Queen Mary
University of London

Deformable 3-D Modelling from Uncalibrated Video Sequences

Del Bue, Alessio

For additional information about this publication click this link.
http://qmro.qmul.ac.uk/jspui/handle/123456789/5058

Information about this research object was correct at the time of download; we occasionally make corrections to records, please therefore check the published record when citing. For more information contact scholarlycommunications@qmul.ac.uk

ISSN 1470-5559

Deformable 3-D Modelling from Uncalibrated Video Sequences

Alessio Del Bue

RR-06-11 August 2006 Department of Computer Science

Deformable 3-D Modelling from Uncalibrated Video Sequences

Alessio Del Bue

Submitted for the degree of Doctor of Philosophy

Queen Mary, University of London

2006

Deformable 3-D Modelling from Uncalibrated Video Sequences

Alessio Del Bue

Abstract

The rigidity of a scene observed by a camera is often the fundamental assumption used to infer 3-D information automatically from the images taken by that camera. However, a video sequence of a natural scene often contains objects that modify their topology (for instance, a smiling face or a beating heart) thus violating the rigidity assumption necessary to reconstruct the 3-D structure of the object. In this thesis, we address the challenging problem of recovering the 3-D model of a deforming object and the motion of the camera observing it purely from image sequences, when nothing is known in advance about the observed object, the internal parameters of the camera or its motion.

Previous solutions to this *non-rigid structure from motion* problem have either provided approximate solutions using linear approaches to a problem that is intrinsically non-linear or required strong assumptions about the nature of the 3-D deformations. In this thesis, we propose a non-linear framework based on bundle adjustment to estimate model and camera parameters. We then upgrade the proposed framework to deal with the case of a stereo camera setup. We show that when the deforming object is not performing a significant overall rigid motion a monocular approach leads to poor reconstructions, and only by fusing the information from both cameras can the correct 3-D shape be extracted.

However, the problem of 3-D reconstruction of deformable objects is still fundamentally ambiguous: given a specific camera motion, different non-rigid shapes can be found that fit the observed 2-D image data. In order to reduce this effect, we introduce shape priors based on the observation that often not all the points on a deforming object are moving non-rigidly but some tend to lie on rigid parts of the structure. First, we propose motion segmentation algorithms to divide the scene automatically into the rigid and non-rigid point sets. Secondly, we use this information to provide priors on the degree of deformability of each point. Crucially all the above methods only work under the assumption of orthographic viewing conditions. Perhaps the most valuable contribution of this thesis is to provide a new algorithm to obtain metric reconstructions of deformable objects observed by a perspective camera.

Submitted for the degree of Doctor of Philosophy

Queen Mary, University of London

2006

Contents

List of Figures

List of Tables

Acknowledgements

The fulfilment of this thesis is due to a lot of people who as well had made my stay in this department an unforgettable experience. I would always be grateful I have been offered a Ph.D. with my supervisor, Dr. Lourdes Agapito. I am even more grateful I have accepted this opportunity. Under her guidance I moved my first steps in research and learned important lessons precious for my future. Every achievement I will obtain in this field, it will be always thanks to her teachings during these three years. Besides, she is a person I would be very fortunate to meet outside the strict context of Ph.D. supervision. I am undoubtedly a lucky person. A special thank goes to Prof. Edmund Robinson, our head of the department, for supporting me during these years in the department.

Being in the Vision group at Queen Mary was an excellent experience, especially for the people, and there are many to thank. First, I have to acknowledge the ”old guard” of Ph.D. students who have welcomed me in the department: Andi, Keith, Hayley, Adam, Andrew, Jeff and Lukas. A particular mention to Hayley, thank you for all the origamis, even if they ”strangely” focused on a single animal speciemen. Regarding the ”new guard” of Ph.Ds, I will miss having a chat in our tea room with Alex, Kevin, Dave, Chris, Milan, Tristan, Jun Li and Bushra. Tony was always ready to help me when I had a question (research related or not), Melanie was a very warm person to talk with and Xavier a person with whom I could chat easily about our common research. I thank him as well for being a good friend and for his continuous help during my last months in the department. I thank Sean for always sharing his interesting point of view about research in computer vision. Fabrizio for being more than a colleague (a flatmate as well). Peter for helpful advices and Pengwei for discussions about China and matrix factorization (what a weird combination!).

However, the Vision group is not the whole department. I thank our small (but strong!) Italian community: Pasquale, Dino, Kurt, Bellin, Corrado. We had a good time tasting some coffee at the Gaggia machine. Finally I will certainly miss all the people from the fourth floor: Jose, Tassos, Jie, Ivana, Josh, Lou, George, Jean-Baptiste (the next Vicon master), Theodora,

Gabriella and Shahin. I am grateful to Tim, Matt, Derek, Dave, Collin and Keith for solving all the software (and sometime hardware) problems I have created. For the everyday problems Joan, Gill, Carly, Carla and Sue were fundamental to make my life easier. A special thank to Chris for all the tea.

Outside this island, I am very grateful I had the chance to collaborate with Enrique, Jose and Luis from the department of artificial intelligence. We have conducted exciting research together and, actually, some of the ideas presented in this thesis were inspired from our discussions. Sure I am in debt with you.

Finally, I would like to express my gratitude to my parents who are nearly resigned to have their two sons roaming around the world. I thank them for being so patient and for constantly believing in me. I wish my achievements will make them proud.

Chapter 1

Introduction: Deformable Structure from Motion

One of the central interests of the Computer Vision community in recent years has been the inference of 3-D information about the world directly from image sequences taken from a moving video camera, when the specific details of the camera and its motion are all not known in advance. Such free-form inference can only succeed if certain assumptions are made, the standard one being that the scene observed by the camera is rigid: its geometry is static and the only motion is that of the camera. However, deformations which vary the structure of a shape are, on the other hand, constantly appearing. The human body itself is a remarkable example; muscles and bones stretch and tend the skin of the face to perform an incredible variety of expressions. Even at the organic level shapes are far from being rigid: hearts beat and lungs are continuously inflating and deflating. In this thesis we explore the challenging case of scenes that are not completely rigid, but which have certain degrees of flexibility or deformation.

The problem of 3-D inference from image sequences, generically known as structure from motion, was originally considered in the context of mobile robots, which carry cameras when navigating in cluttered environments and use the data received to build maps of their surroundings and improve their movement estimates. However, the algorithms developed have actually found more immediate demand in areas such as multimedia, the entertainment industry, and medicine. To address the problem of 3-D reconstruction from video sequences of non-rigid scenes, we will relax the previous assumption of a static world and instead aim to recover not only the essential shape of objects but also information about their deformation.

The approach used in this thesis will extend recent work in non-rigid factorization [19, 16,

141], which has demonstrated that it is possible under certain viewing conditions to infer the principal modes of deformation of an object alongside its 3-D shape within a structure from motion estimation framework. The models recovered by these algorithms, can subsequently be used as compact representations of the objects suitable for use in tracking, animation or other analysis. There have been other computer vision systems able to build similar morphable 3-D models of non-rigid objects. However, most of them rely on having additional information — for instance depth estimates available from 3-D scanning devices [151] — or have been refined to represent the specific object under observation: for example physically-based human face models [41]. Crucially, factorization methods work purely from video in an unconstrained case: a single camera viewing an arbitrary 3-D surface which is moving and articulating. Although there are no constraints as to the type of objects that may be modelled, this thesis has focussed mainly on the domain of human motion analysis — in particular 3-D reconstruction of facial motion.

1.1 Structure from motion: the rigid case

A camera is a projective device, which converts incoming rays of light into image positions depending only on the direction of the rays when they strike the lens: no information is gained directly about the depth of the objects viewed. To recover depth information, it is essential to make use of multiple images of an object from different viewpoints: if there is only one camera, it must move relative to the object. If the motion of the camera were known (for example if it were attached to a precisely-driven robot arm) then calculating depth would be a simple matter of triangulation. In most interesting scenarios, however, this is not the case: the camera motion itself is also uncertain. It was shown by [94] that in fact with certain assumptions it is possible to simultaneously estimate both the motion of a camera and the geometry of the scene it views. Structure from motion has since been defined as this problem of combined inference of the 3-D motion of a camera and the geometry of the scene it views solely from a sequence of images.

The underlying assumption which has allowed solutions to structure from motion to be achieved is that of scene rigidity: if objects are known not to change or deform, their shapes are invariant entities of which estimates can be gradually refined. In typical methods, large numbers of well-localised features of high image salience — usually *corner* points — are detected in each image of a video sequence. Postulating that each is associated with a repeatably identifiable 3-D entity in the environment, the features are then matched between each pair of consecutive (or

close) video frames. The assumption of rigidity in the scene [150] translates into mathematical constraints on the parameters describing camera motion, and many feature matches provide sufficient constraint equations such that solutions for both the motion and the locations of the 3-D features may be obtained.

There has been a great deal of work in rigid structure from motion in the last two decades. Of particular importance to its wide application has been the development of techniques which work even when the camera is uncalibrated: the specifics of its focal length and other internal parameters are not known in advance. These self-calibration algorithms, following on from the seminal work of Faugeras et al. [44], provide the flexibility of being applicable even in cases where little is known about the details of image capture. Solutions to the problem of self-calibration have been given in the case where camera motion is general — it exercises all of its degrees of freedom [60, 146] and also to more specific scenarios: where the camera is known only to rotate on the spot [63, 3], only to translate without rotation [105], or even where the camera has a zoom lens [119, 71], all of which call for slightly different algorithms which take account of this extra prior knowledge.

In a certain relatively common scenario — that when the range of depths of scene objects is much smaller than their distance from the camera — a linear approximation to camera geometry known as an affine projection is valid, and in this case a direct linear method for estimating camera motion and scene geometry over long image sequences can be used. Tomasi and Kanade's factorisation algorithm [138], developed in the early 90's, has been one of the most influential works in structure from motion. The algorithm takes a set of image coordinates of a number of features which can be matched in each image of a sequence of arbitrary length, and performs a direct singular value decomposition (SVD) to recover its affine shape and motion components, taking advantage of the bilinear form of the shape and motion parameters. The 2-D matches observed in an image sequence are stacked in an observation matrix which can be shown to have rank 3. It was consideration of such issues of rank which led to the realisation that not only rigid motion, but also a certain class of deformations could be dealt with within the factorisation framework, as will be discussed later.

1.2 Classification of non-rigid shapes

Biological shapes through their inherent nature are non-rigid. Soft tissues like the skin and muscles vary their shape under stress and pressure. This effect is clearly obvious when one examines the rich and complex set of expressions that can be exhibited by a human face simply by actuating different groups of muscles. These combinations of complex muscular actions have been studied and modelled with particular care, not only from a physiological point of view but also with the aim of creating realistic computer generated animations from which impressive results are available nowadays, as shown in figure 1.1.

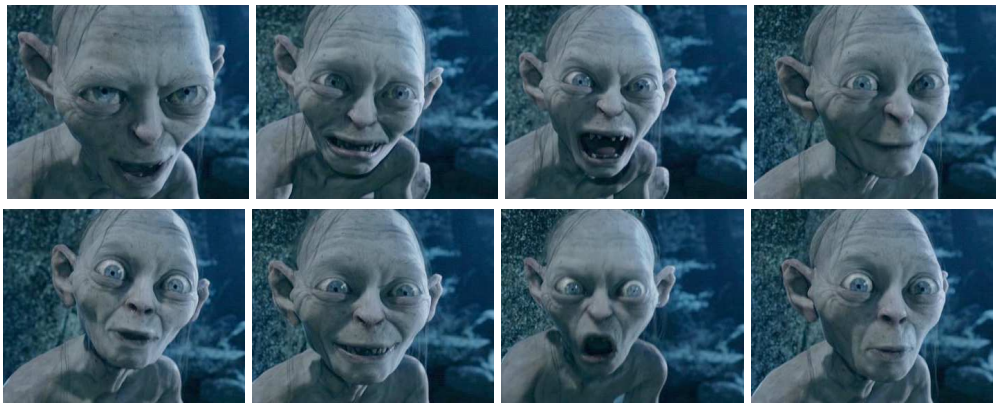


Figure 1.1: A computer animated character performing different facial expressions.

Similarly, non-rigidity and deformation are common properties of biological structures both at the cellular and organic levels. Cells may constantly vary their morphological structure under the effect of physical and chemical interactions. Figure 1.2 shows an example of the temporal evolution of a murine chondrocyte cell. On the other hand, organs may reveal interesting facts about their function with careful analysis of the the deformations that appear in their motion. For instance, anomalies in the heart may be detected by inspecting the repetitive pulse of the cardiac muscles.

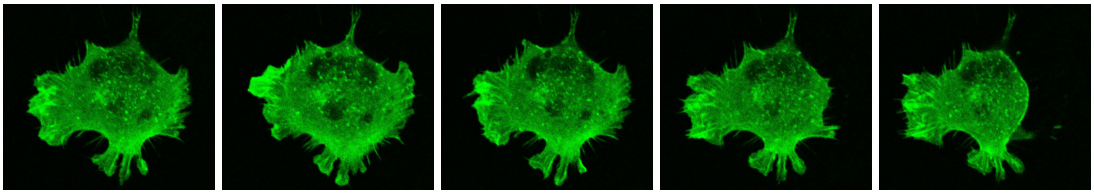


Figure 1.2: A live cell moving and deforming. The sample is taken from an immortilised strain of murine chondrocytes. The purpose of the experiment is to obtain live images of the varying cytoskeleton of the cell [89]. *Courtesy of Dr J. Campbell and Dr M. Knight.*

Given the wide range of possible degrees of non-rigidity present in nature, an effort to classify the type of motion of an object is necessary to understand which model may be applied to efficiently describe the shape variations. A first attempt was presented by Huang [75] and resulted in three generic classes of non-rigid motion:

- Articulated motion appears when an object is made up of a series of connected piecewise rigidly moving parts. A clear example is the collection of articulations of a human body connected by several joints with different degrees of freedom.

- Fluid motion is represented by structures which can freely vary their shape, such a flame, water flow or clouds. They involve strong topological changes in their structure such that they appear not to have any relevant continuity in their deformations.

- Elastic motion is distinguished from fluid motion by a continuity in the deformations that appear in the motion. The shapes presented in figure 1.1 and figure 1.2 show classic examples of elastic motion.

This classification has been further refined by Goldgof et al. [50] and Kambhamettu et al. [81] by introducing specific measures for the non-rigidity of the object. In the scope of this work, we focus particularly on elastic motion that will be referred to more generally as deformable motion throughout this thesis.

Given a deformable motion, our aim is to estimate the underlying 3-D structure of the inspected object using a structure from motion approach to the analysis of the image data. Thus, we seek a description of the visual motion in terms of a deformable 3-D model and the global rigid transformations that affect the shape. Given the complexity of the problem, deformable 3-D models have been studied extensively over the last two decades for the purposes of detecting, tracking and analysing the non-rigid motion appearing in an image. Before introducing the 3-D deformable model used in this thesis, we proceed first with a general description of different non-rigid models that have been proposed within the Computer Vision community.

1.3 Deformable shape models

As previously stated, a deformable object is a shape which varies its topology with continuity. Accordingly, a deformable model of an object is one which has parameters describing not just its shape but also the possible ways that the shape can change. Consider a graphical model of a

human skeleton to be used in animation: by setting the values of a list of parameters corresponding to the angles at each of its joints, it can be put into different configurations. In computer vision, there has been a large amount of work involving deformable models of objects, but with certain restrictions. In many cases the models used have been defined by hand (perhaps benefiting from some automatic refinement) or using sensing systems other than monocular vision, such as 3-D scanning devices [151], structured light, markers, calibrated stereo [49] or multiview reconstruction [115]. Often the models are specialised to represent specific types of objects. As an example, elaborate 3-D face models have been constructed to obtain reliable face tracking systems [32, 72, 116, 125] and, for instance, in the domain of medical image analysis, complex models of the left ventricle of the heart have been applied to diagnose heart conditions [4, 48].

In an attempt to classify deformable shape models, the literature generally identifies three main categories according to the mathematical description used to represent the deforming structure:

- Parametric deformable models. The non-rigid object shape is modelled by a set of parameters which explicitly vary the structure of a contour/surface. Parametric models are generally constructed a priori to suit the specific type of deforming shape (i.e. human faces, hearts, cells, etc).

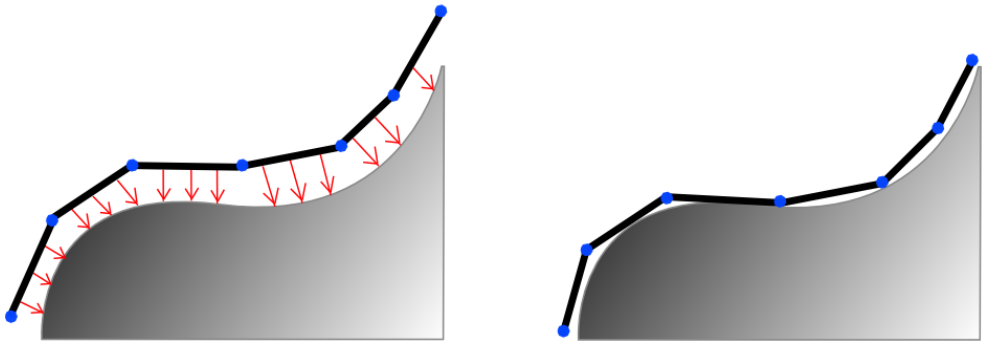
- Implicit deformable models. A specific deformation is represented as a function that is directly estimated from the image data. This function is defined as a level-set of a higher dimensional scalar function whose levels can adapt to a larger range of deformations.

- Generative models. The model is extracted using statistical techniques from a large collection (data-set) of examples showing all the possible changes in topology of the object. The model is therefore a compact description of the given data-set.

These models have been successfully applied to different domains of image analysis, detection, tracking and recognition of deformable shapes. In the following sections we present detailed descriptions and relevant examples of each class.

1.3.1 Parametric deformable models

Kass et al. [86] were the first to introduce 2-D parametric deformable models succesfully in an image analysis domain. The problem addressed was to estimate the shape of a deformable



(a) (b)

Figure 1.3: An example of a *snake* contour (in black) with control points (in blue) which define its shape. A *snake* is moved to match the image contour of the object (in grey) using external forces (red arrows) which attract the model to image edges as shown in (a). Internal forces assure the smoothness of the contour whose position is iteratively estimated until convergence (b).

object using a parameterized planar contour called *snake*. The parameters of the model are estimated such that the snake fits the deforming image contour accurately. In order to compute the parameters, the algorithm gradually iterates to fit the *snake* to the deformable shape under the influence of external image forces (for instance, image edges) and internal forces given by smoothness constraints of the model as shown in figure 1.3. A 2-D contour is easily generalizable to deal with 3-D images, leading to the definition by Cohen [22] of a *balloon*. Further research improved the performance of *snake* curves introducing robustness to the measured image data [136] and specific priors over the modelled objects [123, 165], resulting in a very successful approach for medical applications.

Another class of parametric models which has received considerable attention in the past is the family of shapes called *superquadrics* [114, 56, 103, 45, 15, 104]. Initially introduced in computer graphics by Barr [7], *superquadrics* are essentially derived from the parametric

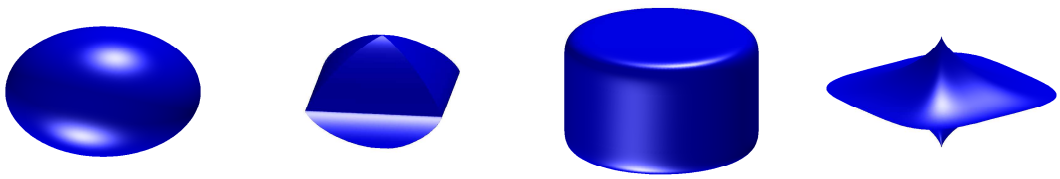


Figure 1.4: An example of different *superquadrics* ellipsoids used to model deforming shapes in images. The different shapes are obtained by varying the parameters of the mathematical model.

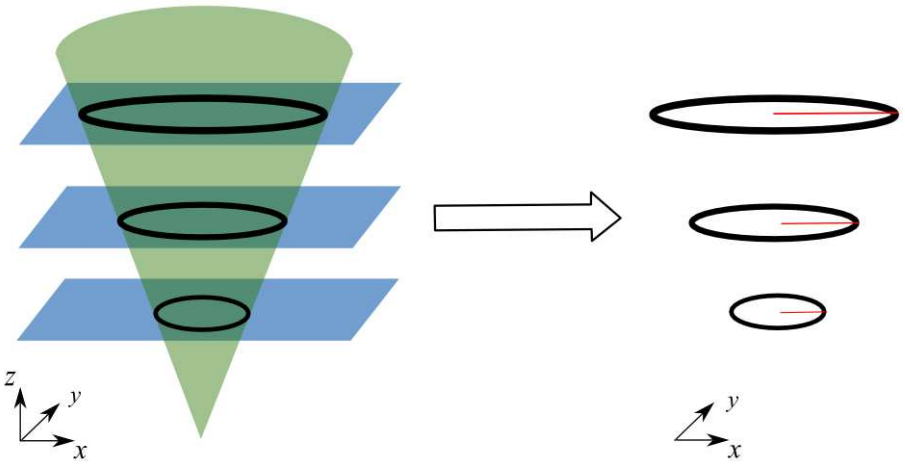


Figure 1.5: A simple level-set parameterisation of a circle contour using a cone as the scalar function. Circles of different radius (right) can be obtained by intersecting a cone (left) with planes (the levels) at different heights.

forms of quadric surfaces (see figure 1.4 for an example), and are used to fit a deformable object globally. However, they are not very accurate in describing natural shapes since the quadric surfaces may result in too coarse an approximation of the real shape.

1.3.2 Implicit deformable models

A crucial drawback of parametric deformable models is their difficulty to adapt to unexpected changes in the modes of variation in the given image data. The contour of a *snake* model has to be constructed accurately to be able to cope with all possible deformations of an object. However, in some cases, complete knowledge of all the possible shape variations is not available in advance. For instance in the medical domain, diseases like a tumor may change the structure of organs and cells unexpectedly. If the parameterized model does not account for these deformations, the result of the fitting procedure will be inaccurate.

A formulation of deformable models without an explicit parametrization of the shape was introduced by Osher and Sethian [111] using front propagation. In this approach, the deforming shape (or contour) is considered as a particular level-set of a scalar function. Thus each level corresponds to a particular deforming surface/contour which has to be fitted to the image data (see figure 1.5 for an example). Since the level-set approach does not rely on a fixed set of parameters but on a family of curves, the representation power of an implicit deformable model is higher than that of a parametric one.

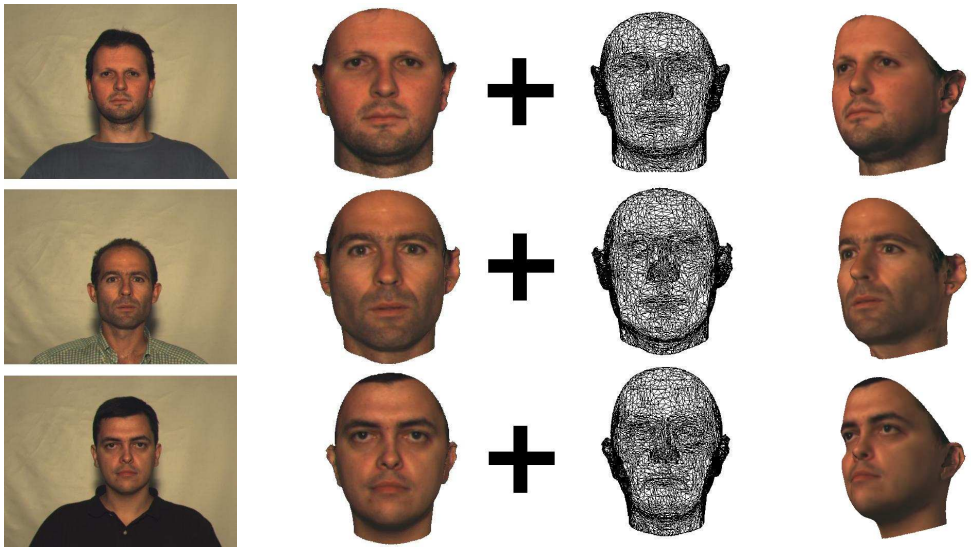
However, the computations necessary to estimate the level-set are costly and user interaction, if required, is problematic to implement since there are no evident parameters to select for driving the convergence. Nevertheless, implicit deformable models have been successfully applied in stereo vision [43], detection and tracking [112] and computer graphics [167].

1.3.3 Generative models

Generative or statistical models are obtained from a large set of observations of the deformations appearing in the inspected object. For instance, 2-D deformable models of faces have been generated from large training sets of images of different people with a range of expressions. These models, determined for example via principal component analysis (PCA) [27] take advantage of the fact that a head-on view of a face is reasonably approximated as a linear combination of the learned basis components. Such linear models have since been extended to cope with the non-linearities introduced by significant variations of face orientations and self-occlusion [52] and with local deformation [23] . However, they still suffer from requiring large amounts of specialised training data and can fail to encode non-linear deformations appearing from very different views. Other learned 2-D deformable models have included models of the outline or contour of moving human figures [11, 124, 9].

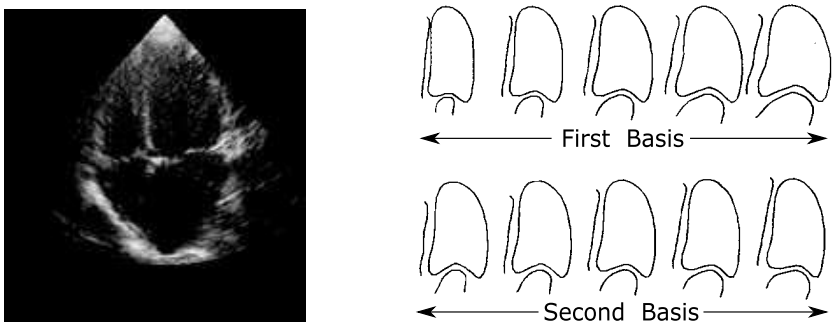
Specifically tuned to facial analysis, Vetter and Blanz [13, 12] have introduced elaborate techniques to create photorealistic 3-D morphable models. The shape and texture face model is derived using hundreds of 3-D laser scans of subjects of different age, sex and ethnic origin. After a preprocessing stage which cleans and aligns the mesh and texture information, the model is extracted from the collection of data using PCA producing a statistical description of the data-set in terms of linear basis shapes which represent the principal modes of variation of the model. The 3-D shape and texture model can then be applied to fit a new subject using a single image as input, as shown in figure 1.6.

Similarly, *active shape models* (ASM) [91, 25, 28, 90, 31] parameterize the 2-D shape variations of a deforming object. Each object is represented using a set of feature points which usually corresponds to key points on the object (such as the corners of the mouth or the eyebrows in the case of facial analysis). The shape is then described as a set of 2-D basis shapes which are fitted to obtain the principal modes of deformation of the large set of training image data (see figure 1.7 for an example). An advantage of such models is that they are obtained directly from images (there is no need for expensive instrumentation such as laser scanner). However, 2-D



(a) (b) (c)

Figure 1.6: The figure shows how 3-D morphable models can be used to recover the 3-D structure and texture of faces in images. The image data (a) is fitted using the linear basis description for the 3-D shape and 2-D texture in the model. The result (b) is the extracted 2-D texture from the image plus a 3-D mesh of the face which are combined to obtain the final 3-D fit of the face (c). *Courtesy of E. Muñoz, Dr J. M. Buenaposada and Dr L. Baumela*.



(a) (b)

Figure 1.7: An example of active shape models (ASM) (b) used to model the 2-D shape variations of the left ventricle of the heart in an echocardiogram (a). The left ventricle is located at the top right of figure (a). The ASM model consists of a set of 2-D basis shapes whose linear combination describes the deforming shape. Figure (b) presents the first two bases and their variations with respect to the mean shape of the ventricle. *Courtesy of Dr T. Cootes*.

models have difficulties in coping with strong pose variations which reduces the applicability of the approach. The shape of the non-rigid object appearing in a new image can then be fitted by computing the weights assigned to each basis which result in the best approximation of the object contour. Of more recent introduction, *active appearance models* (AAM) [24, 100] are a generalisation of active shape models which also incorporate the texture of the deforming shape to obtain a statistical description of the object's 2-D shape and appearance.

The advantage of statistical models compared to parametric ones is that they are derived solely from real observations and thus encapsulate the deformations that appear in the non-rigid object. Often an a priori model of a surface such as a *superquadric* has less representation power since it is not able to accurately describe a real world object. As a drawback, large data-sets from which to extract a comprehensive statistical model are not easy to collect.

1.4 A linear model for 3-D deformable shapes

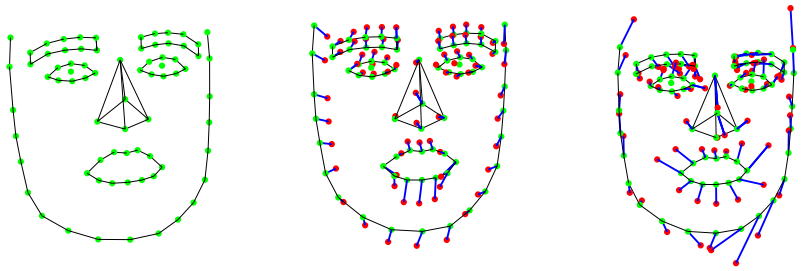
As we have stated in the previous section, it is possible to generate accurate statistical models from either 2-D or 3-D large collections of data. The approach followed in this thesis is of similar nature but differs in a fundamental aspect: given a set of 2-D image measurements extracted from an uncalibrated video sequence, we seek to obtain a full 3-D deformable model of the scene. Thus, the problem is not only restricted to the statistical inference of the 3-D non-rigid model from 2-D data but also to the estimation of the camera matrices which project the non-rigid object onto the image plane.

The shape at each time instance is formulated as a linear combination of a set of basis shapes which describe the principal modes of deformation of the 3-D structure. The model parameters, which we will refer to as configuration weights, are given by a set of scalars that provide the appropriate weight for each basis. In a geometric form, the 3-D shape is represented as a cloud of points lying over the deforming surface. Mathematically, the 3-D shape is represented as a matrix $\mathtt{S}$ which contains the 3-D coordinates for each point of the object. The deforming shape $\mathtt{S}$ at a certain frame is given by the linear combination of the basis shapes $\mathtt{S}_d$ weighted by the configuration weights l_d such that:

$$\mathtt{S} = \sum_{d=1}^{D} l_d \mathtt{S}_d \qquad \mathtt{S}, \mathtt{S}_d \in \Re^{3\times P} \quad l_d \in \Re \tag{1.1}$$

where D is the number of basis shapes and P the number of points in the model.

Each basis shape describes a particular mode of deformation of the object. For instance, in the face modelling domain, the basis shapes may represent specific facial expressions like surprise or a grin as presented in figure 1.8. Models created as a combination of a set of bases have been previously used in many applications ranging from facial analysis [26, 151], tracking [109, 49] and biomedical domains [79].



First Basis Second Basis Third Basis

Figure 1.8: An example of the linear pointwise model used in this thesis. The model is composed of a set of 3-D basis shapes which are defined by a collection of 3-D point coordinates. A deformation is represented as a linear weighted combination of the set of bases. The first basis usually represents a mean 3-D description of the shape (in green). The second and third bases are showed in the figure as a 3-D displacement (blue lines) from the mean component. The resulting structure given the displacement for each basis (red points) usually refers to dominant facial expressions (for instance, surprise and grin).

1.5 A factorization approach to 3-D deformable modelling

In this thesis we are interested in models which represent the full 3-D geometry of a deformable object, but in particular in acquiring these models automatically and only from images rather than having to use prior information or specialised sensors — a model free 3-D approach. The nature of this problem leads us back to the original structure from motion question: what can be determined about the motion of a camera and the 3-D non-rigid shape of the scene when no information about the camera or the structure is available?

Recent results have started to open up this research direction [19, 141, 16] proving that 2-D point tracks in an image sequence are sufficient to recover 3-D non-rigid shape and motion under the same affine viewing conditions in which Tomasi and Kanade's algorithm proved successful in the rigid case. This novel non-rigid factorization approach assumes that the 3-D non-rigid shape

can be represented by the linear model described in the previous section. Their insight was that since this representation is linear, it fits naturally into the factorisation framework. Once more the underlying geometric constraints are expressed as a rank constraint which is used to factorise the measurement matrix into two lower dimensional matrices that encode the motion and the shape of the object using singular value decomposition: $\mathtt{W} = \mathtt{M}\,\mathtt{S}$.

However, in common with all factorization methods, the result is not unique and there exists a full rank transformation matrix $\mathtt{Q}$ that gives the following alternative reconstruction: $\mathtt{W} = \mathtt{M}\mathtt{Q}\;\mathtt{Q}^{-1}\mathtt{S} = \tilde{\mathtt{M}}\tilde{\mathtt{S}}$. The fundamental problem is to find the transformation $\mathtt{Q}$ that imposes the correct structure on the camera matrices encoded in $\mathtt{M}$ and removes the ambiguity upgrading the reconstruction to a metric one. Whereas in the rigid case the problem of computing the transformation matrix $\mathtt{Q}$ to upgrade the reconstruction from affine to metric can be solved linearly [138], in the non-rigid case it results in a non-linear problem.

1.6 Motivations for this thesis

Existing non-rigid factorization methods are very promising and do indeed produce models from scratch that can be useful for tracking or animation in many domains, but there are various limitations which have led to interesting avenues of research in this thesis and have motivated our work. The improvements we have proposed to some of the outstanding problems constitute the main contribution of the work presented here. The three main issues which we have addressed in this thesis are: the non-linearity of the non-rigid structure from motion problem, its inherent ambiguous nature and the extension of the method to deal with perspective imaging conditions.

Firstly, previous solutions to the non-rigid structure from motion problem have either provided approximate solutions using linear approaches [19, 141, 16] to a problem that is intrinsically non-linear or required strong assumptions [17, 159, 161] about the nature of the 3-D deformations. The **non-linearity** of the problem stems from the fact that the parameters modelling the camera motion and the 3-D deformations are strongly coupled. Moreover, in order to obtain a valid solution, orthogonality constraints have to be forced on the rotational component of the motion, thus introducing a further degree of non-linearity. In this thesis, we propose a non-linear framework based on bundle adjustment to estimate model and camera parameters. The advantage of this method is that it provides a maximum likelihood estimate in the presence of Gaussian noise, and prior knowledge on any of the model parameters can easily be incorporated

into the cost function in the form of penalty terms. The proposed framework is then upgraded to deal with the case of a stereo camera setup. We show that when the deforming object is not performing a significant overall rigid motion, a monocular approach leads to poor reconstructions, and only by fusing the information from both cameras can the correct 3-D shape be extracted.

Secondly, non-rigid structure from motion continues to be an **inherently ambiguous** problem since the contribution to the image motion caused by the deformations and rigid motion are often difficult to disambiguate. Given a specific configuration of points on the image plane, different 3-D non-rigid shapes and camera motions can be found that fit the measurements. To solve this ambiguity prior knowledge on the shape and motion should be used to constrain the solution. Recently, Xiao et al. [159] proved that the orthogonality constraints were insufficient to disambiguate rigid motion and deformations. They identified a new set of constraints on the shape bases which, when used in addition to the rotation constraints, provide a closed form solution to the problem of non-rigid structure from motion. However, their solution requires that there be D frames (where D is the total number of basis shapes) in which the shapes are known to be independent.

In this thesis we propose an alternative approach based on the observation that often not all the points on a moving and deforming surface – such as a human face – are undergoing non-rigid motion. Frequently some of the points are on rigid parts of the structure – for instance the nose – while others lie on deformable areas. Intuitively, if a segmentation of points into rigidly moving and deforming ones is available, the rigid points can be used to estimate the overall rigid motion and to constrain the underlying mean shape by estimating the local deformations exclusively with the parameters associated to the non-rigid component of the 3-D model.

Finally, all the methods cited previously rely on affine imaging conditions in which the objects viewed are relatively flat and distant from the camera — they cannot cope with the **projective distortions** which become significant when the scene is closer (and focal lengths are shorter), as may often be the case with PC-mounted "webcam" devices viewing users' faces. Xiao and Kanade [161] were the first to develop a two step factorization algorithm for reconstruction of 3-D deformable shapes under the full perspective camera model. In this thesis we present an alternative approach to non-rigid shape and motion recovery under the full perspective camera model. Once more, the solution is based on the assumption that the scene contains a mixture of rigid and non-rigid points. First rigid and non-rigid motion segmentation is performed on the

image data to separate both types of motion under perspective imaging conditions. To obtain the metric upgrade information we perform self-calibration on the rigid set of points which provides estimates for the camera intrinsic parameters, the overall rigid motion and the mean shape. We then formalise the problem of non-rigid shape estimation as a constrained non-linear minimization using the estimates given by the self-calibration algorithm as the starting point for the minimization and providing priors on the degree of rigidity of each of the points.

1.7 Contributions of this Thesis

In the following section we describe the main contributions of this thesis, in accordance with the motivations exposed in the previous section:

- We propose a framework for non-linear estimation of the geometric parameters of the deformable model based on an adaptation of bundle adjustment techniques [38, 37] to the non-rigid scenario. The non-linear optimization method is able to refine the motion and shape estimates by minimizing image reprojection error, imposing the correct structure on the motion components by choosing an appropriate parameterisation.

- The non-linear framework can easily be modified to include views taken from different cameras. We have extended existing non-rigid factorization algorithms to the stereo camera case and presented an algorithm to decompose the measurement matrix into the motion of the left and right cameras and the 3-D shape [34, 33]. The added constraints in the stereo camera case are that both cameras are viewing the same structure and that the relative orientation between both cameras is fixed. Our focus is on the recovery of flexible 3-D shape rather than on the correspondence problem.

- We have proposed two methods for automatic rigid and non-rigid motion segmentation in the case of orthographic [35] and perspective [36] viewing conditions. In the affine case, our method follows a *sequential backward selection strategy* by initially considering all the trajectories in the measurement matrix and iteratively deleting the points that exhibit the most non-rigid motion. As the stop criterion for the classification task, the rank of the measurement matrix of the remaining points is computed, which will become 3 when only the rigid trajectories are left.

  In the case where perspective distortions affect the measurements, our approach is based on

the fact that rigid points will satisfy the epipolar geometry while the non-rigid points will give a high residual in the estimation of the fundamental matrix between pairs of views. We use a RANSAC algorithm to estimate the fundamental matrices and to segment the scene into rigid and non-rigid points. Additionally, we exploit a measure of the degree of deformability of a point to infer a prior distribution of the probability of a trajectory being rigid or non-rigid given that measure. These distributions are then used as priors to perform guided sampling over the set of trajectories and lower the number of random samples needed to be drawn from the data.

- The advantage of performing a prior segmentation of the image points into rigid and non-rigid trajectories is that this information can be used to constrain the solution of the shape and motion recovery. Firstly, the rigid points can be used to obtain an accurate initial estimate of the overall rigid rotations, translations and mean shape. Secondly, the knowledge that some points on the object do not deform can be used to impose priors on the non-rigid shape. Our prior expectation is that the points detected as being rigid have a zero non-rigid component and can therefore be modelled entirely by the first basis shape. We define linear and non-linear methods to impose these priors [35] and we show that it is possible to obtain exact reconstructions with noiseless data and improved reconstructions and a higher rate of convergence with real data.

- Finally, this thesis presents a novel approach for the 3-D Euclidean reconstruction of deformable objects observed by a full perspective camera [36, 93]. Given an automatic segmentation of the scene into rigid and non-rigid point sets, using the algorithm mentioned above, the set of rigid points is used to estimate the internal camera calibration parameters and the overall rigid motion. The problem of non-rigid shape estimation is then formalised as a constrained non-linear minimization adding priors on the degree of deformability of each point.

The contributions here exposed are presented in the thesis as follows. Chapter 2 is a literature review of the factorization framework for structure from motion recovery and its application to the case of rigid and non-rigid structure recovery under different viewing conditions. Chapter 3 describes our framework for non-linear estimation of the deformable model and camera parameters. The framework can easily deal with the case of two or more cameras as presented in

chapter 4. Chapter 5 describes the use of shape priors for deformable modelling in the case of affine viewing conditions. First we propose an automatic rigid/non-rigid motion segmentation algorithm. The results of the segmentation are then used to derive priors on the degree of deformability of each point in the 3-D object. Such priors are used to drive the inference of the parameters of the deformable model. In chapter 6 we propose a new solution to the problem of metric structure recovery from perspective images. A new rigid/non-rigid motion segmentation algorithm is derived which can deal with projective distortions. The structure and motion recovery is then formulated as a two step process where the metric upgrade transformation is computed first using the rigid points and the deformable structure is then estimated using a non-linear optimization approach. Chapter 7 ends this dissertation presenting aspects of the proposed methods which may lead to future improvements and further avenues of research in the domain of deformable modelling.

Chapter 2

Factorization methods for Structure from Motion

The geometry between two views taken either by a moving camera or by two different cameras is nowadays a well understood concept. The fundamental matrix is the mathematical tool that relates image coordinates between a pair of views [94, 95]. Similarly, three views are related by the trifocal tensor [130, 156, 61], which allows to transfer a point in the first and second view into the third view and, similarly, with lines. The constraints arising from four views of the same scene are encapsulated by the quadrifocal tensor [144].

These multi-view tensors are used as a first step to obtain an initial projective reconstruction of the 3-D shape of an object. However, while these inter-image relations are able to describe the constraints between views of the same scene, they are not always of practical use. A wide-baseline between views is necessary for the estimate of the multi-view tensors to be accurate. On the other hand, matching image points from very different views is a complex task that can easily lead to outliers in the data used for estimation.

Matching image features becomes relatively simple when the images are taken from closely spaced views. However, the overall small baseline affects the depth estimation of the structure negatively. In order to avoid critical configurations of views, the only possible solution is to have a large number of views for which the overall baseline is wide enough to allow an accurate 3-D reconstruction.

The described tradeoff is crucial for the 3-D reconstruction of generic objects observed from a video sequence. If we restrict the problem to the case of a single camera, the multiple views are given by a temporal sequence of images taken by a moving camera or by a fixed camera and

a moving object or by a combination of both. As a result, the distinguishable visual cue is the motion of the projected 3-D object in the sequence of images. In this case, exclusively using the information of two, three or four views may give poor results as previously noted. Thus, a solution which uses the whole information of the entire sequence is always preferable.

2.1 A factorization approach to Structure from Motion

In the early 90's Tomasi and Kanade [139] found an elegant and simple solution to this problem by analyzing the image measurements observed from different views using a weak perspective camera model. Since the motion of each point is globally described by a precise geometric model, the position of their projection on the image plane is constrained. As a result, if all the measurements (i.e. the image coordinates of all the points in all the views) are collected in a single matrix, the point trajectories will reside in a certain sub-space. The dimension of the sub-space in which the image data resides is a direct consequence of two factors: the type of camera that projects the scene (for instance, affine or perspective) and the nature of the inspected object (for instance, rigid or non-rigid).

The crucial advantage of this technique is in the fact that it provides an initial linear and global solution to the problem simply by factorizing the image measurements into the relative motion and 3-D structure using the aforementioned sub-space property of these measurements. This solution by factorization is given by the whole information of the measurements and solved using linear methods.

Given the success and flexibility of Tomasi and Kanade's bilinear formulation of the shape and motion components, we now describe the factorization approach and its application to different models of camera projection and types of object structure. Finally, we focus on existing non-rigid factorization approaches and point to some unsolved issues.

2.1.1 The factorization framework: *motion* and *3-D structure*

The rigid factorization method introduced by Tomasi and Kanade [139] is simple but powerful. It provides a description of the 3-D structure of a rigid object in terms of a set of feature points extracted from salient image features (for instance, image corners). After tracking the points throughout all the images composing the temporal sequence, a set of trajectories is available (see figure 2.1 for an example). These trajectories are constrained globally at each frame by the rigid

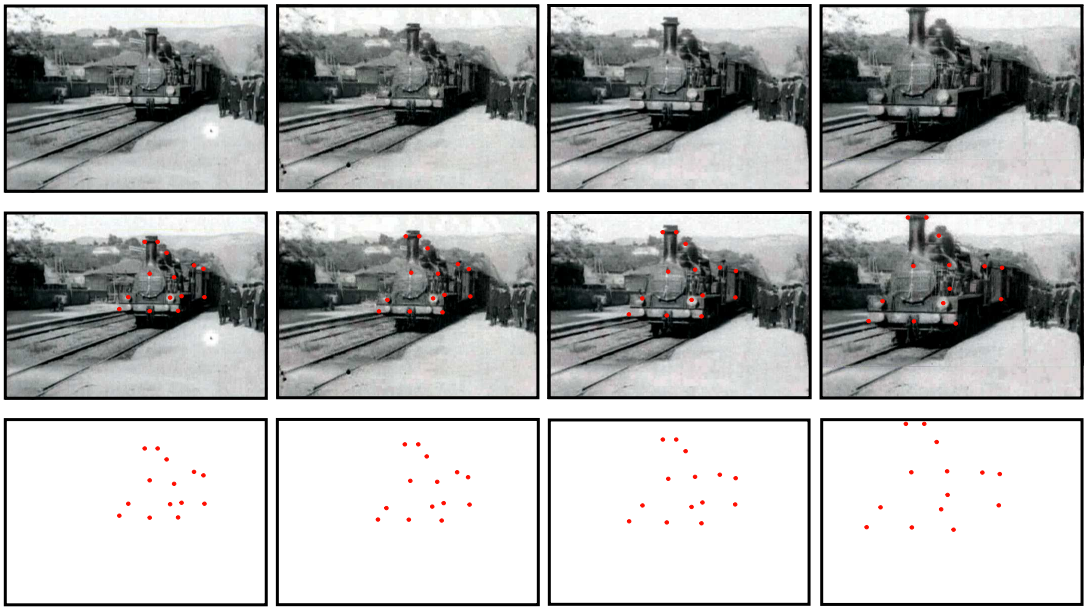


Figure 2.1: The figure shows how point trajectories are extracted from a video sequence. First row: four frames of the movie *L'Arrivée d'un train à la Ciotat* (1895, directed by Lumière brothers). Second row: the image points (in red) are extracted in the first frame and successively tracked in the following frames. Third row: the measured image data is shown on the image plane. Each point is defined by two image coordinates. The collection of the points at each frame composes a trajectory in time which describes the motion of the rigid point.

transformation which the shape is undergoing. Rigid factorization techniques directly factorize or decompose the complete collection of image trajectories into the bilinear components of *motion* and *3-D structure*. The role of the *motion* components is to project the *3-D structure* on the image plane for each frame using a particular camera model.

In order to describe the framework in detail, we need to introduce the formalised mathematical description of the trajectories that will subsequently be factorized. Once a trajectory is extracted, the location of a point j in a certain frame i can be defined as a non-homogeneous 2-vector $\mathbf{w}_{ij} = (u_{ij}\ v_{ij})^T$ or as a homogeneous 3-vector $\bar{\mathbf{w}}_{ij} = (u_{ij}\ v_{ij}\ 1)^T$ where u_{ij} and v_{ij} are the horizontal and vertical image coordinates respectively.

A compact representation of these elements can be expressed collecting all the non-homogeneous

coordinates in a single matrix, called the *measurement matrix* $\mathtt{W}$, such that:

$$\mathtt{W} = \begin{bmatrix} \mathbf{w}_{11} & \cdots & \mathbf{w}_{1P} \\ \vdots & \ddots & \vdots \\ \mathbf{w}_{F1} & & \mathbf{w}_{FP} \end{bmatrix} \tag{2.1}$$

$\mathtt{W}$ is a $2F \times P$ matrix where F is the number of image frames and P the number of trajectories extracted. Ideally, the measurement matrix should contain perfect information about the object being tracked. However, in practice the measurements are corrupted by noise and outliers given by mismatched points. Additionally, some elements of $\mathtt{W}$ may not be available for some points in particular frames due to occlusions. Nevertheless, we continue the factorization problem assuming there are no missing entries in $\mathtt{W}$.

It is possible to decompose $\mathtt{W}$ into the product of two matrices as:

$$\mathtt{W} = \mathtt{M}\,\mathtt{S} \tag{2.2}$$

where $\mathtt{M}$ and $\mathtt{S}$ are respectively the *motion* and *3-D structure* components of the measurement matrix. The matrices $\mathtt{M}$ and $\mathtt{S}$ can be further decomposed such that:

$$\mathtt{M} = \begin{bmatrix} \mathtt{M}_1 \\ \mathtt{M}_2 \\ \vdots \\ \mathtt{M}_F \end{bmatrix} \qquad \mathtt{S} = \begin{bmatrix} \mathtt{S}_1 & \mathtt{S}_2 & \cdots & \mathtt{S}_P \end{bmatrix} \tag{2.3}$$

where $\mathtt{M}_i$ with $i = 1 \ldots F$ is the camera matrix that projects the 3-D metric shape onto image frame i. The size and structure of $\mathtt{M}_i$ generally depends on the type of camera that projects the scene. The component $\mathtt{S}_j$ with $j = 1 \ldots P$ defines the 3-D structure for each point j and its size depends on the shape properties (for instance, whether it is rigid or non-rigid). The framework is such that the product $\mathbf{w}_{ij} = \mathtt{M}_i \mathtt{S}_j$ defines the projection of the point j onto the image frame i.

2.1.2 The rank of the measurement matrix

An interesting property of the measurement matrix is that it is rank-deficient and resides in a lower dimensional space. In fact the dimension is given by the size of the motion and structure matrices $\mathtt{M}$ and $\mathtt{S}$. This property was first used by Tomasi and Kanade [138] who first observed and exploited the rank deficiency of measurement matrices storing image trajectories extracted from a body undergoing a rigid transformation. Also known as the *rank constraint of a matrix*,

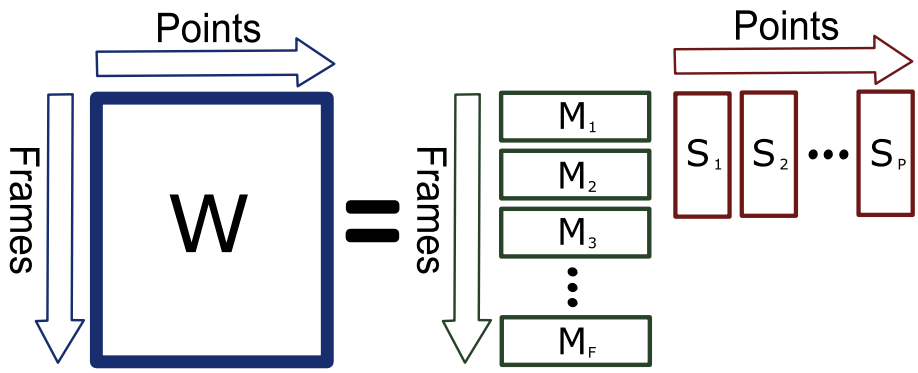


Figure 2.2: The measurement matrix $\mathtt{W}$ is decomposed into the product of the *motion* matrix $\mathtt{M}$ and the *3-D structure* matrix $\mathtt{S}$. The matrix $\mathtt{M}$ contains the parameters of the model that vary frame-wise (i.e. object motion and camera parameters) while $\mathtt{S}$ contains the parameterisation of the 3-D structure for each point. The sizes of $\mathtt{M}$ and $\mathtt{S}$ depend respectively on the camera model and the 3-D point parameterisation.

this property may be exploited by using common techniques for matrix factorization (see section 2.1.3) to reduce the dimensionality of the matrix $\mathtt{W}$ and factorize it into the product of $\mathtt{M}$ and $\mathtt{S}$.

Further studies of the factorization framework have shown that trajectories belonging to different deforming objects show similar rank constraints. Different ranks would be obtained depending on the model used for the camera models observing the scene [145, 132], considering different rigid objects moving independently [29], dealing with non-rigid objects [19] or articulated structures [143, 163, 107]. Moreover the rank constraint has been applied successfully in the work presented by Irani [77] to obtain an estimate of multi-frame optical flow for different camera models and types of motion.

2.1.3 Singular Value Decomposition (SVD) and factorization

The rank-constraint can be efficiently used to obtain a decomposition of $\mathtt{W}$ in terms of motion and structure. SVD is a rank revealing matrix decomposition algorithm that factorises a generic $H \times L$ matrix $\mathtt{W}$ into a product of 3 matrices:

$$\mathtt{W}_{H\times L} = \mathtt{U}_{H\times L}\, \Sigma_{L\times L}\, \mathtt{V}^{T}_{L\times L} \tag{2.4}$$

where Σ is a diagonal matrix whose entries are the singular values of $\mathtt{W}$, $\mathtt{U}$ is an $H \times L$ orthogonal matrix such that $\mathtt{U}\mathtt{U}^{T} = \mathtt{I}_{H\times H}$ and $\mathtt{V}$ is a square and orthogonal matrix such that $\mathtt{V}^{T}\mathtt{V} = \mathtt{V}\mathtt{V}^{T} = \mathtt{I}_{L\times L}$. The number of singular values different from zero reveals the actual rank of the data

stored in the measurement matrix, and they are ordered from largest to smallest according to their magnitude.

If the L columns of $\mathtt{W}$ are linearly dependent, each column can be obtained as a linear combination of a subgroup of r columns with $r \leqslant \min\{H, L\}$. The value r is also called the rank of a matrix and this property is directly related to the singular values in Σ such that:

$$d_{ii} = 0 \quad \forall\, i > r \tag{2.5}$$

where d_{ii} are the diagonal entries of Σ and $i = 1 \ldots L$. As a consequence of the zero-entries in Σ, equation (2.4) can be rewritten as:

$$\mathtt{W}_{H\times L} = \mathtt{U}_{H\times r}\, \Sigma_{r\times r}\, \mathtt{V}^{T}_{L\times r} \tag{2.6}$$

Here, $\mathtt{U}$ and $\mathtt{V}$ are orthogonal matrices defining respectively the *range* and *null space* of $\mathtt{W}$. By using the SVD, obtaining the closest rank-r matrix in terms of the Frobenius norm to the original matrix is guaranteed, if the noise contaminating $\mathtt{W}$ is isotropic and Gaussian [51].

2.2 Rigid factorization

A object moving rigidly enforces a rank constraint over the measurements extracted from the image sequence capturing the motion of the object. The given rank depends on the camera model used to project the 3-D structure in the image plane. The following sections show how factorization methods can extract 3-D structure from sequences viewed with orthographic and perspective cameras.

2.2.1 Rigid Structure under orthographic projection

The first use of the rank constraints to solve multi-view problems was introduced by Tomasi and Kanade [138] to deal with the case of rigid objects under orthographic camera projection. In this scenario, the measurement matrix consists of trajectories extracted from a single object undergoing rigid rotations and translations as showed in figure 2.1. For a single frame i, the measurements can be represented in matrix form such that:

$$\mathtt{W}_i = \begin{bmatrix} \mathbf{w}_{i1} & \ldots & \mathbf{w}_{iP} \end{bmatrix} \tag{2.7}$$

It is possible to obtain the measurement matrix $\mathtt{W}$ by stacking the $\mathtt{W}_i$ for all F frames:

$$\mathtt{W} = \begin{bmatrix} \mathtt{W}_1 \\ \mathtt{W}_2 \\ \vdots \\ \mathtt{W}_F \end{bmatrix} \tag{2.8}$$

A single point j belonging to a 3-D object in a generic frame i can be projected using an orthographic camera such that:

$$\mathbf{w}_{ij} = \begin{bmatrix} r_{i1} & r_{i2} & r_{i3} \\ r_{i4} & r_{i5} & r_{i6} \end{bmatrix} \begin{pmatrix} X_j \\ Y_j \\ Z_j \end{pmatrix} + \begin{pmatrix} t_{ui} \\ t_{vi} \end{pmatrix} = \mathtt{R}_i \mathbf{X}_j + \mathbf{t}_i \tag{2.9}$$

where $\mathtt{R}_i$ contains the first two rows of a rotation matrix, $\mathbf{X}_j$ is a 3-vector containing the metric coordinates of the 3-D point, and $\mathbf{t}_i$ is a vector representing the translation component. Every point belonging to the rigid structure shares the same rotation and translation. Thus, the previous expression is valid for every point in the generic frame i:

$$\mathtt{W}_i = \begin{bmatrix} \mathbf{w}_{i1} & \dots & \mathbf{w}_{iP} \end{bmatrix} = \begin{bmatrix} r_{i1} & r_{i2} & r_{i3} \\ r_{i4} & r_{i5} & r_{i6} \end{bmatrix} \begin{bmatrix} X_1 & X_2 & \cdots & X_P \\ Y_1 & Y_2 & \cdots & Y_P \\ Z_1 & Z_2 & \cdots & Z_P \end{bmatrix} + \mathtt{T}_i \tag{2.10}$$

where $\mathtt{T}_i$ is a $2 \times P$ matrix with the replicated translation vector $\mathbf{t}_i$ for each point. It is possible to rewrite the expression in a compact matrix form as:

$$\mathtt{W}_i = \mathtt{R}_i \mathtt{S} + \mathtt{T}_i \tag{2.11}$$

Stacking the rows of $\mathtt{W}_i$ for every frame we obtain the complete measurement matrix:

$$\mathtt{W} = \begin{bmatrix} \mathtt{W}_1 \\ \mathtt{W}_2 \\ \vdots \\ \mathtt{W}_F \end{bmatrix} = \begin{bmatrix} \mathtt{R}_1 \\ \mathtt{R}_2 \\ \vdots \\ \mathtt{R}_F \end{bmatrix} \begin{bmatrix} X_1 & X_2 & \cdots & X_P \\ Y_1 & Y_2 & \cdots & Y_P \\ Z_1 & Z_2 & \cdots & Z_P \end{bmatrix} + \begin{bmatrix} \mathtt{T}_1 \\ \mathtt{T}_2 \\ \vdots \\ \mathtt{T}_F \end{bmatrix} = \mathtt{M}\mathtt{S} + \mathtt{T} \tag{2.12}$$

where $\mathtt{W}$ is the $2F \times P$ measurement matrix, $\mathtt{M}$ is the $2F \times 3$ collection of F rotation matrices, $\mathtt{S}$ is the $3 \times P$ structure matrix containing the 3-D coordinates of all the world points and $\mathtt{T}$ is a $2F \times P$ matrix with the translation for each frame.

It is easy to eliminate the translation component by determining the centroid of the image points for every frame and subtracting it from the image coordinates. In this case, the components

$\mathtt{M}$ and $\mathtt{S}$ are matrices of at most rank 3, thus Tomasi and Kanade's algorithm obtains an initial decomposition by performing a truncated SVD with $r = 3$ such that:

$$\mathtt{W}_{2F\times P} = \mathtt{U}_{2F\times 3}\,\Sigma_{3\times 3}\,\mathtt{V}^T_{P\times 3} \tag{2.13}$$

It is then possible to rearrange the 3 products to obtain an initial affine estimation of the motion and structure components such that:

$$\tilde{\mathtt{M}} = \mathtt{U}\sqrt{\Sigma} \;\text{ and }\; \tilde{\mathtt{S}} = \sqrt{\Sigma}\mathtt{V}^T \tag{2.14}$$

One important aspect that should be emphasized is that rank revealing numerical techniques, such as SVD, do not provide the solution to the 3-D reconstruction problem [85]. The reason is that the rank-3 decomposition is not unique, but up to a generic affine transformation. Any non-singular 3×3 full rank matrix $\mathtt{Q}$ and its inverse may be inserted in the decomposition giving an equivalent result:

$$\mathtt{W} = (\tilde{\mathtt{M}}\mathtt{Q})(\mathtt{Q}^{-1}\tilde{\mathtt{S}}) = \tilde{\mathtt{M}}(\mathtt{Q}\mathtt{Q}^{-1})\tilde{\mathtt{S}} = \mathtt{M}\,\mathtt{S} \tag{2.15}$$

The matrix product leads to the same measurement matrix, but the structure of $\mathtt{M}$ and $\mathtt{S}$ has clearly changed. This ambiguity may easily be eliminated by enforcing orthonormality of the rotation matrices comprising $\tilde{\mathtt{M}}$ (i.e., imposing the metric constraint) and, thus, upgrading the decomposition from affine to metric.

Computing the transformation $\mathtt{Q}$

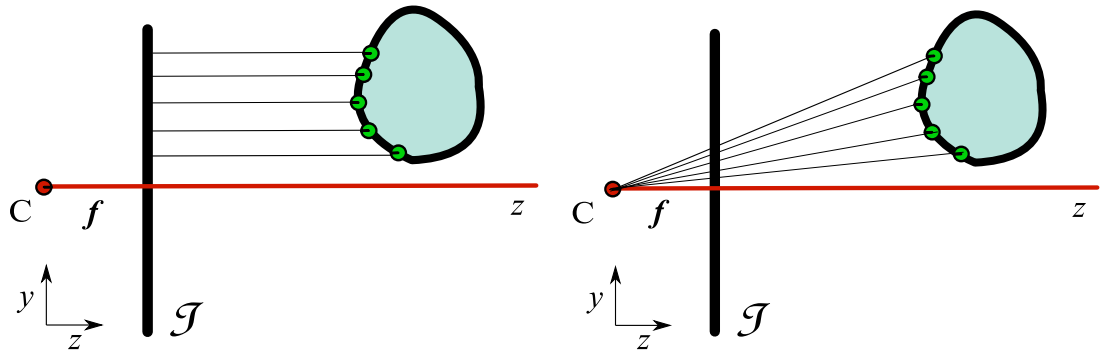
A generic orthographic camera matrix at frame i can be expressed in vector form as:

$$\mathtt{R}_i = \begin{bmatrix} \mathbf{r}^T_{i1} \\ \mathbf{r}^T_{i2} \end{bmatrix} \tag{2.16}$$

Taking into account every $i = 1\ldots F$, it is possible to write the following over-constrained system of equations:

$$\begin{aligned} \mathbf{r}^T_{i1}\mathtt{Q}\mathtt{Q}^{\mathtt{T}}\mathbf{r}_{i1} &= 1 \\ \mathbf{r}^T_{i2}\mathtt{Q}\mathtt{Q}^{\mathtt{T}}\mathbf{r}_{i2} &= 1 \\ \mathbf{r}^T_{i1}\mathtt{Q}\mathtt{Q}^{\mathtt{T}}\mathbf{r}_{i2} &= 0 \end{aligned} \tag{2.17}$$

which expresses the orthonormality of the rows of $\mathtt{R}_i$. The equations are quadratic in the unknowns which are the elements of $\mathtt{Q}$. In order to solve the system linearly, Tomasi and Kanade define the 3×3 symmetric matrix $\mathtt{B} = \mathtt{Q}\mathtt{Q}^{\mathtt{T}}$, solve the system for the 6 unknowns in $\mathtt{B}$ and then



(a) (b)

Figure 2.3: (a) An orthographic camera projects the 3-D points lying on the object surface onto image plane $\Im$. Orthographic projection assumes the object being far from the image plane such that the projecting rays are all parallel to the optical axis and perpendicular to the image plane $\Im$. (b) A full perspective camera projects the 3-D points with rays passing through the optical center C of the camera. The coordinates projected onto the image plane have different image positions depending on the depth of the pointsand the internal parameters of the camera (such as the focal length f).

extract $\mathtt{Q}$ using Cholesky decomposition. Finally, the correct matrix structure for the factorization of rigid shapes is obtained by applying the transformation to the affine solution computed via SVD:

$$\mathtt{M} = \tilde{\mathtt{M}}\mathtt{Q} \text{ and } \mathtt{S} = \mathtt{Q}^{-1}\tilde{\mathtt{S}} \tag{2.18}$$

which ensures that $\mathtt{M}$ contains F rotation matrices as shown in equation (2.12).

The orthographic camera is typically a good approximation when the object's depth is small in comparison to the distance from the camera. In this case depth recovery is difficult and may be sensitive to noise, so an orthographic model is more reliable. Nevertheless, the method has been extended to more general affine camera models, such as the weak perspective [82] and paraperspective [117].

2.2.2 Perspective factorization

If we now assume a perspective projection model for the camera (see figure 2.3 for a comparison with the orthographic case), a 3-D homogeneous point $\bar{\mathbf{X}}_j$ will be projected onto image frame i according to the equation:

$$\bar{\mathbf{w}}_{ij} = \frac{1}{\lambda_{ij}} \mathtt{P}_i \bar{\mathbf{X}}_j \tag{2.19}$$

where $\bar{\mathbf{w}}_{ij}$ and $\bar{\mathbf{X}}_{ij}$ are both expressed in homogeneous coordinates (i.e. $\bar{\mathbf{w}}_{ij} = [u_{ij}\ v_{ij}\ 1]^T$ and $\bar{\mathbf{X}}_j = [X_j\ Y_j\ Z_j\ 1]^T$), $\mathtt{P}_i$ is the 3×4 projection matrix and λ_{ij} is the projective depth for that point. Scaling the image coordinates of all the points in all the views by their corresponding projective depth gives a $3F \times P$ measurement matrix:

$$\bar{\mathtt{W}} = \begin{bmatrix} \lambda_{11}\bar{\mathbf{w}}_{11} & \dots & \lambda_{1P}\bar{\mathbf{w}}_{1P} \\ \vdots & & \vdots \\ \lambda_{F1}\bar{\mathbf{w}}_{F1} & \dots & \lambda_{FP}\bar{\mathbf{w}}_{FP} \end{bmatrix} = \begin{bmatrix} \mathtt{P}_1 \\ \vdots \\ \mathtt{P}_F \end{bmatrix} \mathtt{S} = \mathtt{MS} \tag{2.20}$$

where $\bar{\mathtt{W}}$ is the rescaled measurement matrix, $\mathtt{S} = [\bar{\mathbf{X}}_1 \dots \bar{\mathbf{X}}_P]$ is a $4 \times P$ shape matrix which contains the homogeneous coordinates of the P 3-D points and $\mathtt{M}$ contains the perspective cameras for each frame. In the case of rigid structure, $\mathtt{M}$ and $\mathtt{S}$ are at most rank 4. Therefore, the rank of the scaled measurement matrix $\bar{\mathtt{W}}$ is constrained to be $r \leq 4$.

If the true projective depths λ_{ij} were known it would be possible to factorize the measurement matrix into two rank-4 matrices, $\tilde{\mathtt{M}}$ and $\tilde{\mathtt{S}}$ using SVD. Similarly to the orthographic case, the result of the factorization would not be unique since any invertible 4×4 matrix $\mathtt{Q}$ and its inverse can be inserted in the decomposition, leading to the alternative camera and shape matrices $\tilde{\mathtt{M}}\mathtt{Q}$ and $\mathtt{Q}^{-1}\tilde{\mathtt{S}}$. Therefore, without assuming any additional constraints on the cameras or on the scene the reconstruction can be calculated up to an overall projective transformation. In general, the true projective depths λ_{ij} are unknown so the essence of projective factorization methods is to deal with the estimation of projective depths λ_{ij} in order to obtain a measurement matrix which could be decomposed into camera motion and shape in 3-D projective space using the rank constraint described above. Variants of the projective factorization method have been proposed so far for the case of scenes with rigid objects.

The first work to extend Tomasi and Kanade's algorithm to the perspective camera case was by Sturm and Triggs [132] who proposed a non-iterative factorization method for uncalibrated cameras. The method solves for the projective depths by calculating the fundamental matrices and epipoles between pairs of views. The overall accuracy of the algorithm depends greatly on the estimation of the epipolar geometry, as large errors in the fundamental matrix would affect the measurement matrix and result in errors in the shape and motion. On the other hand, Han and Kanade [57] perform a projective reconstruction using a bilinear factorization algorithm without calculating the fundamental matrices. Heyden's method [68] uses a different iterative approach. It relies on using sub-space constraints to perform projective structure from motion. Ueshiba and

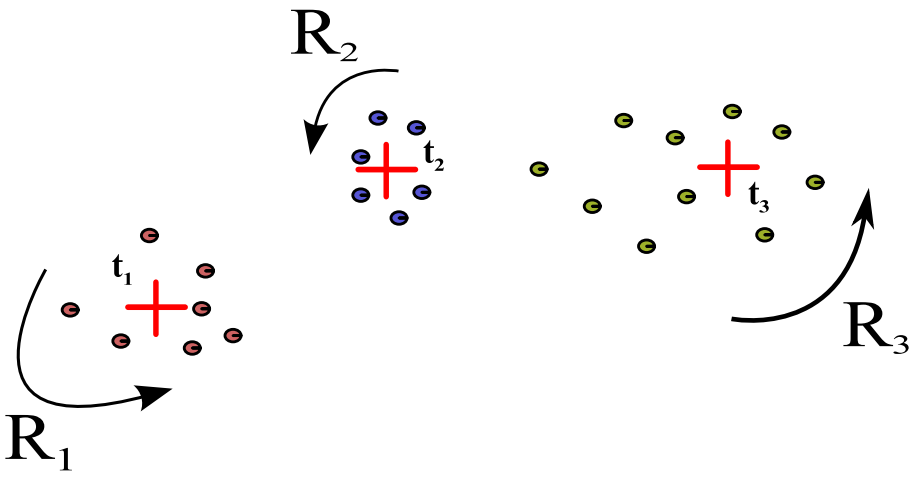


Figure 2.4: Three independent objects are represented in an image by a cluster of feature points. The motion of each object is defined on the image plane by the 2-D coordinates of its centroid ($\mathbf{t}_1$, $\mathbf{t}_2$ and $\mathbf{t}_3$) and the rotation matrices ($\mathtt{R}_1$, $\mathtt{R}_2$ and $\mathtt{R}_3$) which project it onto the image plane.

Tomita [149] presented a method by which the projective depths are iteratively estimated so that the measurement matrix is made close to rank 4. The authors also derived metric constraints for a perspective camera model to upgrade the structure to Euclidean when the internal camera parameters are known. Recently, Tang and Hung [137] proposed an iterative algorithm for projective reconstruction based on minimizing an approximation of the 2-D reprojection errors using weighted least-squares. The iterative nature of these algorithms leads them to be prone to falling into local minima. Additionally, slow convergence rates are also reported, especially in the case of image noise affecting the trajectories.

2.3 Non-rigid factorization

The dimensionality of the sub-space in which the image trajectories lie does not only depend on the camera model that projects the 3-D structure. The rank may also vary depending on the specific structure of the scene; for instance the object may change its shape or the scene could be composed of different objects moving independently.

2.3.1 Multi-body factorization

Given multiple independently moving objects in a scene, it is possible to reformulate the factorization framework to model the 3-D structure and motion components of each object separately. In this case, the measurement matrix contains trajectories belonging to different objects (see fig-

ure 2.4). This scenario was extensively studied in the work of Costeira and Kanade [29]. Briefly, N independent objects are present in a scene and, as a result, each element is modelled by a specific 3-D structure $\mathtt{S}^{(n)}$ of size $4 \times P_n$ where $n = 1 \ldots N$ and $\sum_{n=1}^{N} P_n = P$. Each independent shape can be arranged in a single structure using a block matrix $\mathtt{S}$ such that:

$$\mathtt{S} = \begin{bmatrix} \mathtt{S}^{(1)} & 0 & \ldots & 0 \\ 0 & \mathtt{S}^{(2)} & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \mathtt{S}^{(N)} \end{bmatrix} \tag{2.21}$$

Notice that in this case, the 3-D coordinates are homogeneous and consequently $\mathtt{S}$ is of size $4N \times P$, yielding the following structure for each generic independent shape n:

$$S^{(n)} = \begin{bmatrix} X_1^{(n)} & X_2^{(n)} & \cdots & X_P^{(n)} \\ Y_1^{(n)} & Y_2^{(n)} & \cdots & Y_P^{(n)} \\ Z_1^{(n)} & Z_2^{(n)} & \cdots & Z_P^{(n)} \\ 1 & 1 & \cdots & 1 \end{bmatrix} \tag{2.22}$$

Note that in this case the coordinates cannot be registered to a common centroid since there are multiple objects and the overall centroid will not be preserved by orthographic projection. Thus, the 2×4 motion component $\mathtt{M}_i^{(n)}$ for each shape contains the rotation and translation parameters for frame i:

$$\mathtt{M}_i^{(n)} = \begin{bmatrix} r_{i1}^{(n)} & r_{i2}^{(n)} & r_{i3}^{(n)} & t_{ui}^{(n)} \\ r_{i4}^{(n)} & r_{i5}^{(n)} & r_{i6}^{(n)} & t_{vi}^{(n)} \end{bmatrix} \tag{2.23}$$

The overall motion matrix $\mathtt{M}$ can now be written as:

$$\mathtt{M} = \begin{bmatrix} \mathtt{M}_1^{(1)} & \mathtt{M}_1^{(2)} & \ldots & \mathtt{M}_1^{(N)} \\ \mathtt{M}_2^{(1)} & \mathtt{M}_2^{(2)} & \ldots & \mathtt{M}_2^{(N)} \\ \vdots & \vdots & \ddots & \vdots \\ \mathtt{M}_F^{(1)} & \mathtt{M}_F^{(2)} & \ldots & \mathtt{M}_F^{(N)} \end{bmatrix} \tag{2.24}$$

This formulation implies that each trajectory has already been assigned to the correct object. By grouping together structure, motion and measurement matrices we obtain:

$$\left[\begin{array}{c|c|c|c} \mathtt{W}^{(1)} & \mathtt{W}^{(2)} & \ldots & \mathtt{W}^{(N)} \end{array}\right] = \begin{bmatrix} \mathtt{M}_1^{(1)} & \mathtt{M}_1^{(2)} & \ldots & \mathtt{M}_1^{(N)} \\ \mathtt{M}_2^{(1)} & \mathtt{M}_2^{(2)} & \ldots & \mathtt{M}_2^{(N)} \\ \vdots & \vdots & \ddots & \vdots \\ \mathtt{M}_F^{(1)} & \mathtt{M}_F^{(2)} & \ldots & \mathtt{M}_F^{(N)} \end{bmatrix} \begin{bmatrix} \mathtt{S}^{(1)} & 0 & \ldots & 0 \\ 0 & \mathtt{S}^{(2)} & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \mathtt{S}^{(N)} \end{bmatrix} \tag{2.25}$$

where:

$$\mathtt{W} = \left[\begin{array}{c|c|c|c} \mathtt{W}^{(1)} & \mathtt{W}^{(2)} & \ldots & \mathtt{W}^{(N)} \end{array} \right] \tag{2.26}$$

is the $2F \times P$ measurement matrix obtained by putting together the trajectories belonging to the N independent objects.

Provided an initial grouping of the trajectories is given, it is possible to find a transformation $\mathtt{Q}$ that forces the particular structure of $\mathtt{M}$ and $\mathtt{S}$ for the multi-body factorization scenario. In this case, the rank of $\mathtt{W}$ is constrained to be $4N$ since the measurements are a product of full rank matrices such that $\mathtt{W} = \mathtt{M}_{2F\times 4N}\ \mathtt{S}_{4N\times P}$. The task of correctly segmenting different objects by observing their 2-D motion is not trivial [166], and this problem, essential for a correct 3-D reconstruction, has triggered an active stream of research on motion segmentation. A complete investigation of these issues is postponed until chapter 5 (section 5.2).

2.3.2 Articulated factorization

The dimensionality of the sub-space in which image trajectories reside increases by a quantity proportional to the number of independently moving objects present in the scene. However, if the objects share a dependency such as a joint or a common rotational axis (see figure 2.5) the rank varies with the interdependency between the 3-D shapes.

When two independent objects are considered, the resulting rank of the measurement matrix is $r = 8$. However, if for instance the shapes have a joint between them, the sub-space representing the trajectories will decrease by 1 or 2 dimensions depending on the properties of the joint. This means that the sub-spaces of the two shapes intersect and the rank of $\mathtt{W}$ will reduce respectively to $r = 7$ or $r = 6$. Therefore, if the trajectories of the first object are stored in $\mathtt{W}^{(1)}$ and for the second in $\mathtt{W}^{(2)}$, and no degeneracies are present, it follows that:

$$rank(\mathtt{W}^{(1)}) = 4 \ \text{ and } \ rank(\mathtt{W}^{(2)}) = 4 \tag{2.27}$$

However, by merging the data together into a single measurement matrix, the following rank property holds:

$$rank\left(\left[\begin{array}{c|c} \mathtt{W}^{(1)} & \mathtt{W}^{(2)} \end{array} \right] \right) \leq 8 \tag{2.28}$$

showing that a relation between the two shapes is present. Recent work on articulated factorization describes two types of joints: the universal joint and the hinge joint [143, 163].

When two objects are linked by a *universal joint* (see figure 2.5) the distance between the two centers of mass is constant (for instance, the head and the torso of a human body) but they

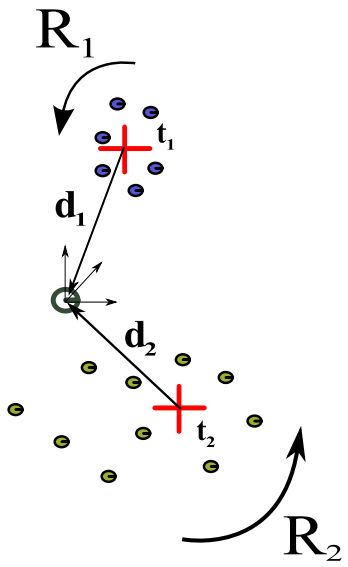


Figure 2.5: An articulated object composed of two shapes connected by a *universal joint* (represented by a circle). Shape (1) and shape (2) are projected onto the image plane and they are composed of a set of feature points whose centroids are indicated respectively by the 2-D-vectors $\mathbf{t}^{(1)}$ and $\mathbf{t}^{(2)}$. Each object has independent rotational components $\mathtt{R}^{(1)}$ and $\mathtt{R}^{(2)}$, while the 3-vectors $\mathbf{d}^{(1)}$ and $\mathbf{d}^{(2)}$ specify the translation between the centroid of each shape and the *universal joint*.

have independent rotation components. At each frame the shapes connected by a joint satisfy the following relation:

$$\mathbf{t}^{(1)} + \mathtt{R}^{(1)}\mathbf{d}^{(1)} = \mathbf{t}^{(2)} + \mathtt{R}^{(2)}\mathbf{d}^{(2)} \tag{2.29}$$

where $\mathbf{t}^{(1)}$ and $\mathbf{t}^{(2)}$ are the 2-D image centroids of the two objects, $\mathtt{R}^{(1)}$ and $\mathtt{R}^{(2)}$ the 2×3 orthographic camera matrices and $\mathbf{d}^{(1)}$ and $\mathbf{d}^{(2)}$ the 3-D displacement vectors of each shape from the central joint. The constraint expressed in equation (2.29) is the reason for the reduced dimensionality of the joint sub-spaces. It is possible to refer the articulated motion to a common reference frame centered on the centroid of the first object thus simplifying the shape matrix $\mathtt{S}$ such that:

$$\mathtt{S} = \begin{bmatrix} \mathtt{S}^{(1)} & \mathbf{d}^{(1)} \\ 0 & \mathtt{S}^{(2)} - \mathbf{d}^{(2)} \\ 1 & 1 \end{bmatrix} \tag{2.30}$$

where $\mathtt{S}$ is a full rank-7 matrix. The motion for a frame i has to be arranged accordingly to satisfy equation (2.29) as:

$$\mathtt{M}_i = \begin{bmatrix} \mathtt{R}_i^{(1)} & \mathtt{R}_i^{(2)} & \mathbf{t}_i^{(1)} \end{bmatrix} \tag{2.31}$$

Finally, we can write the full expression for the image coordinates of two objects linked by a *universal joint* for the frame *i* as:

$$\mathtt{W}_i = \left[\begin{array}{c|c} \mathtt{W}_i^{(1)} & \mathtt{W}_i^{(2)} \end{array} \right] = \left[\begin{array}{cc} \mathtt{R}_i^{(1)} & \mathtt{R}_i^{(2)} \quad \mathbf{t}_i^{(1)} \end{array} \right] \left[\begin{array}{cc} \mathtt{S}^{(1)} & \mathbf{d}^{(1)} \\ \mathtt{0} & \mathtt{S}^{(2)} - \mathbf{d}^{(2)} \\ 1 & 1 \end{array} \right] \tag{2.32}$$

The image coordinates for every frame can then be stacked to form the general structure of the factorization framework and, once more, the problem is to fit the multi-view data to the model expressed in equation (2.32).

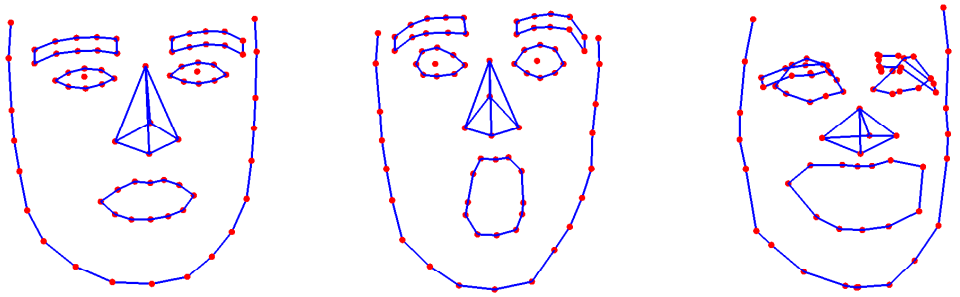
In order to find the reduced sub-space for the measurements in $\mathtt{W}$, a truncated SVD is used to compute the initial solution for $\tilde{\mathtt{M}}$ and $\tilde{\mathtt{S}}$. In the case of a *universal joint*, the task is to find the correct transformation $\mathtt{Q}_{7\times 7}$ that determines the exact factorization in accordance with equation (2.32). Similarly to the rigid and multi-body case, this problem can be neatly solved with a linear system. Further details can be found in [143], alongside a description of additional joint models.

2.4 Deformable factorization methods: a review

In the case of deformable objects, a single object varies its 3-D structure with respect to a set of deformation modes. The specific number of modes used to define the shape has the effect of forcing a specific rank-constraint over the image trajectories stored in $\mathtt{W}$. Thus, by imposing the correct rank, it is possible to carry out an approach similar to those discussed in the previous sections for other factorization problems.

The main issue to be solved is the computation of the transformation matrix $\mathtt{Q}$ that upgrades the structure and motion to metric space. In addition, the simultaneous estimation of motion and deformable shape is often ambiguous. Given a particular motion there may be various non-rigid shapes that fit the measurements. Special care needs to be taken regarding the type of information provided to the system to allow disambiguation.

Deformable shapes are the central interest of this work, thus we will dedicate the next sections to describing the non-rigid factorization methods in the literature before presenting our own contributions.



First basis Second basis Third basis

Figure 2.6: Three basis shapes obtained from the 3-D reconstruction of a human face taking on different facial expressions. The first basis generally represents the mean structure of the object, in this particular case a neutral expression. The second and third basis shapes shows a surprise and exaggerated grin expression and they are obtained by summing the first basis (the mean component) with the second and third (i.e., $\tilde{\mathtt{S}}_2 = \mathtt{S}_1 + \mathtt{S}_2$ and $\tilde{\mathtt{S}}_3 = \mathtt{S}_1 + \mathtt{S}_3$ respectively).

2.4.1 The deformable model

Bregler, Hertzmann and Biermann were the first to propose an extension of Tomasi and Kanade's factorization algorithm to deal with the case of non-rigid deformable 3-D structure [19]. Here, a model is needed to express the deformations of the 3-D shape in a compact way. The chosen representation is a simple linear model where the 3-D shape of any specific configuration $\mathtt{S}$ is approximated by a linear combination of a set of D basis shapes $\mathtt{S}_d$ (see figure 2.6) which represent the principal modes of deformation of the object:

$$\mathtt{S} = \sum_{d=1}^{D} l_d \mathtt{S}_d \qquad \mathtt{S}, \mathtt{S}_d \in \Re^{3\times P} \quad l_d \in \Re \tag{2.33}$$

where each basis shape $\mathtt{S}_d$ is a $3 \times P$ matrix which contains the 3-D locations of P object points for that particular mode of deformation. A perfectly rigid object would correspond to the situation where $D = 1$.

Similarly to Tomasi and Kanade, Bregler et al. also assumed a scaled orthographic projection model for the camera. In this case, the coordinates of the 2-D image points observed at each frame i are related to the coordinates of the 3-D points according to the following equation:

$$\mathtt{W}_i = \begin{bmatrix} \mathbf{w}_{i1} & \dots & \mathbf{w}_{iP} \end{bmatrix} = \mathtt{R}_i \left(\sum_{d=1}^{D} l_{id} \mathtt{S}_d \right) + \mathtt{T}_i \tag{2.34}$$

where

$$\mathtt{R}_i = \begin{bmatrix} r_{i1} & r_{i2} & r_{i3} \\ r_{i4} & r_{i5} & r_{i6} \end{bmatrix} \tag{2.35}$$

is the 2×3 matrix containing the first two rows of a rotation matrix and l_{id} the configuration weight for basis d at frame i. When the image coordinates are registered to the object's centroid, equation (2.34) can be rewritten in matrix form as follows:

$$\mathtt{W}_i = \begin{bmatrix} l_{i1}\mathtt{R}_i & \dots & l_{iD}\mathtt{R}_i \end{bmatrix} \begin{bmatrix} \mathtt{S}_1 \\ \vdots \\ \mathtt{S}_D \end{bmatrix} = \begin{bmatrix} \mathtt{M}_{i1} & \dots & \mathtt{M}_{iD} \end{bmatrix} \begin{bmatrix} \mathtt{S}_1 \\ \vdots \\ \mathtt{S}_D \end{bmatrix} = \mathtt{M}_i\mathtt{S} \tag{2.36}$$

If these P points can be tracked throughout an image sequence, the point tracks may be stacked into the $2F \times P$ measurement matrix $\mathtt{W}$ and we may write:

$$\mathtt{W} = \begin{bmatrix} l_{11}\mathtt{R}_1 & \dots & l_{1D}\mathtt{R}_1 \\ \vdots & & \vdots \\ l_{F1}\mathtt{R}_F & \dots & l_{FD}\mathtt{R}_F \end{bmatrix} \begin{bmatrix} \mathtt{S}_1 \\ \vdots \\ \mathtt{S}_D \end{bmatrix} = \begin{bmatrix} \mathtt{M}_{11} & \dots & \mathtt{M}_{1D} \\ \vdots & & \vdots \\ \mathtt{M}_{F1} & \dots & \mathtt{M}_{FD} \end{bmatrix} \begin{bmatrix} \mathtt{S}_1 \\ \vdots \\ \mathtt{S}_D \end{bmatrix} = \mathtt{MS} \tag{2.37}$$

Since $\mathtt{M}$ is a $2F \times 3D$ matrix and $\mathtt{S}$ is a $3D \times P$ matrix, the rank of $\mathtt{W}$ when no noise is present must be $r \leq 3D$. Note that, in relation to rigid factorization, in the non-rigid case the rank is incremented by three with every new mode of deformation. The goal of factorization algorithms is to exploit this rank constraint to recover the 3-D pose and shape (basis-shapes and deformation coefficients) of the object from the correspondence points stored in $\mathtt{W}$.

In order to obtain a solution for $\mathtt{M}$ and $\mathtt{S}$, it is possible to perform a truncated SVD to rank $3D$ similarly to the rigid case. However, the result of the factorization of $\mathtt{W}$ is not unique; any invertible $3D \times 3D$ matrix $\mathtt{Q}$ and its inverse can be inserted into the decomposition leading to the alternative factorization $\mathtt{W} = (\tilde{\mathtt{M}}\mathtt{Q})(\mathtt{Q}^{-1}\tilde{\mathtt{S}})$. The problem is to find a transformation matrix $\mathtt{Q}$ that imposes the replicated block structure on the motion matrix $\tilde{\mathtt{M}}$ shown in equation (2.37) and that removes the affine ambiguity upgrading the reconstruction to a metric one. Whereas in the rigid case the problem of computing the transformation matrix $\mathtt{Q}$ to upgrade the reconstruction to a metric one can be solved linearly (see section 2.2); in the non-rigid case, imposing the appropriate repetitive structure and forcing the metric constraint to the motion matrix $\tilde{\mathtt{M}}$ results in a non-linear problem.

Various methods have been proposed so far in the literature [19, 16, 141, 159, 17] and they will be discussed in the following sections. It is important to note that while the block structure

of the motion matrix $\mathtt{M}$ is not required if we only wish to determine image point motion, it is crucial for the recovery of metric 3-D shape and motion which is the main goal of our work.

2.4.2 Bregler et al.'s method

Bregler et al. [19] introduced the non-rigid factorization framework and suggested a solution for the computation of the matrix $\mathtt{Q}$. The main problem addressed in their work is the separation of the configuration weights l_{id} from the rotation matrices $\mathtt{R}_i$. The solution proposed is called *sub-block factorization* and it considers F sub-blocks derived from a row-wise partition of $\mathtt{M}$ such that the equation of each sub-block $\mathtt{M}_i$ is given by:

$$\mathtt{M}_i = \begin{bmatrix} l_{i1}\mathtt{R}_i & \dots & l_{iD}\mathtt{R}_i \end{bmatrix} \tag{2.38}$$

The entries of each sub-block are then rearranged as a rank-1 outer product of 2 vectors giving a $D \times 6$ matrix $\check{\mathtt{M}}_i$ which can be expressed as:

$$\check{\mathtt{M}}_i = \begin{bmatrix} l_{i1}\mathbf{r}_i^T \\ \vdots \\ l_{iD}\mathbf{r}_i^T \end{bmatrix} = \begin{bmatrix} l_{i1} \\ \vdots \\ l_{iD} \end{bmatrix} \begin{bmatrix} r_{i1} & r_{i2} & r_{i3} & r_{i4} & r_{i5} & r_{i6} \end{bmatrix} \tag{2.39}$$

where $\mathbf{r}_i = [r_{i1}, .., r_{i6}]^T$ are the coefficients of the rotation matrix $\mathtt{R}_i$. Thus, Bregler et al.'s approach extracts configuration weights and rotation components by performing F SVDs truncated to rank 1 and then stacking each component into a $2F \times 3$ matrix $\bar{\mathtt{R}}$.

Since the individual elements r_{ik} for $k = 1 \dots 6$ obtained from the decomposition do not form orthonormal matrices, a further orthogonalization is required to upgrade the model to a metric one. This can be done simply by applying the metric constraints to the matrix $\bar{\mathtt{R}}$ and computing the correcting transformation $\check{\mathtt{Q}}_{3\times 3}$ as in section 2.2.1. Finally, it is possible to compute the full $3D \times 3D$ transformation $\mathtt{Q}$ as:

$$\mathtt{Q} = \begin{bmatrix} \check{\mathtt{Q}} & \mathtt{0} & \dots & \mathtt{0} \\ \mathtt{0} & \check{\mathtt{Q}} & \dots & \mathtt{0} \\ \vdots & & \ddots & \vdots \\ \mathtt{0} & \mathtt{0} & \dots & \check{\mathtt{Q}} \end{bmatrix} \tag{2.40}$$

a block-diagonal matrix that upgrades the structure to metric.

Discussion

The method presents a significant weakness; the rank-1 SVD used to factorize $\mathtt{M}_i$ in equation (2.39) is a coarse numerical approximation when the measurements are affected by noise. Thus, the second and further singular values retain a considerable contribution to the solution. Additionally, the true transformation matrix $\mathtt{Q}$ is usually dense in the off-diagonal values and so the block diagonal approximation $\check{\mathtt{Q}}$ can only be a solution for a sub-group of all the possible transformations. Only very simple deformations may be solved using this approach.

Furthermore, the solution for $\mathtt{R}_i$ and l_{id} is computed exclusively from the motion matrix $\tilde{\mathtt{M}}$ obtained directly after performing the initial rank-3D SVD on the measurement matrix. This first decomposition redistributes the structure and motion components randomly between $\tilde{\mathtt{M}}$ and $\tilde{\mathtt{S}}$ as pointed out by Brand [16]. However, Bregler et al.'s method assumes that all the components referring to configuration weights and camera parameters are fully contained in $\tilde{\mathtt{M}}$. This assumption does not hold in principle and a transformation able to reorder the components should be carried out before the *sub-block factorization*.

Solutions to this problem are proposed in [141] by using an iterative optimization and in [16] by using a flexible factorization approach. The following sections describe these approaches in more detail.

2.4.3 Torresani et al.'s approach

Torresani et al. [141] define an optimization method to correct the inaccurate solution proposed by Bregler et al. described in the previous section. After obtaining an approximate solution with *sub-block factorization*, their approach optimises the following non-linear cost function:

$$\chi_i = \mathtt{W}_i - \mathtt{R}_i \sum_{d=1}^{D} l_{id}\mathtt{S}_d \tag{2.41}$$

with $i = 1 \ldots F$ and $d = 1 \ldots D$. This optimization is performed by alternatively minimizing three different least-square problems in the three classes of model parameters: $\mathtt{R}_i$, l_{id} and $\mathtt{S}_d$. While each class of parameters is estimated, the other two are assumed to remain constant. This procedure is also known as Alternating Least Squares (ALS) [157] and it has the advantage that it may converge to a solution without the complexity of using a full non-linear approach.

Torresani et al. report that an appropriate initialisation can be obtained using an initial guess of the camera matrices $\mathtt{R}_i$ which they compute by applying Tomasi and Kanade's rigid factorization over the non-rigid measurements in $\mathtt{W}$. Differently, the configuration weights l_{id} are ini-

tialised randomly and this permits to obtain the first estimate of $\mathtt{S}_d$ with the ALS approach. Note that, to obtain more robustness, the rotation components of the orthographic camera model are parameterized using the Rodriguez formula instead of considering each of the 6 elements in $\mathtt{R}_i$. A regularization of the shape matrix is also used during the iterative optimization stage to prevent ill-conditioned values on the shape when there is not enough out of plane rotation.

While this method does preserve the replicated block structure of the $\mathtt{M}$ matrix it minimizes an algebraic cost function rather than a geometrically meaningful criterion. As a further drawback, occasionally the algorithm presents a slow convergence to the solution given by the zig-zagging behavior of the minimization in the different parameter spaces (for a general analysis of the behavior of ALS methods in the SfM domain please refer to [20]).

2.4.4 Brand's orthonormal decomposition and flexible factorization

Brand proposed an alternative algorithm called *flexible factorization* [18], where a solution for $\mathtt{Q}$ is achieved without computing the second series of rank-1 SVDs. The method recovers the camera matrices and configuration weights using an alternative technique called *orthonormal decomposition*.

The strategy is to minimize the deformations (encoded in the $D-1$ basis shapes stored in $\mathtt{S}$) with respect to the mean basis component $\mathtt{S}_1$ computed from the three most significant singular values. The reason for forcing this constraint is based on the observation that most of the motion of the object can be explained by the rigid component.

Flexible factorization

Concisely, the algorithm consists of an initialisation step where an approximate transformation $\mathtt{Q}$ is computed estimating the matrix in a band around the diagonal values. The approach proposed by Brand [16] corrects each column-triple independently applying the rigid metric constraint to each $(2F \times 3)$ $\tilde{\mathtt{M}}_d$ vertical block in $\tilde{\mathtt{M}}$ as shown here:

$$\tilde{\mathtt{M}} = \begin{bmatrix} \tilde{\mathtt{M}}_1 & \dots & \tilde{\mathtt{M}}_D \end{bmatrix} = \begin{bmatrix} \tilde{\mathtt{M}}_{11} & \dots & \tilde{\mathtt{M}}_{1D} \\ \vdots & & \vdots \\ \tilde{\mathtt{M}}_{F1} & \dots & \tilde{\mathtt{M}}_{FD} \end{bmatrix}$$

Since each 2×3 $\tilde{\mathtt{M}}_{id}$ sub-block is a scaled rotation (truncated to dimension 2 for weak perspective projection) a 3×3 matrix $\mathtt{Q}_d$ (with $d = 1 \dots D$) can be computed to correct each vertical block $\tilde{\mathtt{M}}_d$ by imposing orthogonality and equal norm constraints to the rows of each $\tilde{\mathtt{M}}_{id}$. Each $\tilde{\mathtt{M}}_{id}$ block

contributes one orthogonality and one equal norm constraint to solve for the elements in $\mathtt{Q}_d$.

Each vertical block is then corrected in the following way: ($\hat{\mathtt{M}}_d \leftarrow \tilde{\mathtt{M}}_d \mathtt{Q}_d$). The overall $3D \times 3D$ correction matrix $\mathtt{Q}$ will therefore be a block diagonal matrix with the following structure:

$$\mathtt{Q} = \begin{bmatrix} \mathtt{Q}_1 & \mathtt{0} & \dots & \mathtt{0} \\ \mathtt{0} & \mathtt{Q}_2 & \dots & \mathtt{0} \\ \vdots & & \ddots & \vdots \\ \mathtt{0} & \mathtt{0} & \dots & \mathtt{Q}_D \end{bmatrix} \tag{2.42}$$

Unlike the method proposed by Bregler et al. [19] (where the metric constraint was imposed only on the rigid component, so that $\mathtt{Q}_d = \check{\mathtt{Q}}$ for each $d = 1 \dots D$) this provides a different corrective transform for each column-triple of $\tilde{\mathtt{M}}$. The 3-D structure matrix is then corrected appropriately using the inverse transformation: $\mathtt{S} \leftarrow \mathtt{Q}^{-1}\tilde{\mathtt{S}}$.

Brand included a final minimization scheme in his *flexible factorization* algorithm [16]: the deformations in $\tilde{\mathtt{S}}$ should be as small as possible relative to the mean shape. The idea here is that most of the image point motion should be explained by the rigid component. This is similar to the shape regularization used by other authors [141, 2].

This final stage re-estimates the transformation matrix $\mathtt{Q}$ starting from the corrected $\hat{\mathtt{M}} = \tilde{\mathtt{M}}\mathtt{Q}$ by minimizing the following cost function:

$$tr\left\{(\hat{\mathtt{M}}\mathtt{Q} - \tilde{\mathtt{M}})^T(\hat{\mathtt{M}}\mathtt{Q} - \tilde{\mathtt{M}})\right\} + tr\left\{\tilde{\mathtt{S}}^T\mathtt{Q}^T\mathtt{Z}\mathtt{Q}\tilde{\mathtt{S}})\right\} \tag{2.43}$$

where $\mathtt{Z}$ is a matrix such that:

$$\mathtt{Z}\mathtt{S} = \begin{bmatrix} \mathtt{S}_2 \\ \vdots \\ \mathtt{S}_D \end{bmatrix} \tag{2.44}$$

Thus, a global solution is achieved by taking into account both the motion and 3-D structure matrices and strengthening the mean motion component with respect to the deformations contained in the $(D-1)$ basis shapes.

Orthonormal decomposition

The final step in the non-rigid factorization algorithm deals with the factorization of the motion matrix $\tilde{\mathtt{M}}$ into the 2×3 rotation matrices $\mathtt{R}_i$ and the deformation weights l_{id}. Brand proposed an alternative method to factorize each 2 row sub-block of the motion matrix $\tilde{\mathtt{M}}_i = \mathbf{l}_i^T \otimes \mathtt{R}_i$ (where $\otimes$ indicates the tensor product and $\mathbf{l}_i^T = [l_{i1} \dots l_{iD}]$ is a D-vector containing the configuration weights for each frame i).

Following equation (2.39), each motion matrix sub-block $\tilde{\mathtt{M}}_i$ (see [18] for details) is rearranged such that $\tilde{\mathtt{M}}_i \rightarrow \check{\mathtt{M}}_i^T$. The motion matrix $\check{\mathtt{M}}_i^T$ of size $6 \times D$ is then post-multiplied by the $D \times 1$ unity vector $\mathbf{c} = [\, 1 \ \ldots \ 1 \,]^T$ thus obtaining:

$$\mathbf{a}_i = k\mathbf{r}_i = \check{\mathtt{M}}_i^T \mathbf{c} \tag{2.45}$$

where $k = l_{i1} + l_{i2} + \ldots + l_{iD}$ (the sum of all the deformation weights for that particular frame i). A matrix $\mathtt{A}_i$ of size 2×3 is built by re-arranging the coefficients of the column vector $\mathbf{a}_i$. The analytic form of $\mathtt{A}_i$ is:

$$\mathtt{A}_i = \begin{bmatrix} kr_{i1} & kr_{i2} & kr_{i3} \\ kr_{i4} & kr_{i5} & kr_{i6} \end{bmatrix} \tag{2.46}$$

Since $\mathtt{R}_i$ is an orthonormal matrix, the equation $\mathtt{A}_i\mathtt{R}_i^T = \sqrt{\mathtt{A}_i\mathtt{A}_i^T}$ is satisfied, leading to $\mathtt{R}_i^T = \sqrt{\mathtt{A}_i\mathtt{A}_i^T}/\mathtt{A}_i$. This allows one to find a linear least-squares fit for the rotation matrix $\mathtt{R}_i$.

In order to estimate the configuration weights the sub-block matrix $\tilde{\mathtt{M}}_i$ is rearranged as equation (2.39) obtaining $\tilde{\mathtt{M}}_i \rightarrow \check{\mathtt{M}}_i$. The configuration weights for each frame i are then derived exploiting the orthonormality of $\mathtt{R}_i$ since:

$$\check{\mathtt{M}}_i\mathbf{r}_i^T = \begin{bmatrix} l_{i1}\mathbf{r}_i\mathbf{r}_i^T \\ \vdots \\ l_{iD}\mathbf{r}_i\mathbf{r}_i^T \end{bmatrix} = 2 \begin{bmatrix} l_{i1} \\ \vdots \\ l_{iD} \end{bmatrix} \tag{2.47}$$

Discussion

The method proposed by Brand consists on estimating the off-diagonal elements of $\mathtt{Q}$ using a least-squares approach to minimize the Frobenius norm of equation (2.43). Essentially, this further step has the effect of forcing a strong prior over the strength of the deformations of the inspected object. By absorbing most of the contribution of the motion into the first basis (also called the mean component or mean basis), Brand observed that small deformations can be irremediably lost. This is also supported by further tests presented in our work which show that the prior introduced in the *flexible factorization* may be too restrictive to be applicable in specific scenarios with varying degrees of non-rigidity. We should also stress the fact that the cost function is strictly an algebraic error without any consideration of the geometrical model describing the 3-D structure and temporal deformations.

2.4.5 Xiao et al.'s closed form solution

The main problem with the previous approaches stems from the fact that deformation and motion are ambiguous. Given a specific configuration of points on the image plane, different 3-D non-rigid shapes and camera motions can be found that fit the measurements. To solve this ambiguity prior knowledge about the shape and motion could be used to constrain the solution.

Recently, this approach was adopted by Xiao et al. [159] by introducing the concept of *basis constraints*, a set of linear constraints which, when used in addition to the rotation constraints, uniquely determine a closed-form solution to the non-rigid factorization problem.

The basis constraints

In the rigid case ($D = 1$), it is possible to solve for the transformation $\mathtt{Q}$ by linearly imposing the metric constraint on the rotation matrices (see section 2.2.1). However in the deformable case, imposing only the constraints derived from the orthographic projection model renders a solution space that contains a set of invalid or degenerate solutions. In order to remove this ambiguity, Xiao et al. introduced a new set of constraints based on prior information over the data and they proved that the added linear equations can solve uniquely for $\mathtt{Q}$.

Xiao et al.'s assumption is that a set of D frames exists in which the basis shapes are independent such that the shape in that frame can be exactly described by a single 3-D basis. This assumption forces a particular structure in the motion matrix $\mathtt{M}$. For convenience, the measurement matrix is arranged such that the D frames corresponding to the independent bases are in the first $2D$ rows of $\mathtt{W}$:

$$\mathtt{W} = \left[\begin{array}{cccc} \mathtt{R}_1 & 0 & \dots & 0 \\ 0 & \mathtt{R}_2 & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & \mathtt{R}_D \\ \hline l_{(D+1)1}\mathtt{R}_{(D+1)} & l_{(D+1)2}\mathtt{R}_{(D+1)} & \dots & l_{(D+1)D}\mathtt{R}_{(D+1)} \\ \vdots & & & \vdots \\ l_{F1}\mathtt{R}_F & l_{F2}\mathtt{R}_F & \dots & l_{FD}\mathtt{R}_F \end{array}\right] \left[\begin{array}{c} \mathtt{S}_1 \\ \vdots \\ \mathtt{S}_D \end{array}\right] = \mathtt{MS} \qquad (2.48)$$

thus, the top $2D \times 3D$ block of the motion matrix $\mathtt{M}$ is a block-diagonal matrix containing the D camera matrices for the independent basis shapes. Xiao et al.'s algorithm obtains a closed-form solution by enforcing this particular structure to $\mathtt{M}$ in (2.48).

The closed form solution

In more detail, the new set of linear equations is used to solve D linear problems over the sub-matrices $\mathtt{Q}_d$ obtaining a column-wise partition of the matrix $\mathtt{Q}$ such that

$$\mathtt{Q} = \left[\begin{array}{c|c|c} \mathtt{Q}_1 & \ldots & \mathtt{Q}_D \end{array} \right] \tag{2.49}$$

where each $\mathtt{Q}_d$ with $i = 1 \ldots D$ is a $3D \times 3$ matrix. In order to solve for the full transformation $\mathtt{Q}$, the basis constraints are applied D times for each sub-transformation $\mathtt{Q}_d$. Then, the problem is to find the set of linear equations that force the exact structure of the motion matrix in (2.48).

In the following, we will show how to build the linear system for a generic column-triple $\mathtt{M}_d$. Having obtained an initial solution $\mathtt{W} = \tilde{\mathtt{M}}\tilde{\mathtt{S}}$ (see section 2.1.3) with a rank $3D$ decomposition, Xiao et al. build a set of linear equations that verifies the following conditions:

$$\mathtt{M}_d = \tilde{\mathtt{M}}\mathtt{Q}_d. \tag{2.50}$$

For instance, if we consider the second transformation $\mathtt{Q}_2$, we would find the solution that transforms $\tilde{\mathtt{M}}$ such that:

$$\tilde{\mathtt{M}}\mathtt{Q}_2 = \mathtt{M}_2 = \begin{bmatrix} \mathtt{0} \\ \mathtt{R}_2 \\ \mathtt{0} \\ \vdots \\ \mathtt{0} \\ l_{(D+1)2}\mathtt{R}_{(D+1)} \\ \vdots \\ l_{F2}\mathtt{R}_F \end{bmatrix} \tag{2.51}$$

By exclusively using the metric constraints, it is not possible to determine enough equations to solve uniquely for $\mathtt{Q}_d$. This is the crucial problem of the previous methods. Xiao et al. introduced their basis constraints defining a new set of $4F(D-1)$ equations which, when combined with the equations given by the metric constraints are enough to solve the linear system. The constraints are quadratic over the unknowns stored in $\mathtt{Q}_d$, hence, Xiao et al. introduce a $3D \times 3D$ symmetric matrix $\mathtt{B}_d$ such that $\mathtt{B}_d = \mathtt{Q}_d\mathtt{Q}_d^T$. The basis constraints are determined such that the structure in equation (2.48) is satisfied for the configuration weights. This will be true if the following equations are satisfied:

$$\begin{aligned} l_{ii} &= 1, \qquad i = 1 \ldots D \\ l_{id} &= 0, \quad i,d = 1 \ldots D, i \neq d \end{aligned} \tag{2.52}$$

These basis constraints lead to a new set of $4F(D-1)$ equations as described in [159]. Notice that a further step is required to extract the corresponding three-column transformation matrix $\mathtt{Q}_d$ from each symmetric matrix $\mathtt{B}_d$. Xiao et al. suggest computing the solution via SVD; since $\mathtt{B}_d$ is, in a noiseless case, a rank-3 constrained matrix. However when noise is present the solution is numerically approximated to the closest solution in the sense of the Frobenius norm.

Discussion

Solving for the transformation matrix by dividing the problem into D linear systems permits finding a closed form solution that upgrades the factorization to the correct structure. A drawback of this approach lies in the independency of the solution; each column-triple $\mathtt{M}_i$ is upgraded separately, since the block structure of $\mathtt{M}$ is not forced in the solution of the systems of equations. As a result, after fixing a reference basis, $(D-1)$ orthogonal transformations need to be computed to align each of the bases using Procrustes analysis [127].

These further computations are very critical, since incorrect solutions would lead to a misalignment of the bases and a violation of the repetitive structure of the motion matrix. Additionally, the alignment procedure proposed by Xiao et al. attains an exact solution only when identical and isotropic Gaussian noise with zero mean affects the measurements (see an analysis on Procrustes methods in [40]). Such a condition rarely occurs in real scenarios, and in fact might occur only in synthetic tests for which the algorithm can obtain exact reconstructions.

Another criticism has been made regarding the sensitivity of the method to wrongly selected independent bases as reported in [17]. Often it is not trivial to find a single set of independent bases in a real sequence of a deforming object. Even though the method may obtain a unique solution, this solution changes with the selection of a different set of bases.

2.4.6 Brand's direct method

Recently, Brand [17] proposed a variation of Xiao et al.'s approach [159] based on the deviation of the computed solution from the orthogonality constraints and on weaker assumptions on the independent basis shapes.

The approach focuses initially on the estimation of the first three-column transformation $\mathtt{Q}_1$ which corrects the rank-$3D$ approximated $\tilde{\mathtt{M}}$ obtaining $\tilde{\mathtt{M}}\mathtt{Q}_1 = \mathtt{M}_1$. This step has the dual effect of estimating the overall motion components $\mathtt{R}_i$ and the first set of configuration weights l_{i1} with $i = 1 \ldots F$ for the mean basis shape. However, differently from Xiao et al.'s method, the

computation of $\mathtt{Q}_1$ is not given by a least-squares estimation: a quasi-Newton method is applied to a non-linear cost function constructed to impose the orthogonality constraints in $\tilde{\mathtt{M}}$. In this way, the rank-3 approximation as described in section 2.4.5 is avoided and the transformation is estimated given the actual $9D$ parameters of $\mathtt{Q}_1$.

A second stage forces the repetitive structure of $\mathtt{M}$ by linearly computing the full transformation $\mathtt{Q}$ that imposes the pre-estimated rotations $\mathtt{R}_i$ to each D triplet in the motion matrix. In the case of no noise, this two-step procedure provides exact results with synthetic data. However, the author reports more erratic behavior in the performance of the algorithm whenever the data is corrupted by noise, since local minima may appear during the quasi-Newton estimation.

In order to counteract this effect, Brand proposes different strategies based on finding multiple solutions for $\mathtt{Q}_1$ and combining them to obtain a better correction for the rotation matrices. A solution is to introduce weaker basis constraints by assuming that there exists a set of D frames for which the D-vectors of configuration weights are orthogonal to each other. In more detail, if we collect in a single $D \times D$ matrix the configuration weights for the D frames in which the bases are independent, we obtain orthogonal matrix:

$$\tilde{\mathtt{L}} = \begin{bmatrix} l_{11} & \cdots & l_{1D} \\ \vdots & \ddots & \vdots \\ l_{D1} & \cdots & l_{DD} \end{bmatrix} \tag{2.53}$$

In the case of Xiao et al's approach [159], the basis constraints force the matrix $\tilde{\mathtt{L}}$ to be an identity matrix such that $\tilde{\mathtt{L}} = \mathtt{I}$.

No justification is provided that supports the use of these constraints but only that the wrong selection of the independent basis proposed by Xiao et al. in equation (2.52) would perform notably worste than the wrongly selected orthogonal condition proposed by Brand. Alternatively, another suggestion is to start the minimization from a different initialisation of the parameters to obtain multiple estimates of $\mathtt{Q}_1$.

Discussion

Two main positive contributions are present in the direct method. Firstly, it directly estimates the parameters of the transformation matrix $\mathtt{Q}_1$ with a quasi-Newton minimization scheme and, thus, it avoids the rank-3 numerical approximation in Xiao et al's method in the case of noise. Secondly, it enforces the repetitive structure of $\mathtt{M}$ without solving D separate basis alignments. On the other hand, a weakness is present whenever noise affects the data. The matrices $\mathtt{R}_i$ are

not reliably estimated after forcing the first transformation $\mathtt{Q}_1$ over the numerically computed $\tilde{\mathtt{M}}$. Multiple solutions have to be found by imposing weaker priors over the structure of the configuration weights and by initializing the quasi-Newton minimization from different points. Nevertheless, the solution proposed by Brand performs better than the previous algorithms and it shows reliable results in real experiments, where as Xiao et al.'s method fails to select reliable independent basis shapes.

2.5 Closure

A factorization approach to structure from motion computation exhibits considerable advantages over alternative methods. In this chapter we have shown that different motion and structure models may be fitted to a set of trajectories obtained from measurements over an image sequence. As a consequence of the global constraints given by the models, the image point trajectories live in a certain sub-space defined numerically by the rank of the measurement matrix $\mathtt{W}$.

From this observation, every method presented here finds an initial solution to the *motion* and *3-D structure* by truncating unnecessary components from $\mathtt{W}$ with SVD, and then by correcting the solution with a transformation matrix $\mathtt{Q}$ that imposes explicit geometric constraints given the specific model. This approach is successful in many cases with some exceptions in the deformable case.

The main issue is in the ambiguous formulation of the problem. For a deformable shape, deformation and motion are strongly coupled elements. Not only in a mathematical sense (since, for instance, $\mathtt{R}_i$ and l_{id} appear multiplying each other inside the motion matrix $\mathtt{M}$) but, as shown by Xiao et al. [159], a solution computed only by forcing constraints over the camera motion may be degenerate and not unique. Moreover, numerical approximations [19, 18, 159] often do not provide good estimates for the geometric parameters of the deformable model.

Thus, a solution that respects the mathematical structure of the factorization framework and the geometric constraints of the camera projecting the scene is desirable. In such a way, the problem should be formulated by expressing the product of $\mathtt{M}$ and $\mathtt{S}$ as a set of non-linear equations. In this case, the full interaction of the model parameters is explicit and the parameters of the deformable model may be estimated using non-linear optimization techniques, as will be explained in the next chapter.

Chapter 3

A non-linear approach to non-rigid factorization

The non-rigid factorization algorithms described in the previous chapter suffer from a series of drawbacks. Most of them (Bregler et al. [19] and Brand [16]) do not respect the replicated block structure of the motion matrix $\mathtt{M}$ expressed in (2.37). It is important to notice that the replicated structure does not affect the estimation of the motion of image points, which makes these factorization algorithms very well suited to non-rigid tracking [141, 16]. The rank constraint imposes that the trajectories of image points lie in a $3D$ dimensional sub-space, where D is the number of basis shapes, and that any new trajectory may be generated as a linear combination of the columns of the motion matrix $\mathtt{M}$. If a point is only tracked in a small sub-set of images in the sequence, this constraint allows to predict its trajectory in the entire sequence, thus permitting to incorporate new tracks. This property is strictly based on the numerical sub-space in which the trajectories resides and not on the geometrical model estimated from the measurement matrix.

However, if the main goal is to recover the camera matrices and the 3-D non-rigid structure then preserving the replicated block structure of the motion matrix $\mathtt{M}$ after factorization becomes crucial. If this is not achieved, it results in an incorrect estimation of the motion which in turn affects the estimate of the 3-D structure. In the experimental section of this chapter we will show results which prove that the 3-D reconstructions and the motion recovered using previous non-rigid factorization methods [16, 19] are not completely satisfactory. In particular, the estimation of the 3-D pose is unstable and this affects the quality of the deformable shape.

3.1 Factorization as a non-linear estimation problem

Most of the algorithms presented so far, rely on the minimization of algebraic cost functions using linear schemes (with the exception of [17]). However, the correct error function to be minimized should be geometrically meaningful and, by construction, strictly non-linear. Therefore, existing methods only provide an approximation of the true solution so when noise affects the measurements their performance is compromised.

Xiao et al.'s work [159] provides an exact closed-form solution. However, it requires information about the independency of the basis shapes that model the object's modes of deformation, and the solution is affected by their incorrect estimation. Additionally, as noticed by Brand [17], the selection of the independent bases is trivial with well-behaved synthetic experiments but it becomes increasingly error prone with real images of deforming objects.

In order to overcome the problems encountered by previous methods, we now introduce a non-linear optimization stage [38, 35] to refine the motion and shape estimates which minimizes the image reprojection error and imposes the correct structure onto the motion matrix by choosing an appropriate parameterisation of the model parameters.

3.1.1 The non-rigid cost function

The goal is to estimate the motion parameters $\mathtt{R}_i$, the 3-D basis shapes $\mathtt{S}_d$ and the deformation weights l_{id} such that the distance between the measured image points $\mathbf{w}_{ij}$ and the reprojection of the estimated 3-D points is minimised. However, the coordinates in $\mathtt{W}$ are extracted by a measurement process and, therefore, they are affected by noise or by a certain degree of uncertainty $\mathbf{n}_{ij}$. The measured coordinates $\mathbf{w}_{ij}$ can be expressed in terms of the exact measurements $\mathbf{x}_{ij}$ such that:

$$\mathbf{w}_{ij} = \mathbf{x}_{ij} + \mathbf{n}_{ij} \tag{3.1}$$

The projection equation for a 3-D point j in image frame i is given by:

$$\mathbf{x}_{ij} = \mathtt{M}_i \mathbf{S}_j = \mathtt{R}_i \sum_{d=1}^{D} l_{id} \mathbf{S}_{dj} \tag{3.2}$$

where $\mathbf{x}_{ij}$ are the image coordinates of the point and $\mathbf{S}_j$ is the $3D \times 1$ parameterisation of the shape basis for a deformable point j such that:

$$\mathbf{S}_j = \begin{bmatrix} \mathbf{S}_{1j} \\ \mathbf{S}_{2j} \\ \vdots \\ \mathbf{S}_{Dj} \end{bmatrix} \tag{3.3}$$

with the 3-vector $\mathbf{S}_{dj}$ defining the d basis component for point j.

Following equation (3.1), the uncertainty over the measurements is obtained from the residual given by $\mathbf{n}_{ij} = \mathbf{w}_{ij} - \mathbf{x}_{ij}$. This residual is generally referred to as the reprojection error of the image coordinates in the literature and it expresses the difference between the image coordinates given the estimated model parameters and the measured data. Hence, it is possible to recast the problem of estimating the non-rigid structure and motion parameters by minimizing the norm of the reprojection error of all the points in all the frames such that:

$$\min_{\mathbf{R}_i\, l_{id}\, \mathbf{S}_j} \sum_{i,j}^{F,P} \| \mathbf{n}_{ij} \|^2 = \min_{\mathbf{R}_i\, l_{id}\, \mathbf{S}_j} \sum_{i,j}^{F,P} \| \mathbf{w}_{ij} - \mathbf{x}_{ij} \|^2 \tag{3.4}$$

Note that the error is a sum of FP quadratic cost functions. Assuming the noise can be modelled with a Gaussian distribution, the minimization of equation (3.4) provides a true Maximum Likelihood (ML) estimate of the parameters.

The definition of this non-rigid cost function could rise two major criticisms. First, the number of parameters can increase dramatically with the number of frames composing the scene and the complexity of the modelled object. This may render the minimization of equation (3.4) computationally unfeasible given the size of the parameter space. Second, the high non-linearity of the cost function is likely to produce multiple minima which would result in a difficult convergence to the global minimum of the function. The solution proposed is a reformulation of bundle-adjustment techniques for deformable structure from motion which we describe in the following sections.

3.2 A bundle-adjustment approach to deformable modelling

The non-linear optimization of the cost function in (3.4) is achieved using a Levenberg-Marquardt [106] iterative minimization scheme modified to take advantage of the sparse block structure of the matrices involved. This method is generically termed bundle-adjustment in the computer

vision [147] and photogrammetry [5] communities and it is a standard procedure successfully applied to numerous 3-D reconstruction tasks [67]. Our main contribution here is an analysis of its applicability to the non-rigid modelling framework.

In the next section, we will review the concepts involved in bundle-adjustment (Levenberg-Marquardt minimization and sparse computation) and reformulate the factorization framework as a non-linear, large-scale minimization problem.

3.2.1 Levenberg-Marquardt minimization

Levenberg-Marquardt methods [92, 99, 106] use a mixture of Gauss-Newton and gradient descent minimization schemes switching from the first to the second when the estimated Hessian of the cost function is close to being singular. An algorithm with mixed behaviors usually obtains a higher rate of success in finding the correct minimum than other approaches. Other similar second-order or quasi-Newton algorithms may be used to minimize the cost function. However, Levenberg-Marquardt techniques have been studied and tested thoroughly in many Computer Vision applications [67] and they have been found to deliver satisfactory results. Examples are mostly given for classical inference problems in Computer Vision such as fundamental matrix computation [8], camera calibration [118] and 3-D sparse reconstruction [55]. However second-order methods have been successfully applied to less conventional geometric problems such as model-based face reconstruction [47], mosaicing [102] and reconstruction of curves [10].

Most of the computational burden of iterative second-order methods is represented by the Gauss-Newton descent step, each iteration of which requires the calculation of the inverse of the Hessian of the cost function C. Specifically to the deformable factorization case, C can be expressed in terms of the N-vector Θ containing the model parameters such that $\Theta = (\Theta_{\mathbf{l}1}, \ldots, \Theta_{\mathbf{l}F}, \Theta_{\mathrm{R}1}, \ldots, \Theta_{\mathrm{R}F}, \Theta_{\mathbf{S}1}, \ldots, \Theta_{\mathbf{S}P})^T$, where $\Theta_{\mathbf{l}i}$, $\Theta_{\mathrm{R}i}$ and $\Theta_{\mathbf{S}j}$ represent respectively the parameters for the configuration weights, orthographic cameras and 3-D basis shapes for each view and each point. Hence, the cost function C can be written as a sum of squared residuals:

$$C(\Theta) = \sum_{i,j}^{F,P} \| \mathbf{n}_{ij} \|^2 \tag{3.5}$$

where the residual for each frame and each point can be expressed as a $2FP \times 1$ vector $\mathbf{n}$ such that $\mathbf{n} = [\mathbf{n}_{11}^T \ldots \mathbf{n}_{FP}^T]^T$. At each iteration t of the algorithm, an update Δ^t is computed in order to descend to the minimum of the cost function such that the new set of parameters is given by $\Theta^{t+1} = \Theta^t + \Delta^t$. By dropping the iteration index t for notation clarity, it is necessary to express

the generic increment Δ in the model parameters as a second order Taylor expansion assuming local linearities in the cost function such that:

$$C(\Theta+\Delta) \approx C(\Theta) + \mathbf{g}^T\Delta + \frac{1}{2}\Delta^T \mathtt{H}\Delta \tag{3.6}$$

where $\mathbf{g} = \mathtt{J}^T\mathbf{n}$ is the $N \times 1$ gradient vector and $\mathtt{H}$ is the $N \times N$ Hessian matrix that can be approximated as $\mathtt{H} = \mathtt{J}^T\mathtt{J}$ (Gauss-Newton approximation of the Hessian matrix; see [147] for details) with $\mathtt{J} = \frac{\partial \mathbf{n}}{\partial \Theta}$ representing the $2FP \times N$ Jacobian matrix in the model parameters. In order to find the increment Δ, the minimum of the quadratic function $e = \mathbf{g}^T\Delta + \frac{1}{2}\Delta^T\mathtt{H}\Delta$ is computed by imposing $\frac{\partial e}{\partial \Delta} = 0$. Thus, the expression of the Gauss-Newton descent step can be finally expressed as:

$$\mathtt{H}\Delta = -\mathbf{g} \tag{3.7}$$

Levenberg-Marquardt algorithms differ from a pure Gauss-Newton method since they apply a *damping* term to equation (3.7) obtaining:

$$(\mathtt{H} + \lambda\mathtt{I})\Delta = -\mathbf{g} \tag{3.8}$$

The added term $\lambda\mathtt{I}$ has a twofold effect in the minimization. Firstly, by modifying the parameter λ, it is possible to control the behavior of the algorithm that can switch between first order (for high values of λ) and second order (low λ) iterations. Secondly, $\lambda\mathtt{I}$ makes the solution of (3.8) numerically stable by forcing that $\mathtt{H} + \lambda\mathtt{I}$ is a full-rank matrix and thus properly invertible.

3.2.2 Sparse structure of the Jacobian

Solving for the normal equations in equation (3.7) is a problem of complexity $O(N^3)$ and this step has to be repeated at each iteration. In order to render the computation feasible as the number of parameters increases, it is possible to exploit the sparse structure of the Jacobian $\mathtt{J}$.

Motion components (configuration weights and camera parameters) are unrelated between different views and, similarly, structure components are unrelated between different point trajectories. As a result, the Jacobian matrix contains a large number of entries for which the partial derivatives are zero, as we show in the graphical representation of its structure in figure 3.1.

It is possible to solve for the increment Δ in (3.7) efficiently by calculating the inverse of $\mathtt{H}$ using the sparse structure of $\mathtt{J}$. Standard approaches for sparse computation are described in [147] and [67]. Notice that, again, this property is valid for any rigid and non-rigid factorization model, since the sparseness relation is given by the independency between motion parameters

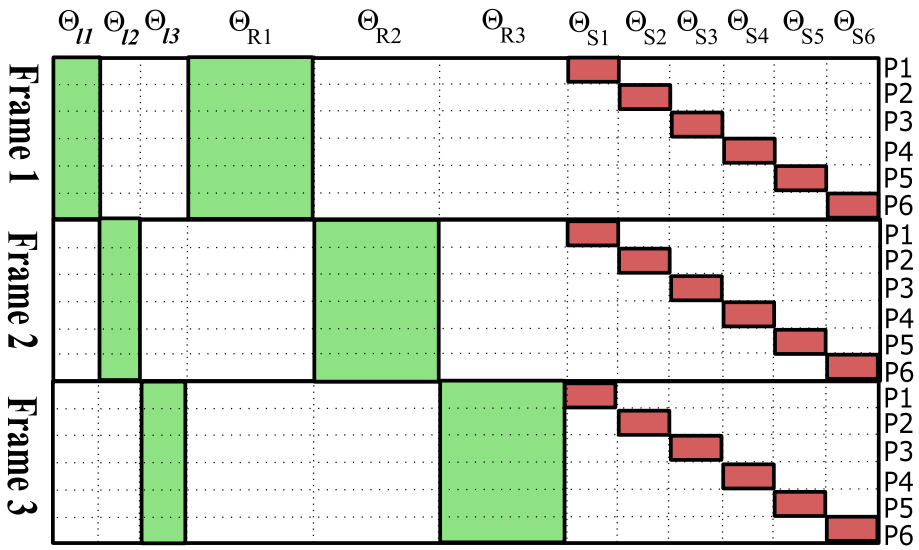


Figure 3.1: Sparse structure of the Jacobian matrix. We show an example for 3 frames and 6 points ($P1, P2, P3, P4, P5, P6$). The zero-entries of the matrix are displayed as white blocks. Θ_{l1}, Θ_{l2} and Θ_{l3} represent the configuration weights respectively for frame 1, 2 and 3. Θ_{R1}, Θ_{R2} and Θ_{R3} are the vectors of the camera components for each frame and Θ_{S1}, Θ_{S2}, Θ_{S3}, Θ_{S4}, Θ_{S5}, Θ_{S6} encode the basis shapes for each deformable point.

(for each frame) and 3-D structure (for each point) in the multi-view cost function and thus independent of the chosen model.

3.2.3 Proposed implementation

The cost function of a deformable object presents more degrees of freedom than in the rigid case, which could lead to the existence of multiple local minima for the motion, deformation and structure components. It is possible to reduce the chance of falling into local minima by carefully designing the algorithm with respect to the following aspects: initialisation, model parameterisation and the use of priors.

Parameterisation

The camera matrices $\mathtt{R}_i$ are parameterised using unit quaternions [74] giving a total of $4 \times F$ rotation parameters, where F is the total number of frames. Quaternions ensure that there are no strong singularities and that the orthonormality of the rotation matrices is preserved by merely enforcing the normality of the 4-vector. This would not be the case with the Euler angle or the rotation matrix parameterisations, where orthonormality of the rotations is more complex to

preserve. The quaternion normalization is directly enforced in the cost function by dividing the quaternion with its norm. Indeed, in an initial implementation the 3-D pose was parameterised using the 6 entries of the rotation matrices $\mathtt{R}_f$, however the use of quaternions led to improved convergence and to much better results for the rotation parameters and the 3-D pose.

The method proposed by Bar-Itzhack [6] in an attitude control context is used to obtain the quaternions from the set of rotation matrices $\mathtt{R}_i$. The algorithm has the main advantage to yield the closest quaternion representation if the constraints of matrix orthonormality are not exactly satisfied. This eventuality usually appears during the initialisation of the non-linear optimization scheme after the first computation of the corrective transform $\mathtt{Q}_{3\times3}$ for the rigid component of the motion. Schematically, the method first define the matrix $\mathtt{B}$ given the singular elements $\{r_{mn}\}$ belonging to a generic 3×3 rotation matrix $\mathtt{R}$:

$$\mathtt{B} = \frac{1}{3}\begin{bmatrix} r_{11}-r_{22}-r_{33} & r_{21}+r_{12} & r_{31}+r_{13} & r_{23}+r_{32} \\ r_{21}+r_{12} & r_{22}-r_{11}-r_{33} & r_{32}+r_{23} & r_{31}-r_{13} \\ r_{31}+r_{13} & r_{32}+r_{23} & r_{33}-r_{22}-r_{11} & r_{12}-r_{21} \\ r_{23}-r_{32} & r_{31}-r_{13} & r_{12}-r_{21} & r_{11}+r_{22}+r_{33} \end{bmatrix} \tag{3.9}$$

The algorithm then follows with the following three steps:

1. Compute the eigenvalues of $\mathtt{B}$.
2. Find the largest eigenvalue λ_{max}.
3. Extract the eigenvector of $\mathtt{B}$ which correspond to λ_{max}.

The given eigenvector is the closest quaternion to the matrix $\mathtt{R}$. In the case of an exact orthonormal matrix we would obtain $\lambda_{max} = 1$.

Finally, the structure is parameterised with the $(3\times D)\times P$ coordinates of the $\mathbf{S}_d$ shape bases and the $D\times F$ deformation weights l_{id}.

Initialisation

A further critical factor is the choice of an initialisation for the parameters of the model. It is crucial, for bundle adjustment techniques to work, that the initial estimate be close to the global minimum to increase the speed of convergence and reduce the chance of being trapped in local minima, particularly when the cost function has a large number of parameters as in this case.

A similar initialisation to the one used by Torresani et al. in their tri-linear optimization scheme [141] is chosen. The idea is to initialize the camera matrices with the motion corresponding to the rigid component, which is likely to encode the most significant part of the motion.

A different initialisation which gives a reasonable starting point is to use the estimates given by Brand's algorithm for both motion and structure [16]. Occasionally, however, we have observed problems with the convergence given this initialisation and generally when the motion associated to the rigid component is used as the initial estimate the minimization reaches the minimum of the cost function in fewer iterations.

Regularization prior

Occasionally, the non-linear optimization leads to a solution corresponding to a local minimum. In particular, at times the 3-D points tend to lie on a plane. To overcome this situation, a prior on the 3-D shape has been added to the cost function. The prior states that the depth of the points on the object's surface cannot change significantly from one frame to the next since the images are closely spaced in time. This is implemented by adding a penalty term C_s that penalizes for strong variations between the shape at frames i and $i+1$ given by:

$$C_s(\Theta) = \| \sum_{d=1}^{D} l_{id} \mathbf{S}_d - \sum_{d=1}^{D} l_{(i+1)d} \mathbf{S}_d \|^2 \tag{3.10}$$

In this way the relief present in the 3-D data is preserved. Similar regularization terms have also been reported in [2, 141].

3.3 Previous work in non-rigid BA

Aanæs and Kahl, also proposed a bundle adjustment solution for the non-rigid scenario [2]. However, their approach differs in some fundamental aspects. Firstly, their initial estimate of the non-rigid shape was obtained by estimating the mean and variance of the 3-D data obtained directly from image measurements. The approach assumes that the cameras are calibrated, and although the authors state that their algorithm would work in the uncalibrated case they do not give experimental evidence. In contrast, we consider a scenario based on uncalibrated data from a generic video sequence. The second main difference is in the parameterisation of the problem. In [2] the camera rotations are parameterised by the six elements of the rotation matrix. We are using quaternions instead which, as will be shown in the experimental section, leads to better behaved results for the motion estimates.

In terms of their experimental evaluation, Aanæs and Kahl do not provide an analysis of the recovered parameters, only some qualitative results of the 3-D reconstruction. In contrast, our quantitative experimental analysis shows that we are able to decouple motion and deformation parameters (see next section for a detailed description).

3.4 Experimental results

In this section we show results of our non-linear optimization approach with synthetic and real image sequences. The quality of the 3-D reconstructions are both evaluated quantitatively with respect to ground truth values and qualitatively over two sequences with a subject performing different facial expressions.

3.4.1 Synthetic data

Xiao et al. [159] showed in recent experiments that previous methods for deformable factorization [19, 16, 141] may fail even for simple deforming objects. Using similar synthetic data sets, the forthcoming tests will shed some light on the efficiency of the proposed non-linear optimization procedure for 3-D reconstruction. The experiments are constructed by generating a random set of D basis shapes whose linear combination creates varying deformable 3-D shapes contained in a cube of $50 \times 50 \times 50$ units (see figure 3.4.1). The set of configuration weights l_{id} are obtained by fitting polynomials to randomly generated values. This was necessary to obtain smoother deformations rather than erratic and unrealistic changes in the 3-D structure at each frame. Notice that the configuration weights and, thus, the temporal evolution of the deformations are as generic as possible. For instance, there is no assumption of independency of the basis shapes as required by the method of Xiao et al. [159] (see section 2.4.5 for a description). Finally, the generated 3-D shapes are projected onto the image plane (of size 640×480) by means of random orthographic cameras $\mathtt{R}_i$. The experimental setup is completed by fixing the number of points to $P = 40$ and frames to $F = 30$.

Two set of tests are presented. First, the number of basis shapes was varied such that $d = 2 \ldots 5$ to verify the algorithm's performance with increasingly complex deformations. A second test is then performed to obtain an evaluation of the quality of the reconstruction in case of varying strength of the deformations but fixing the number of basis shapes to $D = 3$. This

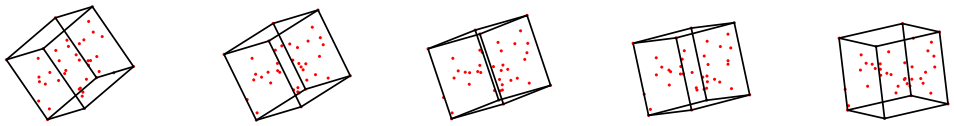
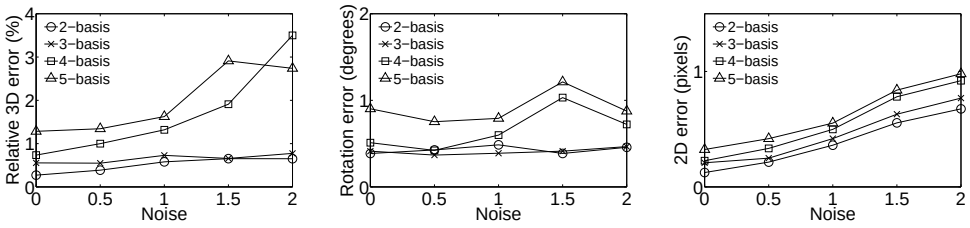


Figure 3.2: Some frames of the cube sequence used for testing the algorithm. The deformable points are sampled inside a cube of $50 \times 50 \times 50$ (wire-frames are added to show the solid contour).

measurement is directly calculated between the ratio of the norm of the rigid components of the 3-D metric shapes and the norm of the 3-D deformable structures (before projection) such that $ratio = \frac{\|\mathbf{S}_{nonrigid}\|}{\|\mathbf{S}_{rigid}\|}$. In order to validate the performance 25 trials were performed for each setup and for different Gaussian noise conditions with variance $\sigma = 0.5, 1, 1.5, 2$.

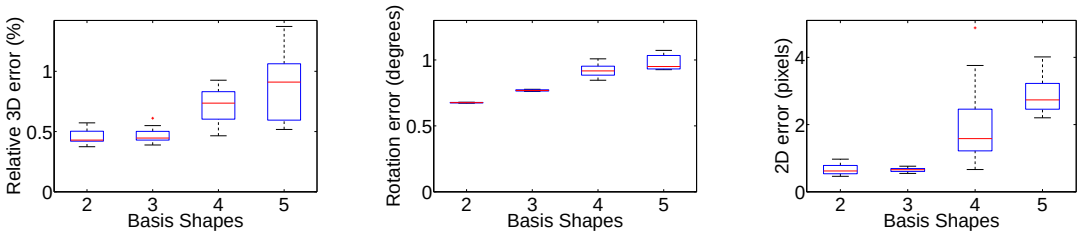
The results are obtained with a MATLAB implementation of non-rigid bundle adjustment using the built-in function `lsqnonlin` for non-linear minimization. The software is designed in a such way that the sparse structure of the Jacobian is automatically computed by calculating the derivatives of the cost function with different number of basis shapes D. Initialisation of the model parameters is as described in the previous section. The stop criteria was fixed for the tolerance over the increment in the model parameters (fixed at 10^{-6}). The minimisation usually converges in a time ranging between 10–30 seconds (on a AMD–Athlon X2 computer clocked at 3800 MHz) for the set of synthetic data considered in the experiments.



3-D Error Rotation Error 2-D Error

Figure 3.3: Relative 3-D error (%), r.m.s. rotation error (deg) and 2-D reprojection error for the synthetic experiments for different basis shapes $d = 2 \ldots 5$ and increasing levels of Gaussian noise. The ratio of non-rigidity is fixed to 40% for all the trials.

Figure 3.3 shows three plots representing the 3-D reconstruction error expressed in percentage relative to the scene size (which it is defined as the maximum of the x, y and z coordinates),



3-D Error Rotation Error 2-D Error

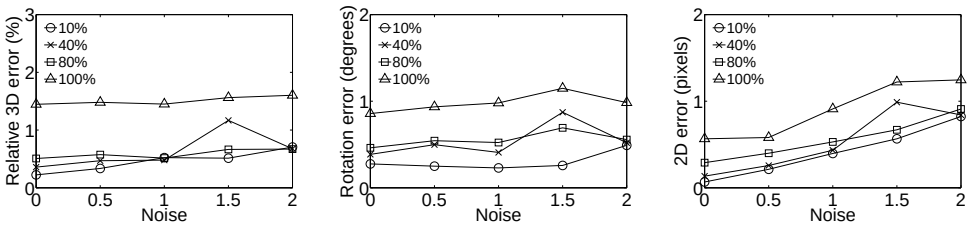
Figure 3.4: Relative 3-D error (%), r.m.s. rotation error (deg) and 2-D reprojection error box-plots for the synthetic experiments for different basis shapes $d = 2 \ldots 5$ and Gaussian noise fixed at $\sigma = 1.5$. The ratio of non-rigidity is fixed to 40% for all the trials.

the absolute rotation error expressed in degrees for varying number of basis shapes and the root mean squared (r.m.s.) 2-D image reprojection error expressed in pixels. The plots of this figure show the mean values corresponding to 25 random trials applied to each level of Gaussian noise. As expected, higher complexity in the degrees of deformation (given by the increasing number of basis) results in worse performance of the algorithm. Note that the increasing levels of Gaussian noise do not affect the estimate of the 3-D structure and rotations strongly.

In order to evaluate more accurately the results, a box-plot in figure 3.4 shows the statistical properties of the errors for the experiment with Gaussian noise level fixed at $\sigma = 1.5$. The plot consists of four blue boxes (one for each number of basis) which lower and upper lines define the 25th and 75th percentiles of the sample. The red line in the middle of the box is the sample median. The black lines extending above and below the box show the range of the rest of the samples. The outliers are shown as red plus signs and they represent a problem in the algorithm convergence to the minimum. Usually this refers to the minimisation being trapped in a local minima.

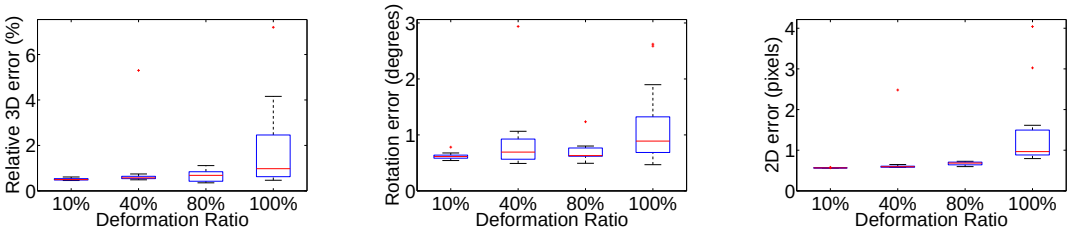
Figure 3.5 shows results of experiments for increasing degrees of non-rigidity of the 3-D structure. Notice that, higher levels of deformity in the shape negatively affect the estimation of the model parameters. An important observation for both experiments is the following: the recovered values for the 3-D reconstruction and rotation errors do not converge to the global minimum in the case of no noise (when perfect data is available).

The box-plots in figure 3.6 reveal a higher rate of outlier errors showing more difficulties in finding the global minima in the case of increasing deformations. In these cases the algorithm showed a tendency to converge to the minimum too slowly or to converge to a local solution.



3-D Error Rotation Error 2-D Error

Figure 3.5: Relative 3-D error (%), r.m.s. rotation error (deg) and 2-D reprojection error for the synthetic experiments for different ratio of deformation (10%, 40%, 80%, 100%) and increasing levels of Gaussian noise.



3-D Error Rotation Error 2-D Error

Figure 3.6: Relative 3-D error (%), r.m.s. rotation error (deg) and 2-D reprojection error box-plots for the synthetic experiments for different ratio of deformation (10%, 40%, 80%, 100%) and fixed Gaussian noise ($\sigma = 1.5$).

This effect is a consequence of the intrinsic ambiguity of the solutions in the case of deformable structure from motion as discussed in Xiao et al. work's [159]. In order to solve this problem, we will introduce our solution based on rigidity priors later in chapter 5.

3.4.2 Experiments with real images and manually tracked data

In this section we compare the results obtained with our bundle-adjustment based 3-D reconstruction algorithm with those obtained using Brand's non-rigid factorization method [16]. A direct comparison with Xiao et al.'s approach is not meaningful since we did not find it possible to extract a set of independent basis shapes that lead to a reasonable reconstruction (a problem already reported in [17] for real data). A real video test sequence shows the face of a subject performing an almost rigid motion for the first 200 frames, moving his head up and down. The subject then changed facial expression with his head facing front for the next 309 frames (see figure 3.7). The point features which appear in figure 3.7 were manually marked throughout the

Figure 3.7: Key frames of the sequence used in the experiments in section 3.4.2, with manually tracked points superimposed. The subject performed an almost rigid motion for the first 200 frames moving the head up and down and then changed facial expression for the next 309 frames.

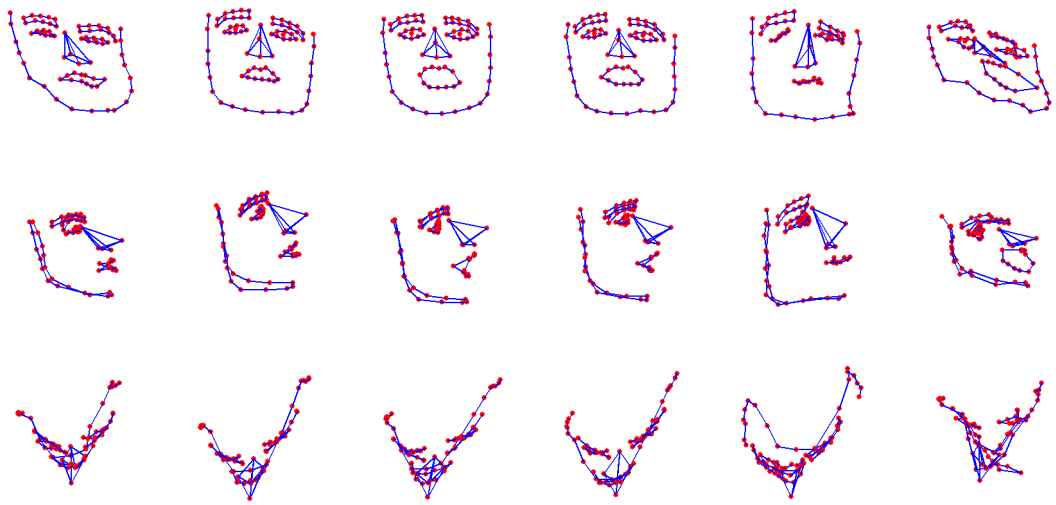
sequence. The number of basis shapes is fixed heuristically to $D = 5$, a compromise between the complexity of the model and the number of captured deformations. The computation time required for the algorithm to convergence is consistently higher (8 minutes approximately) given the number of frames, points and deformations which increases the number of parameters to estimate.

The results of the 3-D reconstructions[1] for some key frames in the sequence obtained using Brand's factorization method are shown in figure 3.8. The front views of the 3-D reconstruction show that the recovered 3-D shape does not reproduce the facial expressions accurately. Besides, depth estimation is not precise, which is evident by inspection of the top views of the reconstruction. Notice the asymmetry of the left and right sides of the face.

In figure 3.9 we show the reconstructed 3-D shape recovered after applying the bundle adjustment refinement step. The facial expressions in the 3-D plots reproduce the original ones reliably: notice for example the motion of the eyebrows in the frowning expression (frame 467) or the opening of the mouth in surprise (frame 358). Finally, the top views show that the overall relief appears to be well preserved, as is the symmetry of the face.

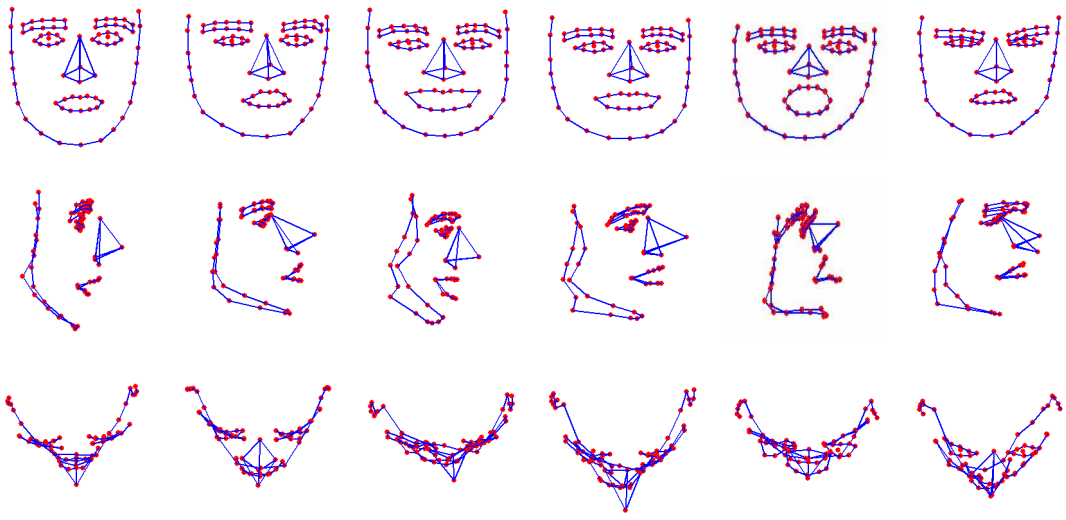
The evolution of the weights l_{id} of the deformation modes can be traced throughout the sequence. In figure 3.10 we show the value of the weight associated with the mean component

[1] Video available at http:/www.bmva.ac.uk/thesis_archive/2006/DelBue1/index.html



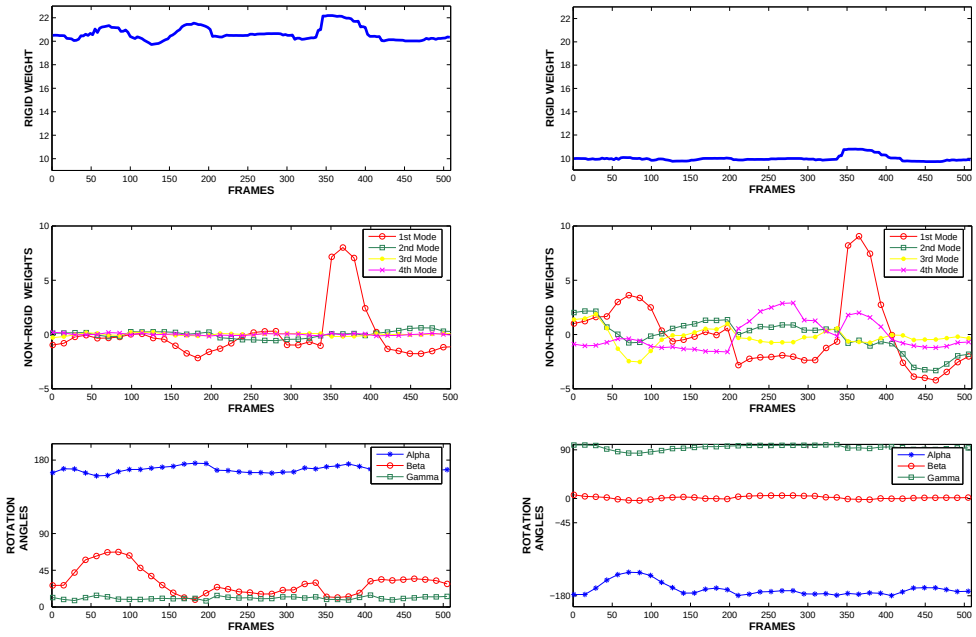
Frame 1 Frame 67 Frame 275 Frame 300 Frame 358 Frame 467

Figure 3.8: Front, side and top views of the 3-D reconstructions obtained from the non-rigid factorization algorithm without bundle adjustment for some of the key frames in the sequence. No ground truth is available in this experiment.



Frame 1 Frame 67 Frame 275 Frame 300 Frame 358 Frame 467

Figure 3.9: Front, side and top views of the 3-D reconstructions obtained after applying non-linear optimization. No ground truth is available in this experiment.



(A) Results from Brand's factorization (B) Results after bundle adjustment

Figure 3.10: Values obtained for the rigid component (top), deformation weights (middle) and rotation angles (bottom) using Brand's approach (A) and bundle adjustment (B) for the sequence in figure 3.7.

(top) $d = 1$ and of those associated with the 4 remaining deformation modes (middle). Results are given for both Brand's flexible factorization (left) and for the bundle adjustment scheme (right). Notice how Brand's flexible factorization has a tendency to suppress weak deformations – the weights associated with the deformation modes for frames with small deformations have a small value. This results in the recovered 3-D shape not reproducing the facial expressions accurately. The weights associated with the deformation modes have higher values in the bundle-adjusted solution. Interestingly, around frame 360 the first non-rigid mode of deformation experiences a large peak, which corresponds to the opening of the mouth in surprise as shown in figure 3.7. This indicates some tendency in the configuration weights to reflect the underlying facial expressions. Although this peak is present also in Brand's solution, it is possible to observe by visual inspection that the corresponding 3-D reconstruction in figure 3.8 is not very accurate.

The results obtained for the motion parameters are shown in the bottom graph of figure 3.10. The rotation angles around the X, Y and Z axes (up to an overall rotation) are recovered for each of the 509 frames in the sequence. In particular, the tilt angle varied smoothly throughout the first 200 frames capturing the up and down tilt of the head of about 50 degrees in total while

the rotation angles around the other 2 axes did not vary significantly throughout the sequence. Notice that both solutions capture this motion correctly. However, the results obtained with the bundle-adjusted solution (right) qualitatively presents less variations in parts of the scene where the subject is not rigidly moving. Using Brand's algorithm (left), it is possible to notice sudden variations of the motion which cannot be observed by visual inspection in the image sequence. This indicates that the estimation of the orthographic camera matrices in Brand's method may be affected by the deformations appearing in the scene.

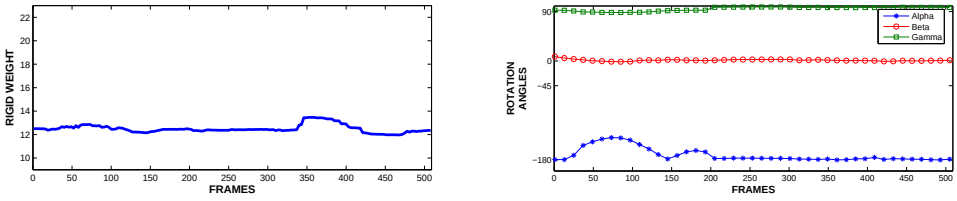


Figure 3.11: Values used for the initialisation of the non-linear minimization algorithm. The value obtained for the rigid component (left) and rotation angles (right) are computed with the motion corresponding to the rigid component.

The non-linear refinement step is initialised using the values of the first configuration weight and the rotation angles associated with the mean component as shown in figure 3.11. Note that the deformable bases and configuration weights are initialized to very small random values. This initialisation was first used by Torresani et al. [141] in their tri-linear optimization stage and it provided reasonable results. It can be observed from the plot that the rigid component of the motion is a good description of the object's rotation, and in fact the bundle-adjustment step does not optimize these parameters much further and focuses on the refinement of the deformation parameters.

3.4.3 Experiments with real images and automatically tracked data

In this section, the behavior of the method is tested with image measurements obtained automatically with a point tracking algorithm [37]. The scope of this test is to show the feasibility of a complete unsupervised system for 3-D deformable reconstruction starting from an uncalibrated video sequence showed in figure 3.12. A ranklet-based tracker [129] specially designed to cope with deforming structures automatically generates the tracks that are input into the non-linear optimization scheme. The system has to cope with a complex 960 frame sequence in which the

Frame 1 Frame 250 Frame 335 Frame 420 Frame 541 Frame 710 Frame 960

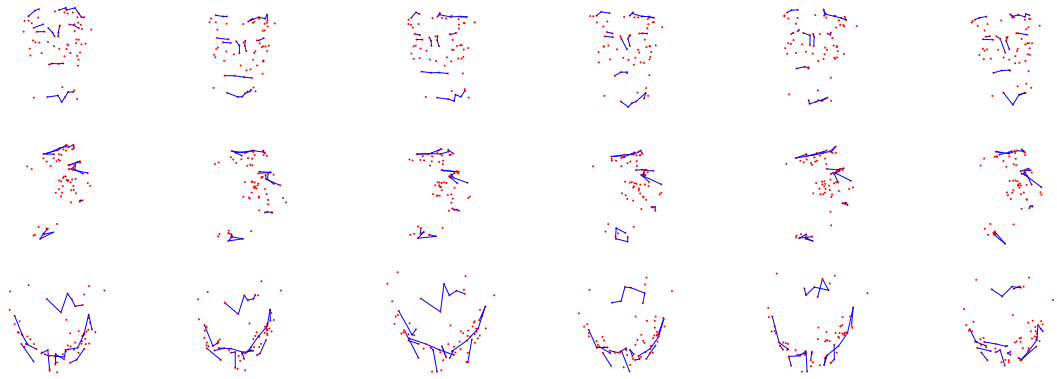
Figure 3.12: Key frames in the sequence used to test the reconstruction of a 3-D deformable shape with automatic tracking of feature points. The subject performed simultaneous rigid and non-rigid motion. Automatically tracked points are superimposed. A set of wireframes outlines the face structure.

subject is undergoing 3-D motion and performing different facial expressions.

A total of 91 points were initialized automatically according to a saliency criterion [38]. The tracker was able to follow a good number of feature points reliably throughout the sequence, even in relatively poorly textured areas such as the subject's cheekbones. Throughout the 960 frame sequence, only 8 points out of the initial 91 were lost. However, a certain number of points initialized on homogeneous texture turned out to be unreliable, and they evidently affect the 3-D shape estimation in those areas.

Figure 3.13 shows the front, top and side views of the 3-D reconstruction of six key frames with different expressions. The number of basis shapes is fixed to $D = 8$ since this value can generate a model which capture most of the deformations appearing in the video sequence. Higher values for D would obtain more accurate models but at the cost of a higher computational time required to minimize the cost function. The initialisation of the non-linear optimization is identical to the one described in section 3.2.3. The overall depth is generally correct: notice the point belonging to the neck relative to the position of the face, and the nose pointing out from the face plane. Face symmetry is generally well preserved, as it is possible to notice from the top views of the reconstruction. Some outliers are obvious in frame 710 in the eyebrow region and generally on the neck area where the tracker performs poorly; such feature points are wrongly reconstructed by our non-rigid model.

Finally, the reconstructed motion and deformation parameters are displayed in figure 3.14. The estimated angles follow the rotation of the subject's head reasonably, with values limited between 10 and -15 degrees for the "beta" angle, while "alpha" and "gamma" show tiny variations. The rigid weight is nearly constant for the whole sequence in accordance with the subject's head



Frame 1 Frame 250 Frame 335 Frame 541 Frame 710 Frame 960

Figure 3.13: Front, side and top views of the 3-D reconstructions obtained by the combined system for some of the key frames in the sequence.

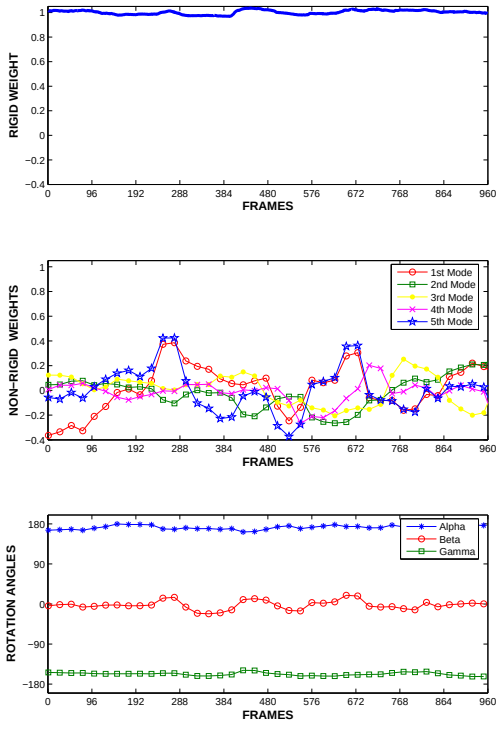


Figure 3.14: Evolution of the rigid weight $(D = 1)$, the first five non-rigid weights $(D = 2, \ldots, 6)$ and the rotation angles (in degrees) throughout the sequence of figure 3.12.

being at the same distance from the camera. The non-rigid configuration weights present more erratic behavior; the two spikes around frame 280 and 670 correspond respectively to a grin and an angry facial expression.

3.5 Summary

Non-linear optimization is applied to obtain a reliable solution for 3-D deformable reconstruction from uncalibrated video sequences. The key features of the approach consist on enforcing the repetitive pattern of the motion matrix $\mathtt{M}$ while at the same time explicitly considering a proper parameterisation for the orthographic cameras using quaternions. Further care is put to render the approach tractable: Levenberg-Marquardt minimization safely descends towards the minimum of the defined cost function and sparse computation efficiently solves for each iteration.

In contrast, the previous linear methods obtained approximate solutions by neglecting the non-linear structure of the framework. The direct consequence is a coupling of motion and deformation components as we have observed in the results using Brand's method [16]. Xiao et al.'s [159] approach avoids the ambiguities but needs to make assumptions about the independency of the 3-D basis shapes.

However, it is shown in the synthetic tests that our non-linear optimization approach not always converges to the global minimum of the cost function. This effect is a consequence of the intrinsic ambiguity of the solutions: local minima are likely to be present if additional information about the 3-D structure of the deforming object is not introduced as previously discussed by Xiao et al. in [159]. In order to solve this problem, we will introduce our solution based on rigidity priors later in chapter 5.

The framework presented here can be easily extended to deal with different types of non-rigid objects (for instance, articulated structures) and of camera models by changing the cost function C accordingly. Additionally, prior information and/or regularization terms may be easily inserted in the minimization by adding quadratic penalty terms in the same way as those introduced in equation (3.10) to ensure the temporal smoothness of the 3-D reconstructions. These terms may help descend towards the global minimum of the cost function and, if applied strictly, can force specific priors on the motion $\mathtt{M}$ and 3-D structure components $\mathtt{S}$.

The expression of the problem as a sum of cost functions for each image point $\mathbf{w}_{ij}$ allows us to deal with missing entries in the measurement matrix $\mathtt{W}$. Hence, if a point becomes occluded

at a certain frame (a likely event in a practical scenario), it is still possible to perform non-linear optimization by not including the cost function related to the lost entry in the minimization.

Although robust estimation is not an issue of this work, point trajectories could have uncertainty information associated with their covariance matrix $\mathtt{C}_{ij}$ derived from the image point tracking algorithm. In this case, it would be possible to define optimal estimates of the parameters given the uncertainties by minimizing the Mahalanobis distance of the quadratic terms $\sum_{i,j}^{F,P} \| \mathbf{n}_{ij} \|^2_{\mathtt{C}_{ij}}$. The covariance values can be easily included in the estimation and may lead to a more robust inference.

In the following chapter, the non-rigid factorization framework will be extended to deal with the information extracted from multiple cameras; a necessary solution when the inspected deformable object undergoes minimal rigid motion.

Chapter 4

Stereo Non-Rigid Factorization

The factorization framework is a flexible tool for modelling data from point trajectories extracted from uncalibrated video sequences. In the case of deformable objects, an aspect of relevant interest is the applicability of the previously described algorithms to the case when the object is viewed by multiple cameras. More specifically, we have formulated the problem for a stereo rig, where the two cameras remain fixed relative to each other throughout the sequence. In this case the measurement matrix requires not only the temporal tracks of points in the left and right image sequences but also the stereo correspondences between left and right image pairs. We have developed a new method to factorize the measurement matrix into the left and right motion matrices and the 3-D non-rigid shape. Note that this method requires both cameras to be synchronized. However, if this were not the case, it could be elegantly solved inside a factorization framework using the solution proposed by Tresadern and Reid [142] for the synchronization of stereo video sequences in an uncalibrated scenario.

4.1 Stereo, motion and structure

Using a calibrated stereo pair is a common and practical solution to obtain reliable 3-D reconstructions (see figure 4.1). In its simpler formulation, once the stereo rig is calibrated, the depth of points in the image is estimated by applying triangulation [148]. In order to obtain accurate depth estimates, the cameras are usually separated from each other by a significant baseline thus creating widely spaced observations of the same object. The disadvantage of this configuration though, is that having a wide baseline makes the matching of features between pairs of view a

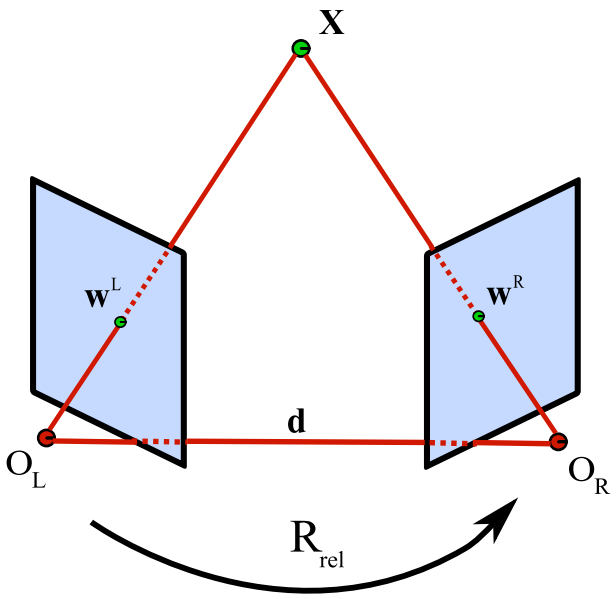


Figure 4.1: A classic stereo setup. The 3-D point **X** is projected into the left and right images with coordinates $\mathbf{w}^L$ and $\mathbf{w}^R$. The camera centers $\mathbf{O}_L$ and $\mathbf{O}_R$ are displaced in 3-D with a baseline **d** and relatively rotated with a 3×3 rotation matrix $\mathtt{R}_{rel}$

more challenging problem.

On the other hand, the task of computing temporal tracks from the single camera sequences is relatively easier since the images are closely spaced in time. As a drawback, disparities may be insufficient to obtain a reliable depth estimation and, as a result, longer sequences are needed to infer the 3-D structure. Particularly, in the case of non-rigid structure, a sufficient overall rigid motion is necessary to allow the algorithms to correctly estimate the reconstruction parameters.

Hence, a question of relevant interest is the feasibility of an approach that efficiently fuses the positive aspects of both methods. The problem of recovering 3-D structure using a stereo-rig moving in time or a stereo rig looking at a moving object has been defined for the rigid case as the *stereo-motion* problem [154, 39, 131, 97] (see figure 4.1). Ho and Chung [73] first formulated this problem within the factorization scenario. Following a similar direction, we introduce a multi-camera motion model that is able to deal with a time-varying shape and to find a linear solution that is subsequently optimized with the non-linear procedure presented in the previous chapter.

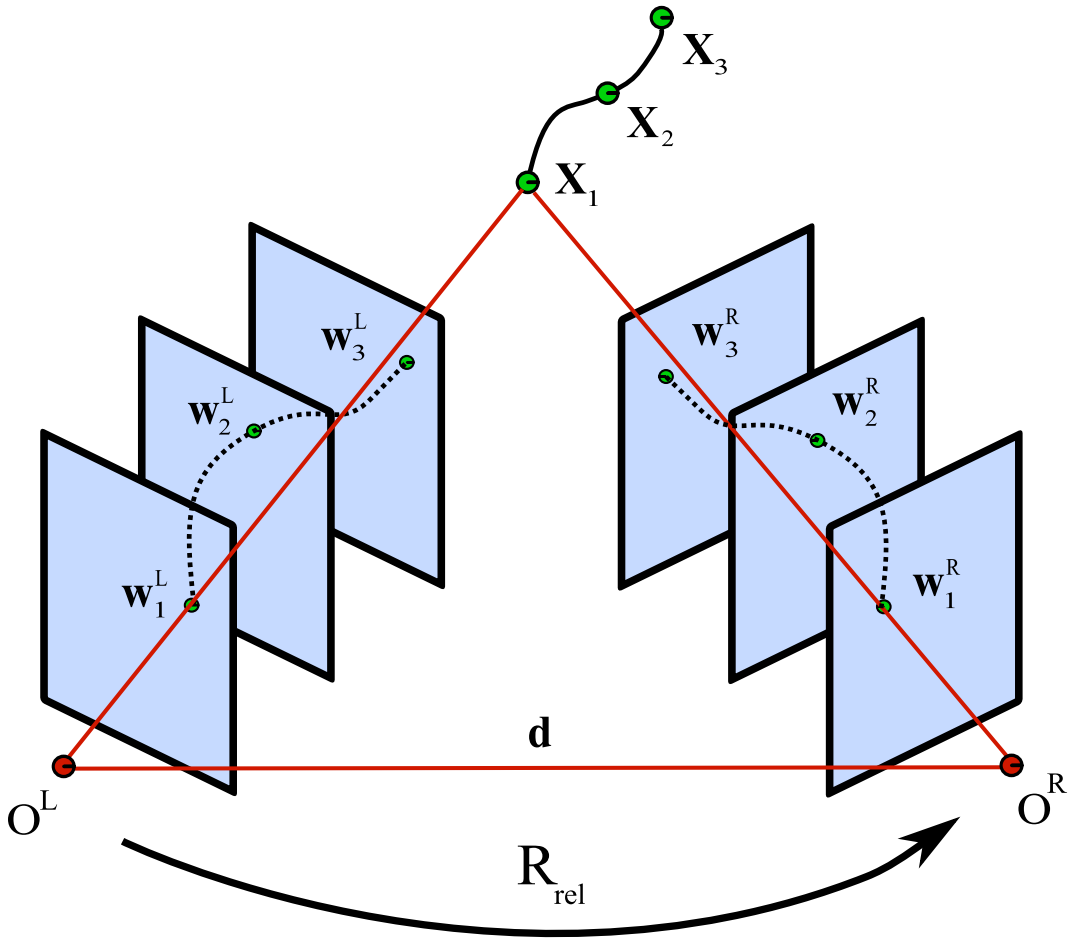


Figure 4.2: A stereo motion setup. A point is moving in space and its position in 3-D is shown for each time instance as $\mathbf{X}_1$, $\mathbf{X}_2$ and $\mathbf{X}_3$. The three points are then projected into the respective image frames obtaining the image coordinates $\mathbf{w}_1^L$, $\mathbf{w}_2^L$ and $\mathbf{w}_3^L$ for the left camera and $\mathbf{w}_1^R$, $\mathbf{w}_2^R$ and $\mathbf{w}_3^R$ for the right one. The dotted lines connecting the points represent the 2-D trajectory in time of the point in the left and right images. Since the position of the cameras is fixed, the relative orientation $\mathtt{R}_{rel}$ and camera displacement $\mathbf{d}$ between the camera centers $\mathbf{O}^L$ and $\mathbf{O}^R$ are considered constants in time.

4.2 The stereo camera case

The main contribution presented here is to extend the non-rigid factorization methods to the case of a stereo rig, where the two cameras remain fixed relative to each other throughout the sequence. However, the same framework could be used in the case of 3 or more cameras. Torresani et. al. [141] first introduced the factorization problem for the multiple camera case but they did not provide an algorithm or any experimental results.

4.2.1 The stereo motion model

When two cameras are viewing the same scene, the measurement matrix $\mathtt{W}$ will contain the image measurements from the left and right cameras resulting in a $4F \times P$ matrix where F is the number of frames and P the number of points. Assuming that not only the single-frame tracks but also the stereo correspondences are known we may write the measurement matrix $\mathtt{W}$ as:

$$\mathtt{W} = \begin{bmatrix} \mathtt{W}^L \\ \mathtt{W}^R \end{bmatrix} \tag{4.1}$$

where for each frame i the stereo correspondences are:

$$\mathtt{W}_i^L = \begin{bmatrix} \mathbf{w}_{i1}^L & \dots & \mathbf{w}_{iP}^L \end{bmatrix} \quad \mathtt{W}_i^R = \begin{bmatrix} \mathbf{w}_{i1}^R & \dots & \mathbf{w}_{iP}^R \end{bmatrix} \tag{4.2}$$

Note that, since we assume that the cameras are synchronized, at each time step i the left and right cameras are observing the same 3-D structure and this results in the additional constraint that the structure matrix $\mathtt{S}$ and the deformation coefficients l_{id} are shared by left and right camera. The measurement matrix $\mathtt{W}$ can be factored into a motion matrix $\mathtt{M}$ and a structure matrix $\mathtt{S}$ which take the following form:

$$\mathtt{W} = \begin{bmatrix} l_{11}\mathtt{R}_1^L & \dots & l_{1D}\mathtt{R}_1^L \\ \vdots & & \vdots \\ l_{F1}\mathtt{R}_F^L & \dots & l_{FD}\mathtt{R}_F^L \\ \hline l_{11}\mathtt{R}_1^R & \dots & l_{1D}\mathtt{R}_1^R \\ \vdots & & \vdots \\ l_{F1}\mathtt{R}_F^R & \dots & l_{FD}\mathtt{R}_F^R \end{bmatrix} \begin{bmatrix} \mathtt{S}_1 \\ \vdots \\ \mathtt{S}_D \end{bmatrix} \tag{4.3}$$

where $\mathtt{R}^L$ and $\mathtt{R}^R$ are the rotation components for the left and right cameras. Once more, we have eliminated the translation for both cameras by registering image points to the centroid in each frame.

Note that the assumption that the deformation coefficients are the same for the left and right sequences relies on the fact that the weak perspective scaling f/Z_{avg} must be the same for both cameras. This assumption is generally true in a symmetric stereo setup where f and Z_{avg} are usually the same for both cameras.

It is also possible to express the stereo motion matrix $\mathtt{M}$ by including explicitly the assumption that a fixed stereo rig is being used. In this case the rotation pair for the left and right cameras can be expressed in terms of the matrix that encodes their relative orientation matrix $\mathtt{R}_{rel}$ such that: $\mathtt{R}^R = \mathtt{R}_{rel}\mathtt{R}^L$. The motion matrix $\mathtt{M}$ in equation (4.3) can be consequently expressed as:

$$\mathtt{M} = \left[\begin{array}{ccc} l_{11}\mathtt{R}_1^L & \dots & l_{1D}\mathtt{R}_1^L \\ \vdots & & \vdots \\ l_{F1}\mathtt{R}_F^L & \dots & l_{FD}\mathtt{R}_F^L \\ \hline l_{11}\mathtt{R}_{rel}\mathtt{R}_1^L & \dots & l_{1D}\mathtt{R}_{rel}\mathtt{R}_F^L \\ \vdots & & \vdots \\ l_{F1}\mathtt{R}_{rel}\mathtt{R}_F^L & \dots & l_{FD}\mathtt{R}_{rel}\mathtt{R}_F^L \end{array}\right] \tag{4.4}$$

4.2.2 Non-rigid stereo factorization

Once more the rank of the measurement matrix $\mathtt{W}$ is at most $3D$ since $\mathtt{M}$ is a $4F \times 3D$ matrix and $\mathtt{S}$ is a $3D \times P$ matrix, where P is the number of points. Assuming that the single frame tracks and the stereo correspondences are all known, the measurement matrix $\mathtt{W}$ may be factorized into the product of a motion matrix $\mathtt{M}$ and a shape matrix $\mathtt{S}$ by truncating the SVD of $\mathtt{W}$ to rank $3D$ (see section 2.4.1):

$$\mathtt{W} = \left[\begin{array}{c} \mathtt{W}^L \\ \mathtt{W}^R \end{array}\right] = \tilde{\mathtt{M}}\tilde{\mathtt{S}} = \left[\begin{array}{c} \tilde{\mathtt{M}}^L \\ \tilde{\mathtt{M}}^R \end{array}\right] \tilde{\mathtt{S}} \tag{4.5}$$

Computing the transformation matrix $\mathtt{Q}$

The result of the factorization is not unique since $(\tilde{\mathtt{M}}\mathtt{Q})(\mathtt{Q}^{-1}\tilde{\mathtt{S}})$ would give an equivalent factorization. We proceed to apply the metric constraint in a similar way as was described for the single camera case in section 2.4.4, correcting each $4F \times 3$ vertical block in $\tilde{\mathtt{M}}$ independently. Note that in this case we have used five constraints per frame: 2 orthogonality constraints (one from each camera) and 3 equal norm constraints (computed from rows $2i-1$, $2i$, $2i+2F-1$, $2i+2F$ of the motion matrix $\tilde{\mathtt{M}}$ where i is a generic frame). Each vertical block will then be corrected as:

$\hat{\mathtt{M}}_d \leftarrow \tilde{\mathtt{M}}_d \mathtt{Q}_d$. The overall transformation $\mathtt{Q}$ is a block diagonal matrix such that:

$$\mathtt{Q} = \begin{bmatrix} \mathtt{Q}_1 & \mathtt{0} & \dots & \mathtt{0} \\ \mathtt{0} & \mathtt{Q}_2 & \dots & \mathtt{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathtt{0} & \mathtt{0} & \dots & \mathtt{Q}_D \end{bmatrix} \tag{4.6}$$

The shape matrix will be corrected with the inverse of the block-diagonal transformation: $\mathtt{S} \leftarrow \mathtt{Q}^{-1}\tilde{\mathtt{S}}$.

Factorization of the motion matrix $\tilde{\mathtt{M}}$

In the stereo case we factorize each $4 \times 3D$ sub-block of the motion matrix (which contains left and right measurements for each frame i) into its truncated 2×3 rotation matrices $\mathtt{R}_i^L$ and $\mathtt{R}_i^R$ and the deformation weights l_{id} using orthonormal decomposition. The structure of the sub-blocks can be expressed as:

$$\begin{bmatrix} \mathtt{M}_{i1}^L & \dots & \mathtt{M}_{iD}^L \\ \mathtt{M}_{i1}^R & \dots & \mathtt{M}_{iD}^R \end{bmatrix} = \begin{bmatrix} l_{i1} \begin{bmatrix} \mathtt{R}_i^L \\ \mathtt{R}_i^R \end{bmatrix} & \dots & l_{iD} \begin{bmatrix} \mathtt{R}_i^L \\ \mathtt{R}_i^R \end{bmatrix} \end{bmatrix} \tag{4.7}$$

The approach used to estimate the rotation components for the left and right cameras is similar to the algorithm described in section 2.4.4. Since now we have 4 rows per frame, we arrange the motion sub-blocks such that:

$$\tilde{\mathtt{M}}_i \rightarrow \breve{\mathtt{M}}_i = \begin{bmatrix} l_{i1} \begin{bmatrix} \mathbf{r}_i^L \\ \mathbf{r}_i^R \end{bmatrix} & l_{i2} \begin{bmatrix} \mathbf{r}_i^L \\ \mathbf{r}_i^R \end{bmatrix} & \dots & l_{iD} \begin{bmatrix} \mathbf{r}_i^L \\ \mathbf{r}_i^R \end{bmatrix} \end{bmatrix} \tag{4.8}$$

where $\mathbf{r}_i^L = [r_{i1}^L \dots r_{i6}^L]^T$ is a column vector which contains the coefficients of the left rotation matrix $\mathtt{R}_i^L$ and similarly for $\mathbf{r}_i^R$. Post-multiplying the rearranged matrix $\breve{\mathtt{M}}_i$ by the $2D$ unity vector $\mathbf{c} = [1 \dots 1]^T$ gives a column vector $\mathbf{a}_i$:

$$\mathbf{a}_i = \breve{\mathtt{M}}_i \mathbf{c} \tag{4.9}$$

which may be rearranged into a 4×3 matrix $\mathtt{A}_i$ with analytic form:

$$\mathtt{A}_i = \begin{bmatrix} kr_{i1}^L & kr_{i2}^L & kr_{i3}^L \\ kr_{i4}^L & kr_{i5}^L & kr_{i6}^L \\ kr_{i1}^R & kr_{i2}^R & kr_{i3}^R \\ kr_{i4}^R & kr_{i5}^R & kr_{i6}^R \end{bmatrix} = \begin{bmatrix} \mathtt{A}_L \\ \mathtt{A}_R \end{bmatrix} \tag{4.10}$$

where $k = l_{i1} + \ldots + l_{iD}$. Since $\mathtt{R}^L$ and $\mathtt{R}^R$ are orthonormal matrices, the following equation is satisfied:

$$\begin{bmatrix} \mathtt{R}_L & 0 \\ 0 & \mathtt{R}_R \end{bmatrix}_{4\times 6} \begin{bmatrix} \mathtt{A}_L^T & 0 \\ 0 & \mathtt{A}_R^T \end{bmatrix}_{6\times 4} = \sqrt{\begin{bmatrix} \mathtt{A}_L\mathtt{A}_L^T & 0 \\ 0 & \mathtt{A}_R\mathtt{A}_R^T \end{bmatrix}_{4\times 4}} \tag{4.11}$$

Therefore, a linear least-squares fit can be obtained for the rotation matrices $\mathtt{R}_L$ and $\mathtt{R}_R$ and the weights l_{id} can be subsequently estimated in a similar way as shown in section 2.4.4. Finally a minimization scheme similar to the one used by Brand [16] in his *flexible factorization* algorithm is applied here (see section 2.4.4).

So far we have presented an extension of non-rigid factorization methods to the case of a stereo camera pair. In particular our algorithm follows the approach by Brand [16]. While this new method improves the quality of the 3-D reconstructions with respect to those using a monocular sequence, it still performs a partial upgrade of the *motion* and *3-D structure* matrices since $\mathtt{Q}$ is computed initially as a block diagonal matrix and then corrected with Brand's *flexible factorization*.

In the next section we will describe a non-linear optimization scheme which renders the appropriate structure to the motion matrix, allowing to properly disambiguate between the motion and shape parameters.

4.2.3 Stereo non-linear optimization

An analogous approach as described in section 3.2 is used to refine the motion and stereo components estimated from the linear method. Similarly to the monocular case, the reprojection error for the stereo rig is defined by rearranging equation (4.4) giving:

$$\mathbf{n}_{ij} = \begin{bmatrix} \mathbf{x}_{ij}^L - \mathtt{R}_i^L \sum_d l_{id}\mathbf{S}_{dj} \\ \mathbf{x}_{ij}^R - \mathtt{R}_{rel}\mathtt{R}_i^L \sum_d l_{id}\mathbf{S}_{dj} \end{bmatrix} \tag{4.12}$$

Optimization of the deformable parameters is performed through the minimization of the cost function $C(\Theta)$ such that:

$$\min_{\mathtt{R}_{rel}\,\mathtt{R}_i^L\,l_{id}\,\mathbf{S}_{dj}} C(\Theta) = \min_{\mathtt{R}_{rel}\,\mathtt{R}_i^L\,l_{id}\,\mathbf{S}_{dj}} \sum_{i,j}^{F,P} \| \mathbf{n}_{ij} \|^2 \tag{4.13}$$

is the minimization of the sum of FP quadratic cost functions for the left and right cameras.

The initial estimate for the constant relative orientation $\mathtt{R}_{rel}$ between the left and right cameras is estimated from the camera matrices $\mathtt{R}_L$ and $\mathtt{R}_R$ (see section 4.2.2) using a least squares estimation. Unit quaternions were used again as the parameterisation and the orthogonality constraint

was enforced by fixing the 4-vector norm to unity such that the solution space is constrained to lie on a hypersphere of dimension 4.

If the internal and external calibration of the stereo rig were known in advance after a process of calibration or self-calibration, an alternative initialisation could be computed by recovering the 3-D structure and performing Principal Component Analysis (PCA) on the data to obtain an initial estimate for the basis shapes and the coefficients. However, our choice was to use an initialisation that does not require a pre-calibration of the cameras.

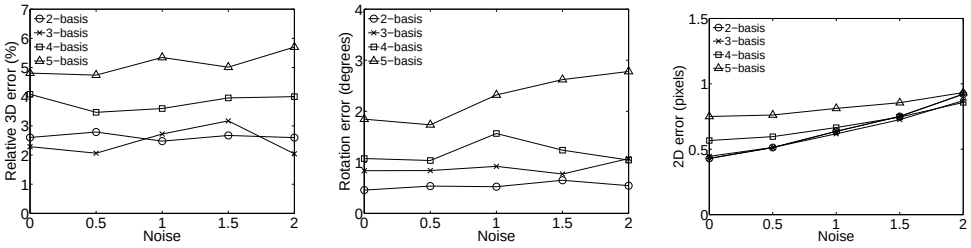
4.3 Experimental results

This section shows the performance of the proposed stereo-motion algorithms. Firstly, synthetic stereo sequences are generated under different Gaussian noise and deformation conditions to assess the validity of the method. A further synthetic test using a computer graphic (CG) generated face model will show the behavior of the configuration weights and motion components when the object in the stereo sequence is static (only deforming). We then carry out some real experiments where the object underwent only a small amount of rigid motion (apart from the deformations) and we will show the improvement of the method by comparing the output of the monocular factorization and the stereo algorithms. Non-linear optimization will follow the computed linear solutions.

4.3.1 Experiments with a synthetic non-rigid cube

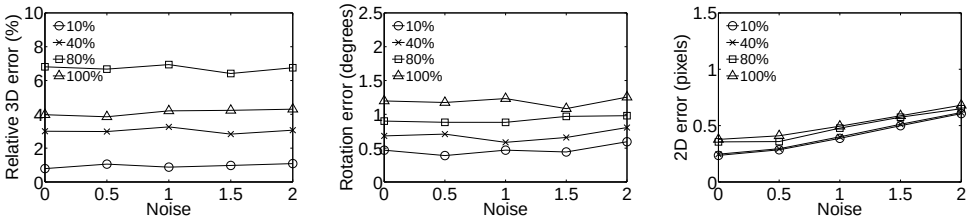
A similar setup as the one used in the monocular case (see section 3.4.1) is used to demonstrate the behavior of the method in the stereo case. A set of deformable points is randomly sampled inside a cube of $50 \times 50 \times 50$ units. A minimal overall rigid motion is introduced to avoid possible ambiguities arising from a completely static object. The 3-D structure computed at each frame is then projected with 2 orthographic cameras displaced by a baseline of 20 units and relatively rotated by 30 degrees about the *y*-axis. Finally, different levels of Gaussian noise ($\sigma = 0.5, 1, 1.5, 2$) are added to the measurements obtained by the stereo pair. Notice that the setup is constructed in such way that the overall rigid motion is not enough to reconstruct the sequences using monocular factorization followed by bundle adjustment. We performed a test and we obtained a relative 3-D reconstruction error of 50% resulting in a meaningless reconstruction.

The results show the plots for the relative 3-D error, rotation error and reprojection error



3-D Error Rotation Error 2-D Error

Figure 4.3: Relative 3-D error (%), r.m.s. rotation error (deg) and 2-D reprojection error for the synthetic experiments with a stereo pair for different basis shapes $d = 2 \ldots 5$ and increasing levels of Gaussian noise. The ratio of non-rigidity is fixed to 40% for all the trials. Relative orientation between the cameras is fixed to 30 degrees with a baseline of 20 pixel units.



3-D Error Rotation Error 2-D Error

Figure 4.4: Relative 3-D error (%), r.m.s. rotation error (in degrees) and 2-D reprojection error for the synthetic experiments for different ratios of deformation (10%, 40%, 80%, 100%) and increasing levels of Gaussian noise.

tested over 25 trials with a 3-D shape deforming with different numbers of basis shapes (figure 4.3) and different degrees of non-rigidity (see figure 4.4) defined as $ratio = \frac{\|\mathbf{S}_{nonrigid}\|}{\|\mathbf{S}_{rigid}\|}$. Notice in this case a higher reconstruction error of the relative 3-D structure compared to the monocular case with higher degrees of deformation.

4.3.2 Synthetic experiments with a CG generated face

In this section we have generated a sequence using a synthetic face model originally developed by Parke et. al. [113]. This is a 3-D model which encodes 18 different muscles of the face. Animating the face model to generate facial expressions is achieved by actuating on the different facial muscles. In particular we have used a sequence where the head did not perform any rigid motion, only deformations a situation where, clearly, monocular algorithms would fail to compute the correct 3-D shape and motion. The sequence was 125 frames long. The model deforms between

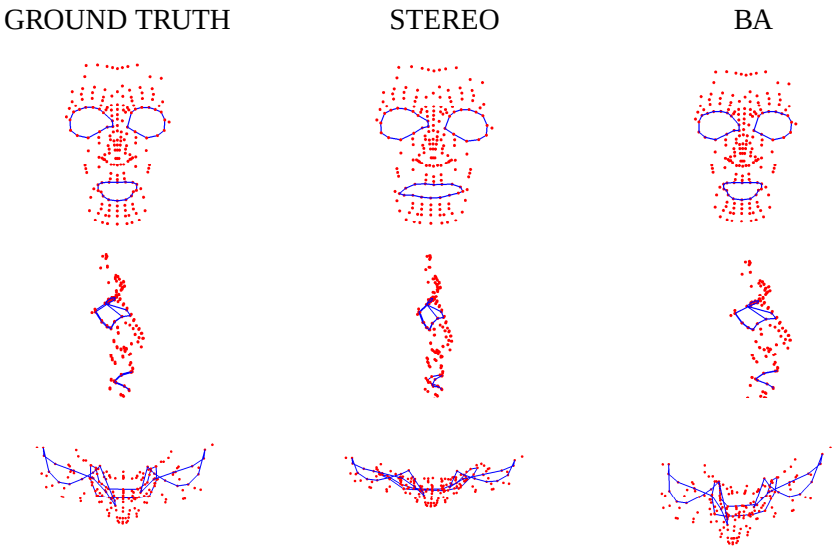


Figure 4.5: Front, side and top views of the 3-D synthetic face for frame 20. The first column shows the shape ground truth while the following two columns present the 3-D reconstructions for the linear and bundle adjustment algorithms. Deformations are present mainly in the mouth region. Notice that the face does not perform rigid motion for the whole sequence.

frames 1 and 50, remains static and rigid until frame 100 and deforms once again between frames 100 and 125.

Once the model was generated we projected synthetically 160 points evenly distributed on the face, onto a pair of stereo cameras. The geometry of the cameras was such that both optical axes were lying on the XZ plane and each pointing inwards by 15 degrees. Therefore the relative orientation of the cameras about the Y axis was 30 degrees and 0 about the X and Z axes. The camera model used to project the points was a projective model however, the viewing conditions were such that the relief of the scene was small compared to the overall depth.

We show in the following figures the comparisons between three key frames of the synthetic sequence providing the 3-D ground truth and the 3-D reconstructions for the linear and bundle adjustment algorithms. Figure 4.5 presents a deformation localised in the mouth region at frame 20. A first visual inspection shows that the result obtained by the bundle adjustment have a qualitative advantage over the stereo linear algorithm. Even if the general mean shape is close to the ground truth, only the optimised solution with bundle adjustment can model properly the deformations. Frame 70 (see Figure 4.6) shows the synthetic face (ground truth) with no deformations appearing. The static pose of the shape permits to compare the 3-D depth reconstructed by the algorithms. Compared to the ground truth, the shape obtained by the stereo algorithm

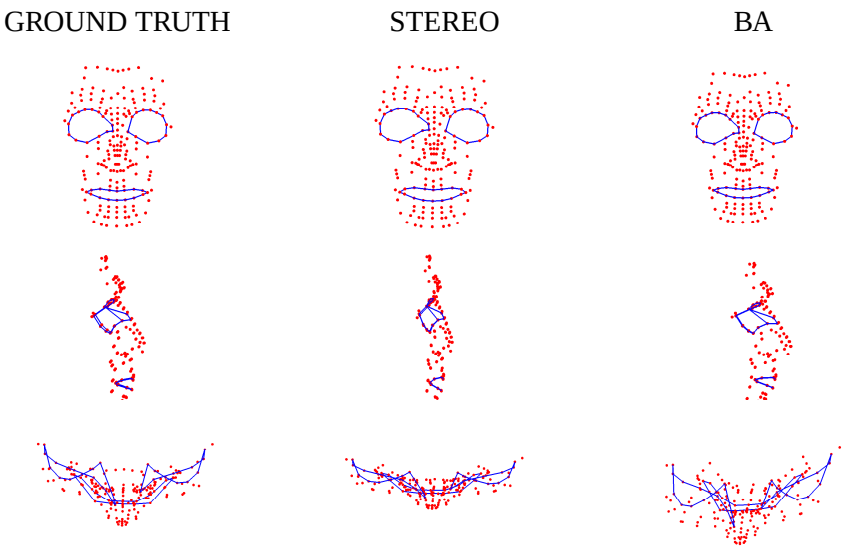


Figure 4.6: Front, side and top views of the 3-D synthetic face for frame 70. The first column shows the shape ground truth while the following two columns present the 3-D reconstructions for the linear and bundle adjustment algorithms. The shape is completely static in this frame.

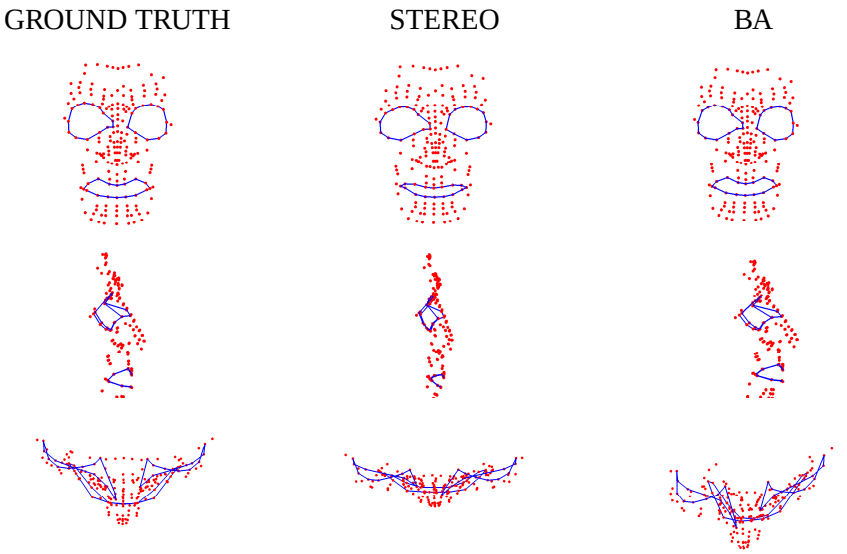


Figure 4.7: Front, side and top views of the 3-D synthetic face for frame 125. The first column shows the shape ground truth while the following two columns present the 3-D reconstructions for the linear and bundle adjustment algorithms. Deformations are localized in the mouth and cheek regions.

shows a good frontal reconstruction but it presents a worst estimation of the relief (see side and top views). The non-linear solution obtains a depth estimate qualitatively closer to the ground truth. Finally figure 4.6 presents the reconstruction obtained for frame 125 where the synthetic face shows consistent deformations in the cheeks and mouth area. The stereo algorithm obtains a reasonable mean 3-D shape but it fails in capturing the deformations appearing in the ground truth.

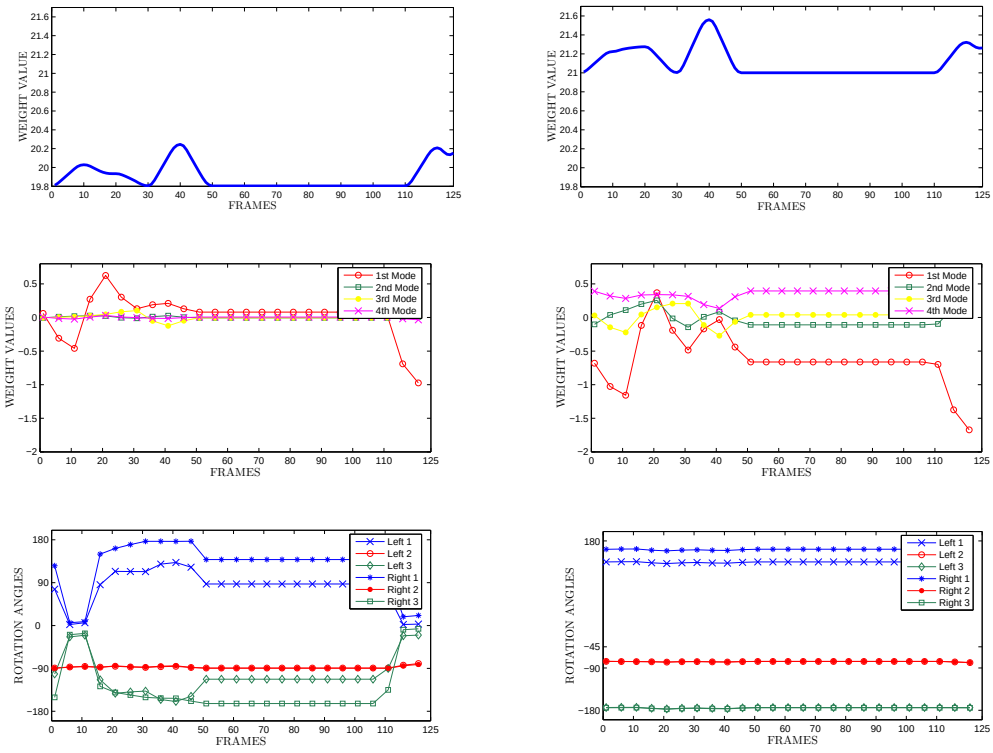
Figure 4.8 shows the results for the estimated rotation angles and configuration weights before and after the non-linear optimization step. The results after bundle adjustment describe fairly accurately the geometry of the cameras and the deformation of the face. In particular, the stereo setup was such that there was no rigid motion of the face (only deformation), the optical axes of the left and right cameras lay on the XZ plane and the relative rotation of the cameras about the Y axis was constant and equal to 30 deg. In this case we have ground truth values for the relative orientation of the cameras since the sequence was generated synthetically. Notice how the values obtained for the rotation angles before bundle adjustment – left – exhibit some problems around frames 10 and 115, when the deformations are occurring. After the bundle adjustment step the the relative rotation about the Y axis is estimated with a final result of 27 deg resulting in a 3 deg error given the ground truth. The relative orientations about the X and Z axes are correctly estimated to 0 deg – notice that the graphs for the left and right angles are superimposed.

Once more, the estimated values for the deformation weights after bundle adjustment have larger values than before the optimization. This explains the fact that the model succeeds to explain the non-rigid deformations accurately. Interestingly, the coefficients remain constant between frames 50 and 110, when no deformations were occurring.

4.3.3 Experiments with real data

Comparison with the monocular solution

In this section we compare the performance of our stereo factorization algorithm – before the non-linear optimization – with Brand's single camera non-rigid factorization method. We present some experimental results obtained with real image sequences taken with a pair of synchronized Fire-i digital cameras with 4,65mm built in lenses. The stereo setup was such that the baseline was 20cm and the relative orientation of the cameras was around 30 deg. Two sequences of a human face undergoing rigid motion and flexible deformations were used: the SMILE sequence (82 frames), where the deformation was due to the subject smiling and the EYEBROW (115



(A) STEREO ALGORITHM (B) BUNDLE ADJUSTMENT

Figure 4.8: Values obtained for the rigid component (top), deformation weights (middle) and rotation angles (bottom) before (A) and after bundle adjustment (B) for the synthetic sequence.

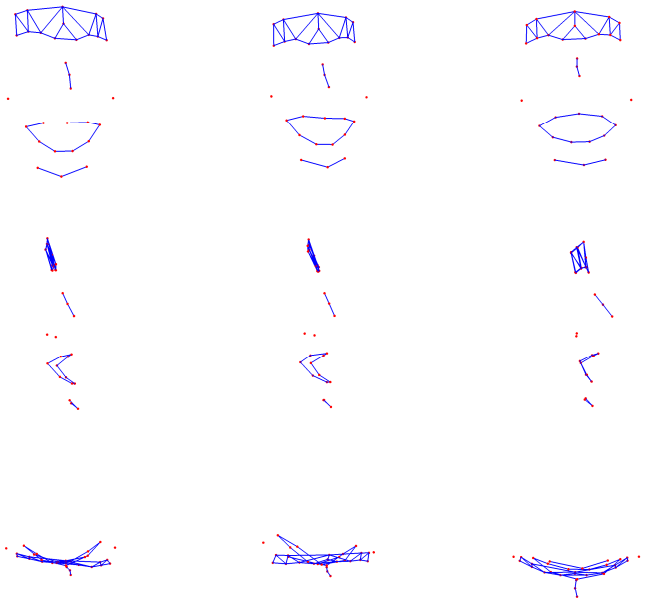
a) SMILE sequence: left view b) EYEBROW sequence: left view

c) SMILE sequence: right view d) EYEBROW sequence: right view

Figure 4.9: Three images from the left (a) and right (c) views of the SMILE sequence and left (b) and right (d) views of the EYEBROW sequence.

frames) sequence where the subject was raising and lowering the eyebrows. Figure 4.9 shows 3 frames chosen from the sequences taken with the left and right cameras.

In order to simplify the temporal and stereo matching the subject had some markers placed on relevant points of the face such as along the eyebrows, the chin and the lips. A simple colour model of the markers using HSV components provided the representation used to track each marker throughout the left and right sequences respectively. The stereo matching was initialized by hand in the first image pair and then the temporal tracks were used to update the stereo matches.

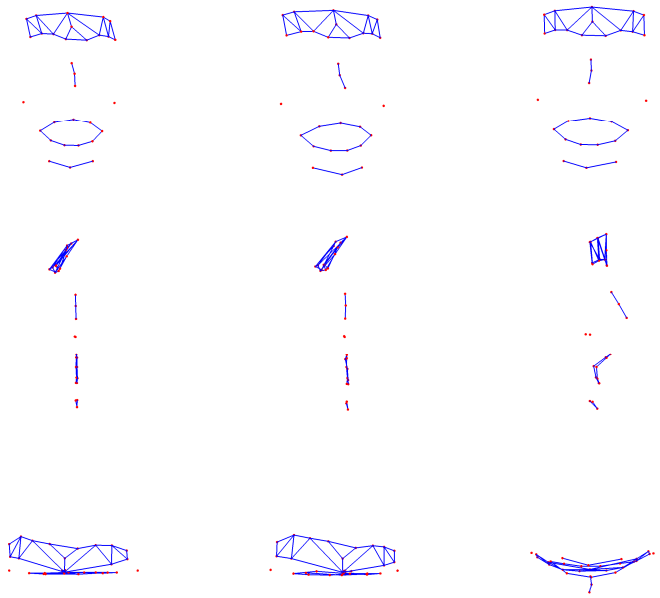


a) Left camera b) Right camera c) Stereo

Figure 4.10: SMILE sequence: Front, side and top views (above, middle, bottom) of the 3-D model for the a) left camera, b) right camera and c) stereo setup for $D = 5$.

Figure 4.10 shows front, side and top views of the 3-D reconstructions obtained for the SMILE sequence. First we applied the single camera factorization algorithm developed by Brand – described in section 2.4.4 – to the left and right monocular sequences. We then applied the proposed stereo algorithm to the stereo sequence. In all cases the number of tracked points was $P = 31$ and the chosen number of basis shapes was heuristically fixed to $D = 5$.

Figure 4.9c shows how the stereo reconstruction provides improved results. The reconstructions obtained using singularly the information from the left and right sequences have worse depth estimates that can be noticed especially in the side and top views. The reconstructed face



a) Left camera b) Right camera c) Stereo

Figure 4.11: EYEBROW sequence: Front, side and top views (above, middle, bottom) of the 3-D model for the a) left camera, b) right camera and c) stereo setup sequences for $D = 5$.

is strongly asymmetric especially in the mouth region and the points on the forehead are almost belonging to a plane. Differently, after merging the data from both sequences in the stereo algorithm, we obtained a symmetric shape and a satisfactory curvature of the forehead.

Figure 4.12(A) shows the front, side and top views of the 3-D reconstructions obtained for frames 16, 58 and 81 of the SMILE sequence. While the 3-D shape appears to be well reconstructed, the deformations are not entirely well modelled. Note how the smile on frame 58 is not well captured. This was caused by the final regularization step proposed by Brand described in section 4.2.2. We found that while this regularization step is essential to obtain good estimates for the rotation parameters it fails to capture the full deformations in the model. This is due to the fact that the assumption is that the deformations should be small relative to the mean shape so that most of the image motion is explained by the rigid component which results in a poor description of the deformations. However, we will see in the following section that the bundle adjustment step resolves the ambiguity between motion and shape parameters and succeeds in modelling the non-rigid deformations.

Figure 4.11 shows the 3-D reconstructions obtained for the EYEBROW sequence. Once more, the single camera factorization algorithm was applied to the left and right sequences and

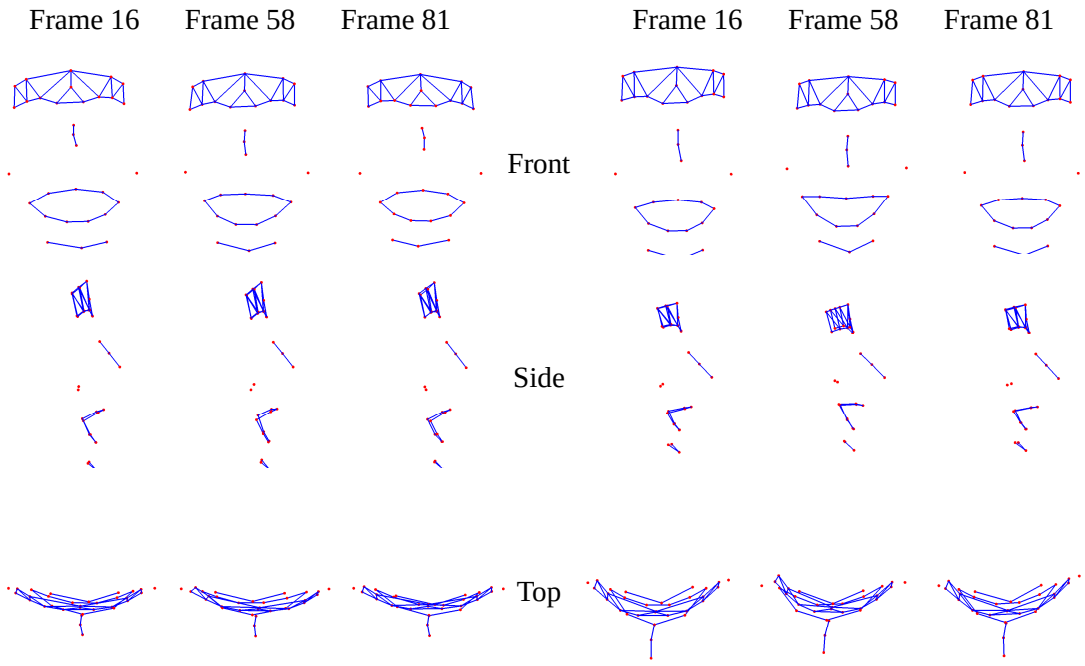


Figure 4.12: Front, side and top views of the reconstructed face for the SMILE sequence using the stereo algorithm (left) and after bundle adjustment (right). Reconstructions are shown for frames 16, 56 and 81 of the sequence.

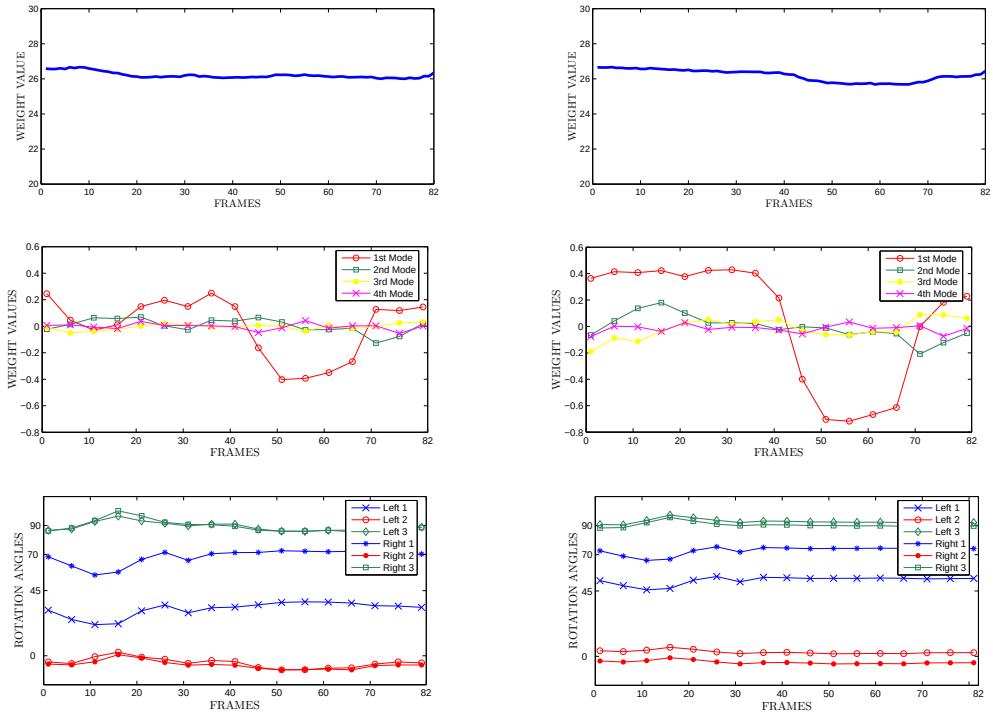
the stereo algorithm was then applied to the stereo sequence. In this sequence the 3-D model obtained using stereo factorization is significantly better than the ones obtained with the left and right sequences. In fact, the left and right reconstructions have very poor quality, particularly the depth estimates. The points belonging to the nose, mouth and chin are almost planar (see side view) while the ones on the forehead have a particularly wrong depth estimate (see top view). Note that there was less rigid motion in this sequence and therefore the single camera factorization algorithm is not capable of recovering correct 3-D information whereas the stereo algorithm provides a good deformable model.

Results after non-linear optimization

In this section we show the results obtained after the final non-linear optimization step.

Figure 4.12 shows the front, side and top views of the 3-D reconstructions before and after the bundle adjustment step for three frames of the SMILE sequence[1]. The initial estimate is shown on the left and the results after bundle adjustment are shown on the right. While the initial estimate recovers the correct 3-D shape, the deformations on the face are not well modelled.

[1]Video available at http://www.bmva.ac.uk/thesis_archive/2006/DelBue1/index.html



(A) STEREO ALGORITHM (B) BUNDLE ADJUSTMENT

Figure 4.13: Values obtained for the rigid component (top), deformation weights (middle) and rotation angles (bottom) before (A) and after bundle adjustment (B) for the SMILE sequence

However, bundle adjustment succeeds to capture the flexible structure – notice how the upper lip is curved first and then straightened.

Figure 4.13 shows the results obtained for the estimated motion parameters and configuration weights using the initial stereo factorization method and the improved results after bundle adjustment. The bottom graphs show the rotation angles about the X, Y and Z axes recovered for each frame of the sequence for the left and right cameras (up to an overall rotation). The recovered angles for the left and right camera after bundle adjustment reflect very well the geometry of the stereo camera setup. This was such that both optical axes lay approximately on the XZ plane – therefore there was no relative rotation between the cameras about the X and Z axes – and the relative rotation about the Y axis was about 15 deg. Note that these values are not ground truth and only approximate as they were not measured accurately. Also note that the rotation matrices for the right camera are calculated as $\mathtt{R}^R = \mathtt{R}_{rel}\mathtt{R}^L$ where $\mathtt{R}_{rel}$ is the estimated relative orientation. Figure 4.13(B) shows how the estimates of the rotations about the X and Z axes (in blue and green) for the left and right views are close to being zero. The relative rotation between left and right cameras about the Y axis (in red) is closer to 15 deg after bundle adjustment than before.

Figure 4.13 also shows the evolution throughout the sequence of the values of the configuration weights associated with the mean component (top) and the 4 modes of deformation (middle). The values appear to be larger after bundle adjustment confirming that the non-linear optimization step has achieved to model the deformations of the face. It is also interesting to note how the first mode of deformation experiences a big change starting around frame 40 until frame 75. This coincides with the moment where the subject started and finished the smile expression.

4.4 Summary

A stereo-motion approach has been presented with the aim to reconstruct the 3-D shape of a deformable object using image sequences extracted from a stereo-pair. As a result, the non-rigid factorization framework has been accordingly updated to accommodate the constraint that trajectories in the left and right camera refer to the same 3-D object.

By construction, the method fuses naturally the advantages of motion and stereo approaches. A global solution for the time varying motion and 3-D structure is obtained from the image tracks without any prior calibration of the stereo pairs. Widely separated stereo views allow a more reliable estimation of motion and deformation parameters even in the absence of rigid motion of the object.

Additionally, non-linear optimization, as presented in the previous chapter, is performed to obtain the correct replicated structure in $\mathtt{M}$. Results show a relevant improvement in the motion and structure estimates and thus the optimization stage is strongly recommended to obtain a correct solution.

The main assumptions of our method are that the cameras must be synchronized and stereo matches be available. Synchronization can be enforced using the method presented in [142] but nowadays it is common to obtain synchronized video from stereo cameras. Stereo matching could be tackled by extending current techniques [73, 110] to deal with the non-rigid case.

Chapter 5

Deformable modelling under affine viewing conditions using shape priors

---

Deformable 3-D shape recovery is an inherently ambiguous problem. Given a specific rigid motion, different non-rigid shapes could be found that fit the measurements. To solve this ambiguity prior knowledge about the shape and motion should be used to constrain the solution. We base our approach [35] on the observation that often not all the points on a moving and deforming surface – such as a human face – are undergoing non-rigid motion. Some of the points are frequently on rigid parts of the structure – for instance the nose – while others lie on deformable areas. First we develop a segmentation algorithm to separate rigid and non-rigid motion. Once this segmentation is available, the rigid points can be used to estimate the overall rigid motion and to constrain the underlying mean shape. We propose two reconstruction algorithms and show that improved 3-D deformable models can be obtained from priors on the shape by using synthetic and real data.

5.1 Motivation

A main issue of factorization approaches for deformable structure stems from the fact that deformation and motion are ambiguous. Intuitively, imagine a deforming object like a sheet of paper floating in the air or a tree bending by the blowing wind; the concepts of motion and deformation are not clearly defined if a notion of global motion is not specified. The deformations that appear in a non-rigid object can be defined as the deviation of the shape from the global

motion. This observation is supported by recent studies on the notion of shape average by Yezzi and Soatto [164] where the authors precisely separate motion and deformation components for robust matching, registering and tracking of deformable objects. Improved results are obtained by explicitly defining the mean component of the object first and then calculating deformations in an active contours domain.

Our approach is slightly different, we realize that the rigid component of the structure carries useful information about the overall non-rigid shape. Our main assumption is that some of the points are frequently on rigid parts of the structure while others lie on deformable areas. For the set of rigid points, multi-frame rigidity constraints hold [150] and these can be appropriately enforced in reconstruction algorithms to obtain reliable camera motion estimates. On the other hand, if a rigid 3-D structure is correctly identified, the rigid points can be used to constrain the underlying mean shape. The deformations can then be estimated as local deviations from this mean shape in a further refinement step.

The approach introduced in this chapter requires an initial information or prior over which of the point trajectories stored in the measurement matrix W are rigid and which non-rigid. Notice that, similar priors were required to obtain an exact solution for the case of independently moving (section 2.3.1) and articulated objects (section 2.3.2), where trajectories belonging to the different parts of the object have to be identified to obtain a proper reconstruction. Thus, we first need to introduce methods and techniques to perform a reliable segmentation of point trajectories into rigid and non-rigid components.

Once the points have been segmented into the rigid and non-rigid sets we recover the overall rigid motion from the rigid set and we formalise the problem of non-rigid shape estimation as a constrained minimization adding priors on the degree of deformability of each point. We perform experiments on synthetic and real data which validate the approach and show that the addition of priors on the rigidity of some of the points improves the motion estimates and the 3-D reconstruction.

5.2 Motion segmentation from image trajectories: previous work for rigid scenes

The assumption that a scene observed by a camera contains a single rigid object is often not realistic. For instance, when both the camera and the observed object are moving, the motion of the background (usually degenerate since most often it can be approximated as a planar object)

and the one of the inspected object represent two distinguishable visual cues. Similarly, often there will be more than one independently moving object in the scene (for instance, a traffic scene containing different vehicles). In these cases it is crucial to be able to segment the trajectories belonging to the respective object so that exact reconstructions can be obtained.

A first approach to segmenting N purely rotating objects was given by Boult and Brown [14] using bi-partite graphs to cluster the image trajectories. Starting from an efficient estimate of the rank robust to noise, the method performs a rank-constrained SVD on the measurement matrix $\mathtt{W}$ giving $\mathtt{W} = \mathtt{U}\Sigma\mathtt{V}^T$ and assigning points to motion clusters by selecting the most significant columns of $\mathtt{V}^T$. The process is repeated iteratively until the N sets of rank-3 measurements are successfully detected. Motion dependencies and degeneracies are not explicitly modelled so these could affect the convergence of the method.

Costeira and Kanade [30] first proposed the use of the *shape interaction matrix* $\mathtt{G}$, defined as $\mathtt{G} = \mathtt{V}\mathtt{V}^T$ where $\mathtt{V}$ is the matrix of right singular vectors. In the presence of independent motions and noiseless data the following condition for the matrix $\mathtt{G}$ holds:

$$\mathtt{G}_{mn} = \begin{cases} 1 & \text{if trajectories } m \text{ and } n \text{ correspond to the same motion} \\ 0 & \text{otherwise} \end{cases} \tag{5.1}$$

where $m = 1 \ldots P$ and $n = 1 \ldots P$ with P being the number of trajectories. Hence, each element $\mathtt{G}_{mn}$ specifies whether a pair of trajectories belongs to the same motion or not. However, in the presence of noise the conditions in equation (5.1) will not be satisfied exactly. A proof of the properties of $\mathtt{G}$ is given by Kanatani [85] using the properties of independent motion sub-spaces. A procedure that optimizes the energy of the entries of $\mathtt{G}$ is used by Costeira and Kanade [29] to cluster the N sets of trajectories such that the matrix $\mathtt{G}$ is block diagonal (see figure 5.1). Neither a priori knowledge of the number of shapes nor an estimate of the rank is required. A known drawback [76] of this method is that noise and outliers affecting the measurements modify the conditions in equation (5.1). In this case, the approach is likely to obtain a sub-optimal solution. Motion dependencies [166] are also a known weakness of the approach if not explicitly modelled.

In order to improve the performance under noise conditions, Ichimura [76] proposed a discriminant criterion that drives the clustering by choosing the trajectories with the most useful information for grouping. The approach relies on an initial computation of the *shape interaction matrix*. This may lead to inaccurate application of the discriminant criterion if the estimated $\mathtt{G}$ is unreliable. However, the overall performance of the algorithm is superior compared to Costeira

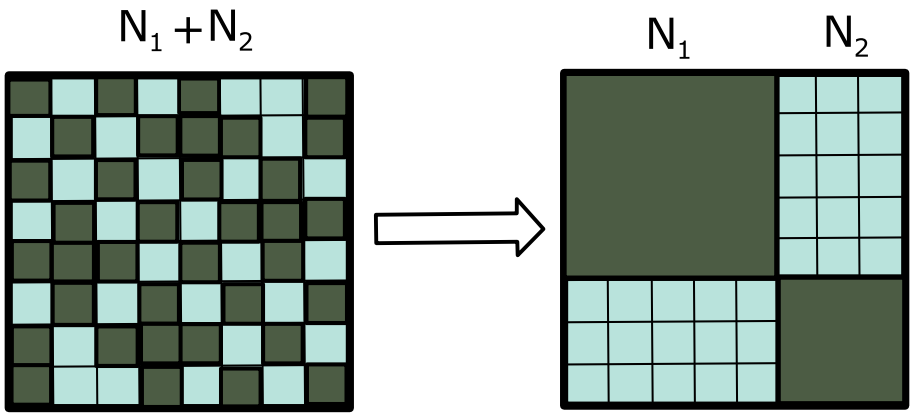


Figure 5.1: Example of *shape interaction matrix* $\mathtt{G}$ obtained from two ($N = 2$) rigid objects with $P_1 = 5$ and $P_2 = 3$. A dark square represents a pair of trajectories belonging to the same motion. Figure (a) represents a sparse $\mathtt{G}$ that is given before ordering of the trajectories into the two clusters of independent motions. Figure (b) shows $\mathtt{G}$ after computing the permutation which arranges the measurement matrix such that $\mathtt{W} = [\mathtt{W}_1|\mathtt{W}_2]$ with $\mathtt{W}_1$ and $\mathtt{W}_2$ containing the trajectories for the first and second object respectively.

and Kanade's approach.

Wu et al. [158] initially compute an approximated over-segmentation of the number of independent motions using Ichimura's method. The method then computes a robust distance measure for the points belonging to each object based on the orthogonality properties of the sub-spaces of the independent shapes and it reduces the over segmented motions to the correct number of sets. As a result, the metric proposed is robust to the noise distribution since the orthogonality condition between sub-spaces still holds with corrupted data.

Kanatani [83] drops the concept of the *shape interaction matrix* in favor of directly fitting the trajectories taken from the independent objects to the related sub-spaces. Model selection [84] is used to infer the number of independent motions and outlier rejection [133] strengthens the approach in the case of outlying image trajectories. The estimation of different motions is performed in a framework similar to the Expectation-Maximization (EM) algorithm and, thus, it is prone to local solutions. The method, however, is inserted in a sound statistical framework with particular robustness to noise. A further improvement introduced by Sugaya and Kanatani [135] permits to deal with degeneracies given 2-D planar motions in the scene. An approach using

the EM algorithm is also presented in the work of Gruber and Weiss [54] where factorization is formulated as a factor analysis problem [53] with the interesting possibility of forcing known priors over the motion and structure components of the objects.

Of broader applicability, the approach of Vidal and Hartley [152] may fit data with motion degeneracies and missing entries in the measurement matrix using a combined method with generalized principal component analysis (GPCA) [153] and Powerfactorization [66]. Briefly, an initial rank-5 decomposition of $\mathtt{W}$ is performed via Powerfactorization that allows to deal with missing data. This initial decomposition preserves the structure of the motion clusters while reducing the dimensionality of the problem. Motion sub-spaces are then fitted with a 5-degree polynomial over the decomposed set of trajectories using least-squares (GPCA). Spectral clustering [155] is finally applied over a similarity matrix constructed over the differentiation of the 5-degrees polynomial. Validation over synthetic experiments is not presented but the algorithm can deal successfully with degenerate and independent motion for measurement matrices with up to 30% of missing entries. Notice that a known drawback is that the GPCA methods need a number trajectories that grows exponentially with the number of motions.

Specifically designed for articulated structures (see section 2.3.2), the approach of Yan and Pollefeys [162] separates dependent motions connected by joints. Their method (with some similarities to the algorithm we propose in section 6.4.1) employs RANdom SAmple Consensus [46] (RANSAC) to assign the trajectories to each articulated part. Given the random nature of the algorithm, a sampling prior is assigned to increase the chance of selecting pairs of image trajectories that are most likely to belong to the same group. The sampling prior is computed with a distance measure obtained from the *shape interaction matrix* of the articulated object. Given the known sensitivity of the *shape interaction matrix* to image noise, this approach could lead to inaccuracies in the computation of the prior.

5.3 Rigid and non-rigid motion segmentation

We now consider the problem of segmenting the rigid and non-rigid motion of a single deforming shape which contains a sub-set of rigid points. In this case, the image trajectories composing the measurement matrix $\mathtt{W}$ are given by two contributions: the overall rotation and translation which the object is globally undergoing and the local deformations of each non-rigid point. Both sets of rigid and non-rigid points share the same rigid transformation and consequently this renders

the straight application of the algorithms for independent motion segmentation presented in the last section less effective.

For instance, if we consider Kanatani's sub-space technique [83] for motion segmentation, the aim would be to assign every rigid trajectory to a sub-space of dimension 3 and the non-rigid trajectories to a sub-space of dimension $3D$. However, the rigid points could be understood as non-rigid points with only one basis shape, and therefore the sub-space for the non-rigid points would completely include the one for the rigid points. Thus, the method would tend to classify every trajectory as being non-rigid. To the best of our knowledge there is no other work able to separate rigid and non-rigid trajectories belonging to a single object.

5.3.1 Our approach

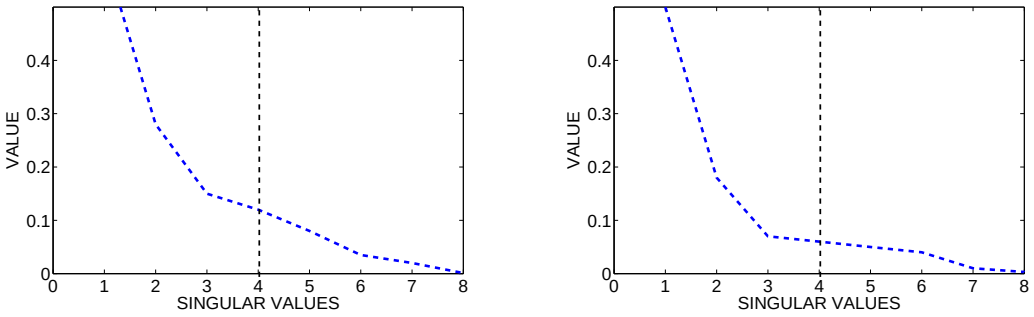
Our approach instead consists in the application of a sub-set selection method on the non-rigid component of the point trajectories encoded in the measurement matrix $\mathtt{W}$. Sub-set selection is a technique commonly used in feature selection problems where a group of features is extracted to obtain a robust solution to a particular estimation problem [80].

Under the factorization framework, features are represented by their image point trajectories stored in $\mathtt{W}$. Our goal is to find the set of features whose motion can be modelled exactly as a rigid motion. In this case we formulate the segmentation problem as finding a sub-set of trajectories $\mathtt{W}_{rigid}$ within the measurement matrix such that the following condition is satisfied:

$$rank(\mathtt{W}_{rigid}) = 3. \tag{5.2}$$

The segmentation algorithm follows a *sequential backward selection strategy* [88] by initially considering all the trajectories in the measurement matrix and iteratively deleting one by one those which are contributing most to the rank of the matrix, i.e., the points that exhibit the most non-rigid motion. As the stop criterion for the classification task, we compute the rank of the measurement matrix of the remaining points which will become 3 when only the rigid trajectories are left.

Obviously the rank of the rigid points will not be exactly equal to 3 in the presence of noise as it can be observed in figure 5.2. Instead, we have used an automatic method to determine the deformability index of a set of trajectories described in the work of Roy Chowdhury [126]. This method estimates the value of D – the number of basis shapes needed to describe the non-rigid motion – automatically in a non-iterative way.



Face Sequence Box Sequence

Figure 5.2: The plots show the values of the singular values ordered in descending order and extracted from different measurement matrices containing rigid points affected by noise. The rigid points are extracted from a face (left) and a deforming box (right). A completely rigid object has a rank-3 measurement matrix (i.e. the fourth singular value is equal to zero). Denoise techniques are necessary to remove the noise component so that the rank-3 condition can be used to detect measurements belonging to a rigid object.

5.3.2 Estimation of the degree of deformability

The approach is based on a reinterpretation of the deformable factorization problem in a stochastic framework. In this way, provided a statistic description of the noise corrupting the image measurements, it is possible to compute a whitened measurement matrix from which the value of the rank and, thus, the number of basis shapes can be extracted.

In more detail, the image coordinates for a frame i are first arranged into a $2P$-vector such that $\mathbf{y}_i = [u_{i1}, \ldots, u_{iP}, v_{i1}, \ldots, v_{iP}]^T$. Now the projection of the deformable points onto the image plane may be expressed as:

$$\mathbf{y}_i^T = \mathbf{m}_i^T \mathtt{s} = \begin{bmatrix} l_{i1}\mathtt{R}_i^{(1)} & \ldots & l_{iD}\mathtt{R}_i^{(1)} & l_{i1}\mathtt{R}_i^{(2)} & \ldots & l_{iD}\mathtt{R}_i^{(2)} \end{bmatrix} \begin{bmatrix} \mathtt{S}_1 & \mathtt{0} \\ \vdots & \vdots \\ \mathtt{S}_D & \mathtt{0} \\ \mathtt{0} & \mathtt{S}_1 \\ \vdots & \vdots \\ \mathtt{0} & \mathtt{S}_D \end{bmatrix} \tag{5.3}$$

with $\mathtt{s}$ containing the re-arranged $6D \times 2P$ structure matrix. $\mathtt{R}_i^{(1)}$ and $\mathtt{R}_i^{(2)}$ denote respectively the first and second row of the orthographic camera matrix $\mathtt{R}_i$ arranged in the $2P$-vector $\mathbf{m}_i$. The noise component $\mathbf{n}_i$ is considered additive and obtained from a zero-mean random process giving

$\tilde{\mathbf{y}}_i^T = \mathbf{m}_i^T \mathtt{s} + \mathbf{n}_i$.

As a further step, the method computes the $2P \times 2P$ correlation matrix for each image trajectory such that:

$$\mathtt{C}_{\tilde{\mathbf{y}}} = \frac{1}{F}\sum_{i=1}^{F} \mathbf{y}_i \mathbf{y}_i^T = \mathtt{s}^T \left(\frac{1}{F}\sum_{i=1}^{F} \mathbf{m}_i \mathbf{m}_i^T \right) \mathtt{s} + \mathtt{C}_{\mathbf{n}} \tag{5.4}$$

where $\mathtt{C}_{\mathbf{n}}$ is the covariance of the noise affecting the measurements. An exact estimate of $\mathtt{C}_{\mathbf{n}}$ is required which can be inferred from the measurement process that obtains the image coordinates stored in $\mathtt{W}$ (for instance such information can be obtained from a point tracking algorithm such as the Kanade-Lucas-Tomasi (KLT) tracker [128]).

In the case of no noise, the correlation matrix $\mathtt{C}_{\tilde{\mathbf{y}}}$ has a rank equal to $6D$. However, the additive contribution of $\mathtt{C}_{\mathbf{n}}$ increases the overall rank by an unknown value. The problem is to find a transformation which can remove the contribution of the noise. In order to find a solution, the noise covariance is firstly diagonalised using SVD:

$$\mathtt{C}_{\mathbf{n}} = \mathtt{U}\Sigma\mathtt{U}^T \tag{5.5}$$

where the matrix Σ has L non-zero diagonal elements with $L > 6D$. It is possible to compute the rank reduced factors for $\mathtt{C}_{\mathbf{n}}$ such that:

$$\mathtt{C}_{\mathbf{n}} = \tilde{\mathtt{U}}_{2P\times L}\ \tilde{\Sigma}_{L\times L}\ \tilde{\mathtt{U}}^T_{2P\times L} \tag{5.6}$$

The noise can then be transformed into an independent and identically distributed (IID) process by pre-multiplying equation (5.3) with the factor $\left(\tilde{\mathtt{U}}\tilde{\Sigma}^{\frac{1}{2}}\right)^{-1}$ giving:

$$\hat{\mathbf{y}}_i^T = \left(\tilde{\mathtt{U}}\tilde{\Sigma}^{\frac{1}{2}}\right)^{-1} \mathbf{m}_i^T \mathtt{s} + \left(\tilde{\mathtt{U}}\tilde{\Sigma}^{\frac{1}{2}}\right)^{-1} \mathbf{n}_i = \hat{\mathbf{m}}_i^T \mathtt{s} + \hat{\mathbf{n}}_i \tag{5.7}$$

Therefore, the correlation for the transformed coordinates $\hat{\mathbf{y}}_i$ is given by:

$$\mathtt{C}_{\hat{\mathbf{y}}} = \frac{1}{F}\sum_{i=1}^{F} \hat{\mathbf{y}}_i \hat{\mathbf{y}}_i^T = \mathtt{s}^T \left(\frac{1}{F}\sum_{i=1}^{F} \hat{\mathbf{m}}_i \hat{\mathbf{m}}_i^T \right) \mathtt{s} + \mathtt{I} \tag{5.8}$$

where $\mathtt{I}$ is a $L \times L$ identity matrix. After applying SVD on $\mathtt{C}_{\hat{\mathbf{y}}}$, it can be observed that the number of basis shapes D can be obtained simply by counting the number of singular values over 1 and dividing the result by 6:

$$D = \frac{\text{number of singular values} > 1}{6} \tag{5.9}$$

This method provides a fixed threshold for comparing the singular values of the matrix to determine the deformability index D. For the case of a 3-D rigid body the deformability index D is equal to 1 while in the case of a non-rigid body the index is $D > 1$, therefore this provides a good selection criterion to separate both sets of trajectories in the presence of noise.

5.3.3 The complete segmentation algorithm

Our approach uses the deformability index measure described in the previous section as a stopping criteria to detect when the set of points gives an index $D = 1$, meaning that the remaining points are rigid. The complete algorithm is detailed below:

- Initialize $\mathtt{W}_{rigid} = \mathtt{W}$
- Determine the initial deformability index D for $\mathtt{W}_{rigid}$

1. Compute $\mathtt{W}_{rigid} \simeq \mathtt{U}\Sigma\mathtt{V}^T$ with SVD and truncate to rank $3D$.
2. Define $\mathtt{S} = \Sigma^{1/2}\mathtt{V}^T$
3. Extract the non-rigid component of the shape matrix $\tilde{\mathtt{S}}_{3(D-1)\times P} = \begin{bmatrix} \tilde{\mathbf{S}}_1 & \dots & \tilde{\mathbf{S}}_P \end{bmatrix}$ where each $\tilde{\mathbf{S}}_j$ is a $3(D-1)\times 1$ vector which contains the 3-D coordinates of the j^{th} 3-D point associated to the $D-1$ non-rigid bases such that:

$$\mathbf{S}_j = \begin{bmatrix} \mathbf{S}_{1j} \\ \mathbf{S}_{2j} \\ \vdots \\ \mathbf{S}_{Dj} \end{bmatrix} \quad \text{and} \quad \tilde{\mathbf{S}}_j = \begin{bmatrix} \mathbf{S}_{2j} \\ \vdots \\ \mathbf{S}_{Dj} \end{bmatrix}$$

4. Determine the maximum vector norm: $\tilde{\mathbf{S}}_t = \max\{\|\tilde{\mathbf{S}}_1\|, \dots, \|\tilde{\mathbf{S}}_P\|\}$.
5. Remove the selected trajectory t from $\mathtt{W}_{rigid}$ and determine the new deformability index D.
6. If $D = 1$ stop the iteration.
7. Else, go to step 1.

Algorithm 1.

We have obtained successful rigid and non-rigid motion segmentations on synthetic sequences using this algorithm. The results will be discussed in the experimental section. Note that the method converges to the right solution only if there is a unique set of rigid points such that $D = 1$. In the case where different groups of features satisfy the rank condition (for instance, in the case of multiple or articulated objects) the algorithm could converge to the wrong set.

5.4 The proposed shape prior

Once we have segmented the scene into rigid and non-rigid points, we can use the information on the rigidity of the points to constrain the shape estimation. First we define the constraints that arise based on the observation that a generic shape is composed by points with different degrees of deformation. Kim and Hong [87] defined the *degree of non-rigidity* of a point as its degree of deviation from the average shape to classify points into three classes: rigid, near-rigid and non-rigid (for a more detailed description refer to section 6.4.1). Based on this measure they proposed a method to estimate average shape using the degree of non-rigidity to weight the contribution of each point in an iterative certainty re-weighted factorization scheme. In contrast, we use the knowledge that some points of the scene are rigid to construct specific linear constraints which will in turn eliminate the inherent ambiguities present in non-rigid shape estimation.

5.4.1 Rigidity constraint

Definition (rigid point). *If the motion of a point j is completely rigid for the entire sequence, the structure referring to the point can be expressed entirely by the first basis ($D = 1$) called the rigid basis.*

It follows from this definition that a completely rigid point p is entirely parameterized by:

$$\mathbf{S}_j = \begin{bmatrix} \mathbf{S}_{j1} \\ \mathbf{0} \end{bmatrix} \tag{5.10}$$

where $\mathbf{S}_{j1}$ is a 3-vector which contains 3-D coordinates of the rigid component and $\mathbf{0}$ is a $3(D-1)$ vector of zeros. Following the segmentation of the scene into rigid and non-rigid points, it is possible to re-order the measurement matrix by defining the permutation matrix $\mathtt{P}$ such that:

$$\mathtt{WP} = \left[\begin{array}{c|c} \mathtt{W}_{rigid} & \mathtt{W}_{nonrigid} \end{array}\right] = \begin{bmatrix} l_{11}\mathtt{R}_1 & \dots & l_{1D}\mathtt{R}_1 \\ \vdots & & \vdots \\ l_{F1}\mathtt{R}_F & \dots & l_{FD}\mathtt{R}_F \end{bmatrix} \left[\begin{array}{c|c} \mathtt{S}_{rigid} & \mathtt{S}_{nonrigid} \\ \mathtt{0} & \end{array}\right] \tag{5.11}$$

where $\mathtt{S}_{rigid}$ is a $3 \times r$ matrix containing the 3-D coordinates of the r rigid points, $\mathtt{S}_{nonrigid}$ is a $3D \times (P-r)$ matrix containing the 3-D coordinates of the D basis shapes for the $(P-r)$ deformable points and $\mathtt{0}$ is a $3(D-1) \times r$ matrix of zeros.

Notice that it is now possible to apply Tomasi and Kanade's rigid factorization on the measurement matrix containing the image trajectories of the rigid points $\mathtt{W}_{rigid}$ and decompose it into

the motion and rigid structure components as:

$$\mathtt{W}_{rigid} = \begin{bmatrix} \mathtt{R}_1 \\ \vdots \\ \mathtt{R}_F \end{bmatrix} \mathtt{S}_{rigid} \tag{5.12}$$

obtaining an initial solution for the orthographic camera matrices for each frame and for the 3-D rigid component of the structure.

5.5 Non-rigid shape and motion estimation using shape priors

In this section we solve for the non-rigid shape and motion given the 2-D image tracks and incorporating the above constraint on the automatically segmented rigid points. Our approach is to minimize image reprojection error subject to the rigidity of the non-deforming points. The cost function being minimised is:

$$\chi = \sum_{i,j} \| \mathbf{w}_{ij} - \mathbf{x}_{ij} \|^2 = \sum_{i,j} \| \mathbf{w}_{ij} - (\mathtt{R}_i \sum_{d=1}^{D} l_{id} \mathtt{S}_d) \|^2 \tag{5.13}$$

where $\mathbf{w}_{ij}$ are the measured image points and $\mathbf{x}_{ij}$ the estimated image points. We propose two alternative solutions to this constrained minimization: a linear alternate least squares approach which incorporates the rigidity constraints using Generalised Singular Value Decomposition and a fully non-linear minimization scheme using priors on the rigid shape parameters in a Maximum A Posteriori estimation.

5.5.1 Linear equality-constrained least squares

First we propose an alternating least squares scheme to minimize the cost function described in equation (5.13). The algorithm alternates between solving for the basis shapes $\mathtt{S}$ and for the configuration weights l_{id}. Note that the algorithm does not solve for the overall rigid motion encoded in the rotation matrices $\mathtt{R}$ since these are calculated before hand by running the rigid factorization algorithm of Tomasi and Kanade on the segmented rigid points. The configuration weights are initialised to random values. The scheme can be summarised as follows:

1. Given $\mathtt{R}_i$ and l_{id} equation (2.36) can be used to estimate $\mathtt{S}$ linearly subject to the constraint $\tilde{\mathbf{S}}_p = \mathbf{0}$ for $p \in \Omega$ with Ω being the set of r points considered to be rigid throughout the sequence.

2. Given $\mathtt{R}_i$ and $\mathtt{S}$ solve for all l_{id} using linear least-squares.

3. Iterate the above two steps until convergence.

Rearranging equation (5.11) the problem of solving for $\mathtt{S}$ subject to the rigidity constraint can be expressed as an unconstrained least squares system of the form:

$$\min \left\| \begin{bmatrix} \mathtt{A} \\ \lambda \mathtt{C} \end{bmatrix} \mathbf{x} - \begin{bmatrix} \mathbf{b} \\ \lambda \mathbf{d} \end{bmatrix} \right\|^2 \tag{5.14}$$

where $\mathtt{A}$ encodes the linear equations, $\mathtt{C}$ the linear constraints and $\mathbf{b}$ and $\mathbf{d}$ are the known observations. It can be shown [51] that for $\lambda \rightarrow \infty$ the final solution lies on the surface defined by $\mathtt{C}\mathbf{x} = \mathbf{d}$ and thus we obtain a linear equality-constrained least squares (LSE) problem:

$$\min \| \mathtt{A}\mathbf{x} - \mathbf{b} \|^2 \tag{5.15}$$

subject to:

$$\mathtt{C}\mathbf{x} = \mathbf{d} \tag{5.16}$$

In our specific case, $\mathbf{x}$ alternatively represents the parameters for the 3-D basis shapes (step 1) or the configuration weights (step 2), $\mathtt{A}$ is the matrix of linear equations given the previously estimated rigid motion components, $\mathbf{b}$ the known observations i.e., the rearranged measurement matrix entries. The matrix $\mathtt{C}$ encodes the linear constraints that enforce the non-rigid component of the basis shapes $\tilde{\mathbf{S}}_j$ being equal to zero.

A method to solve the above LSE problem is to directly factorize both $\mathtt{A}$ and $\mathtt{C}$ using Generalized Singular Value Decomposition (GSVD) (see [58] for details).

5.5.2 Bundle adjustment using priors

An alternative approach to minimize the deformable cost function in equation (5.13) is given by non-linear optimization. One of the main advantages of performing a prior segmentation of rigid and non-rigid motion is firstly that the rigid motion (estimates of the rotation matrices $\mathtt{R}$) can be pre-computed by performing rigid factorization on the rigid points. This provides a reliable initial estimate for the rotation parameters which, coupled with the priors on the 3-D shape, help solve the ambiguities.

The camera parameters $\mathtt{R}_i$ at each frame i are then used to infer the mean basis component of

the deformable points such that:

$$\begin{bmatrix} \mathbf{S}_{1(r+1)} & \dots & \mathbf{S}_{1P} \end{bmatrix} = \begin{bmatrix} \mathtt{R}_1 \\ \vdots \\ \mathtt{R}_F \end{bmatrix}^{+} \mathtt{W}_{nonrigid} \tag{5.17}$$

where $\mathbf{S}_{1(r+1)}$ is the 3-vector which contains the coordinates of the rigid basis for the first non-rigid point (note that there are $(P-r)$ non-rigid points). Finally, the deformable components of the structure (configuration weights and 3-D basis) are initialised to small and random values as already shown in section 3.4.2.

5.5.3 Forcing the prior

Our prior expectation is that a point j detected as being rigid will have a zero non-rigid component and can therefore be modelled entirely by the first basis shape:

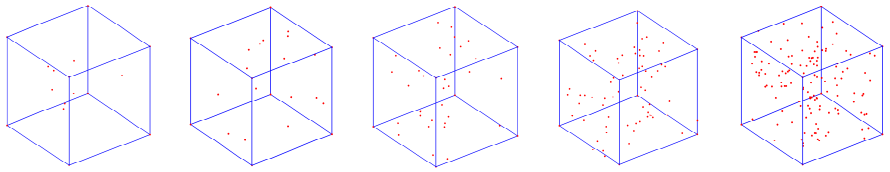
$$\mathbf{S}_j = \begin{bmatrix} \mathbf{S}_{1j} \\ \tilde{\mathbf{S}}_j \end{bmatrix} = \begin{bmatrix} \mathbf{S}_{1j} \\ \mathbf{0} \end{bmatrix} \tag{5.18}$$

where $\tilde{\mathbf{S}}_j = \begin{bmatrix} \mathbf{S}_{2j}^T & \cdots & \mathbf{S}_{Dj}^T \end{bmatrix}^T$. Therefore our expected prior value of the coordinates of the non-rigid bases $\tilde{\mathbf{S}}_j$ is zero in this case. For every rigid point in the scene we model the distribution of $\tilde{\mathbf{S}}_j$ as a Gaussian with a small variance and solve the problem as a Maximum A Posteriori estimation (MAP).

An alternative solution would have been to explicitly parameterise the points only with the rigid component by completely removing in the minimisation the non-rigid bases $\tilde{\mathbf{S}}_j$. However, we expect that the algorithm providing the motion segmentation may be inaccurate. In this case a hard decision given by the complete elimination of the non-rigid bases parameters for the rigid points can negatively affect the estimation process since, in the case of wrong priors, we are trying to infer the wrong model. Differently, a prior enforced as a penalty term can account of inaccuracies in the priors computation as we have shown in section 6.5.3.

5.6 Results

We show results for the proposed segmentation algorithm and the deformable 3-D shape estimation with both linear and non-linear approaches. Synthetic experiments are created especially to test the performance of the algorithms with different ratios of rigid/non-rigid points. The real



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

Figure 5.3: Synthetic sequence. Example of ground truth of the 3-D shape with 8 rigid points (vertices of the cube) and (a) 8, (b) 16, (c) 32, (d) 64 and (e) 128 non-rigid points.

experiments focus on face modelling: a set of trajectories is extracted from a subject performing different facial expressions and then subsequently reconstructed with the non-linear method using priors.

5.6.1 Synthetic data

The synthetic 3-D data consisted of a set of random points sampled inside a cube of size $50 \times 50 \times 50$ units. Five sequences were generated with 8, 16, 32, 64 and 128 non-rigid points sampled inside the cube. Each sequence also included 8 rigid points (the vertices of the cube). Figure 5.3 shows the 3-D data used in each of the five sequences with the rigid points joined up for display purposes. Our aim is to show the performance of our approach under different degrees of non-rigidity. The deformations for the non-rigid points were generated using random basis shapes as well as random deformation weights. Two basis shapes were used and the first basis shape had the assigned configuration weight equal to 1. The data was then rotated and translated over 25 frames and projected onto the images using an orthographic camera model and Gaussian noise was added to the image coordinates. The overall rotation about any axis was 90 degrees at most and the ratio of the norm of the non-rigid and rigid points of the 3-D metric shapes $ratio = \frac{\|\mathbf{S}_{nonrigid}\|}{\|\mathbf{S}_{rigid}\|}$ was fixed to 40%.

Rigid and non-rigid motion segmentation

Figure 5.4 shows results of the motion segmentation algorithm on a sequence using 8 rigid and 32 non-rigid points. The Gaussian noise level for this particular experiment was set to be $\sigma = 1.5$ pixels. The algorithm iteratively classifies points according to the current value of D as shown in Algorithm 1. The $-y$ axis of the graph shows the current value of the deformation index D and the $-x$ axis represents the number of iterations. The first 32 iterations remove non-rigid points as the deformability index D of the remaining set of points is consistently close to 2. When the 33^{rd}

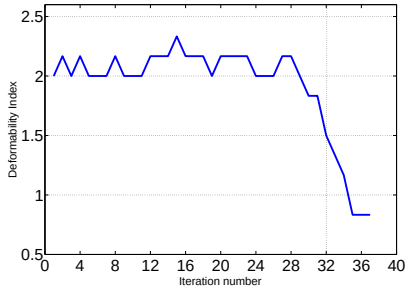


Figure 5.4: Deformability index for the automatic segmentation experiment. The graph shows its sudden decrease upon iteration number 33 which corresponds to the selection of the first rigid point.

iteration is reached, a rigid point is selected and one can observe a sudden drop in the value of D to 1.5 which then tends to 1. This is the cut-off point and the 8 remaining points are correctly classified as being rigid.

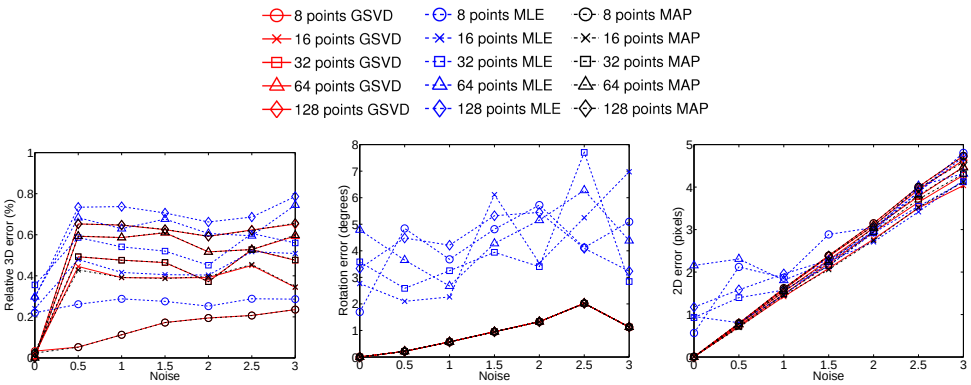
In order to test the algorithm exhaustively, we performed 1000 trials for each configuration when we varied the ratio of rigid/non-rigid points and used 5 different level of Gaussian noise ($\sigma^2 = 0, 0.5, 1, 1.5, 2$ pixels). Results showing the number of misclassified points are displayed in table 5.1. The values refer to the mean number of misclassified points when the $D = 1$ stopping condition becomes true. Notice that the algorithm achieves very low misclassification rates (a maximum of 1 rigid point misclassified as non-rigid) until the trial with 64 non-rigid points and 8 rigid points. For this ratio of rigid/non-rigid points we found the algorithm to fail for levels of noise of 1.5 pixels and above (indicated with a cross in the table) since the given threshold was terminating the iterations prematurely.

3-D reconstruction

We have tested three reconstruction algorithms: the linear GSVD method, bundle adjustment without priors (MLE) and bundle-adjustment incorporating priors on the 3-D structure (MAP). Figure 5.5 shows the relative 3-D reconstruction error, absolute rotation error and 2-D image reprojection error using each of the 3 algorithms, for varying ratios of rigid/non-rigid scene points and different levels of image noise. It becomes clear that GSVD and MAP outperform MLE thus showing the improved performance when prior information on the shape is incorporated. In fact the GSVD and MAP error curves appear superimposed which shows that they converge to the same solution, with the main observable difference being the higher speed of convergence for the

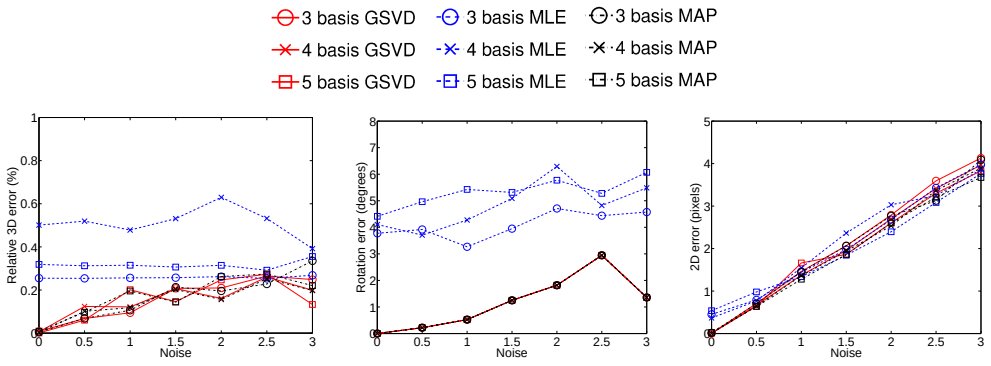
Rigid / Non-rigid	Noise				
	0	0.5	1	1.5	2
8/8	0	0	0.325	0.356	0.313
8/16	0	0.902	0.933	0.989	0.993
8/32	0	0	0.557	0.999	1
8/64	0	0.981	0.976	X	X

Table 5.1: Mean number of misclassified rigid points on 1000 trials for the experiments with 8 rigid points and varying number of non-rigid points $(8, 16, 32, 64)$. A cross indicates a failure of the algorithm to classify the rigid set of points.



3-D Error Rotation Error 2-D Error

Figure 5.5: Relative 3-D error (%), r.m.s. rotation error (in degrees) and 2-D reprojection error (in pixels) for the synthetic experiments for different ratios of rigid/non-rigid points and increasing levels of Gaussian noise.



3-D Error Rotation Error 2-D Error

Figure 5.6: Relative 3-D error (%), r.m.s. rotation error (deg) and 2-D reprojection error for the synthetic experiments for different numbers of basis shapes and increasing levels of Gaussian noise.

MAP approach. Note that the MLE approach is not able to compute a correct 3-D reconstruction even for the noiseless case showing that the added priors are fundamental to avoid local minima given by ambiguous configurations of motion and deformation parameters.

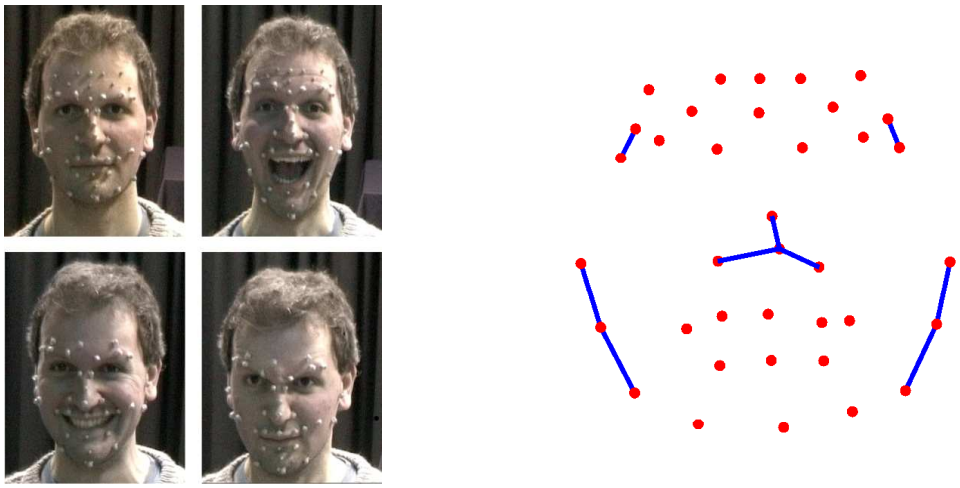
The number of basis shapes was then varied ($d = 3$, 4 and 5) to test the performance of the algorithm with respect to this parameter. Figure 5.6 shows the 2-D image reprojection error, relative 3-D reconstruction error and absolute rotation error obtained with GSVD, MLE and MAP. As expected, the error increases with the number of basis shapes for all 3 algorithms. Once more GSVD and MAP have almost identical performance and provide better results than MLE.

5.6.2 More realistic data

In this experiment[1] we use real 3-D data of a human face undergoing rigid motion – mainly rotation – while performing different facial expressions. The 3-D data was captured using a VICON motion capture system by tracking the subject wearing 37 markers on the face. Figure 5.7 (a) shows four key-frames showing the range of deformations of some expressions in the tested sequence.

The 3-D points were then projected synthetically onto an image sequence 310 frames long using an orthographic camera model and Gaussian noise of variance $\sigma = 0.5$ pixels was added to the image coordinates. In this case the segmentation of points into rigid and non-rigid sets

[1] Video available at http://www.bmva.ac.uk/thesis_archive/2006/DelBue1/index.html



(a) (b)

Figure 5.7: (a) The four frames show a few facial expressions performed by the subject. A VICON motion capture system extracts the 3-D locations from the markers attached to the subject's face (b) Face points used in the real experiment. Points connected with wire-frames show the selected rigid points located on the nose, temples and side of the face.

was done manually. Figure 5.7 (b) shows a frontal view of the face where the 14 rigid points – situated on the nose, temples and the side of the face – are connected with wire-frames.

Figure 5.8 shows the ground truth and reconstructed shape from front, side and top views using the bundle adjustment algorithm incorporating rigidity priors on the non-deforming points. The deformations are very well captured by the model even for the frames in which the facial expressions are more exaggerated.

5.7 Closure

The proposed formulation with shape priors relies on the presence of a set of points on the deforming surface that are only undergoing rigid motion. The priors may be constructed by simply selecting manually the rigid points lying on the object or by automatically finding the points with the motion segmentation algorithm provided in section 5.3. Given a reliable separation of rigid and non-rigid motion, our approach follows with an initial estimation of the rigid components of the 3-D structure and camera motion exclusively from the rigid trajectories by applying Tomasi and Kanade factorization [139]. Notice that at this stage, robust algorithms for rigid factorization such as [1, 78] may be also applied to deliver more accurate reconstructions.

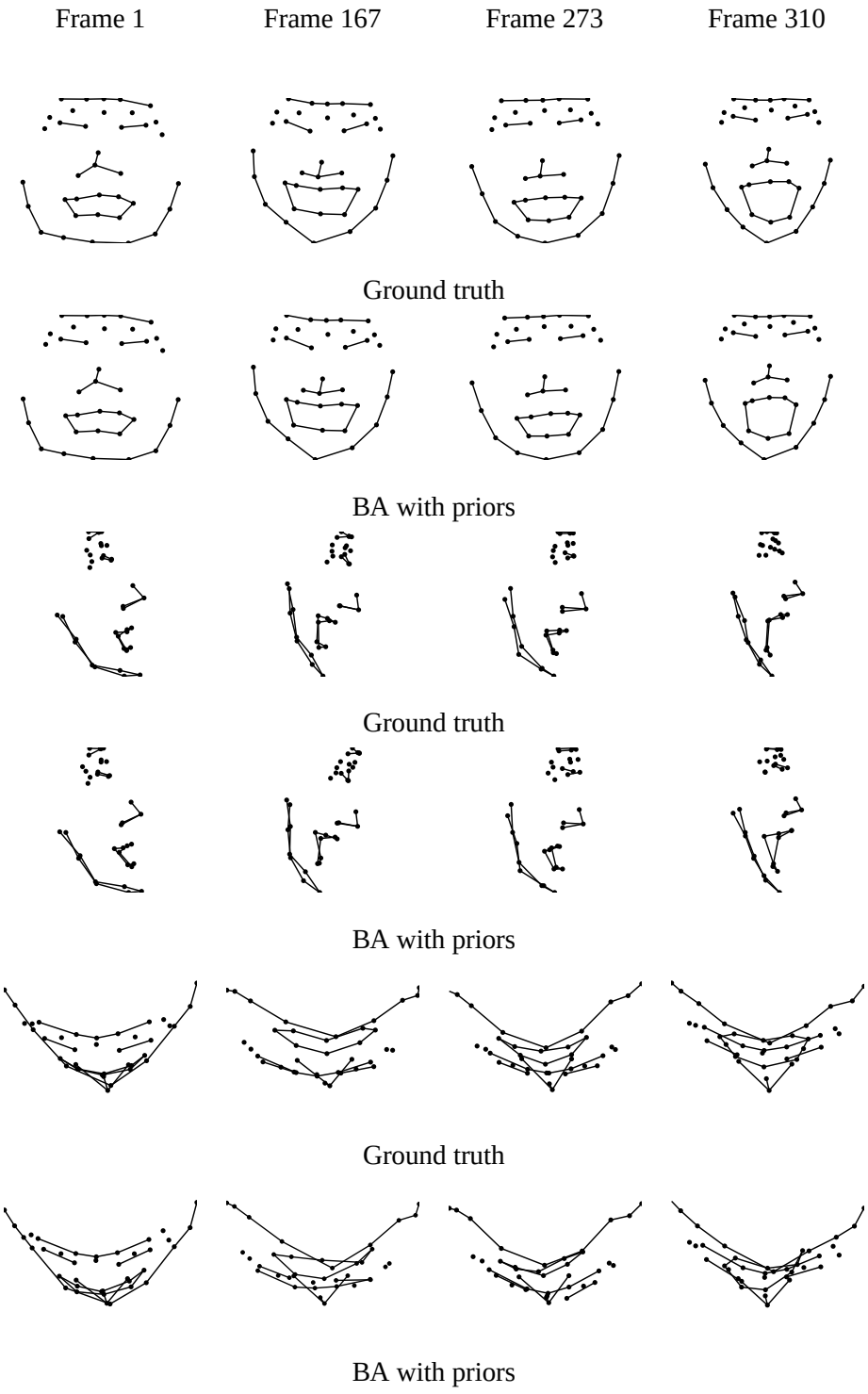


Figure 5.8: Front, side and top views of the ground truth and reconstructed face with priors. Reconstructions are shown for frames 1, 167, 273 and 310.

We then propose to use the extracted rigid component as a strong support for estimating the remaining 3-D deformable structure by designing two different algorithms. Firstly, the non-rigid parameters are estimated using an alternating equality constrained least-squares estimation over the configuration weights and non-rigid 3-D structure components while keeping fixed the orthographic camera parameters previously estimated with the rigid factorization.

Secondly, we include the prior information in the non-linear optimization framework presented in chapter 3. The problem is reformulated as the minimization of a non-linear cost function and, thus, it requires an initialisation close to the global minimum for the rigid and non-rigid parameters of the model. This is reasonably provided by the estimation of the rigid parameters given from the detected rigid points as shown in our experimental section. It is also evident that the introduction of the priors as penalty terms in the cost function gives improved results compared to MLE estimation.

The whole approach relies on the extraction of rigid motion from the image trajectories stored in $\mathtt{W}$. To support the detection of these points, we have introduced a specific method for the segmentation of rigid/non-rigid motion based on the rank constraint properties of rigid shapes. We employ a procedure introduced in [126] that can efficiently estimate the number of basis shapes of the deforming object in the presence of noise.

Provided an accurate estimation of the noise covariance, the algorithm performs well with different ratios of rigid/non-rigid points and different levels of noise affecting the measurements. In real cases, its efficiency can be affected whenever the noise statistics are not correctly provided or whenever the assumption that there is a sub-set of points that is perfectly rigid does not hold.

Finally, notice that in our synthetic experiments we have shown that the approach with priors converges to the global minimum and thus to the exact 3-D structure and camera motion in the case of no noise. Exact results are also obtained by Xiao et al. [159] using priors based on the independency of the basis shapes. A clear advantage of their approach is the proposed closed-form solution that is guaranteed to achieve a unique solution. On the other hand, the method is quite sensitive to the selection of the independent bases (see section 2.4.5 for a discussion) and no study under different levels of noise is given. The advantage of our solution consists on the use of priors extracted from rigid points lying over a deformable surface. Rigidly moving points are intuitively easier to detect, even with manual initialisation, than a set of independent basis shapes. In the next chapter we show that the information provided by the rigid points of a deformable

object can be crucial in the case of projective distortions affecting the image measurements.

Chapter 6

Deformable metric reconstruction from perspective cameras using priors

---

So far, all the algorithms we have presented for deformable factorization, including our non-linear optimization (MLE and MAP) methods, assume the case of images acquired under weak perspective viewing conditions. An extension to more general camera models is required when the inspected shape presents perspective distortion effects. This is the case when images are acquired at closer distances or with a camera with a wide field of view. Given a deformable object and a perspective camera, disambiguating the non-rigidity contributions and the camera distortions is fundamental for obtaining a correct reconstruction.

In this chapter we present a novel approach [36, 93] to the recovery of metric 3-D deformable models from perspective images. The solution proposed is based on the observation that often not all the points on a deformable surface are undergoing non-rigid motion as some of them might lie on rigid parts of the structure. First we use an automatic segmentation algorithm to identify the set of rigid points which in turn is used to estimate the internal camera calibration parameters and the overall rigid motion. We then formalise the problem of non-rigid shape estimation as a constrained non-linear minimization adding priors on the degree of deformability of each point. We perform experiments on synthetic and real data which show firstly that, even when using a minimal set of rigid points, it is possible to obtain reliable metric information and, secondly, that the shape priors help to disambiguate the contribution to image motion caused by deformation and perspective distortion.

6.1 Rigid metric reconstruction from perspective cameras

Affine and orthographic cameras are only an approximation of the real viewing conditions affecting the projection of a rigid body onto the image plane. These models are generally effective when the relief of the object is small compared to the distance from the camera centre. On the other hand, when these assumptions weaken, the use of a perspective camera model is necessary to obtain a correct 3-D reconstruction of the object. However, the introduction of a perspective camera model requires the knowledge of the internal and external parameters of the camera that can be estimated directly from the measured image data using self-calibration methods. We will show in the following section solutions for this problem in the case of rigidly moving objects.

6.1.1 The perspective camera model

In the most restrictive of affine camera model, the orthographic model, the projection of 3-D points is a direct mapping of the 3-D shape coordinates onto the image plane coordinates only up to an overall rotation, translation and scale. A more faithful model of real imaging conditions is given by the perspective camera model (see figure 6.1). Image points are given as the projection of the 3-D structure through a perspective camera $\mathtt{P}_{3\times4}$ defined mathematically as:

$$\mathtt{P}_i = \mathtt{K}_i[\mathtt{R}_i \mid \mathbf{t}_i] \tag{6.1}$$

where the 3×3 rotation matrix $\mathtt{R}_i$ and the translation vector $\mathbf{t}_i$ represent the Euclidean transformation between the camera and the world coordinate system respectively and $\mathtt{K}_i$ is a 3×3 upper triangular matrix which contains the intrinsic camera parameters:

$$\mathtt{K} = \begin{bmatrix} f_x & s & u_x \\ & f_y & v_y \\ & & 1 \end{bmatrix} \tag{6.2}$$

where f_x and f_y represent the focal length divided by the pixel width and height respectively, (u_x, v_y) represents the principal point and s is a factor which is zero in the absence of skew. The intrinsic camera parameters may vary (for instance in the case of a zooming camera) or remain fixed at each frame.

A point $\bar{\mathbf{X}}_j = [X_j\ Y_j\ Z_j\ 1]^T$ in homogeneous 3-D coordinates is projected with a perspective camera $\mathtt{P}_i$ into the image frame i such that the following relation holds:

$$\bar{\mathbf{w}}_{ij} = \frac{1}{\lambda_{ij}}\,\mathtt{P}_i\bar{\mathbf{X}}_j \tag{6.3}$$

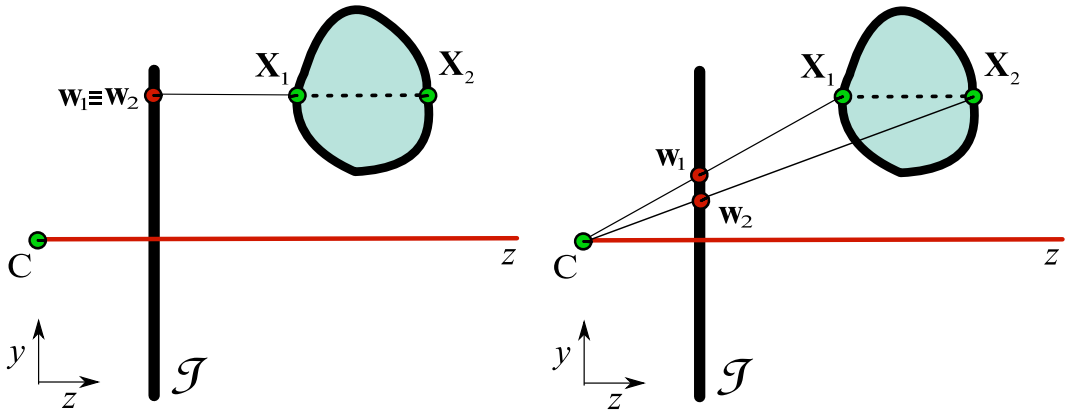


Figure 6.1: Comparison between an orthographic camera (a) and a perspective one (b). The 3-D points $\mathbf{X}_1$ and $\mathbf{X}_2$ are projected on the image plane $\Im$ to give the image coordinates $\mathbf{w}_1$ and $\mathbf{w}_2$ respectively. Orthographic projection (a) assumes the object being far from the image plane such that the projecting rays are all parallel to the optical axis and perpendicular to the image plane $\Im$. As a result, points having the same (x,y) coordinates but different depth z are projected at the same image location. In the perspective case (b), the projected image coordinates $\mathbf{w}_1$ and $\mathbf{w}_2$ have different image position depending on the depths of $\mathbf{X}_1$ and $\mathbf{X}_2$.

with $\bar{\mathbf{w}}_{ij} = [u_{ij}\ v_{ij}\ 1]^T = [\mathbf{w}_{ij}^T\ 1]^T$ representing the 2-D homogeneous image coordinates and λ_{ij} the projective depth of point j at frame i. However, given the 2-D image points $\mathbf{w}_{ij}$ extracted from an object moving rigidly in a perspective image sequence, the value of the correct projective depths λ_{ij} is unknown. In order to obtain a correct solution for the projective cameras $\mathtt{P}_i$ and projective points $\bar{\mathbf{X}}_j$, the extracted measurements need to be properly corrected by the projective weights λ_{ij}.

However, solving for the projective camera matrices $\mathtt{P}_i$ and projective structure $\bar{\mathbf{X}}_j$ is of limited use. The preserved geometrical properties obtained by estimating of projective reconstruction are restricted to the incidence of lines and the cross ratio between points [67]. What we seek is to obtain a metric 3-D structure from the perspective trajectories having an initial solution from the projective camera matrices $\mathtt{P}_i$ and projective structure $\bar{\mathbf{X}}_j$. It is possible to upgrade the estimated projective parameters to metric through a self-calibration process of the camera that solves for the unknown elements in $\mathtt{K}_i$, $\mathtt{R}_i$ and $\mathbf{t}_i$ in equation (6.1).

6.1.2 Self-calibration

Self-calibration is the simultaneous estimation of 3-D structure and camera motion purely from image sequences when no information is available about the internal calibration of the camera, the scene or the specific location of the camera as it moves. Commonly, methods for self-calibration can be distinguished in two classes: stratified [122, 121, 59, 64, 69, 98, 42] and direct [44, 101, 60, 146, 70, 65, 120, 3] approaches. The conceptual difference between the two groups is that stratified approaches work in stages by upgrading sequentially the structure to affine and finally to metric. Differently, direct approaches obtain in one step the full calibration of the camera which upgrades the reconstruction to metric.

Stratified approaches

A stratified method begin by seeking a solution for the perspective camera matrices $\mathtt{P}_i$ and 3-D structure $\bar{\mathbf{X}}_j$. The procedure then upgrades the geometry in two steps: first from perspective to affine and secondly from affine to Euclidean. To upgrade the reconstruction we rely on the estimation of invariant geometric entities in each of the geometric spaces (affine or Euclidean). Obtaining an affine reconstruction requires the location of the location of the plane at infinity – the invariant entity for the affine space. Once an affine reconstruction is obtained, solving for the absolute conic – the invariant for the Euclidean space – upgrades the reconstruction to Euclidean.

The main advantage of a stratified approach is that the solution from affine to metric spaces is linear after the determination of the plane at infinity (for instance, using the method proposed in [60]). However, the computation of the plane at infinity may require to determine specific properties of the scene such as the vanishing points of parallel lines. Another route is using the modulus constraint [122] to compute the coordinates of plane at infinity directly. The method, however, requires solving a set of quartic equations and this may render the algorithm unpractical given the large number of possible solutions.

Direct approaches

Direct methods, on the other hand, solve for the metric structure of the shape directly from the initial estimation of the projective matrices $\mathtt{P}_i$ and the projective 3-D coordinates $\bar{\mathbf{X}}_j$ without going through an affine upgrade of the geometry. The work of Faugeras et al. [44] was the first to analyse this problem, showing that self-calibration was feasible for a camera moving through an unknown scene with constant but unknown intrinsics. The method estimates the camera calibration form pairwise fundamental matrices by introducing the Kruppa equations to

solve for the unknown parameters.

Of more practical use, the method presented by Pollefeys et al. [119] allows to directly impose constraints on the intrinsic camera parameters given an explicit parameterisation of the camera calibration matrix $\mathtt{K}$ (see section 6.4.2 for a detailed description). Different approaches showed later that direct self-calibration is possible also in the case of more specific scenarios: where the camera is known only to rotate on the spot [63, 3], only to translate without rotation [105] or even when the camera has a zoom lens [119, 71].

Finally, note that for both approaches there remains an unsolved ambiguity given by an overall rotation and translation between camera and world coordinates in the Euclidean space. It is not possible to remove this ambiguity unless prior information about the location of the camera is available.

6.2 Projective rigid factorization

In order to perform self-calibration and reconstruct a rigid shape up to an overall similarity transformation (rotation, translation and scale), an initial estimation of the projective matrices $\mathtt{P}_i$ is needed. In a multi-view scenario, we have already discussed the advantages of solving the problem using factorization techniques in the case of a rigid object moving freely and viewed with an orthographic camera (see section 2.2). Similarly, a factorization solution is possible for the perspective case using an extension of Tomasi and Kanade's approach given a set of images taken under perspective viewing conditions. This will provide an initial estimation of the projective matrices $\mathtt{P}_i$ and the structure $\bar{\mathbf{X}}_j$ up to an overall projective transformation that in turn can be upgraded to metric by any of the self-calibration methods presented in the previous section.

Sturm and Triggs [132] firstly introduced projective factorization exploiting the rank constraint of the measurement matrix after the estimation of the weights λ_{ij}. Assuming the values of the projective depths are known and given equation (6.3), it is possible to write:

$$\bar{\mathtt{W}} = \begin{bmatrix} \lambda_{11}\bar{\mathbf{w}}_{11} & \dots & \lambda_{1P}\bar{\mathbf{w}}_{1P} \\ \vdots & & \vdots \\ \lambda_{F1}\bar{\mathbf{w}}_{F1} & \dots & \lambda_{FP}\bar{\mathbf{w}}_{FP} \end{bmatrix} = \begin{bmatrix} \mathtt{P}_1 \\ \vdots \\ \mathtt{P}_F \end{bmatrix} \begin{bmatrix} \bar{\mathbf{X}}_1 \cdots \bar{\mathbf{X}}_P \end{bmatrix} = \mathtt{MS} \tag{6.4}$$

where $\bar{\mathtt{W}}$ is the $3F \times P$ matrix containing the rescaled measurements, $\mathtt{M}$ is a $3F \times 4$ matrix and $\mathtt{S}$ a $4 \times P$ matrix. Thus, after re-weighting the image coordinates $\bar{\mathbf{w}}_{ij}$, the corrected $\bar{\mathtt{W}}$ is a rank-4 matrix. This property is used to perform SVD truncated to the fourth singular value to obtain a

solution for the projective motion and structure. Similarly to the affine case (see section 2.2), the matrices $\mathtt{M}$ and $\mathtt{S}$ are only estimated up to a 4×4 projective transformation matrix $\mathtt{Q}$ such that $\overline{\mathtt{W}} = \tilde{\mathtt{M}}\mathtt{Q}\mathtt{Q}^{-1}\tilde{\mathtt{S}} = \mathtt{M}\mathtt{S}$. The problem of estimating the true perspective depth is fundamental to obtain a correct decomposition and, as already presented in section 2.2.2, many algorithms have been developed in the last decade. Solving the problem in the case of non-rigid objects poses new challenges and the next section is dedicated to the mathematical definition of the problem.

6.3 Deformable metric 3-D reconstruction from perspective images

Given a non-rigid shape, its 3-D structure changes from frame to frame where $\overline{\mathtt{X}}_i = [\overline{\mathbf{X}}_{i1}\ldots\overline{\mathbf{X}}_{iP}]$ is a $(4\times P)$ matrix representing the shape at frame i in homogeneous coordinates. The deformation of a shape can often be explained as a linear combination of a set of D basis shapes $\mathtt{S}_d$ with $d = 1\ldots D$. In the projective case the 3-D vectors are expressed in homogeneous coordinates and so the shape may be written [161] as:

$$\overline{\mathtt{X}}_i = \begin{bmatrix} \sum_{d=1}^{D} l_{id}\mathtt{S}_d \\ \mathbf{1}^T \end{bmatrix} \qquad \overline{\mathtt{X}}_i \in \Re^{4\times P} \quad \mathtt{S}_d \in \Re^{3\times P} \tag{6.5}$$

where $\mathtt{S}_d$ are the $3\times P$ basis shapes, l_{id} are the corresponding deformation coefficients and $\mathbf{1}$ is a P-vector of ones. The projection of the shape at any frame i onto the image is then governed by the projection equation:

$$\overline{\mathtt{W}}_i = \mathtt{P}_i\overline{\mathtt{X}}_i = \mathtt{P}_i \begin{bmatrix} \sum_{d=1}^{D} l_{id}\mathtt{S}_d \\ \mathbf{1}^T \end{bmatrix} \tag{6.6}$$

In matrix form this can be re-written for all frames as:

$$\overline{\mathtt{W}} = \begin{bmatrix} \overline{\mathtt{W}}_1 \\ \vdots \\ \overline{\mathtt{W}}_F \end{bmatrix} = \begin{bmatrix} l_{11}\mathtt{P}_1^{(1:3)} & \ldots & l_{1D}\mathtt{P}_1^{(1:3)} & \mathbf{P}_1^{(4)} \\ \vdots & & \vdots & \vdots \\ l_{F1}\mathtt{P}_F^{(1:3)} & \ldots & l_{FD}\mathtt{P}_F^{(1:3)} & \mathbf{P}_F^{(4)} \end{bmatrix} \begin{bmatrix} \mathtt{S}_1 \\ \vdots \\ \mathtt{S}_D \\ \mathbf{1}^T \end{bmatrix} \tag{6.7}$$

where $\mathtt{P}_i^{(1:3)}$ are the first three columns of the projection matrix, $\mathbf{P}_i^{(4)}$ is the fourth column and $\mathbf{1}$ is a P-vector of ones.

Clearly, the rank of the measurement matrix is at most $3D+1$ for the projective case [161]. Once more, if the projective depths λ_{ij} were known the measurement matrix could be rescaled and decomposed into projective motion and shape matrices using factorization.

6.3.1 Previous work

In their most recent work Xiao and Kanade [161] proposed a new method to estimate the projective depths using the $3D+1$ sub-space constraint and then upgrade the projective reconstruction to a metric one using an extension of their affine closed form solution to the perspective camera case. However, their method still relies on the assumption that there be D frames in which the basis shapes are known to be independent.

Xiao and Kanade's method is a two step approach with similarities to an algorithm presented by Han and Kanade for the rigid case [57]. First, the projective weights λ_{ij} are estimated using the sub-space constraints arising from the $(3D+1)$ rank-constrained *motion* and *3-D structure* matrices. Similarly to the work of Han and Kanade [57] and Mahamud and Hebert [96] the procedure is carried out by performing an alternating minimization over `M` and `S` respectively. Additionally, the weights λ_{ij} are constrained to avoid degenerate solutions (for instance, some of the λ_{ij} can be equal to zero).

The second step is essentially an extension to the non-rigid case of the method proposed by Han and Kanade [57] to recover rigid structure from uncalibrated views with a direct self-calibration approach. However, to avoid degenerate solutions given the deformations, a new set of equations is introduced forcing the constraint that there exists a set of D independent basis shapes as previously introduced by Xiao et al. [159] in the orthographic case.

The aim is to estimate the overall $(3D+1)\times(3D+1)$ transformation matrix `Q` which upgrades the structure to metric and to preserve the repetitive structure of the motion matrix `M`. Similar to the orthographic case, the basis constraints are introduced to solve uniquely for each D column-triple of `Q`. Procrustes analysis is then used to align the structure of the motion matrix `M` to respect the repetitive structure of the factorization framework and to remove the scaling and translation ambiguities.

6.4 Our approach

Once more, our approach is based on the assumption that some of the points are rigid. The method requires three steps. First the image points are segmented into the rigid and non-rigid sets. The rigid points are then used to perform self-calibration and to recover the overall rigid motion and the camera calibration parameters as well as the metric rigid shape. Finally, the non-rigid bases and the deformation coefficients are estimated using a non-linear bundle adjustment

approach initialised using the estimates given by the rigid points. The bundle adjustment step can be seen as a refinement step with priors on the degree of deformability of the points with the aim to avoid ambiguous configurations of motion, perspective distortion and deformation.

6.4.1 Step 1: Segmentation of rigid and non-rigid motion under perspective viewing

In the case of affine cameras the rank of a measurement matrix containing a set of rigid points is constrained to be at most 3. This numerical condition of the measurement matrix $\mathtt{W}$ was used to obtain a reliable segmentation of rigid and non-rigid points using the feature selection strategy as presented in section 5.3. However (see equation (6.4)), when the camera is described by the perspective model, the rank of the measurement matrix increases to 4, provided that the measurement matrix has been rescaled with the correct estimates of the projective depths λ_{ij}. When the points in the measurement matrix are non-rigid the overall rank is $3D+1$ in the projective camera case where D is the number of basis shapes. Unfortunately, the rank constraint cannot be used directly to segment rigid and non-rigid points, since the rigid points could always be explained as non-rigid points with zero configuration weights for the non-rigid basis shapes. Additionally, the segmentation method presented in the previous chapter may misclassify rigid points as being non-rigid since the perspective distortion could be mistaken as a deformation.

Instead, our new approach is based on the fact that rigid points will satisfy the epipolar geometry while the non-rigid points will give a high residual in the estimation of the fundamental matrix between pairs of views. We use a RANSAC algorithm [46] to estimate the fundamental matrices from pairwise frames in the sequence and to segment the scene into rigid and non-rigid points. Therefore, in this case we consider the dominant motion to be the rigid one and the non-rigid points to be the outliers.

However, a well known drawback of random sampling and consensus techniques is the computational cost required to obtain a valid set of points when the percentage of outliers is high, due to the large number of samples needed to be drawn from the data. Unfortunately, this is the most likely scenario in non-rigid structure from motion where we normally deal with a small proportion of completely rigid points. Here we exploit a measure of the degree of deformability of a point to infer a prior distribution of the probability of a trajectory being rigid or non-rigid given that measure. These distributions are then used as priors to perform guided sampling over the set of trajectories in a similar approach to the one proposed by Tordoff and Murray [140] for the stereo matching problem.

Degree of non-rigidity

Kim and Hong [87] introduced the notion of Degree of Non-rigidity (*DoN*) of a point viewed by an orthographic camera as an effective measure of the deviation of the point from the average shape. If the average 3-D shape of a time varying shape $\mathtt{X}_i = [\mathbf{X}_{i1} \dots \mathbf{X}_{iP}]$ (in non-homogeneous coordinates) is given by $\check{\mathtt{X}} = [\check{\mathbf{X}}_1 \dots \check{\mathbf{X}}_P]$ the Degree of Non-rigidity for point j is defined as:

$$DoN_j = \sum_{i=1}^{F} (\mathbf{X}_{ij} - \check{\mathbf{X}}_j)(\mathbf{X}_{ij} - \check{\mathbf{X}}_j)^T \tag{6.8}$$

The 2-D projection $\mathtt{C}_j$ of the *DoN* will be thus given by:

$$\mathtt{C}_j = \sum_{i=1}^{F} \mathtt{R}_i (\mathbf{X}_{ij} - \check{\mathbf{X}}_j)(\mathbf{X}_{ij} - \check{\mathbf{X}}_j)^T \mathtt{R}_i^T = \sum_{i=1}^{F} (\mathbf{w}_{ij} - \check{\mathbf{x}}_j)(\mathbf{w}_{ij} - \check{\mathbf{x}}_j)^T \tag{6.9}$$

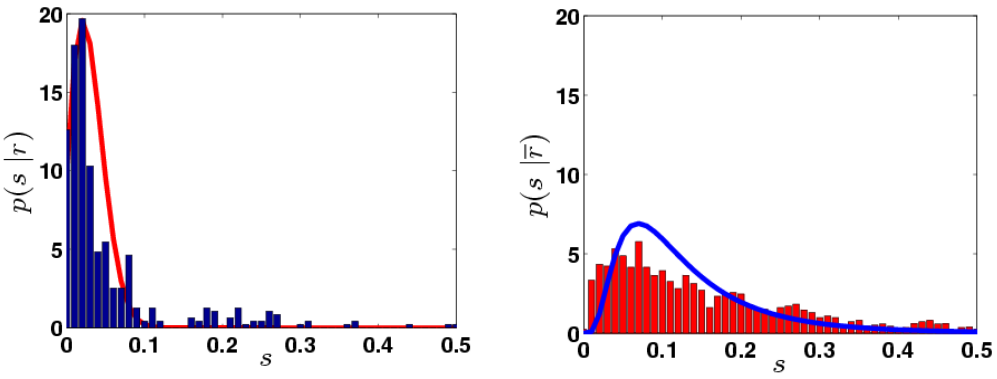
where $\mathbf{w}_{ij}$ are the image coordinates of point j in frame i and $\check{\mathbf{x}}_j$ are the coordinates of its projected mean shape. While the *DoN* cannot be computed without an estimation of the mean 3-D shape (and this implies finding a 3-D deformable reconstruction), the value of its projection can be estimated directly from image measurements.

An approximate estimate of the projected 2-D mean shapes $\check{\mathbf{x}}_j$ can be given simply by the rank-3 approximation of the measurement matrix $\mathtt{W}$ computed using singular value decomposition and given by $SVD_3(\mathtt{W}) = \check{\mathtt{M}}\check{\mathtt{B}}$. The projected deviation from the mean for all the points would then be defined by $\{\mathbf{w}_{ij} - \check{\mathbf{x}}_j\} = \mathtt{W} - \check{\mathtt{M}}\check{\mathtt{B}}$. Kim and Hong computed a more sophisticated estimate of the average shape, but for simplicity we have used the above description which has shown to give a reasonable measure of the degree of deformability.

Notice that the previous definitions all assume affine viewing conditions. However, our trajectories resides in a projective space so we need to re-define the measure of non-rigidity. First, the original measurement matrix must be re-scaled by the estimated projective weights λ_{ij}. We calculate the projective weights λ_{ij} using sub-space constraints [70] and express the rescaled measurement matrix as $\bar{\mathtt{W}} = \{\lambda_{ij}\ [\mathbf{w}_{ij}^T\ 1]^T\}$. Then, we estimate the mean shape as the rank-4 approximation of the rescaled measurement matrix computed using singular value decomposition and given by $SVD_4(\bar{\mathtt{W}}) = \check{\mathtt{M}}\check{\mathtt{S}}$. The projected deviation from the mean would then be defined as before by $\{\bar{\mathbf{w}}_{ij} - \check{\mathbf{x}}_j\} = \bar{\mathtt{W}} - \check{\mathtt{M}}\check{\mathtt{S}}$ and the projection of the *DoN* can finally be computed as:

$$\mathtt{C}_j = \sum_{i=1}^{F} (\bar{\mathbf{w}}_{ij} - \check{\mathbf{x}}_j)(\bar{\mathbf{w}}_{ij} - \check{\mathbf{x}}_j)^T. \tag{6.10}$$

in the form of a 2×2 covariance matrix. Instead of using the full information of $\mathtt{C}_j$, we approximate the score s as the sum of the diagonal values of $\mathtt{C}_j$.



(a) (b)

Figure 6.2: Conditional densities for the score given: (a) that a point is rigid $p(s|r)$ or (b) non-rigid $p(s|\overline{r})$ approximated from the normalised frequency histograms for different synthetic and real sequences with different degrees of perspective distortion, deformation and ratio of rigid/non-rigid points.

Computation of the prior

Tordoff and Murray [140] showed that guided sampling based on knowledge extracted from the images can greatly improve the performance of a random sampling method, especially in the presence of noise or of a high number of outliers. In these cases standard RANSAC becomes computationally prohibitive given the large number of random samples that must be drawn from the data. Here we use the 2-D projection of the *DoN* defined in the previous section to provide the score s for each point trajectory which will be used to build a prior distribution of the conditional probability of each point in the object being rigid or non-rigid given this score.

We have inferred the conditional probability density functions for the score s given that a point is rigid $p(s|r)$ (see figure 6.2(a)) or non-rigid $p(s|\overline{r})$ (see figure 6.2(b)) by computing the normalised frequency histograms over many experimental trials with synthetic and real sequences with different perspective distortions, degrees of deformation and ratios of rigid/non-rigid points. We have then approximated the histograms by fitting appropriate analytical functions. To derive the prior conditional density function of a point being rigid given the non-rigidity score $p(r|s)$ we use Bayes theorem:

$$p(r|s) = \frac{p(s|r)p(r)}{p(s)} \propto \frac{p(s|r)}{p(s|r) + p(s|\overline{r})} \tag{6.11}$$

Figure 6.3 shows an example of a prior obtained from our experiments. Note that although the computation of the score is specific to each method the derivation of the prior given the

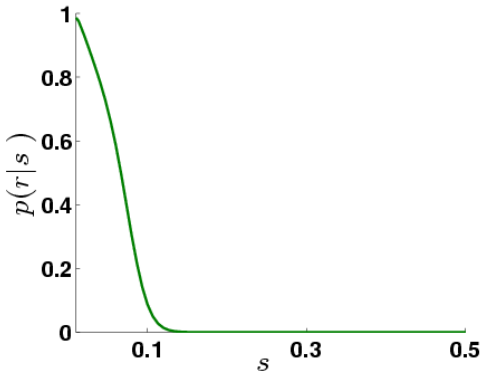


Figure 6.3: Estimated prior given by the estimated densities $p(s|r)$ and $p(s|\overline{r})$.

distribution of the score is general.

Guided RANSAC

We use guided RANSAC to estimate the fundamental matrices between pairs of consecutive views for all the F frames composing the sequence. This process will be used to provide a segmentation of the image trajectories into rigid and non-rigid ones since the non-rigid trajectories will not satisfy the epipolar geometry and will therefore give a high residual in the computation of the pairwise fundamental matrices. In order to speed up the process, we use the prior derived in the previous section to draw the point samples: points with the highest conditional probability of being rigid will be chosen more frequently. The RANSAC with priors procedure is outlined as follows:

1. Compute the score s for each trajectory in $\overline{\mathtt{W}}$.
2. Sample b trajectories given the prior $p(r)$ and the score s.
3. For each sample estimate $(F-1)$ fundamental matrices from each pair of consecutive frames.
4. Calculate the distance of the points from the $F-1$ instantiated models and find the trajectories that are within a threshold t.
5. Repeat N times and determine the largest consensus given a set of trajectories.

Algorithm 2.

The method employed to estimate the fundamental matrix is the standard 8-point algorithm [62] giving $b = 8$. The distance threshold t which decides whether a point is an inlier

or an outlier (rigid or non-rigid in this case) was set empirically to be $t = 4.12$. It was fixed by taking into account the sum of the residuals given by the estimation of F-1 fundamental matrices using normalised coordinates. Notice that we do not consider outliers in the point matching from frame to frame. We show results which asses the performance of the guided sampling RANSAC algorithm applied to the segmentation of rigid and non-rigid points in the experimental section. To notice that a common problem of RANSAC methods is their weakness to clustered outliers, that in our case corresponds to strong deformations affecting a relevant part of the image measurements. Additionally, we assume that the image points are extracted from a single non-rigid body. The algorithm would fail in the presence of articulated structures (for instance, the torso and the hands of a person) which show clustered rigid motions.

Once the scene has been segmented into the rigid and non-rigid point sets we compute metric non-rigid shape in two further steps. First we use the rigid points to estimate the camera intrinsic parameters – which provide the necessary information to upgrade the structure to metric – and the overall rotations and translations. Secondly, we formulate the estimation of metric non-rigid shape as a global non-linear minimization with shape priors over the rigid trajectories.

6.4.2 Step 2: Computing the metric upgrade

In order to obtain a metric upgrade, we first extract a projective reconstruction from the measurement matrix given the rigid set of points using Heyden's [68] sub-space method. The upgrade to metric space is then obtained using Pollefeys et al.'s approach for direct self-calibration which provides estimates for the camera intrinsic parameters, the overall rigid motion and the rigid shape.

Perspective reconstruction

Given the segmentation of the trajectories into rigid and non-rigid, we may now write:

$$\overline{\mathtt{W}} = \left[\begin{array}{c|c} \overline{\mathtt{W}}_{rigid} & \overline{\mathtt{W}}_{nonrigid} \end{array} \right] \tag{6.12}$$

where $\overline{\mathtt{W}}_{rigid}$ and $\overline{\mathtt{W}}_{nonrigid}$ are respectively the $3F \times r$ and $3F \times (P-r)$ matrices containing the r rigid and $(P-r)$ deformable image points. Following the projective approach outlined in section 6.1.1, we initially extract the projective 3-D shape and motion using the sub-space method of

Heyden [68] obtaining:

$$\bar{\mathtt{W}}_{rigid} \longrightarrow \begin{bmatrix} \mathtt{P}_1 \\ \vdots \\ \mathtt{P}_F \end{bmatrix} \begin{bmatrix} \bar{\mathbf{X}}_1 \cdots \bar{\mathbf{X}}_r \end{bmatrix} = \tilde{\mathtt{M}}_{rigid} \tilde{\mathtt{S}}_{rigid} \tag{6.13}$$

with $\tilde{\mathtt{M}}_{rigid}$ and $\tilde{\mathtt{S}}_{rigid}$ containing respectively the projective matrices $\mathtt{P}_i$ with $i = 1 \ldots F$ and the homogeneous coordinates for the rigid 3-D points $\bar{\mathbf{X}}_j$ with $j = 1 \ldots r$. Note that the method computes the projective weights λ_{ij} and decomposes $\bar{\mathtt{W}}_{rigid}$ into the rigid motion and shape matrices. Once more, the decomposition of $\bar{\mathtt{W}}_{rigid}$ is up to an unknown 4×4 projective transformation $\mathtt{Q}$ such that $\bar{\mathtt{W}}_{rigid} = \tilde{\mathtt{M}}_{rigid} \mathtt{Q}\, \mathtt{Q}^{-1} \tilde{\mathtt{S}}_{rigid}$. We solve uniquely for $\mathtt{Q}$ and then upgrading the rigid structure to metric by performing self-calibration over the projective matrices stored in $\tilde{\mathtt{M}}_{rigid}$.

From perspective to metric: self-calibration

In our specific case, we have used the well-known self-calibration method proposed by Pollefeys et al. [120]. The main advantage of this direct method is that it allows to impose different constraints on each of the camera intrinsic parameters (focal length, principal point and aspect ratio) since the camera calibration matrix is parameterized explicitly in terms of them. Each of the parameters may be considered to be known, unknown but constant between views or unknown and varying.

The projection matrix $\mathtt{P}_i$ for frame i is a rank 3 matrix which may be decomposed as $\mathtt{P}_i = \mathtt{K}_i[\mathtt{R}_i \mid \mathbf{t}_i]$, where the rotation $\mathtt{R}_i$ and the translation $\mathbf{t}_i$ represent the Euclidean transformation between the camera and the world coordinate systems and $\mathtt{K}_i$ is an upper triangular as already shown in equation (6.2).

The basic idea of this method (for a detailed description see [118]) consists on parameterizing the *dual image absolute conic* ω_i^* in such a way that it enforces the constraints on the calibration parameters using the equation:

$$\omega_i^* = \mathtt{K}_i \mathtt{K}_i^T \propto \mathtt{P}_i \Omega^* \mathtt{P}_i^T = \mathtt{P}_i \mathtt{Q} \mathtt{Q}^T \mathtt{P}_i^T \tag{6.14}$$

where $\mathtt{K}_i$ encodes the intrinsic parameters of the camera, $\mathtt{P}_i$ are the projective camera matrices and Ω^* is the *absolute quadric* for which a minimum parameterisation of 8 parameters is used. Note that constraints on the intrinsic camera parameters $\mathtt{K}_i$ are translated to constraints on the absolute quadric. As suggested by Pollefeys et al. the solution of the problem can be obtained

through non-linear least squares minimizing:

$$\min \sum_{i=1}^{F} \left\| \frac{\mathtt{K}_i \mathtt{K}_i^T}{\| \mathtt{K}_i \mathtt{K}_i^T \|} - \frac{\mathtt{P}_i \mathbf{\Omega}^* \mathtt{P}_i^T}{\| \mathtt{P}_i \mathbf{\Omega}^* \mathtt{P}_i^T \|} \right\|_F^2 \tag{6.15}$$

where initial estimates are obtained by means of a linear method.

After performing self-calibration, it is possible to obtain $\mathtt{Q}$ which allows to upgrade the camera matrices $\mathtt{P}_i$ and structure to metric space. So $\bar{\mathtt{W}}_{rigid}$ may be expressed as:

$$\bar{\mathtt{W}}_{rigid} = \begin{bmatrix} \mathtt{K}_1[\mathtt{R}_1|\mathbf{t}_1] \\ \vdots \\ \mathtt{K}_F[\mathtt{R}_F|\mathbf{t}_F] \end{bmatrix} \begin{bmatrix} \mathbf{S}_{11} & \cdots & \mathbf{S}_{1r} \\ 1 & \cdots & 1 \end{bmatrix} \tag{6.16}$$

The matrix $\mathtt{S}_{rigid}$ given by the collection of the 3-vectors such that $\mathtt{S}_{rigid} = [\mathbf{S}_{11} \ldots \mathbf{S}_{1r}]$ is the $3 \times r$ rigid basis of the deformable 3-D structure.

6.4.3 Step 3: Non-linear optimization

Following the approach presented in chapter 3, we solve for the non-rigid shape and motion given the 2-D image reprojection error. The cost function being minimised is the geometric distance between the measured image points and the estimated reprojected points $\chi = \sum_{i,j} \| \bar{\mathbf{w}}_{ij} - \bar{\mathbf{x}}_{ij} \|^2 = \sum_{i,j} \| \bar{\mathbf{w}}_{ij} - \mathtt{P}_i \bar{\mathbf{X}}_{ij} \|^2$ where $\mathtt{P}_i$ is the projection matrix in the Euclidean frame i and $\bar{\mathbf{X}}_{ij}$ is the 4-vector that encodes the homogeneous metric 3-D coordinates of point j in frame i. In order to ensure good numerical conditioning we work with normalised image coordinates as described in [67].

We parameterize the projection matrices in terms of the calibration matrices $\mathtt{K}_i$, the rigid rotation matrices $\mathtt{R}_i$ using quaternions and the translation vectors $\mathbf{t}_i$. The coordinates of the non-rigid points $\bar{\mathbf{X}}_{ij}$ are parameterized in terms of the basis shapes $\mathbf{S}_{dj}$ and the deformation coefficients l_{id}. We may now write the non-linear minimization scheme as:

$$\arg \min_{\mathtt{K}_i \mathtt{R}_i \mathbf{t}_i \mathbf{S}_{dj} l_{id}} \sum_{i,j} \left\| \mathbf{w}_{ij} - \Pi \left(\mathtt{K}_i[\mathtt{R}_i|\mathbf{t}_i] \begin{bmatrix} \sum_{d=1}^{D} l_{id} \mathbf{S}_{dj} \\ 1 \end{bmatrix} \right) \right\|^2 \tag{6.17}$$

where Π is a function such that:

$$\Pi \begin{pmatrix} a \\ b \\ c \end{pmatrix} = \begin{pmatrix} \frac{a}{c} \\ \frac{b}{c} \end{pmatrix} \tag{6.18}$$

We then impose the priors on the rigid points (zero value for the non-rigid component) as we explained for the orthographic case in section 5.5.2. If the motion of a point j is completely rigid

for the entire sequence, the structure referring to that point can be expressed entirely by the first basis ($d = 1$) called the rigid basis. From this it follows that for a rigid point $\mathbf{S}_{dj} = \mathbf{0} \quad \forall \; d > 1$ where $\mathbf{S}_j = [\mathbf{S}_{1j}^T, \ldots, \mathbf{S}_{Dj}^T \; 1]^T$.

Note that $\mathbf{S}_j$ is a $3D+1$ vector which encodes the D basis shapes for point j and $\mathbf{S}_{dj}$ is a 3-vector which contains 3-D coordinates of basis shape d for point j. Notice that this forces $3(D-1)$ zeros in each column of the shape matrix corresponding to a rigid point. We write these expectations as priors on the coordinates of the basis vectors $\mathbf{S}_{dj}$ and solve the problem as a Maximum A Posteriori (MAP) estimation.

Note that the final expression for the *motion* and *3-D structure* matrices is as follow:

$$\bar{\mathtt{W}} = \begin{bmatrix} l_{11}\mathtt{K}_1\mathtt{R}_1 & \ldots & l_{1D}\mathtt{K}_1\mathtt{R}_1 & \mathtt{K}_1\mathbf{t}_1 \\ \vdots & & \vdots & \\ l_{F1}\mathtt{K}_F\mathtt{R}_F & \ldots & l_{FD}\mathtt{K}_F\mathtt{R}_F & \mathtt{K}_F\mathbf{t}_F \end{bmatrix} \left[\begin{array}{c|c} \bar{\mathtt{S}}_{rigid} & \bar{\mathtt{S}}_{nonrigid} \end{array} \right] \tag{6.19}$$

where the $(3D+1) \times r$ rigid component of the 3-D structure $\bar{\mathtt{S}}_{rigid}$ is given in homogeneous coordinates by:

$$\bar{\mathtt{S}}_{rigid} = \begin{bmatrix} \mathtt{S}_{rigid} \\ \mathtt{0} \\ \mathbf{1}^T \end{bmatrix} \tag{6.20}$$

with $\mathtt{0}$ being a $3(D-1) \times r$ matrix of zeros and $\mathbf{1}$ a r-vector of ones. The $(3D+1) \times (P-r)$ matrix $\bar{\mathtt{S}}_{nonrigid}$ contains the deformable bases for the non-rigid points in homogeneous coordinates such that:

$$\bar{\mathtt{S}}_{nonrigid} = \begin{bmatrix} \mathtt{S}_{nonrigid} \\ \mathbf{1}^T \end{bmatrix} \tag{6.21}$$

where $\mathbf{1}$ is a $(P-r)$-vector of ones.

Initialisation

Non-linear optimization requires an initialisation of the parameters to minimize. The metric rigid component of the shape and structure given by self-calibration is used to obtain a reliable initialisation of the intrinsic and extrinsic parameters for the camera and the metric structure for the rigid points. Now it is possible to estimate the first basis $\mathtt{S}_1$ for the deformable model $\mathtt{S}$ given

the projection matrices $\mathtt{P}_i = \mathtt{K}_i[\mathtt{R}_i|\mathbf{t}_i]$ using the expression:

$$\begin{bmatrix} \mathbf{S}_{1(r+1)} & \dots & \mathbf{S}_{1P} \\ 1 & \dots & 1 \end{bmatrix} = \begin{bmatrix} \mathtt{P}_1 \\ \vdots \\ \mathtt{P}_F \end{bmatrix}^{+} \bar{\mathtt{W}}_{nonrigid} \tag{6.22}$$

The coordinates of the rest of the basis shapes which encode the $D-1$ non-rigid components $\mathtt{S}_d$ with $d = 2,\dots,D$ are initialised to small random values [141, 38]. The configuration weights associated with the mean shape l_{i1} are initialised to 1 while the rest of the weights l_{id} are initialised to small values.

6.5 Experimental results

This experimental section validates the methods for rigid/non-rigid segmentation and 3-D metric reconstruction with synthetic and real experiments. The synthetic tests are designed in such a way as to verify the performance of the method in case of different ratios of rigid/non-rigid points and with two different setups of perspective distortions. Additionally, the quality of the 3-D reconstruction is tested with cameras with constant and varying intrinsics.

Finally, three experiments present the performances of the approach in the case of real deforming objects. Two tests use image measurements obtained from a Vicon system which provides as well the ground truth for comparing the 3-D reconstructions. The remaining test is performed over less accurate measurements extracted by an image point tracker (KLT).

6.5.1 Synthetic data

The 3-D data consists of a set of random points sampled inside a cube of size $100 \times 100 \times 100$ units. Several sequences were generated using different ratios of rigid/non-rigid points. In particular, we used a fixed set of 10 rigid points while using 10 and 50 non-rigid points. The deformations for the non-rigid points were generated using random basis shapes as well as random deformation weights. The first basis shape had the largest weight equal to 1. We also created different sequences varying the number of basis shapes ($D = 3$ and $D = 5$) for both ratios of rigid/non-rigid points. Finally, in order to evaluate different levels of perspective distortion we used 2 different camera setups in which we varied the distance of the object to the camera and the focal length (Setup 1: z=250, f=900; Setup 2: z=200, f=600). The 3-D data was then projected onto 50 images applying random rotations and translations over all the axes. Gaussian noise of

Experiments	Noise				
	0	0.5	1	1.5	2
Exp1: $D = 5$, $10/10$, setup 1	0.28	0.48	0.55	0.72	0.77
Exp2: $D = 5$, $10/50$, setup 1	0.31	0.38	0.46	0.55	0.72
Exp3: $D = 3$, $10/10$, setup 1	0.95	1.36	1.53	1.60	1.54
Exp4: $D = 3$, $10/50$, setup 1	2.19	2.38	2.33	2.78	2.51
Exp5: $D = 5$, $10/10$, setup 2	0.95	1.36	1.53	1.6	1.54
Exp6: $D = 5$, $10/50$, setup 2	0.3	0.34	0.39	0.51	0.58
Exp7: $D = 3$, $10/10$, setup 2	0.65	0.94	1.27	1.42	1.45
Exp8: $D = 3$, $10/50$, setup 2	2.09	2.37	2.28	2.31	2.27

Table 6.1: Mean misclassification error for different levels of Gaussian noise with variance $\sigma =$ 0.5, 1, 1.5, 2 pixels. The eight experimental setups use different number of bases ($D = 3,5$), ratios of rigid/non-rigid points ($10/10$, $10/50$) and camera parameters (Setup 1: z=250, f=900; Setup 2: z=200, f=600). The mean error is computed over 100 tests for each setup and level of noise.

increasing levels of variance was added to the image coordinates.

6.5.2 Motion segmentation results

The experimental setup described beforehand was first used to obtain an indication of the performance of our segmentation approach presented in section 6.4.1. Firstly, the sampling prior $p(r)$ was generated from a larger set of synthetic and real data. Secondly, tests using the guided RANSAC approach were performed over the synthetic experiments described above. Eight different experimental setups were tested with varying number of rigid/non-rigid points ($10/10$, $10/50$), basis shapes ($D = 3,5$) and camera parameters (Setup 1, Setup 2).

The RANSAC procedure was tested over 100 trials for each setup and for each level of noise. The number of samples randomly chosen over the prior distribution was fixed to 2500. At each new trial the motion components (rotation and translation) of the objects are randomly generated obtaining a 50 frames long sequence. The results in table 6.1 show the rate of non-rigid points being classified as rigid for the different setups. Better performances are obtained for higher ratios of rigid/non-rigid points and for more complex deformations (i.e., more basis shapes).

Parameter	Noise				
	0	0.5	1	1.5	2
mean f	0	0.49	0.98	1.34	2.54
std. dev. f	0.02	0.66	1.42	1.46	2.62
max. f	0.09	3.43	8.56	5.97	10.02
mean p_u	0.01	0.72	1.19	1.61	2.14
mean p_v	0.01	0.77	1.18	1.63	2.26

Table 6.2: Mean, standard deviation and maximum relative error (%) for the focal length, and absolute mean error for the principal point (p_u, p_v) for the different levels of Gaussian noise. Results obtained when the intrinsic parameters were constant.

Experiments 4 and 8 obtain the worse results achieving a mean misclassification rate of more than 2 points.

Notice also a better algorithmic behavior in the case of stronger perspective distortion compared to weaker ones since the effects of perspective distortions and deformations are less ambiguous in such cases.

6.5.3 3-D reconstruction results with constant intrinsics

For the first set of experiments we assumed that all the camera parameters: focal length, aspect ratio, principal point and skew (equal to 0) remained constant over the sequence. We then applied our 3-D reconstruction algorithm to all the experimental setups described before. The results are summarized on the first row of figure 6.4 where we show the 3-D metric reconstruction error expressed in percentage relative to the scene size, the absolute rotation error expressed in degrees and the r.m.s. 2-D image reprojection error expressed in pixels. The plots in this figure show the mean values of 5 random trials applied to each level of Gaussian noise.

Our proposed algorithm appears to perform well in the presence of image noise. The 3-D reconstruction error is low even for large perspective distortions and for a large proportion of non-rigid versus rigid points. The 2-D error is also small and it appears to be of the same order as the image noise. Figure 6.4 also illustrates that the rotations are correctly estimated. Reliable estimates for the internal camera parameters (focal length and principal point) are obtained even in the presence of noise and they are summarized in table 6.2.

As expected, less accurate results were obtained in the presence of outliers (i.e. non-rigid

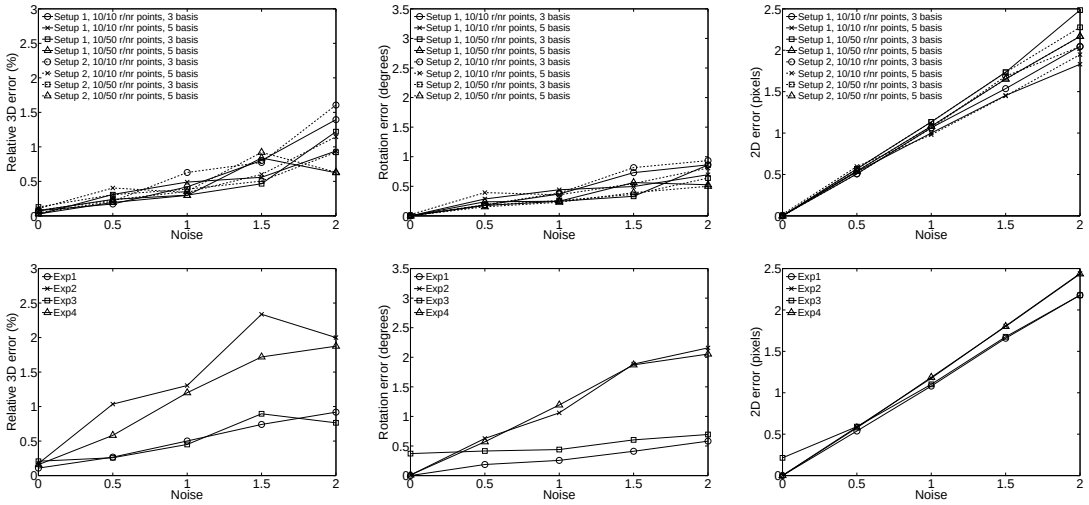


Figure 6.4: 3-D error, rotation error and 2-D error curves. First row: results obtained when the focal length (f) was constant. Second row: results obtained for the 4 experiments with varying intrinsics (see text for description).

points) in the original set of rigid points as shown in figure 6.5. This is due to the fact that outliers introduce errors in the initial estimates obtained by the projective rigid factorization and self-calibration. However, after applying bundle adjustment the results improved, providing acceptable motion and structure estimates.

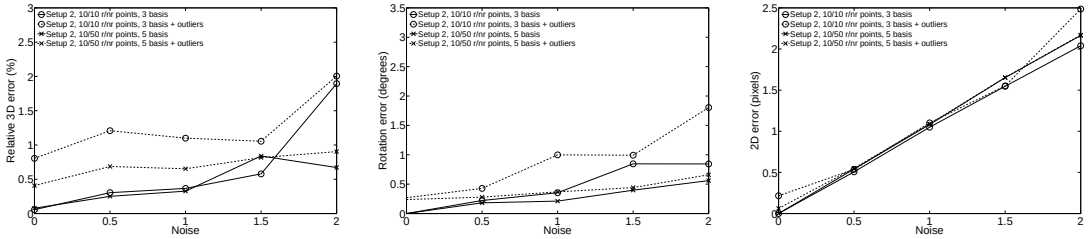


Figure 6.5: 3-D error, rotation error and 2-D error plots in the presence of two non-rigid points in the set of rigid points. Setup 2 is used in two experiments with varying number of rigid/non-rigid points. Results show the effect of outliers compared to the case with correct data.

6.5.4 3-D reconstruction results with varying intrinsics

We then performed a set of experiments in which some of the internal parameters of the camera were varyied throughout the sequence. We designed 4 different experiments using camera setup 2 ($Z = 200$, $f = 600$), a ratio of 10 rigid to 50 non-rigid points and 5 basis shapes. For Experiment 1 the focal length of the camera varied linearly throughout the sequence while the rest of the internal parameters remain constant. In the optimization algorithm we considered the focal

length unknown and allowed it to vary during the minimization while the principal point was considered to be unknown but fixed throughout the sequence and the aspect ratio and skew were considered known ($r = 1$ and $s = 0$). Experiment 2 had the same experimental setup but during the optimization process we allowed both the focal length and the principal point to vary in the minimization. In Experiment 3 the focal length and the principal point both varied throughout the sequence. In the minimization we considered the focal length unknown and allowed it to vary but the principal point was assumed to be fixed but unknown. Finally in Experiment 4 we used the same setup as in Experiment 3 but allowed both the focal length and the principal point to vary in the minimization.

The results for all 4 experiments are illustrated on the second row of figure 6.4. The results obtained for the internal camera parameters are summarised in table 6.3. Note that for the noisy cases in which the real principal point was varying better estimates were obtained assuming the principal point constant during the minimization.

Finally, we performed another experiment in order to show that inclusion of priors is fundamental to avoid local minima and to improve the reconstruction results. We chose the same set of experiments in which only the real focal length was varying and aspect ratio and principal point were assumed constant during the minimization. Results with and without using priors are illustrated on figure 6.6.

As expected, better results are obtained when priors are used. This can be clearly seen in the case of no noise where the use of priors allows to improve the convergence to the global minimum. Notice that the minimizations with and without priors were initialized with the same values thus showing that the inclusion of the additional penalty terms increases the reliability of the reconstruction and the convergence of the algorithm.

6.5.5 Real experiments

In these experiments we tested our method using real 3-D data of a human face and of a scene with deforming and rigidly moving objects. We present three experiments; in the first two we test our method compared using ground truth given by accurate measurements obtained from a VICON motion capture system. The final test shows the 3-D reconstruction results with measurements automatically generated by a point tracking algorithm (KLT).

	Parameter	Noise				
		0	0.5	1	1.5	2
Exp1	mean f	0	0.56	1.68	1.69	3.90
	std. dev. f	0	0.18	1.26	0.94	1.99
	max. f	0	0.83	3.49	3.22	7.16
	mean p_u	0	0.59	1.48	1.29	6.03
	mean p_v	0	0.91	2.43	2.50	3.46
Exp2	mean f	0.01	2.93	5.14	10.28	10.97
	std. dev. f	0.01	0.79	2.92	6.96	4.33
	max. f	0.02	3.91	8.36	20.12	14.92
	mean p_u	0.09	11.17	18.01	26.68	27.50
	mean p_v	0.08	6.66	14.80	22.93	28.91
Exp3	mean f	0.69	1.04	1.16	3.10	2.92
	std. dev. f	0.27	0.50	0.38	2.58	1.15
	max. f	1.04	1.75	1.81	5.96	4.47
	mean p_u	2.97	2.96	3.01	3.77	3.97
	mean p_v	3.49	3.34	3.47	5.88	3.79
Exp4	mean f	0.05	2.11	4.93	10.40	10.38
	std. dev. f	0.04	1.05	3.51	2.92	4.66
	max. f	0.09	3.60	8.80	14.27	14.17
	mean p_u	0.10	5.95	12.71	16.01	16.31
	mean p_v	0.07	3.49	10.61	14.34	15.54

Table 6.3: Mean, standard deviation and maximum relative error (%) of the focal length and absolute mean error (pixels) of the principal point (p_u, p_v) for different levels of Gaussian noise.

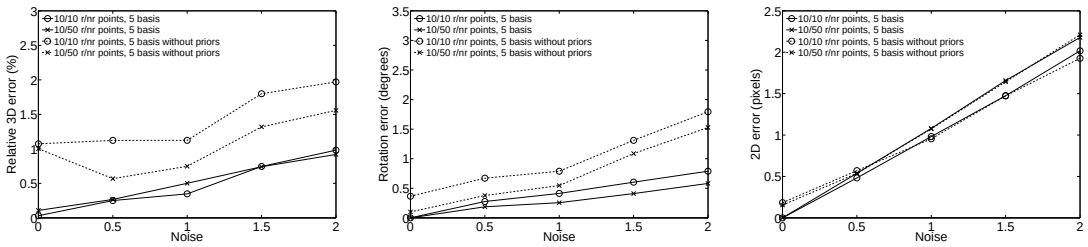


Figure 6.6: Obtained results with and without using shape priors. 3-D error, rotation error and 2-D error curves for the set of experiments obtained with camera setup 2, 5 basis shapes and $10/10$, $10/50$ rigid/non-rigid points. Focal length was varying while aspect ratio and principal point were constant.

Human face

In the first real experiment, 37 trajectories are generated from a human face that is undergoing rigid motion while performing different facial expressions. The 3-D points reconstructed by the motion capture system are then projected synthetically onto an image sequence 74 frames long using a perspective camera model. The size of the face model was $169 \times 193 \times 102$ units and the camera setup was such that the subject was at a distance of 300 units from the camera and the focal length was 600 pixels so the perspective effects are significant.

In this case the segmentation of points into rigid and non-rigid sets was done by manually selecting 14 points situated on the nose, temples and the side of the face. These points are highlighted on the frontal view of the first frame of figure 6.7. This figure shows the ground truth (squares) and reconstructed shape (crosses) from front, side and top views. The 2-D reprojection error was 0.67 pixels, the absolute 3-D error was 2.24 units and the focal length was 595.12 pixels. The results are satisfactory even considering that the selected rigid points were not perfectly rigid during all the sequence. Note that the deformations are very well captured by the model even for the frames in which the facial expressions are more exaggerated.

Pillow scene

The scene consisted of a set of 12 rigid points (9 on two boxes and 3 over a chair) and a set of 20 deformable points situated on a pillow which was deforming during the sequence (see first row of figure 6.8). The 3-D points were then projected synthetically onto an image sequence 75 frames long using a perspective camera model. Gaussian noise of 0.5 pixels was added to the image coordinates. The size of the scene was $61 \times 82 \times 53$ units and the camera setup such that

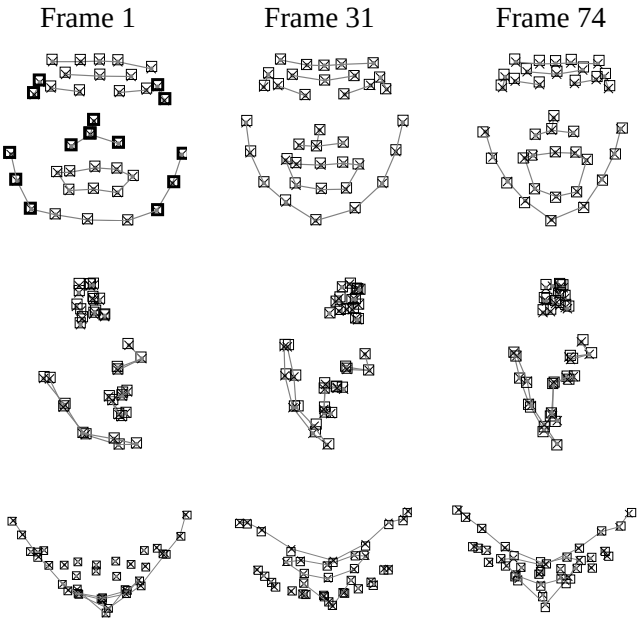


Figure 6.7: Front, side and top views of the reconstructed face. Reconstructions are shown for frames 1, 31 and 74. Cross marks are used to indicate the reconstruction while square marks refer to the ground truth. Highlighted marks on the frontal view of frame 1 indicate rigid points.

the scene was at a distance of 150 units from the camera and the focal length was 900 pixels and constant during the sequence. Figure 6.8 shows the ground truth (squares) and reconstructed shape (crosses) from two different viewpoints. The 2-D reprojection error was 0.95 pixels, the absolute 3-D error was 1.34 units and the absolute rotation error was 2.11 degrees. The focal length was estimated to be 899 pixels. The same experiment was repeated but varying the real focal length from 700 to 1000 during the sequence. In this case the 2-D reprojection error was 0.96 pixels, the absolute 3-D error was 1.65 units and the rotation error 2.77 degrees while the mean focal length error was 34.84 pixels (see table 6.4).

Cushion scene: automatically tracked data

In this experiment we show qualitative results with measurements obtained from a KLT tracker[1]. Some key frames of the sequence are presented in figure 6.9 (a) showing the object rigidly rotating (frames 1 and 160) and three deformations (frames 340, 410 and 490). The 560 frame long video sequence is captured with a Fire-i digital camera with $4,65$mm built in lenses. The tracking algorithm is able to obtain 256 trajectories located on the rigid (60 points over the box) and non-rigid (196 points over the cushion) surfaces of the scene. The trajectories are then sub-

[1]http://www.ces.clemson.edu/∼stb/klt/

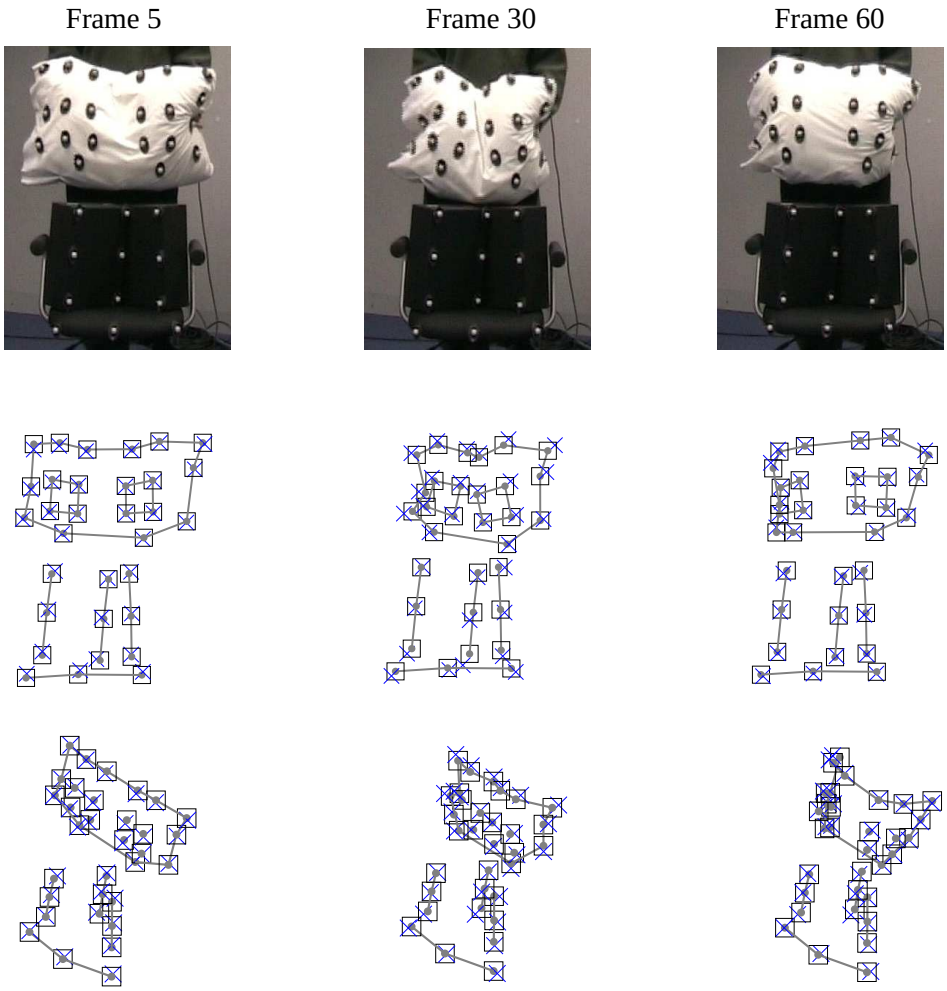


Figure 6.8: Real 3-D data. First row shows examples of the analysed scene. Second and third rows show two views of the reconstructed scene. Cross marks indicate reconstruction while square marks refer to ground truth.

Experiment	Error			
	2-D (r.m.s.)	3-D (%)	Rotation (r.m.s.)	f error
$f = 900$, $d = 150$	0.95 pixel	1.34	2.11 degree	1 pixel
$f = 700 - 1000$, $d = 150$	0.96 pixel	1.65	2.77 degree	34.84 pixel

Table 6.4: Estimated errors for the pillow sequence. Two setups with constant (first experiment) and varying (second experiment) intrinsics are tested and results are showed for the 2-D reprojection, 3-D reconstruction, rotation and focal length errors.

sampled in time to obtain an overall sequence of 112 frames giving a measurement matrix $\mathtt{W}$ of size 224×256. Ground truth reference is not available in this scene, thus we show only the results for the 3-D reconstruction after performing self-calibration using the rigid points and non-linear optimization to correctly model the shape deformations. The camera intrinsic parameters were considered constant for each frame and the aspect ratio and skew were fixed to 1 and 0 respectively. The presented results are obtained after 40 iterations of the non-linear optimization algorithm with a number of basis shapes set to $D = 5$.

Figure 6.9 shows front, side and top views of the 3-D reconstructions[2] for the deforming object. Frames 1 and 160 show only rigid displacements of the object and thus the 3-D structure is correctly not deforming for the two frames. The last three frames show the cushion bending and the box structure remaining rigid. Note in the top views of figure 6.9 (d) the preserved orthogonality of the two reconstructed planes belonging to the box.

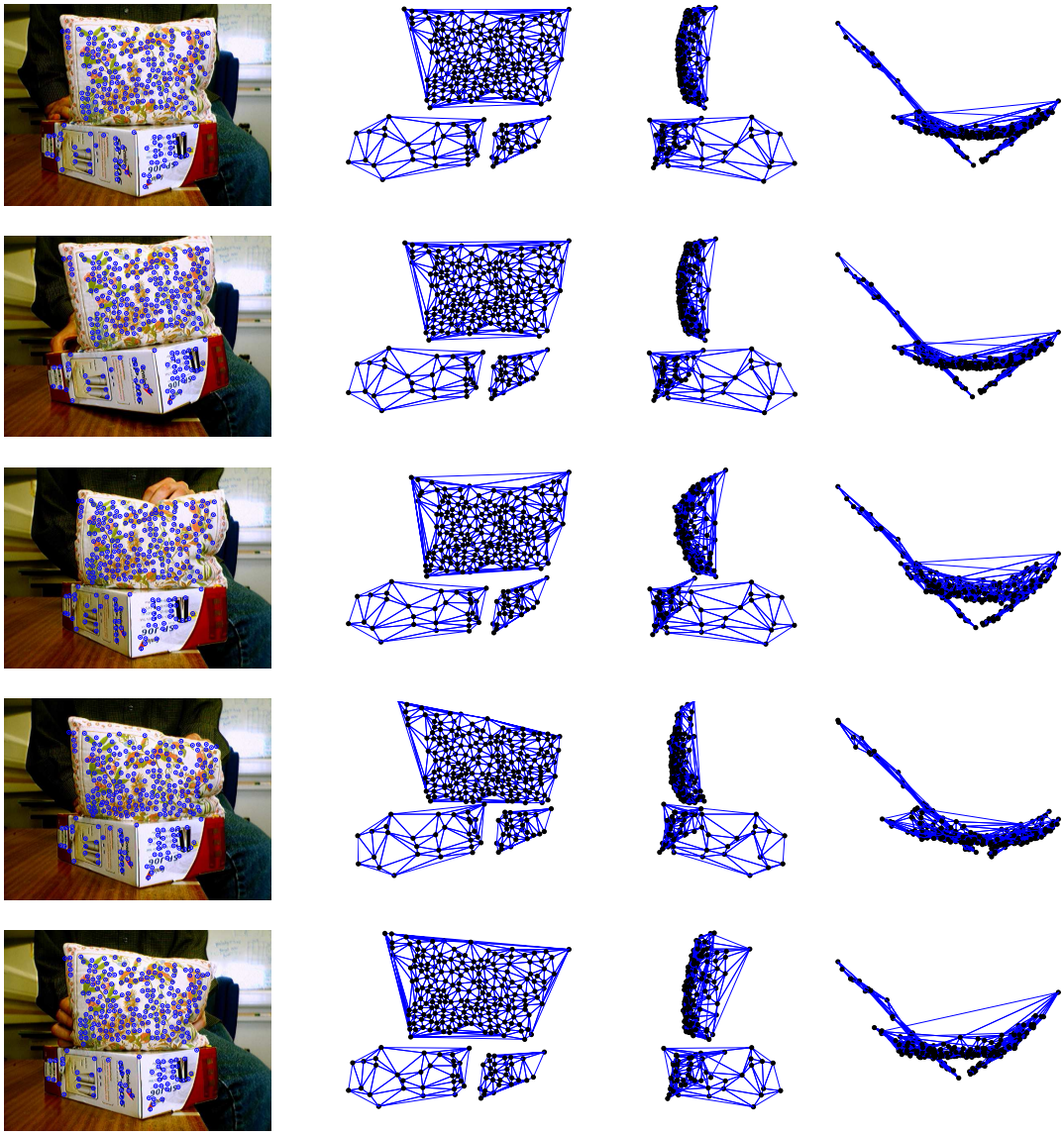
6.6 Closure

The proposed approach for the estimation of Euclidean non-rigid shape from a sequence of uncalibrated images takes advantage of an initial segmentation of the scene points into rigid and non-rigid from which self-calibration can be used to extract the metric rigid structure and the internal camera parameters. Then, a non-linear optimization stage globally solves and refines the estimates for the deformable components of the inspected object.

Motion segmentation of rigid and non-rigid points under perspective viewing conditions is required to define the priors for the specific object. The approach presented is based on a RANSAC technique whose convergence is aided using sampling priors over the complete set of trajectories in $\mathtt{W}$. The discriminant for separating the two classes of motion is given by the consistency of a set of trajectories with the epipolar geometry obtained by estimating fundamental matrices between pairs of view.

The construction of shape rigidity priors has a twofold effect. Firstly, estimating the internal camera parameters allows to upgrade the structure from projective to metric space. Secondly, as a computational aspect, the introduction of the priors in the non-linear optimization shows relevant increments in the convergence ratio to the global minimum. Xiao and Kanade's algorithm [161] – based on prior knowledge about the independency of the shape bases – performs well in the case

[2]Video available at http://www.bmva.ac.uk/thesis_archive/2006/DelBue1/index.html



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

Figure 6.9: Five key frames of the sequence with automatically tracked data for frames 1, 160, 340, 410 and 490. The first column (a) shows the tracked points (blue dots) lying over the rigid and non-rigid parts of the scene. Note the perspective distortion affecting the rigid box. The remaining columns shows front (b), side (c) and top (d) views of the cushion and box 3-D reconstructions.

of no noise but has a slow convergence ratio when Gaussian noise corrupts the image coordinates. Our non-linear minimization, on the other hand, converges fast (usually less than 30 iterations) regardless of the level of Gaussian noise in our synthetic experiments.

Xiao and Kanade's algorithm – based on prior knowledge about the independency of the shape bases – performs well in the case of no noise but has a slow convergence ratio when noise corrupts the image coordinates.

The experiments on synthetic and real data have shown firstly that even when using a minimal set of rigid points and when varying the intrinsic camera parameters it is possible to obtain reliable metric information and secondly that the shape priors are fundamental to avoid local minima given by ambiguous configurations of motion, perspective distortion and deformation. Notice that the method can successfully recover from situations in which a few points are misclassified as rigid even when the deformations are strong. The segmentation stage obtains reasonable results for the configuration of basis, cameras and points tested, however we noticed a higher misclassification ratio with weak perspective effects and higher proportion of non-rigid points. A further observation is that points that are semi-rigid (being rigid only for a part of the sequence), may appear undetected since they conform with the epipolar geometry only for a subset of frames.

Chapter 7

Conclusions

This thesis has dealt with different aspects of the problem of modelling deformable shapes from uncalibrated video sequences. We have reviewed and discussed the methods proposed so far in the literature and described their strengths and weaknesses. In the following sections we summarise our proposed solutions to some of the shortcomings of current methods and point out directions for future research and improvements to our framework.

7.1 Non-linear optimization for non-rigid structure from motion

Three dimensional reconstruction of deformable shapes is intrinsically a non-linear problem due to the fact that the parameters modelling the camera motion and the 3-D deformations are strongly coupled. The linear solutions proposed in the literature, which impose orthogonality constraints on the camera matrices, fail to provide accurate reconstructions. Recently, Xiao et. al. [159, 160] proved that the orthonormality constraints on the camera rotations are not sufficient to solve the ambiguities and they proposed a new set of constraints on the shape bases. Their work proves that when both sets of constraints are imposed, a closed-form solution to the problem of non-rigid structure from motion exists. However, their approach requires that each basis shape in the deformable model be observed independently in at least one view. Their method has been proved to break down with noisy data or when the number of basis shapes is not correctly estimated.

In this thesis we have proposed an alternative approach using a non-linear optimization scheme which preserves the correct geometric structure of the motion and structure matrices by minimizing a non-linear cost function which expresses the image reprojection error in the

model parameters. This minimization presents two main challenges. Firstly it is large-scale in essence since the number of parameters to estimate increases with the number of views and with the number of basis shapes. In this sense, a careful choice of the parameterization of the problem has proven to improve the results. Secondly the high non-linearity of the cost function introduces possible local minima which may prevent the algorithm from converging to the real solution. In order to render the minimization tractable, we have reformulated bundle-adjustment techniques, which take advantage of the sparseness of the jacobian matrix, to deal with the case of deforming objects.

7.2 Stereo non-rigid factorization

Given the same non-linear estimation framework, we have shown that it is possible to extend the method to include measurements from different cameras, to extract reliable 3-D reconstructions, and to compute the relative orientation of the cameras in an uncalibrated scenario. In this thesis we have concentrated on the stereo camera case. The use of two or more cameras is necessary when the object is only deforming since structure from motion methods require a significant component of rigid motion to obtain accurate depth estimates. Our experiments show that the reconstructions obtained with monocular views are of poor quality. As expected, including different camera views solves for the model parameters.

7.3 Non-rigid 3-D modelling using shape priors

A non-rigid object can be thought of as an underlying rigid body undergoing global rotations and translations while suffering some local non-rigid deformations. For this reason, non-rigid 3-D shape recovery is an inherently ambiguous problem. Given a specific rigid motion, different non-rigid shapes can be found that fit the measurement. To solve this ambiguity we propose to exploit prior knowledge on the 3-D structure such as the rigidity of some of the observed points. We have focused on the observation that often not all the points on a moving and deforming surface are undergoing non-rigid motion. Some of the points are frequently on rigid parts of the structure while others lie on deformable areas. Intuitively, if a segmentation is available, the rigid points can be used to estimate the overall rigid motion and to constrain the underlying mean shape by estimating the local deformations exclusively with the parameters associated to the non-rigid component of the 3-D model. We have showed that improved estimates can be achieved

when these priors are used. However, an algorithm able to perform automatic segmentation of rigid and non-rigid motion is required for our approach to be viable.

7.4 Motion segmentation of rigid/non-rigid points

Rigid and non-rigid motion segmentation is not a trivial task since rigid points can always be understood as non-rigid points which can be described by a single basis shape. However, we have shown that it is possible to separate rigid points in both the orthographic and the full perspective case by exploiting the constraints arising from the rigidity of the structure. In the case of orthographic cameras, we have introduced a segmentation method based on a feature selection strategy. Trajectories which are highly non-rigid are selected first and removed from the measurement matrix until W reaches the rank-3 condition that corresponds to the remaining trajectories moving as a rigid body.

In order to segment points in the projective case, we used a different property to disambiguate rigid and non-rigid trajectories: rigid trajectories give small residuals when used to estimate fundamental matrices between pairs of views. We have introduced a RANSAC method which randomly selects sets of trajectories until the best candidate is found. To aid the sampling procedure, we have proposed to assign a sampling prior given a measure of deformability of a point which increases the likelihood to select rigid trajectories.

7.5 Metric 3-D reconstruction of non-rigid shape from perspective images

In the case of perspective viewing conditions, once the scene has been segmented into rigid and non-rigid point sets, the rigid trajectories can be used to obtain an estimate of the mean rigid shape, the overall rigid motion and the camera calibration parameters (which allow to upgrade the structure to Euclidean space). This supporting rigid structure and motion can be used as the initial estimate for a non-linear estimation framework of the overall non-rigid structure where the non-rigid basis shapes and configuration weights are estimated as local variations from the mean rigid shape. The fact that image motion is a consequence of three different contributions: perspective distortions, rigid motion and local deformations is a source of possible ambiguities between the parameters of the model. However, we show that these ambiguities may be avoided by incorporating priors on the degree of deformability of each point in the minimization process. In particular, our expectation is that the rigid points will be fully described by the first basis

shape. These priors can be incorporated within a maximum a posteriori estimation framework.

7.6 Future work

One of the fundamental observations we have used in this thesis is the fact that it is often reasonable to assume that not all the points on the surface of a non-rigid object are deforming: while some of the points might be undergoing pure rigid motion others might deform at the same time. This constraint has proved very valuable both to provide priors on the degree of deformability of each point and to allow the computation of the metric upgrade transformation in the case of perspective viewing conditions. However, the assumption that a point is completely rigid throughout a long image sequence can become too restrictive. A class of trajectories we have not dealt with in this thesis is the class of points which have a semi-rigid behavior. Semi-rigid points are points that are rigid for some frames of the sequence but that occasionally deform with respect to the mean shape. These points could also provide valuable priors to be used in 3-D shape estimation while relaxing our assumptions. Notice, however, that the automatic segmentation algorithms described above would have to be modified to cope with the detection of points that have a mixed behaviour.

A further interesting aspect is the extension of our framework to deal with different types of non-rigid objects. In this thesis, we have restricted ourselves to the case of a single deforming object but often, in a generic and unknown video sequence, image tracks may belong to a structure with higher complexity. For instance, in the case of the human body, the face is obviously deformable but trajectories could also be extracted from the torso, arms and legs which are connected as articulated shapes. Image trajectories lying on different rigid, deforming and articulated parts would have to be associated to the correct model describing the inter-dependency of each body. The problem of associating (i.e. segmenting) the entries of $\mathtt{W}$ correctly to the appropriate object part would be extremely challenging, especially if the only information available are the image tracks taken from an uncalibrated sequence without any user-defined priors.

Of more direct practical use, a future avenue of research is the extension of the framework to deal with missing entries in the measurement matrix $\mathtt{W}$. It is a rather optimistic assumption to believe it is possible to identify the coordinates of all the feature points in all the views, particularly when dealing with long sequences. Besides, this poses restrictions on the types of object motion permitted: there cannot be so much rotation for instance that some of the features

go out of view. Note that this is not a problem for the non-linear minimization framework since if an entry $\mathbf{w}_{ij}$ in $\mathtt{W}$ is missing, the quadratic term referring to $\mathbf{w}_{ij}$ is not included in the cost function. However, the non-linear methods require initialisation which we perform using one of the linear approaches for which the complete measurement matrix is required. Additionally, our proposed motion segmentation algorithms require no missing entries in $\mathtt{W}$. A possible solution would be to adapt methods already designed for the rigid case to deal with deformable structure such as the sub-spaces technique of Sugaya and Kanatani [134].

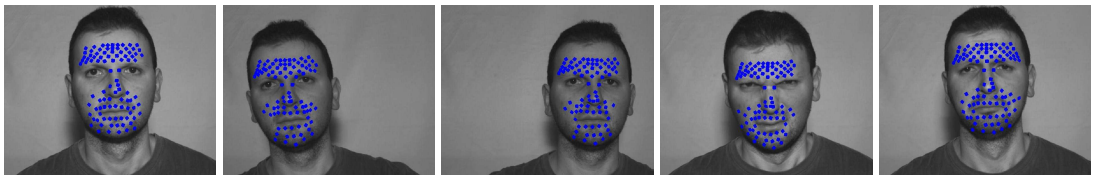


Figure 7.1: Tracking faces with deformable models. The methods presented in this thesis can be used to generate a 3-D deformable model which can then be used efficiently to track in real-time a face performing various expressions [108]. *Courtesy of E. Muñoz*.

7.7 Applications

From the applications point of view, we plan to exploit the generated geometric models in several computer vision systems. Our deformable models obtained automatically from an uncalibrated image sequence have already been shown to obtain promising results for face tracking [109, 108] (see figure 7.1 for an example). A new avenue to explore is their application to medical images. In this case the use of priors may help to model and register deformable biological shapes given the rigidity of some parts of the structure. For instance, in the case of diagnosis of heart conditions it would be possible to detect possible anomalies in the motion of the organ by having an accurate deformable model of a heart.

A real-time tracking algorithm which uses deformable models could be used to drive an avatar as demonstrated by Buenaposada et al. [21] and as shown in figure 7.2. In this case, the face was modelled as a set of 2-D planar patches. If the tracking algorithm can successfully describe the deformations appearing in the object, this in turn can be used directly to animate a synthetic object or 3-D avatar without requiring strong post processing efforts by the user. Moreover, the introduction of 3-D basis shapes to this scenario will ease the task of animating shapes with large

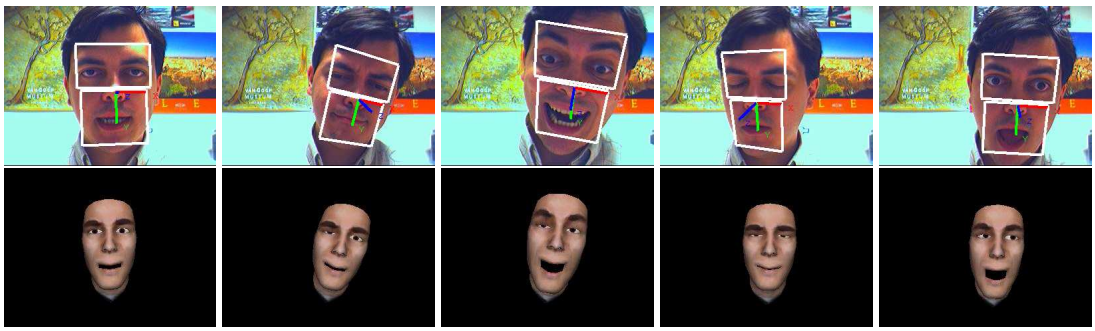


Figure 7.2: A system [21] for real time face tracking (upper row) and automatic animation of a synthetic 3-D face (lower row). The tracking system is based on a 2-D (planar) statistical description of the image appearance. The extracted parameters that describe the deformations in the image are used to pilot the animation of the 3-D face. *Courtesy of Dr J. M. Buenaposada*.

pose variations particularly when they suffer strong deformations.

Finally, our proposed techniques for rigid and non-rigid motion segmentation could be applied to cases in which the deviation of a set of object points from the overall rigid configuration is indicative of a harmful situation. For instance in the medical domain, the growth of a tumor could be detected when some of the points on the surface begin to behave as non-rigid.

Bibliography

[1] H. Aanæs, R. Fisker, K. Åström, and J. M. Carstensen. Robust factorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(9):1215–1225, 2002.

[2] H. Aanæs and F. Kahl. Estimation of deformable structure and motion. In *Workshop on Vision and Modelling of Dynamic Scenes, Copenhagen, Denmark*, 2002.

[3] L. Agapito, E. Hayman, and I. Reid. Self-calibration of rotating and zoooming cameras. *International Journal of Computer Vision*, 45(2):107–127, August 2001.

[4] A.A. Amini, R. Owen, L. Staib, P. Anandan, and J. Duncan. *Non-rigid motion models for tracking the left ventricular wall*. Lecture notes in computer science: Information processing in medical images. Springer-Verlag, 1991.

[5] K. B. Atkinson, editor. *Close Range Photogrammetry and Machine Vision*. Engineering and Science. Whittles Publishing, 1996.

[6] I. Y. Bar-Itzhack. New method for extracting the quaternion from a rotation matrix. *Journal of Guidance, Control and Dynamics*, 23(3):1085–1087, 2000.

[7] A. H. Barr. Superquadrics and angle-preserving transformations. *IEEE Computer Graphics and Applications*, 1(1):11–23, January 1981.

[8] A. Bartoli and P. Sturm. Nonlinear estimation of the fundamental matrix with minimal parameters. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(4):426–432, 2004.

[9] A. Baumberg and D. Hogg. Generating spatiotemporal models from examples. In *Proc. 6th British Machine Vision Conference, Birmingham*, volume 2, pages 413–422, 1995.

[10] R. Berthilsson, K. Åström, and A. Heyden. Reconstruction of general curves, using factorization and bundle adjustment. *International Journal of Computer Vision*, 41(3):171–182, 2001.

[11] A. Blake and M. Isard. *Active contours*. Springer-Verlag, 1998.

[12] V. Blanz and T. Vetter. A morphable model for the synthesis of 3d faces. In *SIGGRAPH '99: Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, pages 187–194, 1999.

[13] V. Blanz and T. Vetter. Face recognition based on fitting a 3d morphable model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(9):1063–1074, 2003.

[14] T. Boult and L. Brown. Factorization-based segmentation of motions. In *Proceedings of the Visual Motion Workshop*, pages 179–186, Princeton, NJ, November 1991.

[15] T. E. Boult and A. D. Gross. Recovery of superquadrics from depth information. In *Proc. of the 1987 Workshop on Spatial Reasoning and Multi-Sensor Fusion*, pages 128–137, St. Charles, IL, 1987.

[16] M. Brand. Morphable models from video. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition, Kauai, Hawaii*, volume 2, pages 456–463, December 2001.

[17] M. Brand. A direct method for 3d factorization of nonrigid motion observed in 2d. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition, San Diego, California*, pages 122–128, 2005.

[18] M. Brand and R. Bhotika. Flexible flow for 3d nonrigid tracking and shape recovery. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition, Kauai, Hawaii*, pages 315–22, December 2001.

[19] C. Bregler, A. Hertzmann, and H. Biermann. Recovering non-rigid 3d shape from image streams. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition, Hilton Head, South Carolina*, pages 690–696, June 2000.

[20] A. M. Buchanan and A. Fitzgibbon. Damped newton algorithms for matrix factorization with missing data. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition, San Diego, California*, volume 2, pages 316–322, 2005.

[21] José M. Buenaposada. *Facial expressions analysis using Computer Vision*. PhD thesis, Computer Science School, Technical University of Madrid, February 2005.

[22] L. D. Cohen. On active contour models and balloons. *CVGIP: Image Understanding*, 53(2):211–218, 1991.

[23] T. Cootes and C. J. Taylor. Combining elastic and statistical models of appearance variation. In *Proc. 6th European Conference on Computer Vision, Dublin, Ireland*, pages 149–163, May 2000.

[24] T. F. Cootes, G. J. Edwards, and C. J. Taylor. Active appearance models. In *Proc. 5th European Conference on Computer Vision, Freiburg, Germany*, volume 2, pages 484–498, 1998.

[25] T. F. Cootes and C. J. Taylor. Active shape models. In *Proc. British Machine Vision Conference*, pages 265–275, 1992.

[26] T. F. Cootes and C. J. Taylor. Active shape models – smart snakes. In *Proc. British Machine Vision Conference*, pages 266–275, 1992.

[27] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham. Active shape models — their training and application. *Computer Vision and Image Understanding*, 61(1):38–59, 1995.

[28] T. F. Cootes, C. J. Taylor, A. Lanitis, D. H. Cooper, and J. Graham. Building and using flexible models incorporating grey-level information. In *Proc. 4th International Conference on Computer Vision, Berlin*, pages 242–246, 1993.

[29] J. Costeira and T. Kanade. A multi-body factorization method for motion analysis. Technical Report CMU-CS-TR-94-220, Carnegie Mellon University, 1994.

[30] J. Costeira and T. Kanade. A multi-body factorization method for motion analysis. In *Proc. 5th International Conference on Computer Vision, Boston*, pages 1071–1076, 1995.

[31] I. Craw and P. Cameron. Parameterizing images for recognition and reconstruction. In *Proc. British Machine Vision Conference*, pages 367–370, 1991.

[32] D. DeCarlo and D. N. Metaxas. Optical flow constraints on deformable models with applications to face tracking. *International Journal of Computer Vision*, 38(2):99–127, 2000.

[33] A. Del Bue and L. Agapito. Non-rigid 3d shape recovery using stereo factorization. *Asian Conference of Computer Vision*, 1:25–30, January 2004.

[34] A. Del Bue and L. Agapito. Stereo non-rigid factorization. *International Journal of Computer Vision*, 66(2):193–207, February 2006.

[35] A. Del Bue, X. Lladó, and L. Agapito. Non-rigid face modelling using shape priors. In S. Gong W. Zhao and X. Tang, editors, *IEEE International Workshop on Analysis and Modelling of Faces and Gestures*, volume 3723 of *Lecture Notes in Computer Science*, pages 96–107. Springer-Verlag, 2005.

[36] A. Del Bue, X. Lladó, and L. Agapito. Non-rigid face modelling using shape priors. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition, New York, NY*, 2006. Accepted for publication.

[37] A. Del Bue, F. Smeraldi, and L. Agapito. Non-rigid structure from motion using non-parametric tracking and non-linear optimization. In *Proceedings of the 2004 Conference on Computer Vision and Pattern Recognition Workshop (CVPRW'04)*, Washington, DC, USA, 2004.

[38] A. Del Bue, F. Smeraldi, and L. Agapito. Non-rigid structure from motion using ranklet–based tracking and non-linear optimization. *Image and Vision Computing*, 2006. Accepted for publication.

[39] F. Dornaika and R. Chung. Stereo correspondence from motion correspondence. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition, Fort Collins, Colorado*, pages 70–75, 1999.

[40] L. Dorst. First order error propagation of the procrustes method for 3d attitude estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(2):221–229, February 2005.

[41] I. Essa and S. Basu. Modeling, tracking and interactive animation of facial expressions and head movements using input from video. In *Proceedings of Computer Animation Conference*, Geneva, Switzerland, June 1996.

[42] O. D. Faugeras. Stratification of 3-dimensional vision: Projective, affine, and metric representations. *Journal of the Optical Society of America*, 12(3):465–484, March 1995.

[43] O. D. Faugeras and R. Keriven. Complete dense stereovision using level set methods. In *Proc. 5th European Conference on Computer Vision, Freiburg, Germany*, volume 1, pages 379–393, 1998.

[44] O. D. Faugeras, Q. Luong, and S. Maybank. Camera self-calibration: Theory and experiments. In *Proc. European Conference on Computer Vision*, LNCS 588, pages 321–334, 1992.

[45] F. P. Ferrie, J. Lagarde, and P. Whaite. Darboux frames, snakes, and super-quadrics: Geometry from the bottom up. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(8):771–784, 1993.

[46] M. A. Fischler and R. C. Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. In M. A. Fischler and O. Firschein, editors, *Readings in Computer Vision: Issues, Problems, Principles, and Paradigms*, pages 726–740. Los Altos, CA., 1987.

[47] P. Fua. Regularized bundle-adjustment to model heads from image sequences without calibration data. *International Journal of Computer Vision*, 38(2):153–171, 2000.

[48] O. Gerard, A. C. Billon, J. M. Rouet, M. Jacob, M. Fradkin, and C. Allouche. Efficient model-based quantification of left ventricular function in 3-d echocardiography. *IEEE Transactions on Medical Imaging*, 21(9):1059–1068, September 2002.

[49] S. Burak Göktürk, J. Y. Bouguet, and R. Grzeszczuk. A data-driven model for monocular face tracking. In *Proc. 8th International Conference on Computer Vision, Vancouver, Canada*, volume 2, pages 701–708, 2001.

[50] D. B. Goldgof, H. Lee, and T. S. Huang. Motion analysis of nonrigid surfaces. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition, Los Alamitos*, pages 375–380, 1988.

[51] G. H. Golub and C. F. Van Loan. *Matrix computations*. Johns Hopkins, 1989.

[52] S. G. Gong, A. Psarrou, and S. Romdhani. Corresponding dinamic appearances. *Image and Vision Computing*, 20:289–300, 2002.

[53] A. Gruber and Y. Weiss. Factorization with uncertainty and missing data: Exploiting temporal coherence. In L. Saul S. Thrun and B. Schölkopf, editors, *Advances in Neural Information Processing Systems 16*. MIT Press, Cambridge, MA, 2004.

[54] A. Gruber and Y. Weiss. Multibody factorization with uncertainty and missing data using the em algorithm. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition, Washington D.C.*, volume 1, pages 707–714, 2004.

[55] N. Guilbert, F. Kahl, K. Åström, M. Oskarsson, M. Johansson, and A. Heyden. Constraint enforcement in structure and motion applied to closing an open sequence. volume 1, Jeju, South Korea, January 2004.

[56] A. Gupta and R. Bajcsy. Volumetric segmentation of range images of 3d objects using superquadric models. *CVGIP: Image Understanding*, 58(3):302–326, November 1993.

[57] M. Han and T. Kanade. Reconstruction of a scene with multiple linearly moving objects. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition, Hilton Head, South Carolina*, pages 542–549, June 2000.

[58] P. C. Hansen. Regularization, gsvd and truncated gsvd. *BIT*, 29(3):491–504, 1989.

[59] R. I. Hartley. Euclidean reconstruction from uncalibrated views. In *Proceedings of the Second Joint European - US Workshop on Applications of Invariance in Computer Vision*, pages 237–256, 1994.

[60] R. I. Hartley. Self-calibration from multiple views with a rotating camera. In *Proc. 3rd European Conference on Computer Vision, Stockholm*, volume 1, pages 471–478, 1994.

[61] R. I. Hartley. A linear method for reconstruction from lines and points. In *Proc. 7th International Conference on Computer Vision, Kerkyra, Greece*, pages 882–888, 1995.

[62] R. I. Hartley. In defense of the eight-point algorithm. *Proc. IEEE Conference on Computer Vision and Pattern Recognition, Puerto Rico*, 19(6):580–593, 1997.

[63] R. I. Hartley. Self-calibration of stationary cameras. *International Journal of Computer Vision*, 22(1):5–23, February 1997.

[64] R. I. Hartley. Self-calibration of stationary cameras. *International Journal of Computer Vision*, 22(1):5–23, 1997.

[65] R. I. Hartley, E. Hayman, L. Agapito, and I. D. Reid. Camera calibration and the search for infinity. In *Proc. 7th International Conference on Computer Vision, Kerkyra, Greece*, pages 510–517, 1999.

[66] R. I. Hartley and F. Schaffalitzky. Powerfactorization: an approach to affine reconstruction with missing and uncertain data. In *In Australia-Japan Advanced Workshop on Computer Vision*, Adelaide, Australia, September 2003.

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

[68] A. Heyden. Projective structure and motion from image sequences using subspace methods. In *Scandinavian Conference on Image Analysis*, pages 963–968, June 1997.

[69] A. Heyden and K. Åström. Algebraic varieties in multiple view geometry. In *Proc. 4th European Conference on Computer Vision, Cambridge*, volume 2, pages 671–682, 1996.

[70] A. Heyden and K. Åström. Euclidean reconstruction from image sequences with varying and unknown focal length and principal point. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 438–443, 1997.

[71] A. Heyden and K. Åström. Flexible calibration: Minimal cases for auto-calibration. In *Proc. 7th International Conference on Computer Vision, Kerkyra, Greece*, pages 350–355, 1999.

[72] K. Hiwada, A. Maki, and A. Nakashima. Mimicking video: real-time morphable 3d model fitting. In *VRST '03: Proceedings of the ACM symposium on Virtual reality software and technology*, pages 132–139, 2003.

[73] P. Ho and R. Chung. Stereo-motion with stereo and motion in complement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(2):215–220, February 2000.

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

[75] T. S. Huang. Modeling, analysis, and visualization of nonrigid object motion. In *Proc. International Conference on Pattern Recognition*, volume 1, pages 361–364, 1990.

[76] N. Ichimura. Motion segmentation based on factorization method and discriminant criterion. In *Proc. 1st International Conference on Computer Vision, London*, pages 600–605, 1991.

[77] M. Irani. Multi-frame optical flow estimation using subspace constraints. In *Proc. 7th International Conference on Computer Vision, Kerkyra, Greece*, pages 626–633, 1999.

[78] M. Irani and P. Anandan. Factorization with uncertainty. In *Proc. 6th European Conference on Computer Vision, Dublin, Ireland*, pages 539–553, 2000.

[79] G. Jacob, J. A. Noble, C. Behrenbruch, A. D. Kelion, and A. P. Banning. A shape-space-based approach to tracking myocardial borders and quantifying regional left-ventricular function applied in echocardiography. *IEEE Transaction on Medical Imaging*, 21(3):226–238, 2002.

[80] A. K. Jain and D. Zongker. Feature selection: Evaluation, application, and small sample performace. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(2):153–158, February 1997.

[81] C. Kambhamettu, D. B. Goldgof, D. Terzopoulos, and T. S. Huang. Nonrigid motion analysis. In *Handbook of Pattern Recognition and Image Processing: Computer Vision*, pages 405–430. Morgan Kaufmann Publishers, 1994.

[82] T. Kanade and D. Morris. Factorization methods for structure from motion. *Philosophical Transactions: Mathematical, Physical and Engineering Sciences*, 356(1740):1153–1173, 1998.

[83] K. Kanatani. Motion segmentation by subspace separation and model selection. In *Proceedings of the 8th International Conference on Computer Vision*, volume 2, pages 301–306, Vancouver, Canada, July 2001.

[84] K. Kanatani. Motion segmentation by subspace separation: Model selection and reliability evaluation. *International Journal of Image and Graphics*, 2(2):179–197, 2002.

[85] K. Kanatani and Y. Sugaya. Factorization without factorization: complete recipe. *Memories of the Faculty of Engineering, Okayama University*, 38(1–2):61–72, 2004.

[86] M. Kass, A. Witkin, and D. Terzopoulos. Snakes: Active contour models. *International Journal of Computer Vision*, 1(4):321–331, January 1988.

[87] T. Kim and H. Hong. On the reconstruction of approximate motion and average shape of deforming objects using a monocular view: application to the nonrigid human face. In *Proc. Asian Conference on Computer Vision*, volume 2, pages 1134–1139, Jeju, South Korea, January 2004.

[88] J. Kittler. Feature selection and extraction. In T. Y. Young and K. S. Fu, editors, *HPRIP*, pages 59–83, Orlando, FL, 1986. Academic Press.

[89] M. Knight, S. Roberts, D. Lee, and D. Bader. Live cell imaging using confocal microscopy induces intracellular calcium transients and cell death. *American Journal of Physiology - Cell Physiology*, 18:1083–1089, 2003.

[90] A. Lanitis, C. J. Taylor, and T. F. Cootes. Automatic face identification system using flexible appearance models. *Image and Vision Computing*, 13(5):393–401, June 1995.

[91] A. Lanitis, C. J. Taylor, and T. F. Cootes. Automatic interpretation and coding of face images using flexible models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):743–756, 1997.

[92] K. Levenberg. A method for the solution of certain problems in least squares. *The Quarterly of Applied Mathematics*, 2:164–168, 1944.

[93] X. Lladó, A. Del Bue, and L. Agapito. Euclidean reconstruction of deformable structure using a perspective camera with varying intrinsic parameters. In *To appear in Proc. International Conference on Pattern Recognition*, Hong Kong, 2006. Accepted for publication.

[94] H. C. Longuet-Higgins. A computer algorithm for reconstructing a scene from two projections. *Nature*, 293:133–135, September 1981.

[95] Q. T. Luong and O. D. Faugeras. The fundamental matrix: Theory, algorithms and stability analysis. *International Journal of Computer Vision*, 1(17):43–76, September 1996.

[96] S. Mahamud and M. Hebert. Iterative projective reconstruction from multiple views. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition, Hilton Head, South Carolina*, volume 2, pages 430–437, June 2000.

[97] R. Mandelbaum, G. Salgian, and H. Sawhney. Correlation-based estimation of ego-motion and structure from motion and stereo. In *Proc. 7th International Conference on Computer Vision, Kerkyra, Greece*, pages 544–550, 1999.

[98] R. A. Manning and C. R. Dyer. Stratified self calibration from screw-transform manifolds. In *Proc. 7th European Conference on Computer Vision, Copenhagen, Denmark*, pages 131–145, 2002.

[99] D. Marquardt. An algorithm for least-squares estimation of nonlinear parameters. *Journal of the Society for Industrial and Applied Mathematics*, 11(2):431–441, June 1963.

[100] I. Matthews and S. Baker. Active appearance models revisited. *International Journal of Computer Vision*, 60(2):135–164, November 2004.

[101] S. Maybank and O. D. Faugeras. A theory of self-calibration of a moving camera. *Int. J. Comput. Vision*, 8(2):123–151, 1992.

[102] P. F. McLauchlan and A. Jaenicke. Image mosaicing using sequential bundle adjustment. *Image and Vision Computing*, 20(9):751–759, 2002.

[103] D. Metaxas and S. J. Dickinson. Integration of quantitative and qualitative techniques for deformable model fitting from orthographic, perspective, and stereo projections. *Proc. 4th International Conference on Computer Vision, Berlin*, pages 641–649, May 1993.

[104] D. Metaxas and D. Terzopoulos. Constrained deformable superquadrics and nonrigid motion tracking. In *Proc. of the Computer Vision and Pattern Recognition Conference, Lahaina, HI*, pages 337–343, 1991.

[105] T. Moons, L. Van Gool, M. van Diest, and A. Oosterlinck. Affine structure from perspective image pairs under relative translations between object and camera. Technical Report KUL/ESAT/M12/9306, Departement Elektrotechniek, Katholieke Universiteit Leuven, Belgium, 1993.

[106] J. Moré. The Levenberg–Marquardt algorithm: Implementation and theory. *Numerical Analysis, Lecture Notes in Mathematics*, 630:105–116, 1977.

[107] T. Nagasaki, T. Kawashima, and Y. Aoki. Structure estimation of an articulated object by

motion image analysis based on factorization method. *Systems and Computers in Japan*, 32(10):69 – 79, August 2001.

[108] E. Mu noz, J. M. Buenaposada, and L. Baumela. Efficient model-based 3d tracking of deformable objects. In *iccv*, volume 1, pages 877–882. Beijing, China, October 2005.

[109] E. Mu noz, A. Del Bue, J. M. Buenaposada, L. Baumela, and L. Agapito. Automatic modelling and efficient tracking of deformable objects. In *IEE International Conference on Visual Information Engineering*, Glasgow (UK), April 2005.

[110] R. Oliveira, J. Costeira, and J. Xavier. Optimal point correspondence through the use of rank constraints. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition, San Diego, California*, volume 2, pages 1016–1021, 2005.

[111] S. Osher and J. A. Sethian. Fronts propagating with curvature-dependent speed: Algorithms based on Hamilton-Jacobi formulations. *Journal of Computational Physics*, 79:12–49, 1988.

[112] N. K. Paragios and R. Deriche. A pde-based level-set approach for detection and tracking of moving objects. In *Proc. 6th International Conference on Computer Vision, Bombay, India*, pages 1139–1146, 1998.

[113] F. I. Parke and K. Waters. *Computer facial animation*. A. K. Peters, Ltd., 1996.

[114] A. P. Pentland. Perceptual organization and the representation of natural form. *Artificial Intelligence*, 28(3):293–331, 1986.

[115] F. Pighin, J. Hecker, D. Lischinski, R. Szeliski, and D. H. Salesin. Synthesising realistic facial expressions from photographs. In *Proceedings of the ACM SIGGRAPH Conference on Computer Graphics*, 1998.

[116] F. H. Pighin, R. Szeliski, and D. Salesin. Resynthesizing facial animation through 3d model-based tracking. In *Proc. 7th International Conference on Computer Vision, Kerkyra, Greece*, pages 143–150, 1999.

[117] C. J. Poelman and T. Kanade. A paraperspective factorization method for shape and motion recovery. In *Proc. 3rd European Conference on Computer Vision, Stockholm*, volume 2, pages 97–108, 1994.

[118] M. Pollefeys. *Self-Calibration and Metric 3D Reconstruction from Uncalibrated Image Sequences*. PhD thesis, Katholieke Universiteit Leuven, Belgium, 1999.

[119] M. Pollefeys, R. Koch, and L. Van Gool. Self calibration and metric reconstruction in spite of varying and unknown internal camera parameters. In *Proc. 6th International Conference on Computer Vision, Bombay, India*, pages 90–96, 1998.

[120] M. Pollefeys, R. Koch, and L. Van Gool. Self-calibration and metric reconstruction in spite of varying and unknown internal camera parameters. *International Journal of Computer Vision*, 32(1):7–25, 1999.

[121] M. Pollefeys and L. Van Gool. A stratified approach to metric self-calibration. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition, Puerto Rico*, pages 407–414, 1997.

[122] M. Pollefeys and L. Van Gool. Stratified self-calibration with the modulus constraint. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(8):707–724, 1999.

[123] S. Ranganath. Contour extraction from cardiac MRI studies using snakes. *IEEE Transactions on Medical Imaging*, 14(2):328, June 1995.

[124] D. Reynard, A.P. Wildenberg, A. Blake, and J. Marchant. Learning dynamics of complex motions from image sequences. In *Proc. 4th European Conference on Computer Vision, Cambridge*, pages 357–368, Cambridge, England, April 1996.

[125] R. Roussel and A. Gagalowicz. A hierarchical face behavior model for a 3d face tracking without markers. In *11th International Conference on Computer Analysis of Images and Patterns (CAIP)*, pages 854–861, September 2005.

[126] A. Roy-Chowdhury. A measure of deformability of shapes with applications to human motion analysis. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition, San Diego, California*, volume 1, pages 398–404, June 2005.

[127] P. H. Schonemann. A generalized solution of the orthogonal procrustes problem. *Psychometrika*, 31:1–10, 1966.

[128] J. Shi and C. Tomasi. Good features to track. *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 593–600, 1994.

[129] F. Smeraldi, A. Del Bue, and L. Agapito. Tracking points on deformable objects with ranklets. In *IEEE International Conference in Image Processing*, volume 3, pages 121–124, Genoa, Italy, September 2005.

[130] M. Spetsakis and J. Aloimonos. Structure from motion using line correspondences. *International Journal of Computer Vision*, 4(3):171–183, 1990.

[131] G. Stein and A. Shashua. Direct estimation of motion and extended scene structure from a moving stereo rig. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition, Santa Barbara*, pages 211–218, 1998.

[132] P. Sturm and B. Triggs. A factorization based algorithm for multi-image projective structure and motion. In *Proc. 4th European Conference on Computer Vision, Cambridge*, pages 709–720, April 1996.

[133] Y. Sugaya and K. Kanatani. Outlier removal for motion tracking by subspace separation. *IEICE Transactions on Information and Systems*, E86-D(6):1095–1102, 2003.

[134] Y. Sugaya and K. Kanatani. Extending interrupted feature point tracking for 3-d affine reconstruction. *IEICE Transactions on Information and Systems*, E87-D(4):1031–1039, 2004.

[135] Y. Sugaya and K. Kanatani. Multi-stage optimization for multi-body motion segmentation. *IEICE Transactions on Information and Systems*, E87-D(7):1935–1942, 2004.

[136] R. Szeliski and D. Terzopoulos. Physically-based and probabilistic modeling for computer vision. *In SPIE, Geometric Methods in Computer Vision*, 1570:140–152, July 1991.

[137] W. K. Tang and Y. S. Hung. A factorization-based method for projective reconstruction with minimization of 2-d reprojection errors. In *Proc. of the 24th DAGM Symposium on Pattern Recognition*, pages 387–394, London, UK, 2002. Springer-Verlag.

[138] C. Tomasi and T. Kanade. Shape and motion from image streams: a factorization method - part 3 detection and tracking of point features. Technical Report CMU-CS-91-132, Computer Science Department, Carnegie Mellon University, Pittsburgh, PA, April 1991.

[139] C. Tomasi and T. Kanade. Shape and motion from image streams under orthography: A factorization approach. *International Journal of Computer Vision*, 9(2):137–154, 1992.

[140] B. J. Tordoff and D. W. Murray. Guided-MLESAC: Faster image transform estimation by using matching priors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(10):1523–1535, October 2005.

[141] L. Torresani, D. Yang, E. Alexander, and C. Bregler. Tracking and modeling non-rigid objects with rank constraints. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition, Kauai, Hawaii*, 2001.

[142] P. Tresadern and I. Reid. Synchronizing image sequences of non-rigid objects. In *Proc. British Machine Vision Conference*, Norwich, 2003.

[143] P. Tresadern and I. Reid. Articulated structure from motion by factorization. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition, San Diego, California*, volume 2, pages 1110–1115, June 2005.

[144] B. Triggs. Matching constraints and the joint image. In *Proc. 5th International Conference on Computer Vision, Boston*, pages 338–343, 1995.

[145] B. Triggs. Factorization methods for projective structure and motion. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition, San Francisco*, pages 845–851, 1996.

[146] B. Triggs. Auto-calibration and the absolute quadric. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition, Puerto Rico*, pages 609–614, 1997.

[147] B. Triggs, P. McLauchlan, R. I. Hartley, and A. Fitzgibbon. Bundle adjustment – A modern synthesis. In W. Triggs, A. Zisserman, and R. Szeliski, editors, *Vision Algorithms: Theory and Practice*, LNCS, pages 298–375. Springer Verlag, 2000.

[148] E. Trucco and A. Verri. *Introductory Techniques for 3-D Computer Vision*. Prentice-Hall, 1998.

[149] T. Ueshiba and F. Tomita. A factorization method for projective and euclidean reconstruction from multiple perspective views via iterative depth estimation. In *Proc. 5th European Conference on Computer Vision, Freiburg, Germany*, volume 1, pages 296–310, 1998.

[150] S. Ullman. *The Interpretation of Visual Motion*. MIT Press, 1979.

[151] T. Vetter and V. Blanz. A morphable model for the synthesis of 3d faces. In *Proceedings of the ACM SIGGRAPH Conference on Computer Graphics*, pages 187–194, 1999.

[152] R. Vidal and R. I. Hartley. Motion segmentation with missing data using powerfactorization and gpca. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition, San Diego, California*, volume 2, pages 310–316, June 2004.

[153] R. Vidal, Y. Ma, and J. Piazzi. A new gpca algorithm for clustering subspaces by fitting, differentiating and dividing polynomials. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition, Washington D.C.*, volume 2, pages 510–517, June 2004.

[154] A. Waxman and J. Duncan. Binocular image flows: steps toward stereo-motion fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 8(6):715–729, 1986.

[155] Y. Weiss. Segmentation using eigenvectors: a unifying view. In *Proc. 7th International Conference on Computer Vision, Kerkyra, Greece*, pages 975–982, 1999.

[156] J. Weng, T. Huang, and N. Ahuja. Motion and structure from line correspondences; closed-form solution, uniqueness, and optimization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(3):318–336, 1992.

[157] H. Wold. Nonlinear estimation by iterative least squares procedures. *Research Papers in Statistics*, 630:411–444, 1966.

[158] Y. Wu, Z. Zhang, T. S. Huang, and J. Y. Lin. Multibody grouping via orthogonal subspace decomposition. In *In Proc. IEEE Conf. Computer Vision and Pattern Recognition*, volume 2, pages 252–257, Kauai, Hawaii, December 2001.

[159] J. Xiao, J. Chai, and T. Kanade. A closed-form solution to non-rigid shape and motion recovery. In *Proc. 8th European Conference on Computer Vision, Copenhaguen, Denmark*, pages 573–587, May 2004.

[160] J. Xiao and T. Kanade. Non-rigid shape and motion recovery: Degenerate deformations. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition, Washington D.C.*, pages 668–675, 2004.

[161] J. Xiao and T. Kanade. Uncalibrated perspective reconstruction of deformable structures. In *Proc. 10th International Conference on Computer Vision, Beijing, China*, October 2005.

[162] J. Yan and M. Pollefeys. Articulated motion segmentation using ransac with priors. *ICCV Workshop on Dynamical Vision*, 2005.

[163] J. Yan and M. Pollefeys. A factorization-based approach to articulated motion recovery. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition, San Diego, California*, volume 2, pages 815–821, June 2005.

[164] A. J. Yezzi and S. Soatto. Deformotion: Deforming motion, shape average and the joint registration and approximation of structures in images. *International Journal of Computer Vision*, 53(2):153–167, 2003.

[165] A. A. Young, D. L. Kraitchman, and L. Axel. Deformable models for tagged MR images: Reconstruction of two- and three-dimensional heart wall motion. In *In Proc. IEEE Workshop on Biomedical Image Analysis*, pages 317–323, June 1994.

[166] L. Zelnik-Manor and M. Irani. Degeneracies, dependencies and their implications in multi-body and multi-sequence factorizations. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, pages 287–293, June 2003.

[167] H. K. Zhao, S. Osher, and R. Fedkiw. Fast surface reconstruction using the level set method. In *Proceedings of the IEEE Workshop on Variational and Level Set Methods (VLSM'01)*, pages 194–202, 2001.