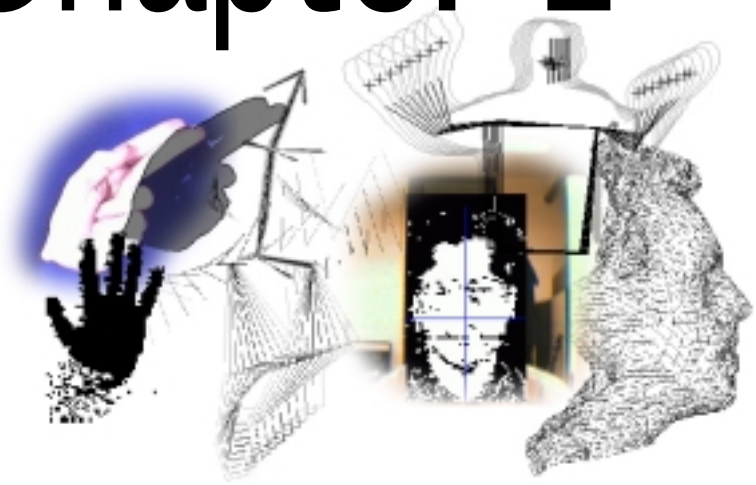
Chapter 1



1 Introduction

The term *Computer Vision* covers a broad field of research encompassing many techniques, applications and disciplines but is commonly summarised as

"the science of making a computer see...."

However, the goal is often to allow the computer to understand what it sees to some extent, and it is here that the science embraces aspects of artificial intelligence. This artificial understanding, or interpretation, of a scene stems from human perception and our attempt to mimic the functionality of the human visual system. It is natural to attempt to emulate the way in which humans perceive or interpret the world and this approach has been instrumental throughout the course of vision research, with developments such as foveal vision systems and stereoscopic depth reconstruction. The most fundamental of such approaches is that of model based vision.

The image plane of a camera is akin to the retina of the eye, and images projected onto it are the 2D projection of the 3D world. This loss of information presents no obstacle for the human brain which interprets the image seamlessly, constantly updating its model of the world. The ability to judge depth through the

disparity of objects falling upon the retinas provides essential clues to the brain about the structure of the real world. However, even when this stereoscopic information is unavailable, the human brain can still interpret the scene and accurately estimate the position and orientation of objects. This is due to the huge knowledge base the brain accumulates about the 3D world, its laws, and the shape and structure of objects and how they project onto the retina.

If the human brain can achieve such feats for millions of objects, then the rationale of providing a similar knowledge of a small subset of objects to a computer is an obvious solution. This is the premise of *model based vision*, where an internal representation of the world or object is provided to a computer allowing it to locate, recognise, track or interact with real world objects. This *a priori* knowledge about objects can be encapsulated and represented in numerous ways.

Probably the simplest form of *model based vision* is that of template matching [Ballard 82]. Given a known object or feature to be located in an image, a template, representing object features, is applied to the image at every location. By formulating template matching with a scoring mechanism, the fit of the model at any location can be assessed and the probable position of objects or features estimated. Although a relatively time consuming approach, template matching algorithms can provide effective object location for constrained applications and have proven invaluable in areas such as industrial inspection. Hardware implementations are commonplace allowing large numbers of templates to be matched in real time.

Industrial inspection has proven a successful application of real time vision systems as the nature of the problems is typically heavily constrained. If the application of biscuits on a conveyor belt is considered, the problem of object location is greatly simplified by the process and nature of the object. The production line produces only biscuits, so the variability of shape is heavily reduced. Biscuits are typically flat and as such can be assumed to be 2D objects, which adhere to ground plane constraints. In addition, lighting inconsistencies and background clutter can be controlled and modelled accurately. Given a

ridgid internal model of an object, probable locations can be identified within the image by matching the features of the object with the extracted features of an image (such as edges or corners). This is often applied as a hypothesise and test procedure, where possible locations of an object are generated and compared to the image. Each hypothesis is then assessed using some metric where the highest scoring hypotheses correspond to the likely location of objects. As more complex objects are considered, techniques such as geometric hashing [Wolfson 92] can be used to allow affine object transformations. However, when real world objects and less constrained environments are considered these tools are insufficient at modelling object variability.

The problems of recognition are compounded when everyday, unconstrained objects are considered. In addition to the variability of lighting, shading and complex scenes containing cluttered backgrounds, even ridgid 3D objects will produce considerably differing views depending upon their position and orientation. Consider a book. The shape of the book projected onto the image frame will vary immensely as its orientation changes. More complex still is the goal of building a generic model of a book where the 3D shape parameters of the object vary immensely between examples. A common solution to this problem is to represent the object in terms of its 3D structure and use the 2D projection of the internal model to match with the 2D projection of the real world object.

Models that bend or articulate introduce further complexity to the task of object recognition and tracking. In addition to the object variation described above, articulated objects also produce variability of shape and structure in the image. Many researchers have tackled this by extending the 3D internal model to that of articulated geometric primitives with tight joint constraints, which closely mimic the movement of the real world object. However, as these types of models are typically hand-coded they do not offer a generic solution that can be applied to all objects.

Deformable objects which can alter their shape to fit an object under some global shape constraints overcome these problems by encapsulating a large amount of an object’s variability into a constrained deformation of a contour or object. By

learning this deformation from a training set of example shapes, they produce a set of tools which allow models to be easily constructed for any number of objects under a multitude of situations.

This thesis is concerned with the construction of generic models of deformation and their application to the recognition and tracking of complex 3D objects. Chapter 2 will present a review of relevant literature to the work and discuss the shortfalls of current formulations. Chapter 3 will introduce linear Point Distribution Models and describe the Active Shape Model approach to object tracking. Chapter 4 will discuss the use of colour in image segmentation and feature extraction. Chapter 5 will present a non-linear approximation technique based upon a piecewise linear model. Chapter 6 will extend the piecewise linear approach to more complex, high dimensional training sets and demonstrate the use of such models in the classification of American Sign Language. Chapter 7 will discuss the addition of temporal constraints. Using motion capture as an example it is shown how time dependent deformation can be both learnt and reproduced from a model. Its is further shown how these temporal constraints can be used to support multiple hypotheses during tracking. Chapter 8 discusses the extension of PDMs into the 3D domain. Chapter 9 presents a new approach to markerless based motion capture which incorporates many of the previously discussed elements to allow the 3D pose and motion of a human body to be extracted from a monoscopic image sequence. Finally a discussion and conclusions are presented.

This manuscript also contains two appendices. Appendix 1 presents the k-means and fuzzy k-means (FCM) algorithms along with associated techniques. Appendix 2 presents a new approach to the surface segmentation of volumetric data. Although this work is extremely relevant to 3D PDM construction it stands as an individual piece of research and hence is consigned to the appendices.