Flexible 3D Models from Uncalibrated Cameras

T.F.Cootes, E.C. Di Mauro, C.J.Taylor, A.Lanitis
Department of Medical Biophysics,
University of Manchester
Manchester M13 9PT
email: bim@sv1.smb.man.ac.uk

Abstract

We describe how to build statistically–based flexible models of the 3D structure of variable objects, given a training set of uncalibrated images. We assume that for each example object there are two labelled images taken from different viewpoints. From each image pair a 3D structure can be reconstructed, up to either an affine or projective transformation, depending on which camera model is used. The reconstructions are aligned by choosing the transformations which minimise the distances between matched points across the training set. A statistical analysis results in an estimate of the mean structure of the training examples and a compact parameterised model of the variability in shape across the training set. Experiments have been performed using pinhole and affine camera models. Results are presented for both synthetic data and real images.

1 Introduction

In many vision problems we study objects which are either deformable in themselves (such as faces) or can be considered examples of a class of variable shapes (such as cars). In order to represent them we must model this variability explicitly. However, the only information we have about the 3D objects is often that contained in one or more 2D images. To understand their 3D shape we must perform some form of reconstruction. When we have accurate camera calibration data we can obtain metric information and recover the euclidean structure of the objects. Unfortunately camera calibration is not robust, is easily lost (by moving or refocusing the camera) and can be difficult to obtain. Recent work has shown that it is possible to obtain relative 3D structure (up to an affine or projective transformation) from uncalibrated images [4–9]. In this paper we combine this work on structure from uncalibrated cameras with methods of building statistical models of shape [1]. We demonstrate how to build flexible models of three dimensional structure from uncalibrated images of examples.

We assume that we have a training set of objects from the class we wish to model, and that for each we have a pair of arbitrary views. From each pair we can reconstruct the structure of the example up to either a projective or an affine transformation, depending upon which camera model is assumed. We give examples for both pinhole and affine camera models. The set of reconstructed examples are aligned into a com-

BMVC 1995 doi:10.5244/C.9.15

mon reference frame by finding the affine or projective transform for each which minimises the total variance across the training set. The mean shape and a compact parameterised model of shape variation are obtained by performing a statistical analysis in this frame for the set of reconstructions.
We show results for synthetic images of cars in which the modes of variation are known, and for real images of a vice and human faces. We discuss practical applications of the method.

2 Background

Cootes *et al* [1] describe 2D statistical models of variable image structure. These models are generated by manually locating landmark points on the structures of interest in each of a set of training images, aligning these training shapes into a common reference frame then performing a statistical analysis to obtain the mean shape and main modes of shape variation. They show how these Point Distribution Models (PDMs) can be used in image search [1,2] by creating Active Shape Models (ASMs). An ASM is analogous to a 'snake' in that it refines its position and shape under the influence of image evidence, giving robust object location.
Hill *et al* [11] show how the PDM/ASM approach can be extended to 3D when volume or range images are available, for example in medical imaging. A review of other deformable models is given in [1]. Shen and Hogg [3] have recently shown how a fairly coarse flexible 3D model can be generated from a set of image sequences.
There is a well established literature on methods of extracting 3D structures from two or more 2D images. Because camera calibration is often inconvenient, and in any case non–robust, recent work has focused on what can be learnt about scene structure from uncalibrated images [6]. Developments in projective geometry [10,12] have led to various constructions which are invariant to camera parameters and pose [16]. Hartley *et al* [4,5], Mohr *et al* [7] and Faugeras [6] have described methods of reconstructing the positions of 3D points up to a projective transformation of 3D space, given their projection into two uncalibrated images.
An alternative approach is to assume an affine camera model, an approximation acceptable when the distance to the subject is large compared to the subject's depth. In this case the Factorisation Method of Tomasi and Kanade [8,9] can be used to reconstruct structure up to an affine transformation of 3D space. This is robust and works well on noisy data from real images.
Most of the work on recovering 3D structure has assumed that the objects viewed are rigid. Sparr describes a framework for dealing with objects which deform in ways which are locally affine [15]. Blake *et al* [13] and Beardesly *et al* [14] have also developed geometric models which allow affine deformations, for tracking objects in image sequences.

3 Overview – Flexible Models from 2D Images

We assume that we have a training set of paired uncalibrated images of objects of interest, in which landmark points representing key points on the objects have been located. Suppose we have N such pairs, each containing n landmark points.
To build a flexible model from this data we must perform three steps :

i) Reconstruct 3D structure (up to an affine or projective transformation) from each paired set of image points.
ii) Align the sets of reconstructed points, applying affine or projective transformations to minimise the distances between equivalent points across the training set - this defines a reference frame for the model.
iii) Apply a Principal Component Analysis to the reconstructed 3D data in the reference frame, to generate the mean shape and main modes of shape variation.

This is analogous to the method of building 2D Point Distribution Models [1], the main difference being in the alignment stage. In the 2D case alignment involves choosing the rotation, translation and scale for each example which minimises the mean squared deviation of the aligned set of shapes. In this case we must choose the most suitable projective (15 parameter) or affine (12 parameter) transformation to align the examples.

4 Reconstructing 3D structure

Given two uncalibrated projective views of a structure represented by a set of points, it is possible to compute the relative positions of the points in three dimensions up to an arbitrary (unknown) affine or projective transformation of 3-space depending on the camera model used. We have investigated the use of both projective and affine camera models.

For a projective camera it is possible to reconstruct structure up to a projective transformation of 3D space [4,6]. Faugeras [6], Hartley *et al* [4,5] and Mohr *et al* [7] all describe reconstruction algorithms.

For an affine camera we can reconstruct the structure up to an affine transformation of 3D space using the factorisation (SVD) method of Tomasi and Kanade [8]. This was developed for shape and motion recovery from image sequences, and gives a robust method of recovering the structure given two or more images of an object. Although the original method assumed an orthographic projection model (later extended to use a paraperspective camera model [9]) it is able to generate structure up to affine transformation if an uncalibrated affine camera model is used.

5 Aligning into a Reference Frame

After reconstruction the sets of points can be considered to each lie in an arbitrary reference frame. Before we can apply any statistical analysis, we must move them all into an appropriate co-ordinate frame; for the statistics to be valid this would ideally be a euclidean co-ordinate frame. This step is analogous to the alignment step used when building Point Distribution Models in 2D (Cootes *et al* [1]). In that case shapes are presented at arbitrary positions with arbitrary orientations and scales. Before a mean shape can be calculated it is necessary to align each example so that they have consistent centres, orientations and scales. This is achieved by minimising the sum of square distances between points after transformation, and applying a constraint to the mean shape. The mean can be constrained by aligning with a set of reference points, usually one of the original examples. We use a similar method for the reconstructed 3D shapes, generalising it to allow affine and projective transformations during the alignment.

The general alignment algorithm has the following steps :

i) Apply a transformation to each set of points to minimise the distance to a reference set

ii) REPEAT :
- Calculate the mean structure
- Align the mean with a reference set
- Re-align each set of points to minimise the distance to the mean

UNTIL change sufficiently small.

The reference set serves two purposes. The first is to ensure that the alignment algorithm converges. Without constraining the mean by re-aligning it with a reference set, the system is underdetermined - for instance the shapes can shrink to a point. We could use any of the reconstructed sets of points as the reference set. Its second purpose is to define a suitable reference frame for shapes in which to perform a statistical analysis of the point positions (see below). After alignment all the examples can be considered to be in the frame of the reference set. Ideally the reference set should be the true euclidean positions of some of the points. If the reference set is severely distorted compared to the true structure, for instance if it is much shorter in one dimension, the statistics of the models will be biased. However, since we are building deformable models, which will only be valid up to an affine or projective transformation, an approximate set of reference points is adequate for most situations.

5.1 Projective Case

In the projective case the alignment consists of calculating the projective transformation which minimises the distance (measured in euclidean space) between original and target points. The transformation has 15 degrees of freedom. An initial estimate can be obtained using the method described in Appendix A. Where necessary this can be optimised using a non-linear optimisation on the elements of the projection matrix to minimise the euclidean distance between points.

Since a projective transformation of 3D space is constrained by knowing 5 point positions, the reference set must define at least 5 points in their approximate euclidean positions. These are only required in this training phase, and in most situations this is fairly easy to do, as we usually know the approximate shape of the objects we are studying.

5.2 Affine Case

Suppose we have two sets of n matched points $\{\mathbf{x}_i = (x_{i1}, x_{i2}, x_{i3}, 1)^T\}$ and $\{\mathbf{y}_i = (y_{i1}, y_{i2}, y_{i3}, 1)^T\}$ which differ by an unknown affine transformation $\mathbf{A}$, ie $\mathbf{y}_i = \mathbf{A}\mathbf{x}_i \ \ (i = 1..n)$.

Let $\mathbf{a}_j^T$ be the jth row of $\mathbf{A}$, (with $\mathbf{a}_4^T = (0\ 0\ 0\ 1)$). Then the first three rows can be obtained by a least squares solution to the n linear equations

$$\mathbf{x}_i^T \mathbf{a}_j = y_{ij}$$

An affine transformation can be constrained by determining 4 points in 3D space, so we could supply the approximate positions of 4 or more points for the reference set. However, unless our cameras have seriously non-square pixels, the factorisation method itself can give a good approximation of the true structure up to scaling [8,9]. Thus we can use one of the reconstructed examples as a reference set.

6 Building a Flexible Model

We build a 3D model using the method of Hill *et al* [11]. We represent a set of 3D points $\{ \mathbf{x}_i = (x_i, y_i, z_i, 1)^T \ (i=1..n) \}$ as a single $3n$ element vector

$$\mathbf{X} = (x_1, \ldots x_n, y_1, \ldots, y_n, z_1, \ldots z_n)^T$$

Thus the set of N reconstructed objects in the reference frame are given by the N 3n element vectors $\{ \mathbf{X}_j \ (j=1..N) \}$. We can calculate the mean of these, $\hat{\mathbf{X}}$, and apply a Principal Component Analysis (PCA) to generate a set of modes of variation. These modes are given by the unit eigenvectors of the covariance matrix of $\mathbf{X}_j$ which correspond to the t largest eigenvalues (See [1] for more details).

A linear model of the landmark points of the objects in the training set is given by

$$\mathbf{X} = \hat{\mathbf{X}} + \Phi \mathbf{b}$$

where Φ is a $3n$ x t matrix of eigenvectors representing the modes,
$\mathbf{b}$ is a t element vector of *shape parameters*.

Limits on the shape parameters can be obtained from the statistics of the training set (typically we use limits of about 3 standard deviations).

We have built models using both synthetic data and real images. We found, however, that our current implementations of the projective reconstruction and alignment are too sensitive to noise to build sensible models from real images. The affine case proved considerably more robust. We describe experiments and give quantitative results using both synthetic and real images.

7 Experiments Using Synthetic Data

As a synthetic example we generated sets of 16 3D points representing the vertices of a car. Figure 1 shows a typical set and indicates the dimensions which we allowed to vary at random. On average the car was 30 units long, 10 units wide and 9 units high. When projected into an image (using a pinhole camera model) we arranged for its projection to be about 200 pixels wide. For this synthetic example we did not hide occluded points - we assumed a wire frame model in which all the points could be located.

7.1 Results for Projective Model

We generated 20 random car structures and projected the points of each into 2 images. In the noise free case reconstructions were perfect up to a projectivity. We aligned the reconstructions as described above, using the known true mean structure as the reference object and built a statistical model of the variations. Figure 2 shows the most significant mode of variation of the model. This modifies the rear of the car, changing it from a saloon to an estate. In addition there is a tapering caused by the projective transformations allowed in the alignment procedure. Figure 3 shows

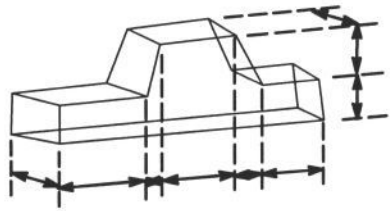


Figure 1 : Synthetic car showing data showing dimensions which vary in the training set

the second mode of variation, a change in the relative height of the bonnet and roof, again with some tapering effects.

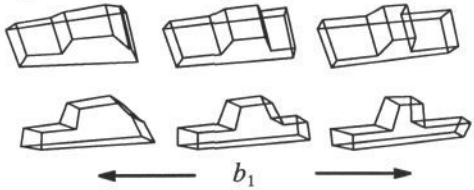


Figure 2 : Effect of varying the first shape parameter (saloon to estate + some tapering)

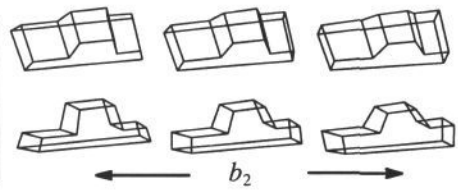


Figure 3 : Effect of varying second shape parameter (relative height change)

7.2 Results for Affine Model

We used similar synthetic data and built models using the affine camera approximation. This proved to be more robust to noise, and we were able to perform quantitative experiments to characterise its performance [18]. The quality of the reconstruction depends on the positional noise in the 2D images, the distance of the cameras from the object relative to the object depth and the angle between the cameras. Figure 4 shows the first three modes of shape variation ($z = 500$, $\theta = 45$) of the model reconstructed from noise free data. There is a small amount of shearing caused by the affine transformation allowed in the alignment phase, but most of the variation is that present in the training set.

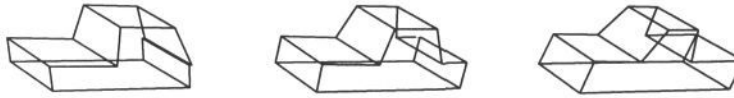
8 Experiments Using Real Data

8.1 Engineers Vice Results

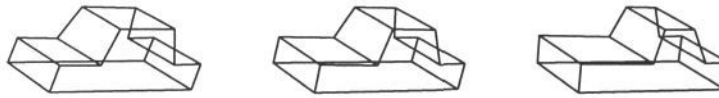
We took 7 pairs of images of a vice with different jaw openings (Figures 5,6). On each we marked 19 landmark points. For these experiments we only used points which were visible in all images. We reconstructed assuming an affine camera and trained a statistical model. The reconstruction from the first shape was used as a reference set for the alignment. The model has only one main mode of variation which represents 91% of the total variance of the training set. This mode, which is illustrated in Figure 7, opens and closes the jaws of the vice – the only degree of freedom affecting the shape present in the training images. Subsequent modes model the noise in the training set.

8.2 Face Image Results

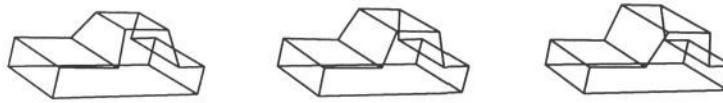
As part of another project [17] we had available images of faces from various individuals with their heads held in different poses (Figure 8). We selected a pair of images



Mode 1 : hatchback < > saloon



Mode 2 : Front bonnet length change



Mode 3: Width of roof change

$b_i = -2\sigma_i$ ← *mean* → $b_i = +2\sigma_i$

Figure 4 : First three modes of shape variation (noise free case), Shape parameters varied between +/- 2 s.d.s observed in training set.

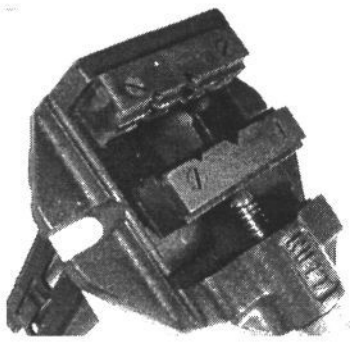


Figure 5 : Example of image of vice used for training a model.

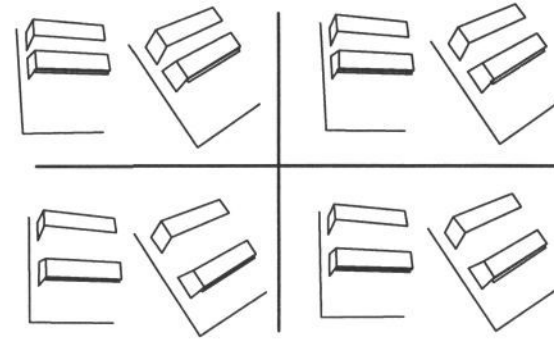


Figure 6 : Examples of shapes used for training

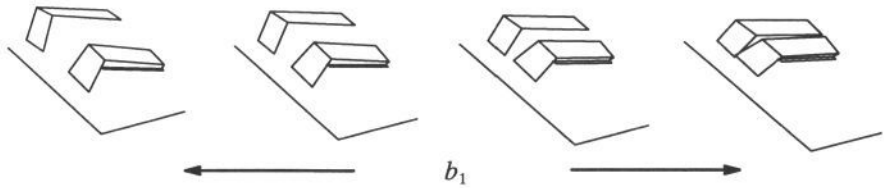


Figure 7 : Most significant mode of vice model shape variation – opening of the jaws.

for each of 12 people, and used an automatic method to locate 144 landmarks on each [17]. We then used these sets of landmarks to build a 3D model, assuming an affine camera model and using the first reconstruction as a reference. Since each subject held their head in a different pose in each of the image pairs, we effectively had two views of the same structure. Figure 9 shows the mean of the reconstructed model, and the effect of varying the first shape parameter. The method has successfully reconstructed the structure, and the shape variation gives the main variation in face structure and expression between individuals in the training set. There is, of course, some noise caused by the errors in the landmark locations and the changes in express-

ion between images (they were taken at different times by a single camera), but the overall structure and shape variation is plausible.

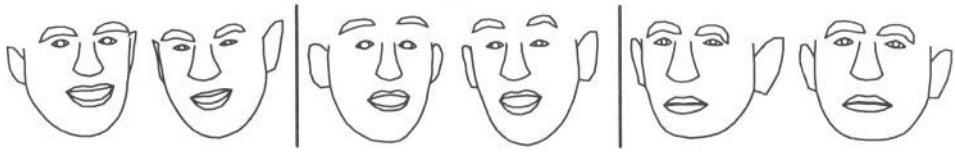


Figure 8 : Examples of pairs of training shapes for face model

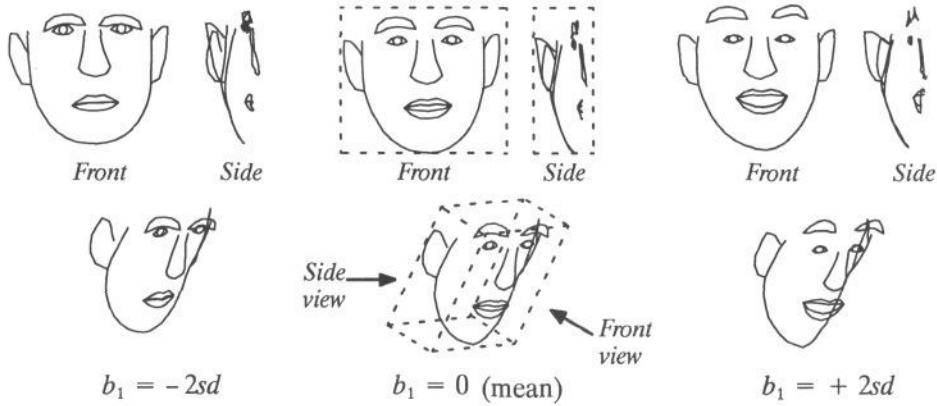


Figure 9 : Different views of the reconstruction of the mean and first mode of shape variation of a face model (smiling mode).

9: Discussion and Conclusions

We have demonstrated that 3D flexible models can be generated from pairs of uncalibrated images. Two camera models have been used. Model building using a projective camera model relies upon a good reconstruction of structure from two uncalibrated images, which can be hard to achieve. For more robust reconstructions many images are required, such as from a sequence.

The factorisation method gives a far more reliable reconstruction, but requires an affine camera model. This is acceptable when the object is far from the camera relative to its depth, but can cause distortions when the object is nearer.

The examples we have given assume all points are visible in all images. If some points are occluded in some image pairs, we can reformulate the methods to allow for this. Weights can be assigned to each point in each example, 1 if it is present in both images of a pair, 0 if not. We can reconstruct from the points, perform a weighted alignment and apply a weighted PCA to obtain the model. As long as there is sufficient overlap in the visible points across different pairs, we should be able to obtain a complete model. Thus we could build up a full model of a 3D object by adding together pairs of views from different orientations. In the affine reconstruction case the full structure of a single object can be recovered from multiple views, each with some occluded points, using the SVD method [8,9].

A long term goal is to construct a flexible 3D model given a training set of *single* uncalibrated images of different examples of a class of objects, taken from arbitrary viewing positions. This would require estimating the mean structure, its allowed variations and the projections required into each image to minimise the errors, an extension of

current optimisation based reconstruction methods. This is likely to be a difficult optimisation problem.

Elsewhere we demonstrate that we can estimate the model shape parameters given the 2D point positions in a new view of a modelled object [18]. The ability to estimate shape parameters from new images will allow classification (such as into different types of car) or certain measurements to be made.

In addition this will allow us to implement a local search strategy similar to Active Shape Models [1]. An initial estimate of the projected model point positions in an image will be refined by locating better estimates nearby in the image and updating the projection and shape parameters accordingly. The approach has proved to be a fast and robust method of finding instances of 2D Point Distribution Models in images, and we anticipate an analogous method will allow us to locate projections of variable 3D objects.

Appendix A : Recovering the Projective Transformation between Sets of Points

Suppose we have two sets of matched points $\{ \mathbf{x}_i = (x_{i1}, x_{i2}, x_{i3}, 1)^T \}$ and $\{ \mathbf{y}_i = (y_{i1}, y_{i2}, y_{i3}, 1)^T \}$ which differ by an unknown projective transformation $\mathbf{H}$, ie $\omega_i \mathbf{y}_i = \mathbf{H}\mathbf{x}_i \quad (i = 1..n)$. We can recover $\mathbf{H}$ as follows :

For the i'th point we have

$$\begin{aligned} \omega_i y_{i1} &= \mathbf{h}_1^T \mathbf{x}_i \\ \omega_i y_{i2} &= \mathbf{h}_2^T \mathbf{x}_i \\ \omega_i y_{i3} &= \mathbf{h}_3^T \mathbf{x}_i \\ \omega_i &= \mathbf{h}_4^T \mathbf{x}_i \end{aligned} \qquad \text{where} \quad \mathbf{H} = \begin{pmatrix} \mathbf{h}_1^T \\ \mathbf{h}_2^T \\ \mathbf{h}_3^T \\ \mathbf{h}_4^T \end{pmatrix}$$

Substituting in for ω_i and assuming that $h_{44} = 1.0$ gives 3 linear constraints on the other 15 elements of $\mathbf{H}$:

$$\mathbf{h}_j^T \mathbf{x}_i^T - y_{ij}(h_{j1}x_{i1} + h_{j2}x_{i2} + h_{j3}x_{i3}) = -y_{ij} \qquad j = 1, 2, 3$$

Thus we have a total of $3n$ equations which can be solved using least squares approaches if necessary. If we have only $m < n$ known matches (if some points are occluded for instance) then we have only $3m$ equations, but a solution can still be obtained if $m \geq 5$.

Note that this method is not directly minimising the sum of euclidean distance errors, but gives an approximation acceptable for our purposes. More accurate estimates require a non-linear optimisation procedure.

Acknowledgements

Tim Cootes is funded by an EPSRC Postdoctoral Fellowship. Enea Di Mauro is funded by an EPSRC Project Grant. Andreas Lanitis is funded by a University of Manchester Research Studentship and an ORS award.

References

[1] T.F.Cootes, C.J.Taylor, D.H.Cooper and J.Graham, Active Shape Models - Their Training and Application.*Computer Vision and Image Understanding* Vol. 61, No. 1, 1995. pp.38-59.

[2] T.F.Cootes , C.J.Taylor, A.Lanitis, Active Shape Models : Evaluation of a Multi-Resolution Method for Improving Image Search, *in Proc. British Machine Vision Conference, (Ed. E.Hancock) BMVA Press* 1994, pp.327-338.

[3] X.Shen, D.Hogg, 3D Shape Recovery Using a Deformable Model. *in Proc. British Machine Vision Conference, (Ed. E.Hancock) BMVA Press* 1994, pp.387-396.

[4] R.Hartley, R.Gupta and T.Chang, Stereo from Uncalibrated Cameras. in *Proc. CVPR'92* IEEE Press, 1992, pp. 761-764.

[5] R.I.Hartley, Projective Reconstruction and Invariants from Multiple Images. *IEEE PAMI* Vol.16, No. 10, 1994, pp. 1036-1041.

[6] O.Faugeras, What can be seen in three dimensions with an uncalibrated stereo rig?, *Proc. European Conference on Computer Vision.* 1992, pp. 563-578

[7] R.Mohr, B.Boubakeur, P.Brand, Accurate Projective Reconstruction. in [16]. pp.257-275.

[8] C.Tomasi, T.Kanade, Shape and Motion from Image Streams under Orthography: a Factorization Method. *IJCV* 9 (Vol.2), pp. 137-154, 1992.

[9] C.J.Poelman, T.Kanade, A Paraperspective Factorization Method for Shape and Motion Recovery. *Proc. ECCV 1994. Lecture Notes in Computer Science, Vol.801,* (ed. J.O. Eklundh), pp. 97-108.

[10] O.Faugeras, Three Dimensional Computer Vision, A Geometric Viewpoint, MIT Press, 1993

[11] A.Hill , A.Thornham, C.J.Taylor. Model Based Interpretation of 3D Medical Images. in *Proc. British Machine Vision Conference* 1993. Vol.2. (Ed. J.Illingworth) BMVA Press. pp. 339-348.

[12] K.Kanatani, Geometric Computation for Machine Vision, Oxford University Press, 1993

[13] A.Blake, R.Curwen, A.Zisserman, Affine-invariant contour tracking with automatic control of spatiotemporal scale. in *Proc. Fourth International Conference on Computer Vision,* IEEE Computer Society Press, 1993, pp.66-75.

[14] P.A.Beardsley, A.Zisserman and D.W.Murray, Navigation using Affine Structure from Motion. *Proc. ECCV 1994 (Vol.2).* Lecture Notes in Computer Science 801, ed. J.O.Eklundh, Springer-Verlag, 1994, pp. 85-96.

[15] G.Sparr, A Common Framework for Kinetic Depth Reconstruction and Motion for Deformable Objects. *Proc. ECCV 1994 (Vol.2).* Lecture Notes in Computer Science 801, ed. J.O.Eklundh, Springer-Verlag, 1994, pp. 471-482.

[16] J.L.Mundy, A.Zisserman, D.Forsyth (Eds.), Applications of Invariance in Computer Vision, Lecture Notes in Computer Science 825, Springer-Verlag, 1993.

[17] A.Lanitis, C.J.Taylor, T.F.Cootes. A Unified Approach to Coding and Interpreting Face Images. *Proc. ICCV 1995.* pp.368-373.

[18] T.F.Cootes. Building Flexible 3D Shape Models from Uncalibrated Camera Images. *Internal Report, Dept. Medical Biophysics, Manchester University, England. March 1995.*