Downloaded from rsta.royalsocietypublishing.org

Geometric motion segmentation and model selection

The Royal Society

Phil. Trans. R. Soc. Lond. A 1998 **356**, 1321-1340
doi: 10.1098/rsta.1998.0224

Email alerting service Receive free email alerts when new articles cite this article - sign up in the box at the top right-hand corner of the article or click **here**

To subscribe to *Phil. Trans. R. Soc. Lond. A* go to: **http://rsta.royalsocietypublishing.org/subscriptions**

This journal is © 1998 The Royal Society

Downloaded from rsta.royalsocietypublishing.org

The Royal Society

Geometric motion segmentation and model selection

By P. H. S. Torr[†]

Robotics Research Group, Department of Engineering Science, University of Oxford, Jenkin Building, 17 Parks Road, Oxford OX1 3PJ, UK

Motion segmentation involves clustering features together that belong to independently moving objects. The image features on each of these objects conform to one of several putative motion models, but the number and type of motion is unknown *a priori*. In order to cluster these features, the problems of model selection, robust estimation and clustering must all be addressed simultaneously. Within this paper I place the three problems into a common statistical framework; investigating the use of information criteria and robust mixture models as a principled way for motion segmentation of images. The final result is a general fully automatic algorithm for clustering that works in the presence of noise and outliers.

Keywords: robust estimation; grouping; epipolar geometry; matching; clustering; degeneracy detection

1. Introduction

Motion is a powerful cue for image and scene segmentation in the human visual system. This is evidenced by the ease with which we see otherwise perfectly camouflaged creatures as soon as they move, and by the strong cohesion perceived when even disparate parts of the image move in a way that could be interpreted in terms of a rigid motion in the scene. Detection of independently moving objects is an essential but often neglected precursor to problems in computer vision.

In robotic vision, motion segmentation turns out to be a most demanding problem and has received considerable attention over the years, a review of which may be found in Torr (1995). Many previous approaches to motion clustering have failed because the motion models that they employ are too restrictive. For instance, if one tries to group purely on similarity of image velocities, then any stream of images from a static scene viewed by a camera undergoing cyclotorsion would be incorrectly segmented. Schemes based on linear variation of the motion flow field will produce false segmentations at depth discontinuities when the camera is translating (Jepson & Black 1993; Black & Anandan 1996). Segmentation under the assumption of orthographic or weak perspective imaging conditions will fragment scenes with strong perspective effects, even if no independent motion is present. Some methods require *a priori* knowledge of camera calibration and motion, this may not always be available. Thus the need for a more general framework is apparent.

The work in this paper stems from the desire to develop a general motion segmentation and grouping algorithm. That is, given two or more views of a scene, we may

[†] Present address: Microsoft Corporation, One Microsoft Way, Redmond, WA 98052-6399, USA (philtorr@microsoft.com).

© 1998 The Royal Society

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

Downloaded from rsta.royalsocietypublishing.org

determine any of the objects within the scene which change their relative dispositions. The motion of both the camera and the object are presumed unknown, as is the camera calibration. Consider figure 3*a*,*b*. A soldier is tracked by a rotating camera, causing apparent motion to the left in the image. The soldier strides, giving the semblance of an upward motion. Given this image pair as input, the ultimate goal would be to identify that the soldier has moved independently of the background. This is a prodigious undertaking and the computational theory and algorithms given in this paper represent analysis of certain geometrical and statistical aspects of the problem. Before developing an algorithm, several key design issues have to be answered: (i) What data primitives should be used to represent the scene? (ii) What decision rule should be used to group the primitives chosen? (iii) Having arrived at (i) and (ii), what algorithm should be used to solve the problem? The first problem is one of data reduction, the second of geometry and the third lies in the domain of computational theory. The main contribution of this paper is in the latter two areas. Returning to the first question, a major hindrance to the analysis of motion across an image is the vast amount of data to be managed. Corner features (Harris & Stephens 1987) are most amenable to geometric and statistical analysis, which is the flavour of this work. Furthermore corner features indicate pixels where both components of image motion might be recovered with reasonable accuracy; providing a strong constraint on the motion model. Unfortunately they only give a sparse representation, and so the models that are estimated from the corner features are used to flesh out the description of the segmentation. With the primitives chosen, a decision rule has to be developed to determine the segmentation. Many previous segmentation algorithms have failed to exploit the geometric reality of the world, which is readily available from image sequences. The approach espoused in this paper is to adopt a decision rule that segments projected features in accordance with the constraints imposed by the assumption that they are rigidly connected in the world. The method of segmentation follows an information theoretic approach using the tools of mixture models and information criteria. Thus within this paper we present a new paradigm that automatically determines (i) how many motions there are in the scene (ii) what type of motion model is appropriate for each motion (iii) the parameters and consistent data for each motion model. Methods that have combined mixtures and model selection include Darrell *et al*. (1990) and Gu *et al*. (1996), the layered representation of Ayer & Sawhney (1995), and in the related field of fitting surfaces to range data (Mirza & Boyer 1992; Leonardis *et al*. 1990), but none of these methods supports completely general models that may be of differing *dimensions*, as will be seen below.

The structure of the paper is as follows. In § 4 several common two-view relations are introduced; being the putative motion models. Maximum likelihood estimation of these models is explained in § 3, showing that the optimal segmentation may be written in terms of a mixture model. The problem is that the use of just maximum likelihood estimation will always lead to the most general model being selected as most likely. Any of the putative motion models may be appropriate for a given set of image correspondences undergoing rigid motion, the question is how to decide which is the best. In § 4 the AIC criterion is introduced. It provides a method of scoring the competing models fitted to the data. But it is not robust and cannot be used to compare data lying on manifolds of different dimensions. Thus Kanatani's (1996) application of the AIC to compare manifolds of differing dimensions and robust AICs (see Hampel *et al*. 1986) are introduced. In this paper these last two ideas are

Downloaded from rsta.royalsocietypublishing.org

combined to produce a robust version of the AIC: GRIC that deals with data from varieties of differing dimension. It is then shown how these can be used to establish the correct posterior distributions for the data to establish a clustering. Finally in § 6 these ideas are all drawn together to give a motion segmentation algorithm, and results are presented on real images. In § 7 some of the deficiencies of the algorithm are discussed, which opens up some future possible avenues of research.

Notation. Noise free (true data) will be denoted by an underscore $\underline{x}$, estimates $\hat{x}$, image points $\boldsymbol{x}$, image matches $\boldsymbol{m}$; the probability density function (PDF) of x given y is $p(x \mid y)$.

2. Putative motion models

Within this section some of the motion models used are described, a complete taxonomy is given in Torr *et al.* (1998). Suppose that the viewed features arise from a three-dimensional (3D) object which has undergone a rotation and non-zero translation. After the motion, the set of homogeneous image points $\{\boldsymbol{x}_i\}, i = 1, \dots , n$, is transformed to the set $\{\boldsymbol{x}'_i\}$, where $\boldsymbol{x}_i = (x_i, y_i, \zeta)^{\mathrm{T}}$, and $\boldsymbol{x}'_i = (x'_i, y'_i, \zeta)^{\mathrm{T}}$. The two sets of features are related by $\boldsymbol{x}'^{\mathrm{T}}_i \boldsymbol{F} \boldsymbol{x}_i = \boldsymbol{0}$, where $\boldsymbol{F}$ is the rank 2, 3×3 fundamental matrix. The fundamental matrix encapsulates the epipolar geometry. It contains all the information on camera motion and intrinsic parameters available from image feature correspondences alone.

When there is degeneracy in the data such that we cannot obtain a unique solution for $\boldsymbol{F}$, it is desirable to use a simpler motion model. For small independently moving objects, there may be an insufficient spread of features to enable a unique estimate of the fundamental matrix. Within this paper for brevity I shall consider the detection of just three other models, the affine camera model (Mundy & Zisserman 1992), with linear fundamental matrix $\boldsymbol{F}_A$, and image projectivities and affinities as induced by planar homographies: $\boldsymbol{x}' = \boldsymbol{B}\boldsymbol{x}$. However, it will be seen that the proposed scoring function is completely general; and has been implemented for a wider range of models. The properties of the models are investigated in greater detail in § 4, and the equations of the constraints are summarized in table 3.

3. Maximum likelihood estimation

In the following we make the assumption that the noise in the two images is Gaussian on each image coordinate with zero mean and uniform standard deviation σ (extension to the more general case is not difficult and is described by Kanatani (1996)). Thus given a true correspondence $\underline{\boldsymbol{m}}$ the probability density function of the noise-perturbed data is

$$p(\boldsymbol{m} \mid \mathcal{R}, \underline{\boldsymbol{m}}) = \prod_i \left(\frac{1}{\sqrt{2\pi}\sigma} \right)^4 \mathrm{e}^{-((\underline{x}_i - x_i)^2 + (\underline{y}_i - y_i)^2 + (\underline{x}'_i - x'_i)^2 + (\underline{y}'_i - y'_i)^2)/(2\sigma^2)}, \qquad (3.1)$$

where $\mathcal{R}$ is the appropriate two-view relation, e.g. fundamental matrix or projectivity. The negative log likelihood of all the correspondences $\boldsymbol{m}_i$, $i = 1, \dots , n$, where n is

Downloaded from rsta.royalsocietypublishing.org

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

the number of correspondences, is

$$L = \sum_i \log(\Pr(\boldsymbol{m}_i \mid \mathcal{R})) \tag{3.2}$$

$$= \sum_i ((\underline{x}_i - x_i)^2 + (\underline{y}_i - y_i)^2 + (\underline{x}'_i - x'_i)^2 + (\underline{y}'_i - y'_i)^2), \tag{3.3}$$

discounting the constant term.

Given two views with a known relation; for each correspondence $\boldsymbol{m}$ the task becomes that of finding the maximum likelihood estimate $\hat{\boldsymbol{m}}$ of the true match $\underline{\boldsymbol{m}}$, such that $\hat{\boldsymbol{m}}$ satisfies the relation and minimizes the negative log likelihood

$$l_i^2 = \sum_j ((\hat{x}_i - x_i)^2 + (\hat{y}_i - y_i)^2 + (\hat{x}'_i - x'_i)^2 + (\hat{y}'_i - y'_i)^2). \tag{3.4}$$

If the type of relation $\mathcal{R}$ is known then, observing the data, we can estimate the parameters of $\mathcal{R}$ to minimize this log likelihood. This inference is called 'maximum likelihood estimation' (Fisher 1936). Thus $L = \sum_i l_i^2$ provides the error function for the point matches, and $\mathcal{R}$ for which L is a minimum is the maximum likelihood estimate of the relation (fundamental matrix, or projectivity). The optimally estimated correspondence $\hat{\boldsymbol{m}}$ and its error l may be obtained as the solution of a high order polynomial equation. A computationally efficient first order approximation to these is given in Torr & Zisserman (1997).

If the type of relation $\mathcal{R}$ is unknown then we cannot use maximum likelihood estimation to decide the form of $\mathcal{R}$, as the most general model will always be most likely, i.e. have lowest L. Fisher was aware of the limitations of maximum likelihood estimation and admits the possibility of a wider form of inductive argument that would determine the functional form of the data (Fisher 1936, p. 250); but then goes on to state, 'At present it is only important to make clear that no such theory has been established'. Some suggestion for this wider form of inductive argument are given in § 4; before this the maximum likelihood solution for multiple motion case is considered using a mixture model.

(*a*) *Optimal clustering*

The initial set of matches Z is obtained by cross correlation as described in Beardsley *et al*. (1996). The problem is to optimally group them into sets consistent with the different motion models. This involves finding the most probable underlying interpretation ϕ, being a classification into several motion models together with the parameters of those models. The most likely partition is obtained by maximizing the probability of the interpretation given the data: $\max_\phi \Pr[\phi \mid Z]$, which may be rewritten using Bayes theorem:

$$\Pr[\phi \mid Z] = \frac{\Pr[Z \mid \phi]\Pr[\phi]}{\Pr[Z]}. \tag{3.5}$$

Thus clustering is defined as a Bayesian decision process that minimizes the Bayes risk incurred in choosing a partition of Z. As $\Pr[Z]$ does not depend on ϕ it may be dropped from the exposition, and the problem becomes that of finding

$$\max_\phi \Pr[Z \mid \phi]\Pr[\phi].$$

Downloaded from rsta.royalsocietypublishing.org

A partition ϕ is a set of s clusters $\kappa_j \subset Z$. Each cluster is defined as a motion model $\mathcal{R}_j$ together with a set of matches κ_j such that each match arises from one cluster, $\kappa_1 \cup \kappa_2 \cdots \cup \kappa_s = Z$, and only one cluster, $\kappa_i \cap \kappa_j = 0$. One cluster κ_s is designated as the cluster containing all mismatches and outliers; matches that do not belong to any of the other clusters are assigned to this. Note that the number of clusters may vary from interpretation to interpretation.

Each match $\boldsymbol{m}$ is modelled by a prior mixture model, such that $\boldsymbol{m}$ belongs to one of the $s-1$ clusters with probabilities $\pi_1, \ldots \pi_{s-1}$ such that $\sum_{j=1}^{j=s-1} \pi_j = 1$ and $\pi_j \geqslant 0$. Adopting the notation of McLachlan & Basford (1988), the PDF of any match $\boldsymbol{m}$ given interpretation $\phi = ((\pi_1 \ldots \pi_{s-1}); (\mathcal{R}_1 \ldots \mathcal{R}_{s-1}))^{\mathrm{T}}$ is given by the finite mixture form,

$$p(\boldsymbol{m} \mid \phi) = \sum_{j=1}^{j=s} \pi_j p_j(\boldsymbol{m} \mid \mathcal{R}_j), \tag{3.6}$$

where $p_j(\boldsymbol{m} \mid \mathcal{R}_j)$ is the PDF of the match given $\mathcal{R}_j$.

Once the clusters have been estimated (a method of initialization is given in § 6), estimates of the posterior probabilities of population membership are formed for each match $\boldsymbol{m}_i$ based on the estimated $\hat{\phi}$. The posterior probability $\tau_{ij}(\boldsymbol{m} \mid \phi)$ is given by

$$\tau_{ij}(\boldsymbol{m} \mid \phi) = \Pr[\boldsymbol{m}_i \in \mathcal{R}_j \mid \boldsymbol{m}; \phi] = \pi_j p_j(\boldsymbol{m}_i \mid \mathcal{R}_j) \Big/ \sum_{k=1}^{k=s-1} \pi_k p_k(\boldsymbol{m}_i \mid \mathcal{R}_k). \tag{3.7}$$

A partitioning of $\boldsymbol{m}_1 \ldots \boldsymbol{m}_n$ into s non-overlapping clusters is effected by assigning each $\boldsymbol{m}_j$ to the population to which it has the highest estimated posterior probability $\tau_{ij}(\boldsymbol{m} \mid \mathcal{R}_j)$ of belonging. That is $\boldsymbol{m}$ is assigned to cluster $\mathcal{R}_t$ if

$$\tau_{it}(\boldsymbol{m} \mid \hat{\phi}) > \tau_{ij}(\boldsymbol{m} \mid \hat{\phi}) \qquad (j = 1 \ldots s-1; j \neq t). \tag{3.8}$$

To make this procedure robust a threshold must be made on $p_j(\boldsymbol{m}_i \mid \mathcal{R}_j)$ in order to eliminate outliers. If $-\log(p_j(\boldsymbol{m}_i \mid \mathcal{R}_j)) > l_o$ for $j = 1 \ldots s-1$ then the match is redesignated an outlier. This threshold is arrived at in the next section.

To calculate the complete log likelihood $L_C = \log(\Pr[Z \mid \phi])$ account must be taken of the fact that each match can only belong to one cluster. For this purpose, for each match an s-dimensional vector of unknown indicator variables $\boldsymbol{c}_i = (c_{1i} \ldots c_{si})$ is introduced, where

$$c_{ij} = \begin{cases} 1 & \boldsymbol{m}_i \in \kappa_j, \\ 0 & \boldsymbol{m}_i \notin \kappa_j. \end{cases} \tag{3.9}$$

Then following standard texts (McLachlan & Basford 1988), the complete log likelihood is given by

$$L_C = \sum_{i=1}^{i=n} \sum_{j=1}^{j=s} c_{ij}(\log \pi_j + \log(p(\boldsymbol{m}_i \mid \mathcal{R}_j))). \tag{3.10}$$

This may be maximized by treating the c_{ij} as missing data from the mixture model and using the EM algorithm (Dempster *et al.* 1977), or some suitable gradient descent algorithm.

Thus far there has been no discussion of either how to determine the type of motion model appropriate for each cluster, or the number of such clusters; these topics are

Downloaded from rsta.royalsocietypublishing.org

Table 1. *Mean SSE*

(Mean SSE for 100 matches over 100 trials. Variance of noise on the coordinates: $\sigma^2 = 1$, together with GIC values in parentheses.)

	point motion		
estimated	general	orthographic	rotation
fundamental	114 (728)	92 (706)	92 (706)
affine	399 (1007)	96 (694)	143 (751)
projectivity	493 (909)	452 (868)	195 (611)

related to the final term to be considered in Bayes formula: the prior $\Pr[\phi]$. In order to go about estimating this, the province of model selection must be entered, which is considered in the next section.

4. Model selection

Robotic vision has its basis in geometric modelling of the world, and many vision algorithms attempt to estimate these geometric models from perceived data. Usually only one model is fitted to the data. But what if the data might have arisen from one of several possible models? In this case the fitting procedure needs to fit all the potential models and select which of these fits the data best. This is the task of robust model selection which, in spite of the many recent developments in the application of robust fitting methods within the field of computer vision, has been, by comparison, quite neglected.

One approach might be to fit a model, and then test the hypothesis that it is acceptable. This approach was followed in Torr (1995) where a system of hypothesis testing was developed, testing the hypothesis that the data were degenerate against the hypothesis that the data were non-degenerate. In the Neyman–Pearson theory of statistical hypothesis testing only the probabilities of rejecting and accepting the correct and incorrect hypotheses, respectively, are considered to define the cost of a decision. The problem with this approach is that it is difficult to adapt to a situation where several models might be appropriate, as the test procedure for a multiple-decision problem involves a difficult choice of a number of dependent significance levels. What is needed is a scoring mechanism to rate each model. As seen, maximum likelihood methods will always lead to the most general model being selected; hence the need for a more general method of inductive inference that takes into account the complexity of the model. This has lead to the development of various *information criteria* (see the special issue of *Psychometrika* on information criteria (Takane & Bozdogan 1987)). Foremost amongst these is 'an information criterion' (AIC) (Akaike 1974). It will be seen that AIC fails in the presence of outliers, and so a new cost function, GRIC is devised. I examine the GRIC criterion in a real application to a basic problem in computer vision, developing a statistically based GRIC estimator to determine the relationship between point matches over two views that robustly selects the motion model and detects the presence of outliers.

MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES THE ROYAL A SOCIETY PHILOSOPHICAL TRANSACTIONS OF

Downloaded from rsta.royalsocietypublishing.org

(*a*) *AIC for model selection*

Akaike's information criterion is a useful statistic for model identification and evaluation. Akaike (1974) was perhaps the first to lay the foundations of information theoretic model evaluation. He developed a model selection procedure—for use in auto-regressive modelling of time-series—that chose the model with minimum estimated expected residual, with respect to the model fitted, for future observations as the best fit. The procedure selects the model that minimizes expected error of new observations with the same distribution as the ones used for fitting.† It has the form

$$\mathrm{AIC} = (-2)\log L + 2k, \tag{4.1}$$

where k is the number of parameters in the chosen model, and L is the log likelihood. With a Gaussian error model this is equivalent to the sum of squares of residuals:

$$\mathrm{AIC} = \sum \frac{e_i^2}{\sigma^2} + 2k,$$

plus an additive constant which is discounted from here on. It can be seen that AIC has two terms, the first corresponding to the badness of fit, the second a penalty on the complexity of the model; this can be thought of as analogous to an estimate of the log likelihood of the prior $\log \Pr[\phi]$. When there are several competing models, the parameters within the models are estimated by maximum likelihood and the AIC scores compared to find the model with the minimum value of AIC. This procedure is called the minimum AIC procedure, and the model with the minimum AIC is called the minimum AIC estimate (MAICE), which is chosen as the best model. Therefore the best model is the one with highest information content but least complexity. An advantage of the AIC is its simplicity as it does not require reference to look up tables, it is very easy to calculate AIC once the maximum likelihood estimate of the model parameters is made. Furthermore, there is no problem of specifying an arbitrary significance level at which models should be acceptable, and comparison between two models need not be nested or ordered.

It has been pointed out that there are some problems with AIC. One is that the AIC does not produce an asymptotically consistent (i.e. as the number of data tends to infinity) estimate of the order of the model (Schwarz 1978) as there is no account made in (4.1) for the number of observations. There have been several AIC inspired paradigms all of which provide some sort of scoring mechanism for the models, the model with least score being accepted. Table 2 summarizes some of the better known criteria. The typical form of these scoring criteria is a function of the badness of fit, the number of parameters used, the amount of data, and the information matrix. Schwarz (1978) and Kashyap (1982) work from a Bayesian point of viewpoint, expanding Pr(model | data), the posterior probability of a model given the data, trying to devise prior probabilities of the models based on their complexity. It is interesting to observe that Rissanen (1978) developed a criterion with a similar form to Akaike's from a totally different standpoint. He derived the minimum-bit representation of the data, termed shortest description length (SSD) and minimum description length (MDL) (an approach suggested by the inductive theory of Solomonoff (1964)). Bozdogan (1987) attempts to derive measures that are asymptotically consistent, and experimentally verifies this on 100 trials. Wallace &

† He later demonstrated that AIC was an estimate of the expected entropy (Kullback–Leibler information), showing that the model with the minimum AIC score also minimized the expected entropy.

Downloaded from rsta.royalsocietypublishing.org

Table 2. *Model selection scoring functions*

(Showing some different model evaluation criteria suggested in the literature. $\log L$ is the log likelihood of the model, k the number of parameters, n the number of data, ϕ_k the estimate set of parameters; p_r is the prior probability, Σ is the covariance and $\boldsymbol{J}$ the information matrix of the estimated parameters.)

author	criterion
Mallows's C_p	$-2\log L - n + 2k$
Akaike's AIC	$-2\log L + 2k$
Schwarz	$-2\log L + 2k\log n$
Schwarz KC	$-2\log L - \log p_r + \log\lvert\Sigma\rvert + k\log n$
Rissanen's SSD	$-2\log L + k\log(n+2)/24 + 2\log(k+1)$
Rissanen's MDL	$-2\log L + \frac{1}{2}k\log n$
Bozdogan's CAIC	$-2\log L + k(\log(n)+1)$
Bozdogan's CAICF	$-2\log L + k(\log(n)+2) + \log\lvert\boldsymbol{J}\rvert$
Wallace's MML2	$-2\log L - \log p_r + \frac{1}{2}(\log\lvert\boldsymbol{J} + k)\rvert$

Freeman (1987) develop a very similar criterion to Bozdogan's CAICF following a minimum message length (MML) approach. It is somewhat eerie to observe that AIC, MDL and MML derivations all produce similar criterion for model selection, even though they start from very different premises; Leclerc (1989) points out that this is because, from a Bayesian perspective, model selection comes down to assigning a prior probability to each of the putative underlying interpretations ϕ in (3.5). In Torr (1997) the model scoring criteria are compared for two-view motion model selection, and it is found that Akaike's original AIC performs reasonably, but must be adjusted to cope for differing dimensions of the models. This adjustment is detailed below.

In the *original* derivation of the AIC (Akaike 1974) it is apparent that models can only be compared if they are 'nested', meaning that given the most general model having parameters $(a_0, a_1, a_2, \dots, a_{p-1})$, the less general models are formed by setting some of these coefficients to zero, e.g. models may be formed with parameters (a_0, a_1), or (a_0, a_1, a_2), etc. This is also an assumption common to the clones of AIC given in table 2. This leads to a distinct problem. Consider two models that are non-nested for two-dimensional data: a point and a line. Both have $k = 2$ parameters but a line will always have a lower SSE (thus higher likelihood) than a point. This can be seen in figure 1. The point cannot be described by setting any of the lines coefficients to zero and hence the AIC criterion cannot be used. The problem is that the two models are of different dimensions, the point is a 0-dimensional model and the line is 1 dimensional, in a 2-dimensional space.

How does this relate to image constraints? Each pair of corresponding points $\underline{\boldsymbol{x}}$, $\underline{\boldsymbol{x}}'$ defines a single point $\boldsymbol{m}$ in a measurement space $\mathcal{R}^4$, formed by considering the coordinates in each image. The image correspondences induced by a rigid motion have an associated algebraic variety V in $\mathcal{R}^4$. The fundamental matrix, and affine fundamental matrix for two images are dimension 3 varieties, a projectivity between two images is a dimension 2 variety (Torr *et al.* 1998). The fundamental matrix has seven degrees of freedom, the projectivity has eight, and yet the fundamental matrix is more general. The affine fundamental matrix has four degrees of freedom, affinity 6. These properties are summarized in table 3. The AIC as it stands provides no

Downloaded from rsta.royalsocietypublishing.org

Table 3. *Motion models used*

(A description of the reduced models that are fitted to degenerate sets of correspondences. c is the minimum number of correspondences needed in a sample to estimate the constraint. k is the number of parameters in the model; d is the dimension of the constraint.)

model	c	k	d	constraint	parameters
fundamental matrix	7	7	3	$\boldsymbol{x}'^{\mathrm{T}}\boldsymbol{F}\boldsymbol{x}=0$	$\boldsymbol{F}=\begin{bmatrix} f_1 & f_2 & f_3 \\ f_4 & f_5 & f_6 \\ f_7 & f_8 & f_9 \end{bmatrix}$
affine $\boldsymbol{F}_A$	4	4	3	$\boldsymbol{x}'^{\mathrm{T}}\boldsymbol{F}_A\boldsymbol{x}=0$	$\boldsymbol{F}_A=\begin{bmatrix} 0 & 0 & g_1 \\ 0 & 0 & g_2 \\ g_3 & g_4 & g_5 \end{bmatrix}$
projectivity	4	8	2	$\boldsymbol{x}'=\boldsymbol{B}\boldsymbol{x}$	$\boldsymbol{B}=\begin{bmatrix} b_1 & b_2 & b_3 \\ b_4 & b_5 & b_6 \\ b_7 & b_8 & b_9 \end{bmatrix}$
affinity	3	6	2	$\boldsymbol{x}'=\boldsymbol{A}\boldsymbol{x}$	$\boldsymbol{A}=\begin{bmatrix} a_1 & a_2 & a_3 \\ a_4 & a_5 & a_6 \\ 0 & 0 & a_7 \end{bmatrix}$

mechanism for coping with models of different dimension, which is essential if we are to discern the difference between a dimension 2 and 3 variety. Table 1 shows that the AIC will always lead to the higher dimension varieties being selected regardless of the underlying ground truth. Within the next section Kanatani's (1996) approach to this problem will be described.

(*b*) *Geometric information criterion*

The AIC is developed from the idea that the best model is that which minimizes the expected SSE for future data. Consider the case of fitting a manifold of dimension d to r-dimensional points, in this case the codimension is $r-d$. Kanatani (1996) generalizes AIC to

$$\mathrm{GIC} = -2\log L + 2(dn+k), \tag{4.2}$$

which he claims is an unbiased estimator of the expected SSE. Kanatani's derivation of the GIC is rather drawn out, and the interested reader is referred to his book (Kanatani 1996). In fact Akaike (1987) gave a similar form for the GIC in the case of factor analysis when fitting models of differing dimensions. Rather than present it here, an intuitive interpretation is presented in the next section; in the two-dimensional case $r=2$, fitting a line model $d=1$ and point model $d=0$.

(*c*) *Intuitive interpretation*

Consider equation (4.2), the first term is the usual sum of squares of residuals, divided by their variances, representing the goodness of fit. The next two terms represent the parsimony of the model. The second being a penalty term for the dimensionality of the model, the greater the dimension for the model the greater the penalty. The last term is the usual AIC criterion of adding the number of parameters of the model, to greater penalize models with more parameters.

Downloaded from rsta.royalsocietypublishing.org

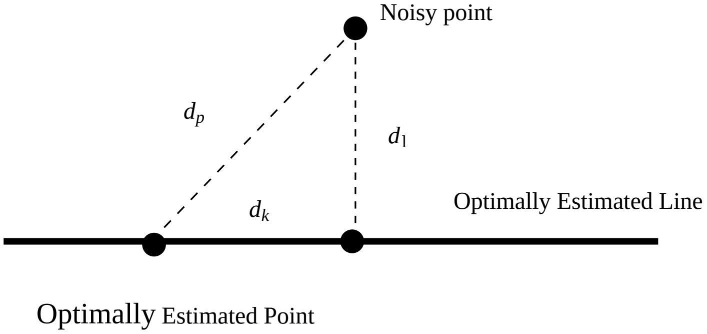

Figure 1. Showing the relationship between the noisy point, the optimally estimated line and the optimally estimated point in the Kanatani scheme.

This is now illustrated by a simple example, consider the two-dimensional example shown in figure 1. Suppose points are generated from a fixed location with added mean zero, unit standard deviation, Gaussian noise in both the x and y coordinates. If a point and a line are fitted separately by minimizing the sum of squared Euclidean distances, the optimally fitted point will lie on the optimally fitted line. Let the sum of squared distances of the points to the line model be d_l and the sum of squared distances of the points to the point model be d_p, then $d_p = d_l + d_k$, where d_k is the 'parallel' sum of squared distances as shown for one point in figure 1. It can be seen that unless the data all lie exactly on a point then d_l is always less than d_p. The GIC for the line model compensates for this bias by the penalty term, which is twice the expectation of the 'parallel' sum of squares (d_k). If the model estimated is a line then the GIC has the form

$$\mathrm{GIC}(\mathrm{line}) = d_l + 2(n + 2), \tag{4.3}$$

as the model has dimension 1, codimension 1 and two degrees of freedom in the parameters. If the number of data is large, the degree of freedom of the model (i.e. the number of the parameters) has little effect because it is a simple constant. What matters is twice the dimension of the model, which is multiplied by the number of data. The dimension equals the 'internal' degree of freedom of the data, which in turn equals the expectation of the 'parallel' (or in a direction on the manifold) sum of squares per datum. Returning to the example, the GIC for a point is

$$\mathrm{GIC}(\mathrm{point}) = d_l + d_k + 4, \tag{4.4}$$

thus a point is favoured if $d_k \leqslant 2n$. Hence the algorithm is equivalent to a test of spread along the line.

Kanatani's method works well for two-view geometric constraints as well, consider the average SSE given in table 1 for 100 data points. These can be turned into GIC by the addition of 614, 608 and 416 ($2(nd + k)$) for $\boldsymbol{F}$, $\boldsymbol{F}_A$ and $\boldsymbol{H}$ respectively. It can be seen that on average the lowest GIC equates to the correct model, although it behaves less well for distinguishing $\boldsymbol{F}$ from $\boldsymbol{F}_A$, than $\boldsymbol{F}$ from $\boldsymbol{H}$. Generally the GIC tends to underestimate the dimension of the data and overestimate the number of motion model parameters; suggesting that a more general form

$$\mathrm{GIC} = -2 \log L + \lambda_1 dn + \lambda_2 k \tag{4.5}$$

might be appropriate with $1 \leqslant \lambda_1 \leqslant 2$ and $\lambda_2 > 2$; experimentation with the form of GIC is beyond the scope of this paper and will be the subject of future work.

Downloaded from rsta.royalsocietypublishing.org

Generally when n is small λ_1 should be set to 2, when n is large λ_1 should be set to 1, but for reasonable data experiments reveal that the solution obtained is fairly stable over a range of values of n, λ_1 and λ_2. The major drawback of Kanatani's work is that it is non-robust, which is dealt with in the next section.

(*d*) *Robust AIC*

Thus far model selection in the case of known error distribution has been considered. Yet it must be realized that there is a big gap between the theoretical results and the practical procedures of identification. This is because the data only approximately conform to postulated theoretical probability distributions; furthermore they may contain *outliers* which correspond to data belonging to a totally different population. Thus the model selection procedure needs to be robust, which means that it will still work even if the theoretical assumptions about the data are violated (such as there being outliers in the data). To illustrate the effects of outliers in the presence of degeneracy, a simple example is furnished. Figure 2 shows four cases of line fitting to two-dimensional data sets. Figure 2*a* shows a set which we might consider non-degenerate, and for which a line model is appropriate. Figure 2*b* demonstrates degenerate data, where there are an infinite number of lines that fit the data equally well. A noise model is essential if this type of degeneracy is to be detected; in the absence of a noise model an arbitrary rescaling of a given axis can make the data look linear. Similarly, in figure 2*a* if the noise were very high relative to the dispersion of the points then this might indeed be a degenerate set. The need for methods that can flag degeneracy in the presence of outliers is demonstrated by figure 2*c*, where even one outlier can effectively mask the degeneracy. From this example it is clear that the detection of outliers and model selection are inextricably linked.

Ronchetti (see Hampel *et al*. 1986) notes that the derivation of the AIC is independent of the distribution assumed for it, and that it can be robustified in much the same way that Huber (1981) extended maximum likelihood estimation to M-estimation. Ronchetti proposed the robust AIC: $\sum_i \rho(e_i^2) + \alpha k$. The decision then becomes one of choosing an appropriate robust error function $\rho(e_i^2)$ and deriving the value of α that correctly weights the two terms of AICR. Correctly determining $\rho(e_i^2)$ entails some knowledge of the outlier distribution; here it is assumed, without *a priori* knowledge, that the outlier distribution is uniform, with negative log likelihood $l_o = \lambda_3$ for error dimension one. For higher-dimensional errors $l_o = (r-d)\lambda_3$ where $r-d$ is the codimension. The minimum of these two log likelihoods defines $\rho(e_i^2)$,

$$\rho(e^2) = \left\{ \begin{array}{ll} e^2/\sigma^2 & e^2/\sigma^2 < \lambda_3(r-d) \\ \lambda_3(r-d) & e^2/\sigma^2 \geqslant \lambda_3(r-d) \end{array} \right\} = \min(e^2/\sigma^2, \lambda_3(r-d)), \qquad (4.6)$$

where σ is the standard deviation of the error on each coordinate. The form of $\rho()$ is derived to behave like a Gaussian for data with low residuals and a uniform distribution for data with high residuals. What value should λ_3 take? Consider an outlier in dimension d space; it is assumed that absolutely nothing about the distribution of the outlier (not even that it of the same probability distribution as the other outliers); hence it is described by d parameters, one for each coordinate, i.e. x, y, x', y'. Thus here I choose $\lambda_3 = 2$, but experiments have shown that the solution remains unchanged for a range of λ_3; in the case that the codimension is 1, this choice incorrectly rejects inliers 10% of the time. In the simple case of fitting a line versus a

Downloaded from rsta.royalsocietypublishing.org

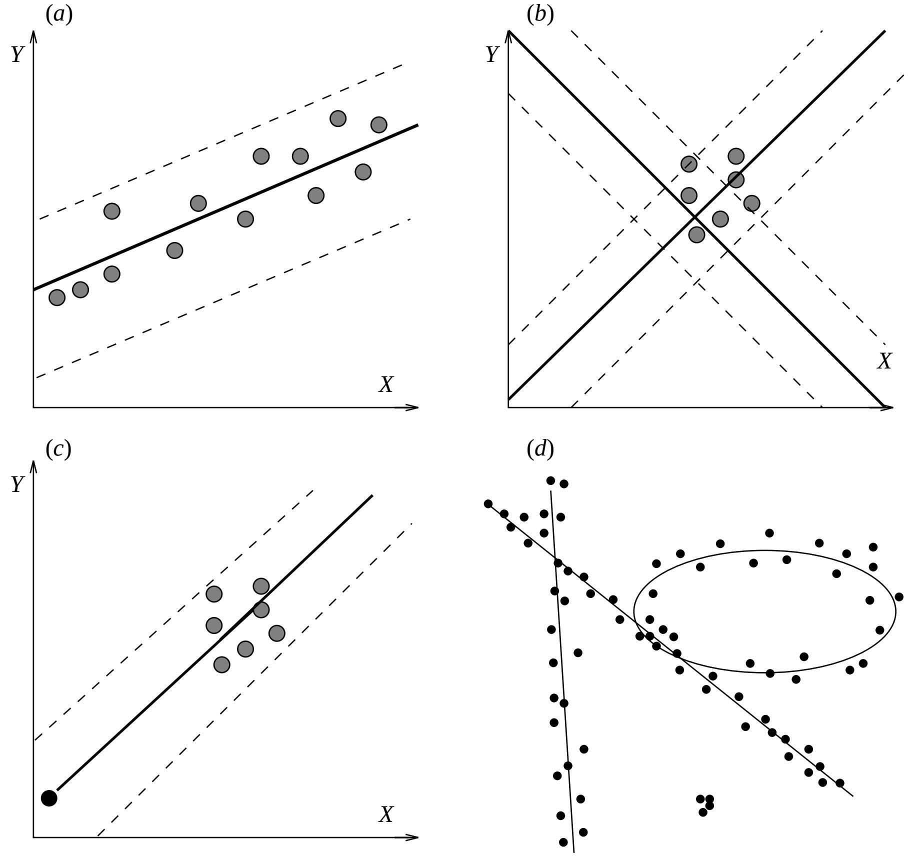

Figure 2. Line fitting to two-dimensional data sets. (*a*) A non-degenerate data set, with no ambiguity in determining the best line fit. (*b*) A degenerate data set. Many solutions will have a similar error. (*c*) A single outlier renders a degenerate data set apparently non-degenerate. (*d*) The case when the data should be modelled by a mixture model.

point, the cost function for a line is $-2\log L + 2(i_l + k) + 2(2o_l)$, where i_l is the number of inliers to a line and o_l the number of outliers, and $k = 2$ for a point $-2\log L + 2(k) + 2(2o_p)$, where o_p is the number of outliers to a point and $k = 2$.

The form of the function given in (4.6) has several advantages. Firstly, it provides a clear dichotomy between inliers and outliers. Secondly, outliers to a given model are given a fixed cost, reflecting that they probably arise from a diffuse or uniform distribution. The fixed penalty is also in the same spirit as robust MDL approaches, where outliers are assigned a fixed cost. Furthermore, if the outliers follow a large uniform distribution, then they will only be incorrectly flagged as inliers a vanishingly small percentage of the time (false positives). Having made a choice for $\rho(e_i^2)$, the question of what to choose for α arises. When the error is Gaussian, Akaike makes a strong case that $\alpha = 2$. Using the robust error function I can also claim that $\alpha \approx 2$, as I am advocating what is in effect a trimmed Gaussian distribution.

The first new research in this paper comes from the simple step of drawing together the geometric AIC developed by Akaike and Kanatani, with the robust AIC expounded above, to produce an information criterion that is both robust and

Downloaded from rsta.royalsocietypublishing.org

capable of dealing with models of different dimensionalities, termed the geometric robust information criterion: GRIC

$$\mathrm{GRIC} = \sum \rho(e_i^2) + \lambda_1 dn + \lambda_2 k. \tag{4.7}$$

This has terms for the error, dimension and number of parameters in the model. In order to evaluate this sum an estimate of the standard deviation of the error σ must be made. This is done *a priori* from the properties of the corner detector. The second new piece of research in this paper comes from adjusting the mixture model of § 3 to cope consistently with data arising from varieties of different dimensions; this is done by developing the posterior probabilities for each datum given the model.

5. Establishing posterior distributions

When calculating the posterior distributions $\tau_{ij}(\boldsymbol{m} \mid \phi)$ given in (3.7), simply using the likelihoods of $p_k(\boldsymbol{m}_i \mid \mathcal{R}_k)$ as given in (3.1) produces a solution which is biased in that data will be more likely to belong to a higher-dimensional than a lower-dimensional model as shown in § 4. This is because, for the dimension 3 varieties, there is only one degree of freedom in the error (3.4) while there are three degrees of freedom in the choice of $\hat{\boldsymbol{m}}$. Whereas for dimension two varieties there are two degrees of freedom as there are only two degrees of freedom in the choice of $\hat{\boldsymbol{m}}$.

Bozdogan (1987) suggests that the AIC is an unbiased estimator of minus twice the mean expected log likelihood, or equivalently $-\frac{1}{2}\mathrm{AIC}$ is asymptotically an unbiased estimator of the mean expected log likelihood. This result suggests that asymptotically a reasonable definition of the likelihood of a model is

$$p_{\mathrm{AIC}}(\phi) = \exp(-\tfrac{1}{2}\mathrm{AIC}). \tag{5.1}$$

This suggests that to cope with data arising from varieties of differing dimensions (3.1) should be altered to

$$p(\boldsymbol{m} \mid \mathcal{R}_j; \mathrm{AIC}) = \prod_i \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^4 \mathrm{e}^{-((\underline{x}_i - x_i)^2 + (\underline{y}_i - y_i)^2 + (\underline{x}'_i - x'_i)^2 + (\underline{y}'_i - y'_i)^2)/(2\sigma^2) + \lambda_1 d_j}, \tag{5.2}$$

where d_j is the dimension of $\mathcal{R}_j$. This is a key new idea: to use the expected log likelihood furnished by the AIC to allow the use of mixture models for data arising from varieties of differing dimensions. Without this compensation factor, I have found that matches are assigned to clusters consistent with fundamental matrices at the expense of those consistent with homographies. The posterior probabilities $\tau_{ij}(\boldsymbol{m} \mid \phi; \mathrm{AIC})$ are now given by

$$\Pr[\boldsymbol{m}_i \in \mathcal{R}_j \mid \boldsymbol{m}; \phi; \mathrm{AIC}] = \pi_j p_j(\boldsymbol{m}_i \mid \mathcal{R}_j; \mathrm{AIC}) \Bigg/ \sum_{k=1}^{k=s-1} \pi_k p_k(\boldsymbol{m}_i \mid \mathcal{R}_k; \mathrm{AIC}). \tag{5.3}$$

6. Motion segmentation

The final cost function to be maximized is

$$L_C(\mathrm{AIC}) = \sum_{i=1}^{i=n} \left(\sum_{j=1}^{j=s} c_{ij} (\log \pi_j + \log(p(\boldsymbol{m}_i \mid \mathcal{R}_j)) + \lambda_1 d_j) \right) + \lambda_3 ro + \lambda_2 k_j, \tag{6.1}$$

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

Downloaded from rsta.royalsocietypublishing.org

Table 4. *Clustering algorithm*

1. Initialize the number of clusters extracted to $j = 0$.
2. RANSAC followed by a robust nonlinear estimator is used to estimate each model $\boldsymbol{F}$, $\boldsymbol{F}_A$, $\boldsymbol{H}$, $\boldsymbol{A}$, from the data Z as described in Torr & Murray (1994, 1997). Information about spatial proximity is used to exploit the spatial cohesion of moving objects, by sampling correspondence sets that are close to each other in the image.
3. Calculate the GRIC score for each model.
4. If the number of inliers for the model with minimum GRIC is lower than a threshold (for details of this threshold see Torr & Murray (1994)), declare all unassigned matches outliers goto step (8).
5. Increment j by one.
6. The model with minimal GRIC is stored as a motion model $\mathcal{R}_j$, together with the its set of inliers κ_j.
7. Remove the matches κ_j from Z and go to step 1.
8. Using the established clusters as input apply the EM algorithm to maximize (6.1).
9. Clusters are pruned by removing one cluster at a time and recomputing (6.1) to see whether the cost function is reduced, if it is the cluster is discarded and the matches reassigned to the other clusters; as described in § 3.
10. Reassign matches $\boldsymbol{m}$ to their optimal estimates $\hat{\boldsymbol{m}}$.
11. Densely segment.

where o is the number of outliers, k_j is the number of parameters and d_j is the dimension of the kth model; e.g. 7 for $\boldsymbol{F}$, 8 for $\boldsymbol{H}$. For the experiments presented in the next section $\lambda_1 = \lambda_2 = \lambda_3 = 2$. It would be too computationally expensive to search every possible ϕ in order to minimize (6.1), hence a random sampling type algorithm is used, in which solutions are grown from minimal subsets of points within the image. This procedure is described in more detail in the next section.

The segmentation algorithm extracts models using RANSAC (Fischler & Bolles 1981) and is described in table 4. Once the clusters are identified, a dense segmentation is made; for each pixel in image 1 its location in image 2 is predicted using each motion model, and the squared difference in image intensity calculated. The pixel is assigned to the cluster which minimizes this difference; matches may be reassigned at this juncture. For $\boldsymbol{F}$ and $\boldsymbol{F}_A$ pixel transfer is not defined. In order to accomplish the transfer, the nearest three optimally estimated matches $\hat{\boldsymbol{m}}$ consistent with the cluster to the pixel are found, and using these three matches and the estimate of $\boldsymbol{F}$ or $\boldsymbol{F}_A$ for the cluster a homography is computed to transfer that pixel from image 1 to 2. Finally, in order to improve the segmentation, morphological operations are applied (Rosin 1998).

(*a*) *Results of the segmentation algorithm*

(i) *Soldier example*

Figures 3*a*, *b* show two images, the first and twelfth, from a sequence of a soldier striding through the jungle. Figure 3*b* shows the disparity vectors for the initial point correspondences. The camera tracks the figure, the background apparently moves to

Downloaded from rsta.royalsocietypublishing.org

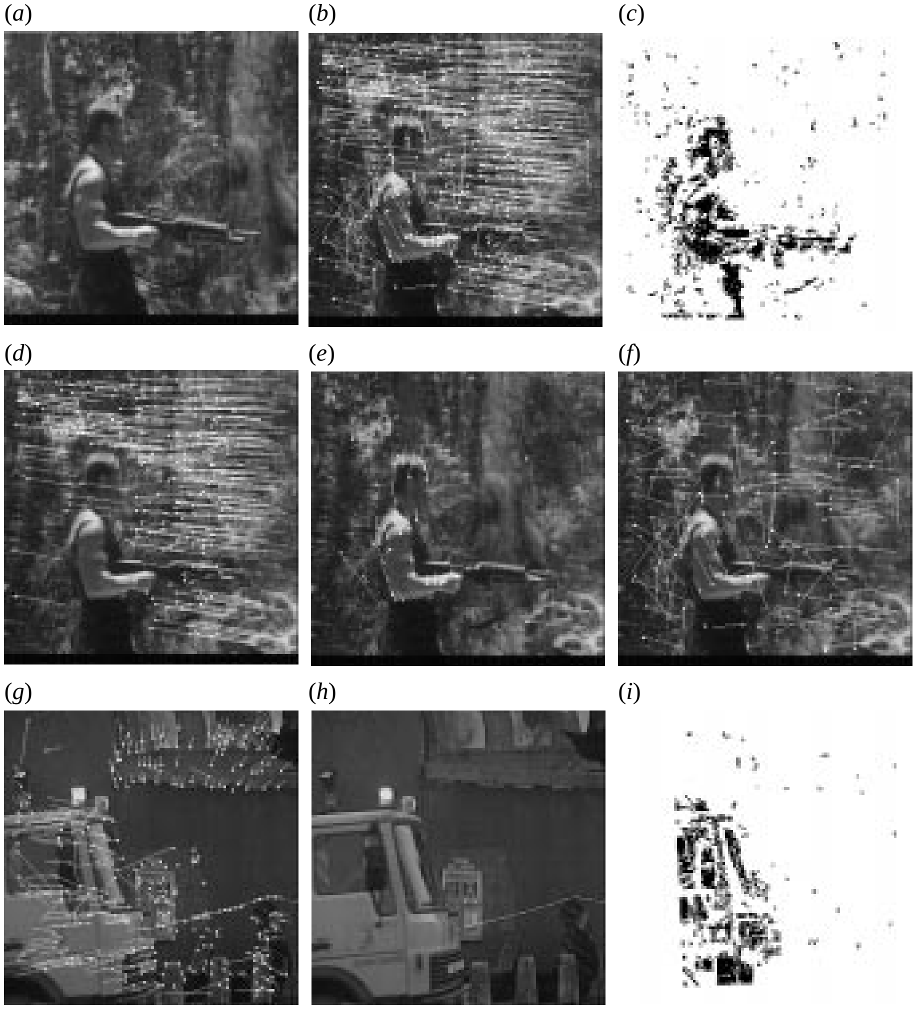

Figure 3. Test data: truck and soldier.

the left while the figure moves down. The background is relatively distant and the motion vectors are all of roughly equal length. The foreground figure apparently moves down towards the foot of the image. The clusters generated by the RANSAC extraction algorithm are shown in figures 3*d*, *e* and the outliers in 3*f*. The GRIC scores calculated at step (iii) for the models are given in table 5, the background model is given as $\boldsymbol{H}$, the foreground (soldier) is $\boldsymbol{A}$; and the dense segmentation gained is shown in figure 3*f*.

(ii) *Lorry*

Figures 3*g*, *h* are two images from a sequence of a lorry translating to the right, as the camera zooms away from it; figures 3*g* shows initial disparity vectors. By chance, the epipolar geometry for the lorry is *locally* similar to that of the background within the region of the image subtended by the lorry, and attempts to segment based solely

Downloaded from rsta.royalsocietypublishing.org

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

Table 5. *Clustering results*

(GRIC values for the images. The model with lowest GRIC is underlined.)

		motion of points				
estimated	n	general	orthographic	homography	affinity	inliers
cluster 2 truck	319	1089	$\underline{1079}$	1110	1089	91
cluster 1 truck	319	2125	2118	2001	$\underline{1997}$	166
cluster 2 soldier	565	1336	1315	1228	$\underline{1225}$	66
cluster 1 soldier	565	3853	3879	$\underline{3299}$	3367	388

on $\boldsymbol{F}$ fail. The motion segmentation algorithm identifies two clusters, the larger consistent with an affinity $\boldsymbol{A}$ corresponding to the background, the smaller to an affine camera corresponding to the lorry. The GRIC scores calculated at step (iii) for the models are given in table 5. Using these models more accurately represents the motion, and the dense segmentation gained, shown in figure 3*c*, is better than that gained by just using fundamental matrices.

Generally we found that the method provided the dimension of the model varieties quite stably, but that their degree was more difficult to ascertain. Thus the decision between $\boldsymbol{F}$ and $\boldsymbol{H}$ was more reliable than that between $\boldsymbol{F}$ and $\boldsymbol{F}_A$ or $\boldsymbol{H}$ and $\boldsymbol{A}$; suggesting there needs to be further work to improve the AIC criterion.

7. Summary and conclusions

There have been three contributions contained within this paper. The first is to provide a new method of motion segmentation using the constraints enforced by rigid motion. This method uses mixture models in an optimal statistical framework to estimate the number and type of motion models as well as their parameters. To achieve this two things needed to be done. The first was to make the AIC robust to outliers. A new general method, GRIC, has been presented for robust model selection, and its application to two-view motion model fitting demonstrated. The method is highly robust and simultaneously flushes outliers and selects the type of model that best fits the data. The second was to use the AIC to estimate the expected likelihoods for data arising from models of differing dimensions; solving the problem of over-fitting of the dimension if just the unadjusted likelihood were used.

The convergence of the EM algorithm is notoriously slow (Redner & Walker 1984) and it may be better to use some sort of conventional numerical optimization technique. Methods such as gradient descent have the added advantage of automatically supplying the covariance matrix as part of the algorithm. Comparison of EM and other algorithms for optimising a motion segmentation is an interesting vein of future work.

The dense segmentation method is still in the initial stage of development. It might prove desirable to adjust the way that the motion models are determined to incorporate information from all pixels, not just highly textured areas. This is the approach followed by Ayer & Sawhney (1995), although they use only projectivities as motion models; and it is not clear how much extra advantage to the estimation of the motion models is gained by including non-textured regions of pixels in the

Downloaded from rsta.royalsocietypublishing.org

minimization. Information from edges could be used to improve the detected motion discontinuities and this would be an interesting avenue for future research.

I gratefully acknowledge W. Triggs, Dr P. McLauchlan, Dr K. Kanatani, A. Fitzgibbon, Dr D. Murray and Dr Andrew Zisserman for comments. ACTS Project VANGUARD provided financial support.

References

Akaike, H. 1974 A new look at the statistical model identification. *IEEE Trans. Automatic Control* **AC-19**, 716–723.

Ayer, S. & Sawhney, H. 1995 Layered representation of motion video using robust maximum-likelihood estimation of mixture models and mdl encoding. In *Proc. 5th Int. Conf. on Computer Vision, Boston*, pp. 777–784. Los Alamitos, CA: IEEE Computer Society Press.

Beardsley, P., Torr, P. H. S. & Zisserman, A. 1996 3D model aquisition from extended image sequences. In *Proc. 4th European Conf. on Computer Vision, Cambridge* (ed. B. Buxton & R. Cipolla), pp. 683–695. Lecture Notes in Computer Science, vol. 1065. Springer.

Black, A. & Anandan, P. 1996 The robust estimation of multiple motions: parametric and piecewise-smooth flow fields. *Computer Vision Image Understanding* **63**(1), 75–104.

Bozdogan, H. 1987 Model selection and Akaike's information criterion (AIC): the general theory and its analytical extensions. *Psychometrika* **52**, 345–370.

Darrell, T., Sclaroff, S. & Pentland, A. 1990 Segmentation by minimal description. In *Proc. 3rd Int. Conf. on Computer Vision, Osaka*, pp. 112–116. IEEE.

Dempster, A. P., Laird, N. M. & Rubin, D. B. 1977 Maximum likelihood from incomplete data via the em algorithm. *J. R. Statist. Soc.* B **39**, 1–38.

Faugeras, O. D. 1992 What can be seen in three dimensions with an uncalibrated stereo rig? In *Proc. 2nd European Conf. on Computer Vision, Santa Margherita Ligure* (ed. G. Sandini), pp. 563–578. Lecture Notes in Computer Science, vol. 588. Springer.

Fisher, R. A. 1936 Uncertain interference. *Proc. Am. Acad. Art Sci.* **71**, 245–258.

Fischler, M. & Bolles, R. 1981 Random sample consensus: a paradigm for model fitting with application to image analysis and automated cartography. *Commun. Ass. Comp. Mach.* **24**, 381–395.

Gu, H., Yoshiaki, S. & Asada, M. 1996 MDL-based segementation and motion modeling in a long image sequence of scene with multiple independently moving objects. *IEEE Trans. Pattern Analysis Machine Intellig.* **18**, 58–64.

Hampel, J. P., Ronchetti, E. M., Rousseeuw, P. J. & Stahel, W. A. 1986 *Robust statistics: an approach based on influence functions.* New York: Wiley.

Harris, C. J. & Stephens, M. 1987 A combined corner and edge detector. In *Proc. Alvey Conf.*, pp. 147–151.

Hartley, R. I. 1992 Estimation of relative camera positions for uncalibrated cameras. In *Proc. 2nd European Conf. on Computer Vision, Santa Margherita Ligure* (ed. G. Sandini), pp. 579–587. Lecture Notes in Computer Science, vol. 588. Springer.

Huber, P. J. 1981 *Robust statistics.* New York: Wiley.

Jepson, A. & Black, M. 1993 Mixture models for optical flow computation. In *CVPR-93*, pp. 760–766. IEEE.

Kanatani, K. 1996 *Statistical optimization for geometric computation: theory and practice.* Amsterdam: Elsevier Science.

Kashyap, R. L. 1982 Optimal choice of AR and MA parts in autoregressive moving average models. *IEEE Trans. Pattern Analysis Machine Intellig.* **4**, 99–104.

Leclerc, Y. G. 1989 Constructing simple stable descriptions for image partitioning. *Int. J. Computer Vision* **3**, 73–102.

Downloaded from rsta.royalsocietypublishing.org

Leonardis, A., Gupta, A. & Bajcsy, R. 1990 Segmentation as the search for the best description of the image in terms of primitives. In *Proc. 3rd Int. Conf. on Computer Vision, Osaka*, pp. 121–125. IEEE.

McLachlan, G. I. & Basford, K. 1988 *Mixture models: inference and applications to clustering.* New York: Marcel Dekker.

Mallows, C. L. 1973 Some comments on c_p. *Technometrics* **15**, 661–675.

Mirza, M. J. & Boyer, K. L. 1992 An information theoretic robust sequential procedure for surface model order selection in noisy range data. In *Proc. CVPR92*, pp. 366–371. Los Alamitos, CA: IEEE Computer Society Press.

Mundy, J. & Zisserman, A. 1992 *Geometric invariance in computer vision.* MIT press.

Redner, R. A. & Walker, H. F. 1984 Mixture densities, maximum likelihood and the EM algorithm. *SIAM Rev.* **26**, 195–239.

Rissanen, J. 1978 Modeling by shortest data description. *Automatica* **14**, 465–471.

Rosin, P. 1998 Thresholding for change detection. In *Proc. ICCV98* (ed. U. Desai), pp. 274–279. Narosa Publishing.

Schwarz, G. 1978 Estimating dimension of a model. *Ann. Stat.* **6**, 461–464.

Solomonoff, R. 1964 A formal theory of inductive inference i. *Informat. Control* **7**, 1–22.

Takane, Y. & Bozdogan, H. (eds) 1987 *Psychometrika* **52**(3). (Special Issue.)

Torr, P. H. S. 1995 Outlier detection and motion segmentation. Ph.D. thesis, Department of Engineering Science, University of Oxford, UK.

Torr, P. H. S. 1997 An assessment of information criteria for motion model selection. In *CVPR97*, pp. 47–53.

Torr, P. H. S. & Murray, D. W. 1994 Stochastic motion clustering. In *Proc. 3rd European Conf. on Computer Vision, Stockholm* (ed. J.-O. Eklundh), pp. 328–338. Lecture Notes in Computer Science, vol. 800/801. Springer.

Torr, P. H. S. & Murray, D. W. 1997 The development and comparison of robust methods for estimating the fundamental matrix. *Int. J. Computer Vision* **24**, 271–300.

Torr, P. H. S. & Zisserman, A. 1997 Performance characterization of fundamental matrix estimation under image degradation. *Machine Vision Applic.* **9**, 321–333.

Torr, P. H. S., Zisserman, A. & Maybank, S. 1998 Robust detection of degenerate configurations for the fundamental matrix. *Computer Vision Image Understanding.* (In the press.)

Wallace, C. S. & Freeman, P. R. 1987 Estimation and inference by compact coding. *J. R. Statist. Soc.* B **49**, 240–265.

Discussion

O. Faugeras (*INRIA, France*). This is very nice, but I would prefer to see F smoothly varying rather than having just several discrete possibilities like projective and affine.

P. H. S. Torr. This can be thought of as model averaging, rather than fitting a line or a conic to some data, we should allow for smooth combination of the two models: the line and the conic. At the moment I am experimenting with Bayes factors as a method to perform this model averaging.

T. Kanade (*Robotics Institute, Carnegie Mellon University, Pittsburgh, USA*). How does Dr Torr's scheme compare with Saranoff's layered method?

P. H. S. Torr. The work of Kumar, Anandan, Ayer and Sawhney does not use general three-dimensional models for segmentation, rather the image is segmented into two-dimensional 'layers'. I am interested in getting a general purpose motion

Downloaded from rsta.royalsocietypublishing.org

segmenter which works with outliers, and for 3D as well as 2D motions (rather than just the motion induced by planes), so this method is more general than theirs, although I liked their work and thought it was very interesting.

N. Hollinghurst (*Olivetti and Oracle Research Laboratory, Cambridge, UK*). Does Dr Torr use correlations over general patches of the images and then choose a model using them?

P. H. S. Torr. To initialize the various motion models I used a point-based process. Once we have an estimate of the motion model we can use a homography, for instance (for example, the background in the Schwarzenegger scene (figure 3)) to transfer each pixel in order to compare with the pixel in the next image. This cannot be done with a fundamental matrix, that's why I use a Delaunay triangulation assuming local planar patches. This is of course, only a first stab at the problem.

J. Lasenby (*Department of Engineering, University of Cambridge, UK*). The signal/image processing community devotes much research effort to motion segmentation; in particular there have been recent algorithms proposed for layered segmentation, which seems to be doing much the same as you are doing. In those models the number of layers would be chosen via some criterion like MDL. Has Dr Torr compared his method with such schemes?

P. H. S. Torr. These methods are not robust, because, as I mentioned before they do not deal with 3D motion. As many of these methods use an instantaneous motion model (as opposed to discrete), they will only be useful for very closely spaced images. The system I have described here is completely automatic from the matching onwards.

M. Sabin (*Numerical Geometry Ltd, Cambridge, UK*). There is a purely geometric technique which came out of the computational geometry world called the α-shapes method which might be useful for this problem. It is essentially a pruning of the Delaunay triangulation in 4-space, estimating the dimension (locally) of the manifold given by a cloud of points.

P. H. S. Torr. These methods are essentially similar to the robust convex hull techniques. I have some experience which such techniques and have found that some care is needed in using them. For example, during fitting, these methods will designate good data as outlying.

A. Fitzgibbon (*Department of Engineering, University of Oxford, UK*). First, a comment on the previous question. I think Dr Torr will find that the α-shapes method will not be able to deal with varying dimensions of the manifold. What you get when you apply it to a 4D point set is a 3D submanifold, but not a 2D or 1D one, as it is simply a subset of the Delaunay simplicization.

Secondly, I would say that the important tuning factor in the AIC method is the noise variance σ. Do you estimate the noise variance from the image data, and do you find that performance is sensitive to the estimate of σ?

P. H. S. Torr. Estimating the noise correctly is a very hard thing to do when there is an unknown number of independently moving objects and outliers. One can use the median of the errors to robustly estimate the noise variance, but if half the data are not inliers to a model, the median may be arbitrary. Therefore, a calibration

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

Downloaded from rsta.royalsocietypublishing.org

phase is required in order to get prior estimates of the thresholds involved. There are several approaches to this: first we could adopt a Kanatani-style approach and work out what the thresholds should be from the properties of the corner matches leading to some threshold t such that we accept only points with error below t pixels. Overall, because of the big difference between the inlier and outlier distributions we have some good flexibility in the choice of threshold, and the methods described here will be robust to misspecification of t by several pixels.