# MINIMUM DESCRIPTION LENGTH AND THE INFERENCE OF SCENE STRUCTURE FROM IMAGES

S.J. Maybank[1] and P.F. Sturm[1]

## 1 Introduction

Model selection is a central task in computer vision: given data obtained from images and given a number of models, which model is most strongly supported by the data? Is it better to have *i)* a simple model fitting the data approximately; or *ii)* a complicated model fitting the data very closely [3],[4], [9],[11]? In many cases the models vary widely in complexity and flexibility, and there is little prior knowledge about the best choice of model.

The Minimum Description Length (MDL) method links model selection to data compression: the best model is the one which yields the largest compression of the data. The general theoretical framework for compression is Kolmogorov complexity: let $x$, $y$ be bit strings, ie. elements of $\Sigma = \{0,1\}^*$. The Kolmogorov complexity $K(x|y)$ of $x$ conditional on $y$ is the length in bits of the shortest program $p$ which outputs $x$, given $y$ as the input. The Kolmogorov complexity of $x$, $K(x)$, is $K(x) = K(x|\epsilon)$, where $\epsilon$ is the empty string. Kolmogorov complexity extends to functions. Let $f$ be a computable function defined on $\Sigma$. Then the complexity, $K(f)$, of $f$ is the length in bits of the shortest program which computes $f$. Further information about Kolmogorov complexity and MDL is given in [3], [4] and applications of MDL to computer vision are given in [2], [5].

Let the models be $M_1, M_2, \ldots$, coded by elements of $\Sigma$ and let $s$ be the code for the data. The best model is defined to be the one with the least value of $K(M_i) + K(s|M_i)$ for $i \geq 1$. The Kolmogorov complexity is not computable, so in practice $K(M_i)$ and $K(s|M_i)$ are replaced by computable approximations.

MDL differs from Bayesian model selection (BMS) in that it is biased against complex probability density functions. Let $M \mapsto Q(M)$ be a prior probability density defined on the models, suppose that each model defines a probability density $s \mapsto P(s|M)$ on the data and let $\alpha(Q,P,M)$ be defined by $\alpha(Q,P,M) = K(Q) + K(P(.|M))$. The connection between MDL and BMS is given by the equation [4]

$$K(M) + K(s|M) = \alpha(Q,P,M) - \log_2(Q(M)) - \log_2(P(s|M)) + O(1) \tag{1}$$

which holds with a high probability provided $s$, $M$ are both chosen randomly with respect to the distributions $Q$ and $P(.|M)$. If $\alpha(Q,P,M)$ is small, then MDL agrees with BMS, at least to a good approximation, because the model $M$ which minimises the left-hand side of (1) also minimises $-\log_2(Q(M)P(s|M))$ which is proportional to $-\log_2(P(M|s))$

The major problem with BMS is the selection or estimation of the prior probability $Q$ for the models. Without $Q$ there is no way of comparing models with different numbers of parameters or with radically different structures. MDL provides a default density for $Q$, namely $M \mapsto c2^{-K(M)}$, where $c$ is a normalising constant. Radically different models are compared by reducing the data to a bit string and using the length of the string as a measure of the effectiveness of the model.

MDL is applied to a model selection problem in computer vision. Two images of the same scene are taken from different viewpoints and corresponding points $q \leftrightarrow q'$ are obtained. Points $q$, $q'$ in different images correspond if they are projections of the same scene point [1]. Algorithms for finding corresponding points usually return only points associated scene structures for which the grey level gradients are high and rapidly varying, for example the corner of an object seen against a contrasting background.

Suppose that the data are the pixel coordinates of a list of corresponding points $q_i \leftrightarrow q'_i$, $1 \leq i \leq n$. There are many possible models, each of which involves assumptions about the relative position of the two cameras or the scene geometry [9], [10]. The following models are considered here.

- $\mathcal{B}$) Background: the image points have no discernable structure.
- $\mathcal{C}$) Collineation: there is a collineation from the first image to the second such that $\omega(q_i) = q'_i$, $1 \leq i \leq n$.
- $\mathcal{A}$) Affine fundamental: there is a $3 \times 3$ matrix $A$ with rank 2 such that $A_{11} = A_{12} = A_{21} = A_{22} = 0$ and $q_i^\top A q'_i = 0$, $1 \leq i \leq n$ [8].
- $\mathcal{F}$) Fundamental: there is a $3 \times 3$ matrix $F$ with rank 2 such that $q_i^\top F q'_i = 0$, $1 \leq i \leq n$.

[1]Department of Computer Science, The University of Reading, Whiteknights, Reading, Berkshire RG6 6AY, UK. Email: (S.J.Maybank,P.F.Sturm)@reading.ac.uk

Model $\mathcal{C}$ applies if the scene points are all coplanar. Model $\mathcal{A}$ applies if the optical axes of the two cameras are parallel and the depth variation of the scene points is small compared to their distances from the camera. Model $\mathcal{F}$ applies to two general images of a scene. It includes $\mathcal{A}$ and $\mathcal{C}$ as special cases.

Let $s$ be the data coded as a bit string. Then $U_M(s)$ is the compressed string, assuming model $M \in \{\mathcal{B}, \mathcal{C}, \mathcal{A}, \mathcal{F}\}$. The length $|U_M(s)|$ is a computable approximation to $K(M) + K(s|M)$.

# 2 Coding the Data

In this section the strings $U_{\mathcal{B}}(s)$, $U_{\mathcal{C}}(s)$, $U_{\mathcal{A}}(s)$, $U_{\mathcal{F}}(s)$ are defined. All the codes are constructed using rational arithmetic, in order to avoid inaccuracies arising from floating point approximations. The algorithms are implemented in Mathematica [12].

The image points are defined for $1 \leq i \leq n$ by $q_i = (x_i, y_i, 1)$, $q_i' = (x_i', y_i', 1)$, where $x_i$, $y_i$, $x_i'$, $y_i'$ are integers. The addition of a 1 as a third coordinate simplifies the notation. If the feature points are located in each image to an accuracy of 1 pixel, then the $x_i$, $y_i$ etc. are the pixel coordinates. If feature points are located with subpixel accuracy, then the $x_i$, $y_i$ etc. are scaled pixel coordinates.

The code $c : Z^n \to \Sigma$ used in this section is defined below in §3, and $Z^n$ is the set of $n$-tuples of integers, positive or negative.

## 2.1 Background $\mathcal{B}$

Let $x, y, x', y' \in Z^n$ be the vectors with respective components $x_i$, $y_i$, $x_i'$, $y_i'$. The code for $q_i \leftrightarrow q_i'$, $1 \leq i \leq n$ under the background model $\mathcal{B}$ is $U_{\mathcal{B}}(s) = c(x).c(y).c(x').c(y')$.

## 2.2 Collineation $\mathcal{C}$

The collineation $\omega$ is a $3 \times 3$ matrix[1]. Let $a_i$, $b_i$ be defined for $1 \leq i \leq n$ by $\omega(q_i) = (a_i, b_i, 1)$ and let $\epsilon_i$, $\delta_i$ be defined by $(\epsilon_i, \delta_i) = (\lfloor x_i' - a_i + 0.5 \rfloor, \lfloor y_i' - b_i + 0.5 \rfloor)$. The point $q_i'$ can be recovered exactly from $\omega$, $q_i$ and the integers $\epsilon_i$, $\delta_i$.

Let $\epsilon, \delta \in Z^n$ be the vectors with respective coordinates $\epsilon_i$, $\delta_i$ and let code($\omega$) be a coding of $\omega$. The $q_i$, $q_i'$, $1 \leq i \leq n$ are coded by the string

$$t = c(x).c(y).\text{code}(\omega).c(\epsilon).c(\delta) \tag{2}$$

If $\omega$ is a good fit to the data, then $\epsilon$, $\delta$ are small and compression is achieved.

How should code($\omega$) be constructed? A key issue is the precision needed for the components of $\omega$. If the precision is low, then $|\text{code}(\omega)|$ is small but $|c(\epsilon).c(\delta)|$ is large. If the precision is high, then $|\text{code}(\omega)|$ is large but $|c(\epsilon).c(\delta)|$ is small. The problem of choosing the best precision is circumvented by using RANSAC [10].

Let $u_i \leftrightarrow v_i$, $1 \leq i \leq 4$ be pairs of corresponding points in $\mathbb{P}^2$, such that no three of the $u_i$ are collinear and no three of the $v_i$ are collinear. There is a unique collineation $\omega$ such that $\omega(u_i) = v_i$, $1 \leq i \leq 4$. A coding of the $u_i$, $v_i$, $1 \leq i \leq 4$ yields a coding of $\omega$. Ideally, all quadruples $q_{i_j} \leftrightarrow q_{i_j}'$, $1 \leq j \leq 4$ of corresponding points should be examined to find the quadruple for which

$$|\text{code}(\omega).c(\epsilon).c(\delta)| \tag{3}$$

is a minimum. In practice there are too many quadruples, so a random selection of $N$ quadruples is made and (3) is minimised over the chosen quadruples.

An advantage of RANSAC is that the precision of $\omega$ is appropriate for the data; in addition, the code for $\omega$ is redundant because it includes the points $q_{i_j}$, $1 \leq j \leq 4$ already coded in $c(x).c(y)$. The redundancy is removed and compression achieved by omitting the $q_{i_j}$ from code($\omega$), and instead coding the index of the four-tuple $\iota = (i_1, i_2, i_3, i_4)$, $i_1 < i_2 < i_3 < i_4$ in the list of all ordered four-tuples with distinct entries drawn from $n$. The code length for $\iota$ is at most $\lceil \log(b(n,4)) \rceil + 1$ bits where $b(n,4)$ is the binomial coefficient.

Further compression of $t$ in (2) is achieved by omitting from $\epsilon$, $\delta$ the eight entries known to be zero, yielding the code $U_{\mathcal{C}}(s)$.

## 2.3 Affine fundamental matrix $A$

Let $A$ be an affine fundamental matrix, and let $l$ be the line $l' = q^\top A$. The geometrical interpretation of the equation $q^\top A q' = 0$ is that $q'$ lies on $l'$. If $q$, $A$ are given, then $q'$ can be coded by giving its position on $l'$. Compression is achieved because only one coordinate is needed rather than two.

As with $\mathcal{C}$, RANSAC is used to find a suitable matrix $A$ compatible with the $q_i \leftrightarrow q_i'$, $1 \le i \le n$. Let $u_i \leftrightarrow v_i$, $1 \le i \le 4$ be pairs of corresponding points. In general, there is a unique affine fundamental matrix $A$ such that $u_i^\top A v_i = 0$, $1 \le i \le 4$.

The point $q_i'$ is specified relative to an origin which depends on $i$, the $q_j$ and $A$. In detail, there is a three dimensional family of collineations which preserve the epipolar lines associated with $A$ in that if $\rho$ is any one of the collineations and $l$ is an epipolar line of $A$ in the first image, then $\rho(l)$ is the corresponding epipolar line in the second image [6]. The three dimensional family is spanned the collineations associated with any four linearly independent matrices $H$ for which

$$AH + H^\top A^\top = 0 \tag{4}$$

Let $q_{i_j} \leftrightarrow q_{i_j}'$, $1 \le j \le 4$ be the pairs of corresponding points which define $A$. From the $q_{i_j}$ select the three points $q_{i_j}$, $q_{i_k}$, $q_{i_l}$ which define a triangle with the greatest area. A unique collineation is specified by the matrix $H$ for which $Hq_{i_j} = q_{i_j}'$, $Hq_{i_k} = q_{i_k}'$, $Hq_{i_l} = q_{i_l}'$ and (4) holds.

Let $\nu_i$ be a unit vector in $\mathbb{R}^2$ parallel to $q_i^\top A$, let $\nu_i^\perp$ be a unit vector perpendicular to $\nu_i$, and define $r_i$, $s_i$ by $q_i' - Hq_i = (r_i\nu_i + s_i\nu_i^\perp, 1)$ as shown in Figure 1. Define integers $\epsilon_i$, $\delta_i$ by

$$(\epsilon_i, \delta_i) = (\lfloor 2r_i + 0.5 \rfloor, \lfloor 2s_i + 0.5 \rfloor) \tag{5}$$

The factor 2 on the right hand side of (5) is needed to avoid quantisation errors.

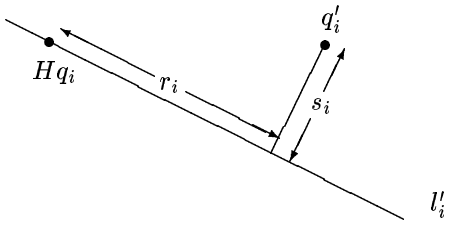


Figure 1. Definition of $r_i$, $s_i$.

The code for the $q_i \leftrightarrow q_i'$, $1 \le i \le n$ is

$$c(x).c(y).\text{code}(A).c(\epsilon).c(\delta) \tag{6}$$

The matrix $A$ is specified by giving the index of $\iota = (i_1, i_2, i_3, i_4)$ in the list of ordered four-tuples of distinct elements drawn from $n$. When coding $\delta$, the four entries known to be zero are omitted.

A random selection of $N$ quadruples is made and the length of the code (6) minimised over the quadruples. The code with the minimum length is $U_{\mathcal{A}}(s)$.

### 2.4 Fundamental matrix $\mathcal{F}$

The coding of $s$ as $U_{\mathcal{F}}(s)$ is similar to the coding as $U_{\mathcal{A}}(s)$, with one significant change, due to the fact that four pairs of image correspondences are not sufficient to specify a unique fundamental matrix. Let $q_{i_j} \leftrightarrow q_{i_j}'$, $1 \le j \le 7$ be seven pairs of corresponding points. There are in general at most three linearly independent rank two matrices $\tilde{F}$ such that $q_{i_j}^\top \tilde{F} q_{i_j}' = 0$, $1 \le j \le 7$. The matrix $\tilde{F}$ is specified by the seven pairs of corresponding points and a two bit code specifying one of the three possible matrices.

The matrix $\tilde{F}$ is replaced by a rational approximation $F$, retaining the constraint $\det(F) = 0$. Let $\tilde{u}$ be the eigenvector of $\tilde{F}^\top \tilde{F}$ with the least eigenvalue, let $u$ be a rational approximation to $\tilde{u}$ and let $G$ be a rational approximation to $\tilde{F}$. The matrix $F$ is defined by $F = G - (u.u)^{-1} Gu \otimes u$.

## 3 Code for Vectors in $Z^n$

If $c_1, \ldots, c_p$ are different codes for vectors $x \in Z^n$, then a new code $c$ can be constructed by first finding the index $j$ such that $|c_j(x)| = \min\{|c_i(x)|, 1 \le i \le p\}$ and then setting $c(x) = d(j, b).c_j(x)$, where $d(j, b)$ is a code of fixed length $b$ for the integer $j$. If $p$ is small and $n$ is large, then $c$ may give shorter average code lengths than any single code $c_i$.

The code $c$ in §2 is constructed from four separate codes $c_1$, $c_2$, $c_3$, $c_4$, which are described in turn.

### 3.1 Codes $c_1$ and $c_2$

Let $C_\sigma$ be defined for $\sigma \in \mathbb{N}$ by

$$C_\sigma = \{x \in Z^n, |x_i| \leq \sigma, 1 \leq i \leq n\}$$

The set $C_\sigma$ contains $(2\sigma+1)^n$ points. The elements of $C_\sigma$ are enumerated by a function $\zeta_\sigma : C_\sigma \to \mathbb{N}^+$ constructed such that $\zeta_0(0) = 1$, and such that $\zeta_\sigma$ is an extension of $\zeta_{\sigma-1}$ for $\sigma \geq 1$. The functions $\zeta_\sigma$, $\sigma \geq 0$ together define a single function $\zeta : Z^n \to \mathbb{N}^+$.

Let $m$ be the median of the $x_i$ and let $v$ be the vector with components $v_i = x_i - m$, $1 \leq i \leq n$. The codes $c_1$, $c_2$ are defined by $c_1(x) = r(\zeta(x))$ and $c_2(x) = r(\mathrm{zton}(m)).c_1(v)$.

### 3.2 Code $c_3$

Let $\sigma = \mathrm{abs}(x_j)$ for some $j$, let $u$ be the vector of components $x_i$ such that $\mathrm{abs}(x_i) \leq \sigma$, and let $v$ be the vector of components $x_i - \mathrm{sign}(x_i)\sigma$ for those $i$ such that $|x_i| > \sigma$. Let $w_\sigma \in \{0,1\}^n$ be defined such that $(w_\sigma)_i = 1$ if $x_i$ is a component of $u$ and $(w_\sigma)_i = 0$ if $x_i - \mathrm{sign}(x_i)\sigma$ is a component of $v$.

The code $d_\sigma$ is defined by $d_\sigma(x) = w_\sigma.c_1(u).c_1(v)$. Let $\theta$ be the value of $\sigma$ at which $|d_\sigma(x)|$ is a minimum over all the distinct values of $\sigma = \mathrm{abs}(x_i)$, $1 \leq i \leq n$, $|d_\theta(x)| = \min_\sigma\{|d_\sigma(x)|\}$. The code $c_3$ is defined by $c_3(x) = d_\theta(x)$.

### 3.3 Code $c_4$

Let $m$ be the median of $x$, let $y_1, \ldots, y_p$ be the distinct integers appearing in the set $\{x_i - m, 1 \leq i \leq n\}$, and let $y_i$ occur $k_i$ times, $1 \leq i \leq p$. Let $S$ be the set of all permutations of $x$. The number $b$ of elements of $S$ is given by the multinomial $b = n!/(k_1! \ldots k_m!)$. The elements of $S$ are ordered in any convenient way. Let $\iota$ be the index of $x$ in the chosen order, let $y$, $k$ be the vectors with the respective components $y_i$, $k_i$, and let $m_k$ be median of $k$. The code $c_4$ is defined by

$$\begin{aligned} f(x) &= e(k_1 - m_k).e(y_1). \ldots .e(k_p - m_k).e(y_p) \\ c_4(x) &= e(m).e(m_k).f(x).d(\iota, b) \end{aligned}$$

If the $x_i$ are independent realisations of a random variable and $n$ is large, then $|c_4(x)|/n$ is, with a high probability, close to the shortest possible expected length of a code word. See for example [4], §1.11.4. In practice the effectiveness of $c_4$ is reduced because of the extra code needed for $y$, $k$.

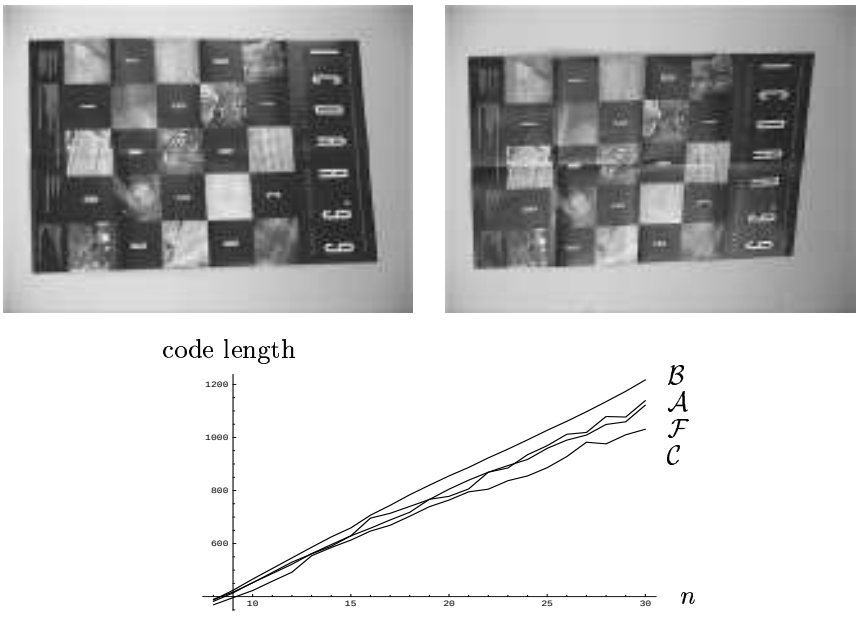


Figure 2. Images and code lengths when the 'true' model is $\mathcal{C}$

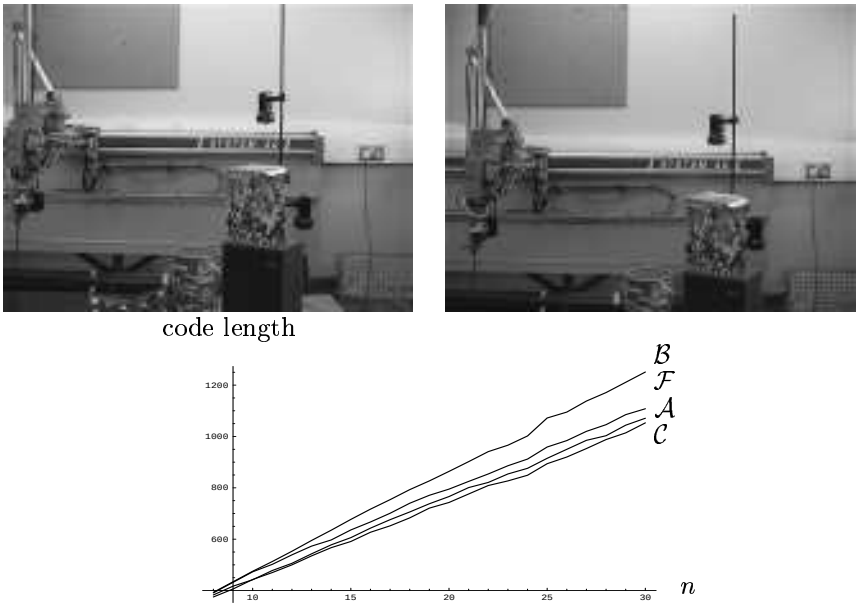


Figure 3. Images and code lengths when the 'true' model is $\mathcal{A}$

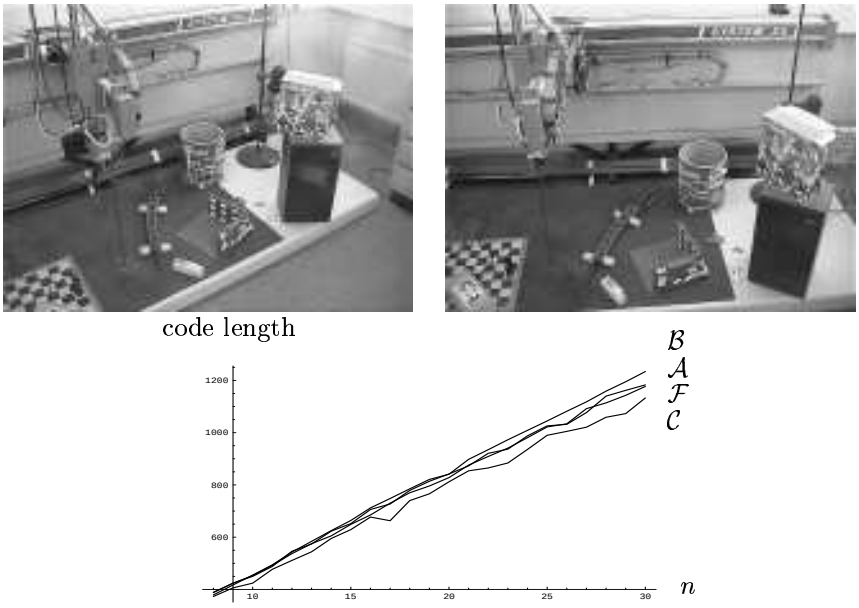


Figure 4. Images and code lengths when the 'true' model is $\mathcal{F}$

# 4 Experiments

Images of a laboratory were taken by a Canon MV-1 Camcorder mounted on a tripod. Typical pairs of images are shown in Figures 2, 4, 6 with the 'true' models shown in the captions. In each case the 'true' model is known, but only because of the prior information available to a human observer. The task of the program is to make the best choice of model using only the data $q_i \leftrightarrow q_i'$, $1 \leq i \leq n$ and the models $\mathcal{B}$, $\mathcal{C}$, $\mathcal{A}$, $\mathcal{F}$. This best choice can and does differ from the 'true' model.

The size of the original images in pixels is $640 \times 480$. Feature points were located in each image and matched to obtain pairs of corresponding points $q_i \leftrightarrow q_i'$, $1 \leq i \leq n$. The graphs of code length against $n$ for $8 \leq n \leq 30$ are shown in Figures 3, 5, 7. The maximum number on the vertical scale is 1200 bits, and the spacing between numbers is 200 bits. The number of random samples in the RANSAC algorithm was $N = 10$.

It is apparent from the graphs that $\mathcal{C}$ is always the preferred model even when the 'true' model is $\mathcal{F}$. The models $\mathcal{A}$, $\mathcal{F}$ show a similar performance, and $\mathcal{B}$ is always the worst model.

# 5 Discussion

The experiments show that the collineation model $\mathcal{C}$ is a good choice even for sets of image correspondences for which the 'true' model is a fundamental matrix. Why does the model $\mathcal{F}$ perform so badly under MDL? The reason can be seen in Figure 1. In the usual methods for assessing the fit of the model $\mathcal{F}$ to the data, the error measure is the sum of the squares of the $s_i$, and the numbers $r_i$, measuring distance along the epipolar lines, are ignored. In MDL the numbers $r_i$ must be included, in order to obtain a loss free coding of the data. The extra code length needed for the $r_i$ reduces the preference for $\mathcal{F}$, so much so that in these experiments $\mathcal{C}$ is almost always preferred.

In practice it is necessary to make a boundary between the models $\mathcal{C}$ and $\mathcal{F}$. If the scene points are close to a plane, then $\mathcal{C}$ is chosen. As the scene points diverge from a coplanar configuration, $\mathcal{C}$ becomes less likely and the model $\mathcal{F}$ becomes more likely. These experiments indicate that the boundary favours $\mathcal{C}$ much more than has been previously supposed and that a very large deviation from coplanarity is needed before $\mathcal{F}$ becomes the best choice.

# References


[1] O.D. Faugeras, *Three-Dimensional Computer Vision: a geometric viewpoint.* MIT Press, 1993.

[2] Y.G. Leclerc, "Constructing simple stable descriptions for image partitioning," *International Journal of Computer Vision*, Vol. 3, pp. 73-102, 1989.

[3] M. Li and P.M.B. Vitányi, "Inductive reasoning and Kolmogorov complexity," *J. of Computer and System Sciences*, Vol. 44, pp. 343-384, 1982.

[4] M. Li and P. M. B. Vitányi, *An Introduction to Kolmogorov Complexity and Its Applications*, Graduate Texts in Computer Science, Springer, 2nd edition, 1997.

[5] T. Lindeberg and M.-X. Li, "Segmentation and classification of edges using minimum description length approximation and complementary junction clues," *Computer Vision and Image Understanding*, Vol. 67, pp. 88-98, 1997.

[6] Q.-T. Luong and T. Viéville, "Canonic representations for the geometries of multiple projective views," In J.-O. Eklundh (ed.) *Computer Vision-ECCV'94, Vol I*, Lecture Notes in Computer Science, Vol. 800, pp. 589-599, Springer, 1994.

[7] S.J. Maybank and R. Fraile, "Minimum description length method for facet matching," *Proc. International Symposium on Multispectral Image Processing, ISMIP'98*, SPIE Vol. 3545, Wuhan, China, pp. 330-335, 1998.

[8] L.S. Shapiro, A. Zisserman and M. Brady "Motion from point matches using affine epipolar geometry". In J.-O. Eklundh (ed.) *Computer Vision - ECCV'94, Vol. II*, Lecture Notes in Computer Science, Vol. 801, pp. 73-84, Springer, 1994.

[9] P. H. S. Torr and A. Zisserman, "Concerning Bayesian motion segmentation, model averaging, matching and the trifocal tensor," In H. Burkhardt and B. Neumann (eds.) *Computer Vision - ECCV'98, Vol. I*, Lecture Notes in Computer Science, Vol. 1406, pp. 511-527, Springer, 1998.

[10] P.H.S. Torr, A. Zisserman and S.J. Maybank, "Robust detection of degenerate configurations whilst estimating the fundamental matrix," *Computer Vision, Graphics, and Image Processing*, Vol. 71, pp. 312-333, 1998.

[11] C.S. Wallace and P.R. Freeman, "Estimation and inference by compact coding," *J. Royal Stat. Soc. Series B*, Vol. 49, pp. 240-265, 1987.

[12] S. Wolfram, *The Mathematica Book*, 3rd Edition, Cambridge University Press, Cambridge, 1996.