journal of Computer Vision*, 61(3):211–231, 2005. 2, 3.2

[BY95] M. J. Black and Y. Yacoob. Tracking and recognizing rigid and non-rigid facial motions usin glocal parametric models of image motion. In *Proceedings of the Fifth International conference on Computer Vision*, pages 374–381, 1995. 2, 2.4.2

[CBF05] K.I. Chang, K.W. Bowyer, and P.J. Flynn. Adaptive rigid multi-region selection for handling expression variation in 3D face recognition. In *IEEE Workshop on Face Recognition Grand Challenge Experiments*, page 157, 2005. 2, 2.4.3, 6.1.3

[CBK03] G.K.M. Cheung, S. Baker, and T. Kanade. Shape-from-silhouette of articulated objects and its use for human body kinematics estimation and motion capture. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition 2003 (CVPR'03)*, volume 1, pages 77–84, 2003. 2, 2.2.1, 6.3.1

[CBM+01] A.J. Calder, A.M. Burton, P. Miller, A.W. Young, and S. Akamatsu. A principal component analysis of facial expressions. *Vision Research*, 41(9):1179–1208, 2001. 2, 2.4.1

[CCS06] A. Colombo, C. Cusano, and R. Schettini. 3D face detection using curvature analysis. *Pattern Recognition*, 39(3):444–455, 2006. 2, 2.4.3

[CDB02] E.S. Chuang, H. Deshpande, and C. Bregler. Facial expression space learning. In *Pacific Graphics*, pages 68–76, 2002. 2, 2.4.2, 2.5, 6.3.1

[CF01] H. Cantzler and R. B. Fisher. Comparison of HK and SC curvature description methods. In *Proceedings of the 3rd International Conference on 3-D Digital Imaging and Modeling*, pages 285–291, 2001. 2, 2.1.1, 4.1, 4.2

[CGC+02] S. Capell, S. Green, B. Curless, T. Duchamp, and Z. Popović. A multiresolution framework for dynamic deformations. In *ACM SIGGRAPH Symposium on Computer Animation 2002*, pages 41–47, 2002. 2, 2.1.2

[CGH00] I. Cohen, A. Garg, and T.S. Huang. Emotion recognition from facial expressions using multilevel HMM. In *Neural Information Processing Systems Workshop on Affective Computing*, 2000. 2, 2.4.2

[CH06] M. Castelan and E.R. Hancock. Acquiring height data from a single image of a face using local shape indicators. *Computer Vision and Image Understanding*, 103(1):64–79, 2006. 2, 2.2.1

[CH07] W. Cheung and G. Hamarneh. N-SIFT: N-Dimensional scale invarient feature transform for matching medical images. In *4th IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, pages 720–723, 2007. 6.3.1

[CHT03] Y. Chang, C. Hu, and M. Turk. Manifold of facial expression. In *IEEE International Workshop on Analysis and Modeling of Faces and Gestures*, pages 28–35, 2003. 2, 2.4.2, 2.5

[CJM03] C. Chu, O.C. Jenkins, and M.J. Matarić. Towards model-free markerless motion capture. In *IEEE Intl. Conf. on Robotics and Automation*, 2003. 2, 2.2.2

[CMG01] I. Cohen, G. Medioni, and H. Gu. Inference of 3D human body posture from multiple cameras for vision-based user interfaces. In *5th World Multi-Conference on Systemics, Cybernetics and Informatics*, 2001. 2, 2.2.1

[CRRM07] D. Cosker, S. Roy, P. L. Rosin, and D. Marshall. Remapping animation parameters between multiple types of facial model. In *Mirage 2007 - Computer Vision/Computer Graphics Collaboration Techniques and Applications*, pages 365–378, 2007. 6.3.4

[CSC+04] D. M. Cash, T. K. Sinha, C. Chen, B. M. Dawant, W. C. Chapman, M. I. Miga, and R. L. Galloway. Identification of deformation using invariant surface information. In *Proc. SPIE*, volume 5367, pages 140–150, 2004. 2, 2.3.1

[CSGE02] J. Cohn, K. Schmidt, R. Gross, and P. Ekman. Individual differences in facial expression: Stability over time, relation to self-reported emotion, and ability to inform person identification. In *Proceedings of the International Conference on Multimodal User Interfaces*, pages 491–497, October 2002. 2, 2.4.1

[CSW+03] M. Chen, D. Silver, A. S. Winter, V. Singh, and N. Cornea. Spatial transfer functions - a unified approach to specifying deformation in volume modeling and animation. In *Proc. Third International Workshop on Volume Graphics*, pages 35–44, 2003. 2

[CT04] T. F. Cootes and C. J. Taylor. Statistical models of appearance for computer vision. Technical report, University of Manchester, UK, 2004. 2, 2.3.1

[Cur99] B. Curless. From range scans to 3D models. *Computer Graphics*, 33(4):38–41, 1999. 2, 2.2.1

[CWP+01] M. K. Chung, K. J. Worsley, T. Paus, C. Cherif, D. L. Collins, J. N. Giedd, J. L Rapoport, and A. C. Evans. A unified statistical approach to deformation-based morphometry. *NeuroImage*, 14(3):595–606, 2001. 2, 2.1.3, 2.5

[Dar98] C. Darwin. *The Expression of the Emotions in Man and Animals*. Oxford University Press, 3rd edition, 1998. 1.1

[DBR00] J. Deutscher, A. Blake, and I. Reid. Articulated body motion capture by annealed particle filtering. In *Proc. Computer Vision and Pattern Recognition*, volume 2, pages 126–133, 2000. 2

[DK06] H.Q. Dinh and S. Kropac. Multi-resolution spin-images. In *Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 863–870, 2006. 6.3.1

[DM96] D. DeCarlo and D. Metaxas. The integration of optical flow and deformable models with applications to human face shape and motion estimation. In *CVPR*, pages 231–238, 1996. 2, 2.4.2

[Ebn07] M. Ebner. *Color Constancy*. John Wiley, 2007. 3.3.1

[EF78] P. Ekman and W. Friesen. *Facial Action Coding System: A Technique for the Measurement of Facial Movement*. Palo Alto: Consulting Pshchologists Press, 1978. 2, 2.4.1

[EP97] I. Essa and A. Pentland. Coding, analysis, interpretation, and recognition of facial expressions. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 19(7):757–763, 1997. 2, 2.4.2

[FDL04] G.D. Finlayson, M.S. Drew, and C. Lu. Intrinsic images by entropy minimization. In *8th European Conference on Computer Vision*, volume 3, pages 582–595, 2004. 3.3

[FF01] P. Faber and R. B. Fisher. Pros and cons of euclidean fitting. In *Proceedings of the Annual German Symposium for Pattern Recognition*, pages 414–420, 2001. 4.3.2, 4.3.2

[FGDP02] P. Fua, A. Gruen, N. D'Apuzzo, and R. Plänkers. Markerless full body shape and motion capture from video sequences. *International Archives of Photogrammetry and Remote Sensing*, 34(5):256–261, 2002. 2, 2.2.2

[FGW06] K. Fife, A. El Gamal, and H.-S.P. Wong. A 3D multi-aperture image sensor architecture. In *Proc. IEEE Custom Integrated Circuits*, pages 281 – 284, 2006. 6.3.1

[FHK+04] A. Frome, D. Huber, R. Kolluria, T. Buelow, and J. Malik. Recognizing objects in range data using regional point descriptors. In *European Conference in Computer Vision (ECCV)*, pages 224–237, 2004. 6.3.1

[Fis89] R.B. Fisher. *From Surfaces to Objects: Computer Vision and Three Dimensional Scene Analysis*. John Wiley and Sons, 1989. 2, 2.1.1

[FP03] D.A. Forsyth and J. Ponce. *Computer Vision: A modern approach*. Prentice Hall, 2003. 3.1

[GB97] P. Golland and A.M. Bruckstein. Motion from colour. *Computer Vision and Image Understanding*, 68(3):346–362, 1997. 2, 2.3.2, 3.3

[GBBS07a] V. Gay-Bellile, A. Bartoli, and P. Sayd. Direct estimation of non-rigid registrations with image-based self-occlusion reasoning. In *ICCV'07 - Proceedings of the Eleventh IEEE International Conference on Computer Vision*, pages 1–6, 2007. 6.3.1

[GBBS07b] V. Gay-Bellile, A. Bartoli, and P. Sayd. Feature-driven direct non-rigid image registration. In *British Machine Vision Conference*, volume 1, pages 42–51, 2007. 6.3.1

[GGS+01] P. Golland, W. E. L. Grimson, M. E. Shenton, , and R. Kikinis. Deformation analysis for shape based classification. In *Proceedings of Information Processing in Medical Imaging*, pages 517–530, 2001. 2, 2.1.3

[GJ97] P.R. Giaccone and G.A. Jones. Spatio-temporal approaches to computation of optical flow. In *British Machine Vision Conference*, volume 2, pages 420–429, 1997. 2, 3.2

[GLW06] R. Godding, Th. Luhmann, and A. Wendt. 4D surface matching for high-speed stereo sequences. In *Proceedings of the ISPRS Commission V Symposium: 'Image Engineering and Vision Metrology'*, September 2006. 2, 2.2.2

[Gor91] G. G. Gordon. Face recognition based on depth maps and surface curvature. In *Proceedings of the SPIE, Geometric Methods in Computer Vision*, volume 1570, pages 234–247, 1991. 2, 2.4.3

[Gor92] G. G. Gordon. Face recognition based on depth and curvature features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 108–110, 1992. 2, 2.4.3, 2.5, 6.1.3

[Gra97] A. Gray. *Modern Differential Geometry of Curves and Surfaces with Mathematica*. Boca Raton, FL: CRC Press, 2nd ed. edition, 1997. 2, 2.1

[GS03] R. V. Garimella and B. K. Swartz. Curvature estimation for unstructured triangulations of surfaces. Technical Report LA-UR-03-8240, Los Alamos National Laboratory, Nov 2003. 2, 2.1.1, 4.3

[GVM03] S. Goldenstein, C. Vogler, and D. Metaxas. Statistical cue integration in dag deformable models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(7):801–813, 2003. 2, 2.3.1

[HBHP03] T.J. Hutton, B.F. Buxton, P. Hammond, and H.W.W. Potts. Estimating average growth trajectories in shape-space using kernel smoothing. *IEEE Transactions on Medical Imaging*, 22(6):747–753, 2003. 2, 2.1.3

[HCFT04] C. Hu, Y. Chang, R. Feris, and M. Turk. Manifold based analysis of facial expression. In *Computer Vision and Pattern Recognition Workshop*, pages 81–81, 2004. 2, 2.4.2, 6.3.1

[HL04] J. Hoey and J.J. Little. Decision theoretic modeling of human facial displays. In *Proceedings of European Conference on Computer Vision (ECCV)*, pages 26–38, 2004. 2, 2.4.2

[Hoe01] Jesse Hoey. Hierarchical unsupervised learning of facial expression categories. In *IEEE Workshop on Detection and Recognition of Events in Video*, pages 99–106, 2001. 2, 2.4.2

[HPA04] T. Heseltine, N.E. Pears, and J. Austin. Three-dimensional face recognition using combinations of surface feature map subspace components. *Image and Vision Computing*, 26(3):382–396, 2004. 2, 2.4.3

[HS81] B.K.P. Horn and B. G. Schunk. Determining optical flow. *Artificial Intelligence*, 17:185–203, 1981. 2, 3.2, 5.3

[HSC02] A. Hilton, J. Starck, and G. Collins. From 3D shape capture to animated models. In *1st International Symposium on 3D Data Processing Visualization and Transmission (3DPVT 2002)*, pages 246–255, 2002. 2, 2.2

[HSE03] C. Hesher, A. Srivastava, and G. Erlebacher. A novel technique for face recognition using range imaging. In *Proceedings Seventh International Symposium on Signal Processing and Its Applications*, volume 2, pages 201–204, 2003. 2, 2.4.3

[HZ00] R.I. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2000. 2, 2.2.1, 3.1

[HZW+04] X. Huang, S. Zhang, Y. Wang, D. Metaxas, and D. Samaras. A hierarchical framework for high resolution facial expression tracking. In *In Proc. 3rd IEEE Workshop on Articulated and Nonrigid Motion (in conjunction with CVPR04)*, pages 22–28, 2004. 2, 2.3.1

[JAP99] T. Jebara, A. Azarbayejani, and A. Pentland. 3D structure from 2d motion. *IEEE Signal Processing Magazine*, 16(3):66–84, May 1999. 2, 2.2.1

[JBS03] X. Ju, T. Boyling, and P. Siebert. A high resolution stereo imaging system. In *Workshop on 3D Modelling*, 2003. 2, 2.2.1, 3.1

[JMS+04] X. Ju, Z. Mao, J.P. Siebert, N. McFarlane, J. Wu, and R. Tillett. Applying mesh conformation on shape analysis with missing data. In *Proceedings of the 2nd International Symposium on 3D Data Processing, Visualization and Transmission (3DPVT)*, pages 696–702, 2004. 2, 2.3.1, 5.4.1

[JPF+02] S. Joshi, S. Pizer, P.T. Fletcher, P. Yushkevich, A. Thall, and J.S. Marron. Multiscale deformable model segmentation and statistical shape analysis using medial descriptions. *IEEE Transactions on Medical Imaging*, 21(5):538–550, 2002. 2, 2.1.3

[KBL04] S. Kemp, V. Bruce, and A. Linney. *Future Face: The Human Face and How We See It*. Profile Books Ltd., 2004. 2, 2.4.1

[KBM+01] M. Kamachi, V. Bruce, S. Mukaida, J. Gyoba, S. Yoshikawa, and S. Akamatsu. Dynamic properties influence the perception of facial expressions. *Perception*, 30(7):875–887, 2001. 2, 2.4.1

[KG06] R. Kehl and L. Van Gool. Markerless tracking of complex human motions from multiple views. *Computer Vision and Image Understanding*, 104:190–209, 2006. 6.3.1

[KHHI05] J.V. Kittler, A. Hilton, M. Hamouz, and J. Illingworth. 3D assisted face recognition: A survey of 3D imaging modelling and recognition approaches. In *Proc. of SafeSecur05*, volume III, pages 114–114, 2005. 6.3.1

[Koe90] J. J. Koenderink. *Solid Shape*. MIT Press, Cambridge, 1990. 2, 2.1.1, 6.1.3

[KPT+06] I. Kakadiaris, G. Passalis, G. Toderici, N. Murtuza, and T. Theoharis. 3D face recognition. In *Proceedings of the British Machine Vision Conference*, volume 3, pages 869–868, 2006. 2, 2.4.3

[KvD86] J. J. Koenderink and A. J. van Doorn. Optic flow. *Vision Research*, 26(1):161–180, 1986. 2, 2.3.2, 6.1.3

[KvD92] J. J. Koenderink and A. J. van Doorn. Surface shape and curvature scales. *Image and Vision Computing*, 10(8):557–565, 1992. 2, 2.1.1, 2.5, 4.1, 6.1.3

[KWT87] M. Kass, A. Witkin, and D. Terzopoulos. Snakes: Active contour models. *International Journal of Computer Vision*, 1(4):321–331, 1987. 2.1.2

[LCJ06] X. Lu, D. Colbry, and A. K. Jain. Matching 2.5D scans for face recognition. In *Proc. of the International Conference on Biometric Authentication*, pages 30–36, 2006. 2, 2.4.3

[LF05] T. C. Lukins and R. B. Fisher. Colour constrained 4D flow. In *Proceedings of the British Machine Vision Conference*, pages 340–348, 2005. 1.3.1

[LF06] T. C. Lukins and R. B. Fisher. Qualitative characterization of deforming surfaces. In *Proceedings of the Third International Symposium on 3D Data Processing, Visualization and Transmission*, pages 287–293, 2006. 1.3.1

[LF08] T. C. Lukins and R. B. Fisher. A mark-up for describing dynamic surfaces. In *BMVA Symposium on 3D Video: Analysis, Display and Applications*, 2008. 1.3.1

[LJ06] X. Lu and A.K. Jain. Deformation modeling for robust 3D face matching. In *Proceedings IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR2006)*, pages 1377–1383, 2006. 2, 2.4.3, 2.5

[LK81] B.D. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. In *7th Int. Conf. on Artificial Intelligence*, pages 674–679, 1981. 2, 2.3.2, 3.2, 5.3.1

[LKCL98] J-J. J. Lien, T. Kanade, J. Cohn, and C-C. Li. Subtly different facial expression recognition and expression intensity estimation. In *IEEE Conference on Computer Vison and Pattern Recogntion*, pages 853–859, June 1998. 2, 2.4.2

[LKSM06] Y. Lee, I. Kim, J. Shim, and D. Marshall. 3D facial image recognition using a nose volume and curvature based eigenface. In *Geometric Modelling and Processing*, pages 616–622, 2006. 6.3.1

[LP01] J. Lang and D.K. Pai. Estimation of elastic constants from 3D range-flow. In *IEEE 3rd International Conference on 3D Digital Imaging and Modeling*, pages 331–338, 2001. 2, 2.3.2

[Mar82] D. Marr. *Vision: a computational investigation into the human representation and processing of visual information*. W. H. Freeman, San Francisco, 1982. 1.1, 2

[MBO06] A.S. Mian, M. Bennamoun, and R. Owens. Automatic 3D face detection, normalization and recognition. In *Third International Symposium on 3D Data Processing, Visualization and Transmission (3DPVT)*, pages 735–742, 2006. 2, 2.4.3

[MBO08] A.S. Mian, M. Bennamoun, and R. Owens. Keypoint detection and local feature matching for textured 3D face recognition. *International Journal of Computer Vision (IJCV)*, pages 1–12, 2008. 6.3.1

[MBT96] T. Molet, R. Boulic, and D. Thalmann. A real-time anatomical converter for human motion capture. In *7th EUROGRAPHICS International Workshop on Computer Animation and Simulation*, pages 79–94. EGCAS, Springer-Verlag Wien, 1996. 2

[MDSB02] M. Meyer, M. Desbrun, P. Schroder, and A. Barr. Discrete differential geometry operators for triangulated 2-manifolds. In *Proc. VisMath*, 2002. 2, 2.1.1

[ME03] B. Ma and R. E. Ellis. Robust registration for computer-integrated orthopedic surgery: Laboratory validation and clinical experience. *Medical Image Analysis*, 7(3):237–250, 2003. 4.3.1

[MG01] T.B. Moeslund and E. Granum. A survey of computer vision-based human motion capture. *Computer Vision and Image Understanding: CVIU*, 81(3):231–268, 2001. 2, 2.2

[MLM01] D. Marshall, G. Lukacs, and R.R. Martin. Robust segmentation of primatives from range data in the presence of geometric uncertainty. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(3):304–314, 2001. 4.3.2

[MS04] P. Monnier and S.K. Shevell. Influence of motion on chromatic detection. *Visual Neuroscience*, 21(3):327–330, 2004. 2.3.2

[MSJV03] A.B. Moreno, A. Sánchez, and F.J. Dí'az J.F. Vélez. Face recognition using 3D surface-extracted descriptors. In *Proc. of the Irish Machine Vision and Image Processing Conference*, 2003. 2, 2.4.3, 2.5

[MSR07] E. Magida, O. Solde, and E. Rivlin. A comparison of gaussian and mean curvature estimation methods on triangular meshes of range image data. *Computer Vision and Image Understanding*, 107(3):139–159, 2007. 6.3.1

[MT95] T. McInerney and D. Terzopoulos. A dynamic finite element surface model for segmentation and tracking in multidimensional medical images with application to cardiac 4D image analysis. *Computerized Medical Imaging and Graphics*, 19(1):69–83, 1995. 2, 2.3.1

[MTL+06] D. Metaxas, G. Tsechpenakis, Z. Li, Y. Huang, and A. Kanaujia. Dynamically adaptive tracking of gestures and facial expressions. In *Proc. International Conference on Computational Science*, pages 554–561, 2006. 5.3.2

[MV96] A. M. McIvor and R. J. Valkenburg. Principal frame and principal quadric estimation. *Image and Vision Computing New Zealand*, pages 55–60, 1996. 2, 2.1.1, 4.3.2, 5.4.2

[NA02] J. Neumann and Y. Aloimonos. Spatio-temporal stereo using multi-resolution subdivision surfaces. *International Journal of Computer Vision*, 47(1/2/3):181–193, 2002. 2, 2.2.2

[NLB+05] R. Ng, M. Levoy, M. Bredif, G. Duval, M. Horowitz, and P. Hanrahan. Light field photography with a hand-held plenoptic camera. Technical report, Stanford University, 2005. 6.3.1

[NM65] J.A. Nelder and R. Mead. A simplex method for function minimization. *Computer Journal*, 7:308–313, 1965. 4.3.2

[NS02] J.-C. Nebel and A. Sibiryakov. Range flow from stereo-temporal matching: Application to skinning. In *IASTED Int. Conf. on Visualization, Imaging, and Image Processing*, 2002. 2, 2.2.2, 3.2

[PFS+05] P.J. Phillips, P.J. Flynn, T. Scruggs, K.W. Bowyer, J. Chang, K. Hoffman, J. Marques, J. Min, and W. Worek. Overview of the face recognition grand challenge. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 947–954, June 2005. 2, 2.4.3

[PH06] N.E. Pears and T. Heseltine. Isoradius contours: New representations and techniques for 3D face matching and registration. In *Third International Symposium*

on 3D Data Processing, Visualization and Transmission (3DPVT), pages 176–183, 2006. 2, 2.4.3

[PHCP03] F. E. Pollick, H. Hill, A. Calder, and H. Paterson. Recognising facial expression from spatially and temporally modified movements. *Perception*, 32(2):813–826, 2003. 2, 2.4.1

[Pig99] F. Pighin. *Modeling and Animating Realistic Faces Images from Images*. PhD thesis, University of Washington, 1999. 5.4.1

[PKS+03] D.L. Page, A. Koschan, S.R. Sukumar, B. Roui-Abidi, and M.A. Abidi. Shape analysis algorithm based on information theory. In *Proceedings of the IEEE International Conference on Image Processing*, volume 1, pages 229–232, 2003. 2, 2.1.1

[PP01] S-I Pai and J. Ponce. On computing structural changes in evolving surfaces and their appearance. *International Journal of Computer Vision*, 43(2):113–131, 2001. 2, 2.1.3

[PR00] M. Pantic and Leon J. M. Rothkrantz. Automatic analysis of facial expressions: The state of the art. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(12):1424–1445, 2000. 2, 2.4.2

[PR04] T. Papatheodorou and D. Rueckert. Evaluation of automatic 4D face recognition using surface and texture registration. In *Proceedings of the Sixth IEEE International Conference on Automatic Face and Gesture Recognition*, pages 321–326, May 2004. 2, 2.4.3

[PWY+06] H. Pottmann, J. Wallner, Y.-L. Yang, Y.-K. Lai, and S.-M. Hu. Principal curvatures from the integral invariant viewpoint. *Computer Aided Geometry and Design*, 24(8-9):428–442, 2006. 6.3.1

[RCMB+06] S. Ruiz-Correa, M. Meila, G. Berson, M.L. Cunningham, and R.W. Sze. Symbolic signatures for deformable shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(1):75–90, 2006. 2, 2.1.3

[RD07] D. Riccio and J.-L. Dugelay. Geometric invariants for 2D/3D face recognition. *Pattern Recognition Letters*, 28(14):1907–1914, 2007. 6.3.1

[RFWA00] C. Robertson, R. B. Fisher, N. Werghi, and A. P. Ashbrook. Fitting of constrained feature models to poor 3D data. In *Proc. Adaptive Computing in Design and Manufacture (ACDM 2000)*, pages 149–160, 2000. 4.3.2

[RG78] V.S. Ramachandran and R.L. Gregory. Does colour provide an input to human motion perception. *Nature*, 275:55–57, 1978. 2.3.2

[SB02] H. Spies and J.L. Barron. Evaluating the range flow motion constraint. In *16th Internatioal conference on Pattern Recognition*, volume 3, pages 517–520, 2002. 2, 2.3.2

[SH05] J. Starck and A. Hilton. Spherical matching for temporal correspondence of non-rigid surfaces. In *IEEE International Conference on Computer Vision (ICCV)*, pages 1387–1394, 2005. 2, 2.3.1

[Sha48] C.E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423 & 623-656, 1948. 3.3

[Sim94] E.P. Simoncelli. Design of multi-dimensional derivative filters. In *IEEE Int. Conf. Image Processing*, volume 1, pages 790–793, 1994. 2, 3.2, 3.3.3

[SJB00] H. Spies, B. Jähne, and J. Barron. Dense range flow from depth and intensity data. In *ICPR*, pages 131–134, 2000. 2, 2.3.2, 3.2, 3.3, 3.3.2

[SJB02] H. Spies, B. Jähne, and J. L. Barron. Range flow estimation. *Computer Vision Image Understanding*, 85(3):209–231, 2002. 2, 2.3.2, 2.5, 3.3.2, 3.4, 6.1.3

[SMH06] J. Starck, G. Miller, and A. Hilton. Volumetric stereo with silhouette and feature constraints. In *Proc. British Machine Vision Conference*, volume 3, pages 1189–1188, 2006. 6.3.1

[SP86] T.W. Sederberg and S.R. Parry. Free-form deformation of solid geometric models. *Computer Graphics*, 20(4):151–159, 1986. 2.1.2

[SSD06] C. Samir, A. Srivastava, and M. Daoudi. Three-dimensional face recognition using shapes of facial curves. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(11):1858–1863, November 2006. 2, 2.4.3

[SU94] J.P. Siebert and C.W. Urquhart. C3D: a novel vision-based 3D data acquisition system. In *Proc. Mona Lisa European Workshop, Combined Real and Synthetic Image Processing for Broadcast and Video Production*, 1994. 2, 2.2.1, 2.5, 3.1

[TCMS03] C. Theobalt, J. Carranza, M. Magnor, and H.-P. Seidel. Enhancing silhouette-based human motion capture with 3D motion fields. In *Proc. IEEE Pacific Graphics*, 2003. 2

[TF00] J. B. Tenenbaum and W. T. Freeman. Separating style and content with bilinear models. *Neural Computation*, 6(12):1247–1283, 2000. 2, 2.4.2, 6.3.3

[TGA05] V. Tsagaris, G. Ghirstoulas, and V. Anastassopoulos. A measure for evaluation of the information content in colour images. In *IEEE International Conference on Image Processng*, volume 1, pages 417–420, 2005. 3.3

[TGS01] L.V. Tsap, D. B. Goldgof, and S. Sarkar. Fusion of physically-based registration and deformation modeling for nonrigid motion analysis. *IEEE Transactions on Image Processing*, 10(11):1659–1669, 2001. 2, 2.1.2

[TK91] C. Tomasi and T. Kanade. Detection and tracking of point features. Technical Report CMU-CS-91-132, Carnegie Mellon University, 1991. 5.3.1, 5.3.1

[TV96] N.F. Troje and T. Vetter. Representations of human faces. Technical Report 41, Max Planck Institute for Biological Cybernetics, 1996. 2, 2.4.3

[VBK02] S. Vedula, S. Baker, and T. Kanade. Spatio-temporal view interpolation. In *Proceedings of the 13th ACM Eurographics Workshop on Rendering*, pages 65–76, 2002. 2, 2.3.2

[VBR+99] S. Vedula, S. Baker, P. Rander, R. Collins, and T. Kanade. Three-dimensional scene flow. In *Int. Conf. on Computer Vision (2)*, pages 722–729, 1999. 3.2

[VBR+05] S. Vedula, S. Baker, P. Rander, R. Collins, and T. Kanade. Three-dimensional scene flow. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(3):475–480, 2005. 2, 2.3.2

[vDK84] A.J. van Doorn and J.J. Koenderink. Spatiotemporal integration in the detection of coherent motion. *Vision Research*, 24(1):47–53, 1984. 2.3.2

[VTS06] F. Vadakkumpadan, Y. Tong, and Y. Sun. Elastic surface registration by parameterization optimization in spectral space. In *Proceedings of Computational Imaging IV*, pages 321–329, 2006. 2, 2.3.1

[Wat00] A. Watt. *3D Computer Graphics*. Addison-Wesley, third edition, 2000. 2, 2.1.1

[WBCB08] C. Wallraven, M. Breidt, D.W. Cunningham, and H.H. Bülthoff. Evaluating the perceptual realism of animated facial expressions. *TAP*, 4(4):1–20, 2008. 6.3.1

[WCP00] C.R. Wren, B.P. Clarkson, and A.P. Pentland. Understanding purposeful human motion. In *Proceedings of the Fourth IEEE International Conference on Automatic Face and Gesture Recognition*, pages 19–25, 2000. 2

[WGZ+05] Y. Wang, M. Gupta, S. Zhang, S. Wang, X. Gu, D. Samaras, and P. Huang. High resolution tracking of non-rigid 3D motion of densely sampled data using harmonic maps. In *Proc. International Conference on Computer Vision*, volume 1, pages 388–395, 2005. 2, 2.3.1, 2.5, 5.5, 6.3.1

[WHL+04] Y. Wang, X. Huang, C.S. Lee, S. Zhang, Z. Li, D. Samaras, D. Metaxas, A. Elgammal, and P. Huang. High resolution acquisition, learning and transfer of dynamic 3D facial expressions. *Eurographics*, 23(3):677–686, 2004. 2, 2.2.2

[WP95] A. Witkin and Z. Popović. Motion warping. *Computer Graphics*, 29:105–108, 1995. 2

[WPWW06] Y. Wang, G. Pan, Z. Wu, and Y. Wang. Exploring facial expression effects in 3D face recognition using partial ICP. In *ACCV06*, volume I, pages 581–590, 2006. 6.3.1

[YB07] P. Yan and K.W. Bowyer. A fast algorithm for ICP-based 3D shape biometrics. *Computer Vision and Image Understanding*, 107(3):195–202, 2007. 6.3.1

[YBBR93] M. Yamamoto, P. Boulanger, J.-A. Beraldin, and M. Rioux. Direct estimation of range flow on deformable shape from a video rate range camera. *Pattern Analysis and Machine Intelligence*, 15(1):82–89, 1993. 2, 2.3.2, 3.2

[YD93] Y. Yacoob and L. Davis. Labeling of human face components from range data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 592–593, 1993. 2, 2.4.3

[YHR04] I. A. Ypsilos, A. Hilton, and S. Rowe. Video-Rate capture of dynamic face shape and appearance. In *IEEE International Conference on Automatic Face and Gesture Recognition*, pages 117–122, 2004. 2, 2.2.2

[YRC+97] A.W. Young, D. Rowland, A.J. Calder, N.L. Etcoff, A. Seth, and D.I. Perrett. Facial expression megamix: Tests of dimensional and category accounts of emotion recognition. *Cognition*, 63(3):271–313, 1997. 2, 2.4.1

[ZSCA04] M. Zhili, J.P. Siebert, W.P. Cockshott, and A.F. Ayoub. Constructing dense correspondences to analyze 3D facial change. In *Proc. 17th International Conference on Pattern Recognition*, volume 3, pages 144–148, 2004. 2, 2.3.1, 2.5, 5.1

[ZSCS04] L. Zhang, N. Snavely, B. Curless, and S.M. Seitz. Spacetime faces: High-resolution capture for modeling and animation. In *ACM SIGGRAPH'04*, pages 548–558, 2004. 2, 2.2.2, 2.5, 3.5