Reconnaissance des Formes et Intelligence Artificielle (RFIA 2004), 28–30 January 2004, Toulouse, France.

Uncertainty Modeling and Geometric Inference

Kenichi Kanatani

Department of Information Technology, Okayama University, Okayama 700-8530 Japan
kanatani@suri.it.okayama-u.ac.jp

Abstract

We investigate the meaning of "statistical methods" for geometric inference based on image feature points. Tracing back the origin of feature uncertainty to image processing operations, we discuss the implications of asymptotic analysis in reference to "geometric fitting" and "geometric model selection". We point out that a correspondence exists between the standard statistical analysis and the geometric inference problem. We also compare the capability of the "geometric AIC" and the "geometric MDL" in detecting degeneracy.

1. Introduction

Statistical inference from images is one of the key components of computer vision research today. Traditionally, statistical methods have been used for recognition and classification purposes. Recently, however, there are many studies of statistical analysis for *geometric inference* based on geometric primitives such as points and lines extracted by image processing operations.

However, the term "statistical" has somewhat a different meaning for such geometric inference problems than for the traditional recognition and classification purposes. This difference has often been overlooked, causing controversies over the validity of the statistical approach to geometric problems in general. In Sec. 2, we take a close look at this problem, tracing back the origin of feature uncertainty to image processing operations. In Sec. 3, we discuss the implications of asymptotic analysis in reference to "geometric fitting" and "geometric model selection". In Sec. 4, we point out that a correspondence exists between the standard statistical analysis and the geometric inference problem. We also compare the capability of the "geometric AIC" and the "geometric MDL" in detecting degeneracy. Sec. 5 presents our concluding remarks.

2. What is Geometric Inference?

2.1 Ensembles for geometric inference

The goal of statistical methods is not to study the properties of observed data themselves but to infer the properties of the *ensemble* from which we regard the observed data as sampled. The ensemble may be a collection of existing entities (e.g., the entire population), but often it is a hypothetical set of conceivable possibilities. When a statistical method is employed, the underlying ensemble is often taken for granted. However, this issue is very crucial for geometric inference based on feature points.

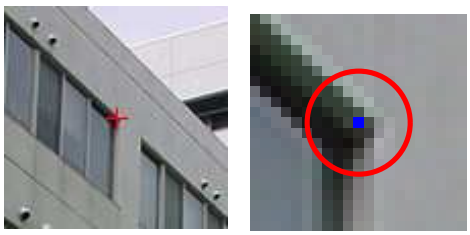
Suppose, for example, we extract feature points, such as corners of walls and windows, from an image of a building and want to test if they are collinear. The reason why we need a statistical method is that the extracted feature positions have uncertainty. So, we have to judge the extracted feature points as collinear if they are sufficiently aligned. We can also evaluate the degree of uncertainty of the fitted line by propagating the uncertainty of the individual points. What is the ensemble that underlies this type of inference?

This question reduces to the question of why the uncertainty of the feature points occurs at all. After all, statistical methods are not necessary if the data are exact. Using a statistical method means regarding the current feature position as sampled from a set of its possible positions. But where else could it be if not in the current position?

2.2 Uncertainty of feature extraction

Many algorithms have been proposed for extracting feature points including the Harris operator [7] and SUSAN [33], and their performance has been extensively compared [3, 28, 32]. However, if we use, for example, the Harris operator to extract a particular corner of a particular building image, the output is unique (Fig. 1). No matter how many times we repeat the extraction, we obtain the same point because no external disturbances exist and the internal parameters (e.g., thresholds for judgment) are unchanged. It follows that the current position is the sole possibility. How can we find it elsewhere?

If we closely examine the situation, we are compelled to conclude that other possibilities should exist because the extracted position is not necessarily correct. But if it is not correct, why did we extract it? Why didn't we extract the correct position in the first place? The answer is: *we cannot*.



(a) (b)

Figure 1: (a) A feature point in an image of a building. (b) Its enlargement and the uncertainty of the feature location

2.3 Image processing for computer vision

The reason why there exist so many feature extraction algorithms, none of them being definitive, is that they are aiming at an intrinsically impossible task. If we were to extract a point around which, say, the intensity varies to the largest degree in such and such a measure, the algorithm would be unique; variations may exist in intermediate steps, but the final output should be the same.

However, what we want is not "image properties" but "3-D properties" such as corners of a building, but the way a 3-D property is translated into an image property is intrinsically heuristic. As a result, as many algorithms can exist as the number of heuristics for its 2-D interpretation. If we specify a particular 3-D feature to extract, say a corner of a window, its appearance in the image is not unique. It is affected by many properties of the scene including the details of its 3-D shape, the viewing orientation, the illumination condition, and the light reflectance properties of the material. A slight variation of any of them can result in a substantial difference in the image.

Theoretically, exact extraction would be possible if all the properties of the scene were exactly known, but to infer them from images is the very task of computer vision. It follows that we must make a guess in the image processing stage. For the current image, some guesses may be correct, but others may be wrong. The exact feature position could be found only by an (non-existing) "ideal" algorithm that could guess everything correctly.

This observation allows us to interpret the "possible feature positions" to be *the positions that would be located by different (non-ideal) algorithms based on different guesses*. It follows that the set of hypothetical positions should be associated with *the set of hypothetical algorithms*. The current position is regarded as produced by an algorithm sampled from it. This explains why one always obtains the same position no matter how many times one repeats extraction using that algorithm. To obtain a different position, one has to sample another algorithm.

Remark 1 We may view the statistical ensemble in the following way. If we repeat the *same* experiment, the result should always be the same. But if we declare that the experiment is the "same" if such and such are the same while other things can vary; those variable conditions define the ensemble. The conventional view is to regard the experiment as the same if the *3-D scene* we are viewing is the same while other properties, such as the lighting condition, can vary. Then, the resulting image would be different for each (hypothetical) experiment, so one would obtain a different output each time, using the same image processing algorithm. The expected spread of the outputs measures the robustness of that algorithm. Here, however, we are viewing the experiment as the same *if the image is the same*. Then, we could obtain different results only by sampling other algorithms. The expected spread of the outputs measures the uncertainty of feature detection from *that image*. We take this view, because we are analyzing the reliability of geometric inference from a particular image, while the conventional view is suitable for assessing the robustness of a *particular algorithm*.

2.4 Covariance matrix of a feature point

The performance of feature point extraction depends on the image properties around that point. If, for example, we want to extract a point in a region with an almost homogeneous intensity, the resulting position may be ambiguous whatever algorithm is used. In other words, the positions that potential algorithms would extract should have a large spread. If, on the other hand, the intensity greatly varies around that point, any algorithm could easily locate it accurately, meaning that the positions that the hypothetical algorithms would extract should have a strong peak. It follows that we may introduce for each feature point its *covariance matrix* that measures the spread of its potential positions.

Let $V[p_\alpha]$ be the covariance matrix of the αth feature point p_α. The above argument implies that we can estimate the qualitative characteristics of uncertainty but not its absolute magnitude. So, we write the covariance matrix $V[p_\alpha]$ in the form

$$V[p_\alpha] = \varepsilon^2 V_0[p_\alpha], \tag{1}$$

where ε is an unknown magnitude of uncertainty, which we call the *noise level*. The matrix $V_0[p_\alpha]$, which we call the (*scale*) *normalized covariance matrix*, describes the relative magnitude and the dependence on orientations.

Remark 2 The decomposition of $V[p_\alpha]$ into ε^2 and $V_0[p_\alpha]$ involves scale ambiguity. We assume that the decomposition is made unique by an appropriate scale normalization such as $\mathrm{tr}V_0[p_\alpha] = 2$. However, the subsequent analysis does not depend on particular normalizations, so we do not explicitly specify it except that it should be done in such a way that ε is

much smaller than the data themselves. Note that mathematically, modeling the covariance matrix by a common scale factor ε^2 and the individual matrix part $V_0[p_\alpha]$ is rather restrictive. However, this model is sufficient for most practical applications, as we describe in the following.

2.5 Covariance matrix estimation

If the intensity variations around p_α are almost the same in all directions, we can think of the probability distribution as isotropic, a typical equiprobability line, known as the *uncertainty ellipses*, being a circle (Fig. 1(b)).

On the other hand, if p_α is on an object boundary, distinguishing it from nearby points should be difficult whatever algorithm is used, so its covariance matrix should have an elongated uncertainty ellipse along that boundary.

However, existing feature extraction algorithms are usually designed to output those points that have large image variations around them, so points in a region with an almost homogeneous intensity or on object boundaries are rarely chosen. As a result, the covariance matrix of a feature point extracted by such an algorithm can be regarded as nearly isotropic. This has also been confirmed by experiments [21], justifying the use of the identity as the normalized covariance matrix $V_0[p_\alpha]$.

Remark 3 The intensity variations around different feature points are usually unrelated, so their uncertainty can be regarded as statistically independent. However, if we track feature points over consecutive video frames, it has been observed that the uncertainty has strong correlations over the frames [34].

Remark 4 Many interactive applications require humans to extract feature points by manipulating a mouse. Extraction by a human is also an "algorithm", and it has been shown by experiments that humans are likely to choose "easy-to-see" points such as isolated points and intersections, avoiding points in a region with an almost homogeneous intensity or on object boundaries [21]. In this sense, the statistical characteristics of human extraction are very similar to machine extraction. This is no surprise if we recall that image processing for computer vision is essentially a heuristic that simulates human perception. It has also been reported that strong microscopic correlations exist when humans manually select corresponding feature points over multiple images [26].

2.6 Image quality and uncertainty

The uncertainty of feature points has often been identified with "image noise", giving a misleading impression as if the feature locations were perturbed by random intensity fluctuations. Of course, we may obtain better results using higher-quality images whatever algorithm is used. However, the task of computer vision is not to analyze "image properties" but to study the "3-D properties" of the scene. As long as the image properties and the 3-D properties do not correspond one to one, any image processing inevitably entails some degree of uncertainty, however high the image quality may be, and the result must be interpreted statistically. The underlying ensemble is the set of hypothetical (inherently imperfect) algorithms of image processing. Yet, the performance of image processing algorithms has often been evaluated by adding *independent Gaussian noise* to individual pixels.

Remark 5 This also applies to *edge detection*, whose goal is to find the boundaries of 3-D objects in the scene. In reality, all existing algorithms seek *edges*, i.e., lines and curves across which the intensity changes discontinuously. Yet, this is regarded by many as an objective image processing task, and the detection performance is often evaluated by adding independent Gaussian noise to individual pixels. From the above considerations, we conclude that edge detection is also a heuristic and hence no definitive algorithm will ever be found.

3. Asymptotic Analysis

3.1 What is asymptotic analysis?

As stated earlier, *statistical estimation* refers to estimating the properties of an ensemble from a finite number of samples, assuming some knowledge, or a *model*, about the ensemble.

If the uncertainty originates from external conditions, as in experiments in physics, the estimation accuracy can be increased by controlling the measurement devices and environments. For internal uncertainty, on the other hand, there is no way of increasing the accuracy except by repeating the experiment and doing statistical inference. However, repeating experiments usually entails costs, and in practice the number of experiments is often limited.

Taking account of this, statisticians usually evaluate the performance of estimation *asymptotically*, analyzing the growth in accuracy as the number n of experiments increases. This is justified because a method whose accuracy increases more rapidly as $n \to \infty$ can reach admissible accuracy *with a fewer number of experiments* (Fig. 2(a)).

In contrast, the ensemble for geometric inference is, as we have seen, the set of potential feature positions that could be located if other (hypothetical) algorithms were used. As noted earlier, however, we can choose only *one* sample from the ensemble as long as we use a particular image processing algorithm. In

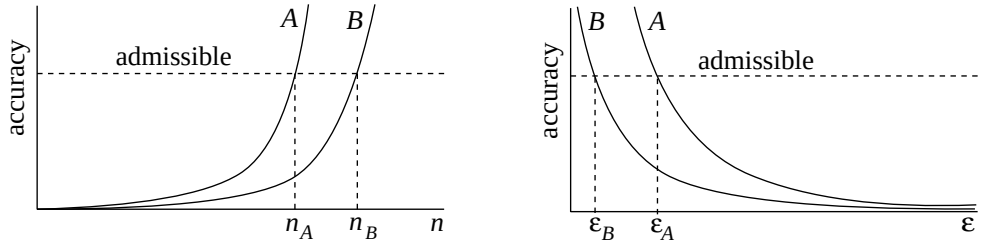


Figure 2: (a) For the standard statistical analysis, it is desired that the accuracy increases rapidly as the number of experiments $n \to \infty$, because admissible accuracy can be reached with a smaller number of experiments. (b) For geometric inference, it is desired that the accuracy increases rapidly as the noise level $\varepsilon \to 0$, because larger data uncertainty can be tolerated for admissible accuracy.

other words, the number n of experiments is 1. Then, how can we evaluate the performance of statistical estimation?

Evidently, we want a method whose accuracy is sufficiently high *even for large data uncertainty*. This implies that we need to analyze the growth in accuracy as the noise level ε decreases, because a method whose accuracy increases more rapidly as $\varepsilon \to 0$ can tolerate larger data uncertainty for admissible accuracy (Fig. 2(b)).

3.2 Geometric fitting

We now illustrate the above consideration in more specific terms. Let $\{p_\alpha\}$, $\alpha = 1, ..., N$, be the extracted feature points. Suppose each point should satisfy a parameterized constraint

$$F(p_\alpha, \boldsymbol{u}) = 0 \tag{2}$$

when no uncertainty exists. In the presence of uncertainty, eq. (2) may not hold exactly. Our task is to estimate the parameter $\boldsymbol{u}$ from observed positions $\{p_\alpha\}$ in the presence of uncertainty.

A typical problem of this form is to fit a line or a curve to given N points in the image, but this can be straightforwardly extended to multiple images. For example, if a point (x_α, y_α) in one image corresponds to a point (x'_α, y'_α) in another, we can regard them as a single point p_α in a 4-dimensional joint space with coordinates $(x_\alpha, y_\alpha, x'_\alpha, y'_\alpha)$. If the camera imaging geometry is modeled as perspective projection, the constraint (2) corresponds to the *epipolar equation*; the parameter $\boldsymbol{u}$ is the *fundamental matrix* [8].

3.2.1 General geometric fitting

The above problem can be stated in abstract terms as *geometric fitting* as follows. We view a feature point in the image plane or a set of feature points in the joint space as an m-dimensional vector $\boldsymbol{x}$; we call it a "datum". Let $\{\boldsymbol{x}_\alpha\}$, $\alpha = 1, ..., N$, be observed data. Their true values $\{\bar{\boldsymbol{x}}_\alpha\}$ are supposed to satisfy r constraint equations

$$F^{(k)}(\bar{\boldsymbol{x}}_\alpha, \boldsymbol{u}) = 0, \quad k = 1, ..., r, \tag{3}$$

parameterized by a p-dimensional vector $\boldsymbol{u}$. We call eq. (3) the (*geometric*) *model*. The domain $\mathcal{X}$ of the data $\{\boldsymbol{x}_\alpha\}$ is called the *data space*; the domain $\mathcal{U}$ of the parameter $\boldsymbol{u}$ is called the *parameter space*. The number r of the constraint equations is called the *rank* of the constraint. The r equations $F^{(k)}(\boldsymbol{x}, \boldsymbol{u}) = 0$, $k = 1, ..., r$, are assumed to be mutually independent, defining a manifold $\mathcal{S}$ of codimension r parameterized by $\boldsymbol{u}$ in the data space $\mathcal{X}$. Eq. (3) requires that the true values $\{\bar{\boldsymbol{x}}_\alpha\}$ be all in the manifold $\mathcal{S}$. Our task is to estimate the parameter $\boldsymbol{u}$ from the noisy data $\{\boldsymbol{x}_\alpha\}$ (Fig. 3(a)).

3.2.2 Maximum likelihood estimation

Let

$$V[\boldsymbol{x}_\alpha] = \varepsilon^2 V_0[\boldsymbol{x}_\alpha] \tag{4}$$

be the covariance matrix of $\boldsymbol{x}_\alpha$, where ε and $V_0[\boldsymbol{x}_\alpha]$ are the noise level and the normalized covariance matrix, respectively. If the distribution of uncertainty is Gaussian, which we assume hereafter, the probability density of the data $\{\boldsymbol{x}_\alpha\}$ is given by

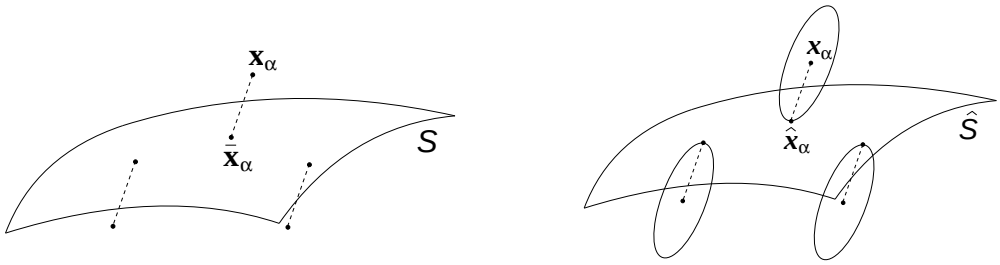
$$P(\{\boldsymbol{x}_\alpha\}) = C \prod_{\alpha=1}^{N} \mathrm{e}^{-(\boldsymbol{x}_\alpha - \bar{\boldsymbol{x}}_\alpha, V[\boldsymbol{x}_\alpha]^{-1}(\boldsymbol{x}_\alpha - \bar{\boldsymbol{x}}_\alpha))/2}, \tag{5}$$

where C is a normalization constant. Throughout this paper, we denote the inner product of vectors $\boldsymbol{a}$ and $\boldsymbol{b}$ by $(\boldsymbol{a}, \boldsymbol{b})$.

Maximum likelihood estimation (*MLE*) is to find the values of $\{\bar{\boldsymbol{x}}_\alpha\}$ and $\boldsymbol{u}$ that maximize the *likelihood*, i.e., eq. (6) into which the data $\{\boldsymbol{x}_\alpha\}$ are substituted, or equivalently minimize the sum of the squared *Mahalanobis distances* in the form

$$J = \sum_{\alpha=1}^{N} (\boldsymbol{x}_\alpha - \bar{\boldsymbol{x}}_\alpha, V_0[\boldsymbol{x}_\alpha]^{-1}(\boldsymbol{x}_\alpha - \bar{\boldsymbol{x}}_\alpha)) \tag{6}$$

subject to the constraint (3) (Fig. 3(b)). The solution is called the *maximum likelihood* (*ML*) *estimator*. If the uncertainty is small, which we assume hereafter, the constraint (3) can be eliminated by introducing Lagrange multipliers and applying first order approximation. After some manipulations, we obtain the



(a) (b)

Figure 3: (a) Fitting a manifold $\mathcal{S}$ to the data $\{\boldsymbol{x}_\alpha\}$. (b) Estimating $\{\bar{\boldsymbol{x}}_\alpha\}$ and $\boldsymbol{u}$ by minimizing the sum of squared Mahalanobis distance with respect to the normalized covariance matrices $V_0[\boldsymbol{x}_\alpha]$.

following form [9]:

$$J = \sum_{\alpha=1}^{N} \sum_{k,l=1}^{r} W_\alpha^{(kl)} F^{(k)}(\boldsymbol{x}_\alpha, \boldsymbol{u}) F^{(l)}(\boldsymbol{x}_\alpha, \boldsymbol{u}). \quad (7)$$

Here, $W_\alpha^{(kl)}$ is the (kl) element of the inverse of the $r \times r$ matrix whose (kl) element is $(\nabla_{\boldsymbol{x}} F_\alpha^{(k)}, V_0[\boldsymbol{x}_\alpha] \nabla_{\boldsymbol{x}} F_\alpha^{(l)})$; we symbolically write

$$\left(W_\alpha^{(kl)}\right) = \left((\nabla_{\boldsymbol{x}} F_\alpha^{(k)}, V_0[\boldsymbol{x}_\alpha] \nabla_{\boldsymbol{x}} F_\alpha^{(l)})\right)^{-1}, \quad (8)$$

where $\nabla_{\boldsymbol{x}} F^{(k)}$ is the gradient of the function $F^{(k)}$ with respect to $\boldsymbol{x}$. The subscript α means that $\boldsymbol{x} = \boldsymbol{x}_\alpha$ is substituted.

Remark 6 The data $\{\boldsymbol{x}_\alpha\}$ may be subject to some constraints. For example, each $\boldsymbol{x}_\alpha$ may be a unit vector. The above formulation still holds if the inverse $V_0[\boldsymbol{x}_\alpha]^{-1}$ in eq. (6) is replaced by the (Moore-Penrose) generalized (or pseudo) inverse $V_0[\boldsymbol{x}_\alpha]^-$ [9]. Similarly, the r constraints in eq. (3) may be redundant, say only r' $(< r)$ of them are independent. The above formulation still holds if the inverse in eq. (8) is replaced by the generalized inverse of rank r' with all but r' largest eigenvalues are replaced by zero [9].

3.2.3 Accuracy of the ML estimator

It can be shown [9] that the covariance matrix of the ML estimator $\hat{\boldsymbol{u}}$ has the form

$$V[\hat{\boldsymbol{u}}] = \varepsilon^2 \boldsymbol{M}(\hat{\boldsymbol{u}})^{-1} + O(\varepsilon^4), \quad (9)$$

where

$$\boldsymbol{M}(\boldsymbol{u}) = \sum_{\alpha=1}^{N} \sum_{k,l=1}^{r} W_\alpha^{(kl)} \nabla_{\boldsymbol{u}} F_\alpha^{(k)} \nabla_{\boldsymbol{u}} F_\alpha^{(k)\top}. \quad (10)$$

Here, $\nabla_{\boldsymbol{u}} F^{(k)}$ is the gradient of the function $F^{(k)}$ with respect to $\boldsymbol{u}$. The subscript α means that $\boldsymbol{x} = \boldsymbol{x}_\alpha$ is substituted.

Remark 7 It can be proved that no other estimators could reduce the covariance matrix further than eq. (9) except for the higher order term $O(\varepsilon^4)$ [9, 12]. The ML estimator is optimal in this sense. Recall that we are focusing on the asymptotic analysis for $\varepsilon \to 0$. Thus, what we call the "ML estimator" should be understood to be a first approximation to the true ML estimator for small ε.

Remark 8 The p-dimensional parameter vector $\boldsymbol{u}$ may be constrained. For example, it may be a unit vector. If it has only p' $(< p)$ degrees of freedom, the parameter space $\mathcal{U}$ is a p'-dimensional manifold in $\mathcal{R}^p$. In this case, the matrix $\boldsymbol{M}(\boldsymbol{u})$ in eq. (9) is replaced by $\boldsymbol{P}_{\boldsymbol{u}} \boldsymbol{M}(\boldsymbol{u}) \boldsymbol{P}_{\boldsymbol{u}}$, where $\boldsymbol{P}_{\boldsymbol{u}}$ is the projection matrix onto the tangent space to the parameter space $\mathcal{U}$ at $\boldsymbol{u}$ [9]. The inverse $\boldsymbol{M}(\hat{\boldsymbol{u}})^{-1}$ in eq. (9) is replaced by the generalized inverse $\boldsymbol{M}(\hat{\boldsymbol{u}})^{-1}$ of rank p' [9].

3.3 Geometric model selection

Geometric fitting is to estimate the parameter $\boldsymbol{u}$ of a a given model. If we have multiple candidate models

$$F_1^{(k)}(\bar{\boldsymbol{x}}_\alpha, \boldsymbol{u}_1) = 0, \quad F_2^{(k)}(\bar{\boldsymbol{x}}_\alpha, \boldsymbol{u}_2) = 0, \quad ..., \quad (11)$$

from which we are to select an appropriate one for the observed data $\{\boldsymbol{x}_\alpha\}$, the problem is (*geometric*) *model selection* [9, 11, 13].

Suppose, for example, we want to fit a curve to given points in two dimensions. If they are almost collinear, a straight line may fit fairly well, but a quadratic curve may fit better, and a cubic curve even better. Which curve should we fit? A naive idea is to compare the *residual* (*sum of squares*), i.e., the minimum value $\hat{J}$ of J in eq. (6); we select the one that has the smallest residual $\hat{J}$. This does not work, however, because the ML estimator $\hat{\boldsymbol{u}}$ is so determined as to minimize the residual $\hat{J}$, and the residual $\hat{J}$ can be made arbitrarily smaller if the model is equipped with more parameters to adjust. So, the only conclusion would be to fit a curve of a sufficiently high degree passing through all the points.

3.3.1 Geometric AIC

The above observation leads to the idea of compensating for the negative bias of the residual caused

by substituting the ML estimator. This is the principle of Akaike's *AIC* (*Akaike information criterion*) [1], which is derived from the asymptotic behavior of the *Kullback-Leibler information* (or *divergence*) as the number n of experiments goes to infinity. Doing a similar analysis to Akaike's and examining the asymptotic behavior as the noise level ε goes to zero, we can obtain the following *geometric AIC* [9, 10]:

$$\text{G-AIC} = \hat{J} + 2(Nd + p)\varepsilon^2 + O(\varepsilon^4). \quad (12)$$

Here, d is the dimension of the manifold $\mathcal{S}$ defined by the constraint (3) in the data space $\mathcal{X}$, and p is the dimension of $\boldsymbol{u}$ (i.e., the number of unknowns). The model for which eq. (12) is the smallest is regarded as the best. The derivation of eq. (12) is based on the following facts [9, 10]:

- The ML estimator $\hat{\boldsymbol{u}}$ converges to its true value as $\varepsilon \to 0$.
- The ML estimator $\hat{\boldsymbol{u}}$ obeys a Gaussian distribution under linear constraints, because the noise is assumed to be Gaussian. For nonlinear constraints, linear approximation can be justified in the neighborhood of the solution if ε is sufficiently small.
- A quadratic form in standardized Gaussian random variables is subject to a χ^2 distribution, whose expectation is equal to its degree of freedom.

3.3.2 Geometric MDL

Another well known criterion for model selection is Rissanen's *MDL* (*Minimum description length*) [29, 30, 31], which measures the goodness of a model by the minimum information theoretic code length of the data and the model. The basic idea is simple, but the following difficulties must be resolved for applying it in practice:

- Encoding a problem involving real numbers requires an infinitely long code length.
- The probability density, from which a minimum length code can be obtained, involves unknown parameters.
- The exact form of the minimum code length is very difficult to compute.

Rissanen [29, 30, 31] avoided these difficulties by quantizing the real numbers in a way that does not depend on individual models and substituting the ML estimators for the parameters. They, too, are real numbers, so they are also quantized. The quantization width is so chosen as to minimize the total description length (the *two-stage encoding*). The resulting code length is evaluated asymptotically as the data length n goes to infinity. If we analyzes the asymptotic behavior of encoding the geometric fitting problem as the noise level ε goes to zero, we obtain the following *geometric MDL* [15]:

$$\text{G-MDL} = \hat{J} - (Nd + p)\varepsilon^2 \log\Big(\frac{\varepsilon}{L}\Big)^2 + O(\varepsilon^2). \quad (13)$$

Here, L is a reference length chosen so that its ratio to the magnitude of data is $O(1)$, e.g., L can be taken to be the image size for feature point data. Its exact determination requires an a priori distribution that specifies where the data are likely to appear (we will discuss this more in Sec. 4.1), but it has been observed that the model selection is not very much affected by L as long as it is within the same order of magnitude [15].

4. Statistical vs. Geometric Inference

We now point out that a correspondence exists between the standard statistical analysis and the geometric inference problem. We also compare the capability of the geometric AIC and the geometric MDL in detecting degeneracy.

4.1 Standard statistical analysis

The asymptotic analysis in Sec. 3 bears a strong resemblance to the standard statistical estimation problem: after observing n data $\boldsymbol{x}_1$, $\boldsymbol{x}_2$, ..., $\boldsymbol{x}_n$, we want to estimate the parameter $\boldsymbol{\theta}$ of the probability density $P(\boldsymbol{x}|\boldsymbol{\theta})$ called the (*stochastic*) *model*, according to which each datum is assumed to be sampled independently.

Maximum likelihood estimation (*MLE*) is to find the value $\boldsymbol{\theta}$ that maximizes $\prod_{i=1}^n P(\boldsymbol{x}_i|\boldsymbol{\theta})$, or equivalently minimizes its negative logarithm $-\sum_{i=1}^n \log P(\boldsymbol{x}_i|\boldsymbol{\theta})$. It can be shown that the covariance matrix $V[\hat{\boldsymbol{\theta}}]$ of the resulting ML estimator $\hat{\boldsymbol{\theta}}$ converges, under a mild condition, to $\boldsymbol{O}$ as the number n of experiments goes to infinity (*consistency*) in the form

$$V[\hat{\boldsymbol{\theta}}] = \boldsymbol{I}(\boldsymbol{\theta})^{-1} + O\Big(\frac{1}{n^2}\Big), \quad (14)$$

where we define the *Fisher information matrix* $\boldsymbol{I}(\boldsymbol{\theta})$ by

$$\boldsymbol{I}(\boldsymbol{\theta}) = n\text{E}[(\nabla_{\boldsymbol{\theta}} \log P(\boldsymbol{x}|\boldsymbol{\theta}))(\nabla_{\boldsymbol{\theta}} \log P(\boldsymbol{x}|\boldsymbol{\theta}))^\top]. \quad (15)$$

The operation $\text{E}[\,\cdot\,]$ denotes expectation with respect to the density $P(\boldsymbol{x}|\boldsymbol{\theta})$. The first term in the right-hand side of eq. (14) is called the *Cramer-Rao lower bound*, describing the minimum degree of fluctuations in all estimators. Thus, the ML estimator is optimal if n is sufficiently large (*asymptotic efficiency*).

If we have multiple candidate models

$$P_1(\boldsymbol{x}|\boldsymbol{\theta}_1), \quad P_2(\boldsymbol{x}|\boldsymbol{\theta}_2), \quad P_3(\boldsymbol{x}|\boldsymbol{\theta}_3), \quad ..., \quad (16)$$

from which we are to select an appropriate one for the observations $\boldsymbol{x}_1$, $\boldsymbol{x}_2$, ..., $\boldsymbol{x}_n$, the problem is (*stochastic*) *model selection*. Akaike's AIC has the following

form:

$$\mathrm{AIC} = -2\sum_{i=1}^{n} \log P(\boldsymbol{x}_i|\hat{\boldsymbol{\theta}}) + 2k + O\Big(\frac{1}{n}\Big). \qquad (17)$$

The model for which this quantity is the smallest is regarded as the best. The derivation of eq. (17) is based on the following facts [1]:

- The maximum likelihood estimator $\hat{\boldsymbol{\theta}}$ converges to its true value as $n \to \infty$ (the *law of large numbers*).
- The maximum likelihood estimator $\hat{\boldsymbol{\theta}}$ asymptotically obeys a Gaussian distribution as $n \to \infty$ (the *central limit theorem*).
- A quadratic form in standardized Gaussian random variables is subject to a χ^2 distribution, whose expectation is equal to its degree of freedom.

The Rissanen's MDL has the following form [30, 31]:

$$\mathrm{MDL} = -\sum_{i=1}^{n} \log P(\boldsymbol{x}_i|\hat{\boldsymbol{\theta}}) + \frac{k}{2}\log\frac{n}{2\pi} + \log\int_{\mathcal{T}} \sqrt{|\boldsymbol{I}(\boldsymbol{\theta})|}\mathrm{d}\boldsymbol{\theta} + O(1). \qquad (18)$$

Here, $\hat{\boldsymbol{\theta}}$ is the ML estimator; the symbol $O(1)$ denotes terms of order 0 in n in the limit $n \to \infty$.

In order that the integration in the right-hand side of eq. (18) exists, the domain $\mathcal{T}$ of the parameter $\boldsymbol{\theta}$ must be compact. In other words, we must specify in the k-dimensional space of $\boldsymbol{\theta}$ a finite region $\mathcal{T}$ in which the true value of $\boldsymbol{\theta}$ is likely to exist. This is nothing but the *Bayesian* standpoint that requires a prior distribution for the parameter to estimate. If it is not known, we must introduce an appropriate expedient to suppress an explicit dependence on the prior. Such an expedient is also necessary for the geometric MDL, i.e., the introduction of the reference length L in eq. (18).

4.2 Dual interpretations

Thus, we have seen that the limit $n \to \infty$ for the standard statistical analysis corresponds to the limit $\varepsilon \to 0$ for geometric inference. For example, the covariance matrix of the ML estimator agrees with the Cramer-Rao lower bound up to $O(1/n^2)$ for $n \to \infty$ (see eq. (14)), while for geometric inference it agrees with the lower bound bound up to $O(\varepsilon^4)$ for $\varepsilon \to 0$ (see eq. (9)). If follows that $1/\sqrt{n}$ for the standard statistical analysis plays the same role as ε for geometric inference.

The same correspondence exists for model selection, too. The unknowns for geometric inference are the p parameters of the constraint plus the N true positions specified by the d coordinates of the d-dimensional manifold $\mathcal{S}$ defined by the constraint. If eq. (12) is divided by ε^2, we have $\hat{J}/\varepsilon^2 + 2(Nd + p) + O(\varepsilon^2)$, which is ($-2$ times the logarithmic likelihood)+2(the number of unknowns), the same form as Akaike's AIC (17). The same holds for eq. (13), which corresponds to Rissanen's MDL (18) if ε is replaced by $1/\sqrt{n}$ [15].

This correspondence can be interpreted as follows. Since the underlying ensemble is hypothetical, we can actually observe only one sample as long as a particular algorithm is used. Suppose we hypothetically sample n different algorithms to find n different positions. The optimal estimate of the true position under the Gaussian model is their sample mean. The covariance matrix of the sample mean is $1/n$ times that of the individual samples. Hence, this hypothetical estimation is equivalent to dividing the noise level ε in eq. (4) by $\sqrt{n}$.

In fact, there were attempts to generate a hypothetical *ensemble of algorithms* by randomly varying the internal parameters (e.g., the thresholds for judgments), not adding random noise to the image [4, 5]. Then, one can compute their means and covariance matrix. Such a process as a whole can be regarded as one operation that effectively achieves higher accuracy.

Thus, the asymptotic analysis for $\varepsilon \to 0$ is equivalent to the asymptotic analysis for $n \to \infty$, where n is the number of hypothetical observations. As a result, the expression $\cdots + O(1/\sqrt{n^k})$ in the standard statistical analysis turns into $\cdots + O(\varepsilon^k)$ in geometric inference.

4.3 Noise level estimation

In order to use the geometric AIC or the geometric MDL, we need to know the noise level ε. If not known, it must be estimated. Here arises a sharp contrast between the standard statistical analysis and our geometric inference.

For the standard statistical analysis, the noise magnitude is a *model parameter*, because "noise" is defined to be *the random effects that cannot be accounted for by the assumed model*. Hence, the noise magnitude should be estimated, if not known, *according to the assumed model*. For geometric inference, on the other hand, the noise level ε is *a constant that reflects the uncertainty of feature detection*. So, it should be estimated *independently of individual models*.

If we know the true model, it can be estimated from the residual $\hat{J}$ using the knowledge that $\hat{J}/\varepsilon^2$ is subject to a χ^2 distribution with $rN - p$ degrees of freedom in the first order [9]. Specifically, we obtain an unbiased estimator of ε^2 in the form

$$\hat{\varepsilon}^2 = \frac{\hat{J}}{rN - p}. \qquad (19)$$

The validity of this formula has been confirmed by many simulations.

One may wonder if model selection is necessary at all when the true model is known. In practice, however, a typical situation where model selection is called for is *degeneracy detection*. In 3-D analysis from images, for example, the constraint (3) corresponds to our knowledge about the scene such as rigidity of motion. However, the computation fails if degeneracy occurs (e.g., the motion is zero). Even if exact degeneracy does not occur, the computation may become numerically unstable in near degeneracy conditions. In such a case, the computation can be stabilized by switching to a model that describes the degeneracy [11, 16, 19, 20, 23, 27, 40].

Degeneracy means *addition* of new constraints, such as some quantity being zero. It follows that the manifold $\mathcal{S}$ degenerates into a submanifold $\mathcal{S}'$ of it. Since the general model still holds irrespective of the degeneracy, i.e. $\mathcal{S}' \subset \mathcal{S}$, we can estimate the noise level ε from the residual $\hat{J}$ of the general model $\mathcal{S}$ using eq. (19).

Remark 9 Eq. (19) can be intuitively understood as follows. Recall that $\hat{J}$ is the sum of the square distances from $\{\boldsymbol{x}_\alpha\}$ to the manifold $\hat{\mathcal{S}}$ defined by the constraint $F^{(k)}(\boldsymbol{x}, \boldsymbol{u}) = 0$, $k = 1, ..., r$. Since $\hat{\mathcal{S}}$ has codimension r (the dimension of the orthogonal directions to it), the residual $\hat{J}$ should have expectation $rN\varepsilon^2$. However, $\hat{\mathcal{S}}$ is fitted by adjusting its p-dimensional parameter $\boldsymbol{u}$, so the expectation of $\hat{J}$ reduces to $(rN - p)\varepsilon^2$.

Remark 10 It may appear that the residual $\hat{J}$ of the general model cannot be stably computed in the presence of degeneracy. However, what is unstable is *model specification*, not the residual. For example, if we fit a planar surface to almost collinear points in 3-D, it is difficult to specify the fitted plane stably; the solution is very susceptible to noise. Yet, the residual is stably computed, since unique specification of the fit is difficult *because all the candidates have almost the same residual*.

Remark 11 Note that the noise level estimation from the general model $\mathcal{S}$ by eq. (19) is still valid even if degeneracy occurs, because degeneracy means shrinkage of the model manifold $\mathcal{S}'$ *within* $\mathcal{S}$, which does not affect the data deviations in the "orthogonal" directions (in the Mahalanobis sense) to $\mathcal{S}$ that account for the residual $\hat{J}$.

4.4 Geometric AIC vs. geometric MDL

We now illustrate the different characteristics of the geometric AIC and the geometric MDL in detecting degeneracy. Consider a rectangular region $[0, 10] \times [-1, 1]$ on the xy plane in the xyz space.

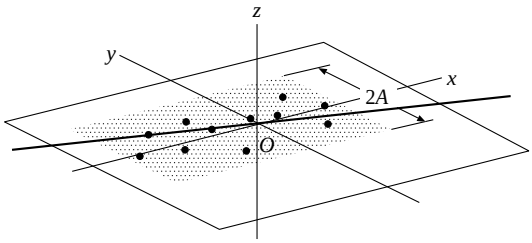


Figure 4: Fitting a space line and a plane to points in space.

We randomly take eleven points in it and magnify the region A times in the y direction. Adding Gaussian noise of mean 0 and variance ε^2 to the x, y, and z coordinates of each point independently, we fit a space line and a plane in a statistically optimal manner (Fig. 4). The rectangular region degenerates into a line segment as $A \to 0$.

A space line is a one-dimensional model with four degrees of freedom; a plane is a two-dimensional model with three degrees of freedom. Their geometric AIC and geometric MDL are

$$\begin{aligned} \text{G-AIC}_l &= \hat{J}_l + 2(N+4)\varepsilon^2, \\ \text{G-AIC}_p &= \hat{J}_p + 2(2N+3)\varepsilon^2, \\ \text{G-MDL}_l &= \hat{J}_l - (N+4)\varepsilon^2 \log\Big(\frac{\varepsilon}{L}\Big)^2, \\ \text{G-MDL}_p &= \hat{J}_p - (2N+3)\varepsilon^2 \log\Big(\frac{\varepsilon}{L}\Big)^2, \end{aligned} \qquad (20)$$

where the subscripts l and p refer to lines and planes, respectively. For each A, we compare the geometric AIC and the geometric MDL of the fitted line and plane and choose the one that has the smaller value. We used the reference length $L = 1$.

Fig. 5(a) shows the percentage of choosing a line for $\varepsilon = 0.01$ after 1000 independent trials for each A. If there were no noise, it should be 0% for $A \neq 0$ and 100% for $A = 0$. In the presence of noise, the geometric AIC has a high capability of distinguishing a line from a plane, but it judges a line to be a plane with some probability. In contrast, the geometric MDL judges a line to be a line almost 100%, but it judges a plane to be a line over a wide range of A.

In Fig. 5(a), we used the true value of ε^2. Fig. 5(b) shows the corresponding result using its estimate obtained from the general plane model by eq. (19). We observe somewhat degraded but similar performance characteristics.

Thus, we can observe that the geometric AIC has a higher capability for detecting degeneracy than the geometric MDL, but the general model is chosen with some probability when the true model is degenerate. In contrast, the percentage for the geometric MDL to detect degeneracy when the true model is really degenerate approaches 100% as the noise decreases. This is exactly the dual statement to the

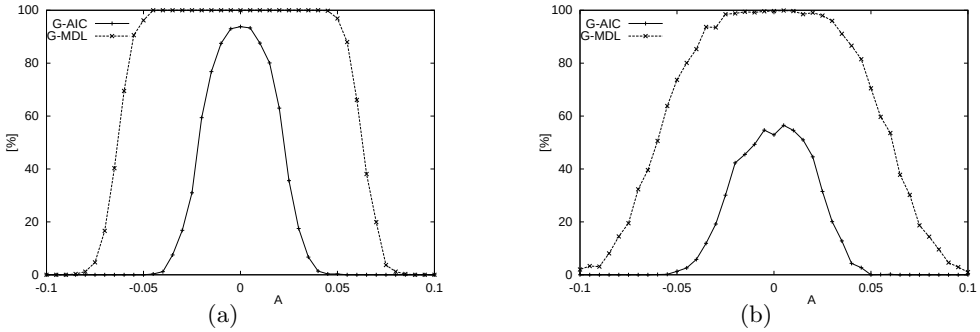


Figure 5: The rate (%) of detecting a space line by the geometric AIC (solid lines) and the geometric MDL (dashed lines) with (a) the true noise level and (b) the estimated noise level.

well known fact, called the *consistency of the MDL*, that the percentage for Rissanen's MDL to identify the true model converges to 100% in the limit of an infinite number of observations. Rissanen's MDL is regarded by many as superior to Akaike's AIC because the latter lacks this property.

At the cost of this consistency, however, the geometric MDL regards a wide range of nondegenerate models as degenerate. This is no surprise, since the penalty $-(Nd+p)\varepsilon^2 \log(\varepsilon/L)^2$ for the geometric MDL in eq. (13) is heavier than the penalty $2(Nd+p)\varepsilon^2$ for the geometric AIC in eq. (12). As a result, the geometric AIC is more faithful to the data than the geometric MDL, which is more likely to choose a degenerate model. This contrast has also been observed in many applications [23, 18].

Remark 12 Despite the fundamental difference of geometric model selection from the standard (stochastic) model selection, many attempts have been made in the past to apply Akaike's AIC and their variants to computer vision problems based on the asymptotic analysis of $n \to \infty$, where the interpretation of n is different from problem to problem [35, 36, 37, 38, 39]. Rissanen's MDL is also used in computer vision applications. Its use may be justified if the problem has the standard form of linear/nonlinear regression [2, 24]. Often, however, the solution having a shorter description length was chosen with a rather arbitrary definition of the complexity [6, 22, 25].

Remark 13 Note that one cannot compare different model selection criteria in general terms, because each is based on its own logic. Not only that, one cannot prove that a particular criterion works at all. In fact, although Akaike's AIC and Rissanen's MDL are based on rigorous mathematics, there is no guarantee that they work well in practice. The mathematical rigor is in their *reduction* from their starting principles (the Kullback-Leibler information and the minimum description length principle), which are beyond proof. What one can tell is which criterion is more suitable for a particular application when used in a particular manner. The geometric AIC and the geometric MDL have shown to be effective in many computer vision applications [14, 17, 18, 19, 20, 23, 27, 40], but other criteria may be better in other applications. The important thing is, however, to understand the underlying logic of each criterion.

5. Conclusions

We have investigated the meaning of "statistical methods" for geometric inference based on image feature points. Tracing back the origin of feature uncertainty to image processing operations, we discussed the implications of asymptotic analysis in reference to "geometric fitting" and "geometric model selection". We also compared the capability of the "geometric AIC" and the "geometric MDL" in detecting degeneracy.

The main emphasis of this paper is on the correspondence between the asymptotic analysis for $\varepsilon \to 0$ for geometric inference and the asymptotic analysis for $n \to \infty$ for the standard statistical analysis, based on our interpretation of the uncertainty of feature detection.

Acknowledgments: This work was supported in part by the Ministry of Education, Culture, Sports, Science and Technology, Japan, under the Grant in Aid for Scientific Research C(2) (No. 15500113), the Support Center for Advanced Telecommunications Technology Research, and Kayamori Foundation of Informational Science Advancement.

References

[1] H. Akaike, A new look at the statistical model identification, IEEE Trans. Autom. Control, **16**-6 (1977), 716–723.

[2] K. Bubna and C. V. Stewart, Model selection techniques and merging rules for range data segmentation algorithms, *Comput. Vision Image Understand.*, **80**-2 (2000), 215–245.

[3] F. Chabat, G. Z. Yang and D. M. Hansell, A corner orientation detector, *Image Vision Comput.*, **17**-10 (1999), 761–769.

[4] K. Cho and P. Meer, Image segmentation form consensus information, *Comput. Vision Image Understand.*, **68**-1 (1997), 72–89.

[5] K. Cho, P. Meer, J. Cabrera, Performance assessment through bootstrap, *IEEE Trans. Patt. Anal. Mach. Intell.*, **19**-11 (1997), 1185–1198.

[6] H. Gu, Y. Shirai and M. Asada, MDL-based segmentation and motion modeling in a long sequence of scene with multiple independently moving objects, *IEEE Trans. Patt. Anal. Mach. Intell.*, **18**-1 (1996), 58–64.

[7] C. Harris and M. Stephens, A combined corner and edge detector, *Proc. 4th Alvey Vision Conf.*, Aug. 1988, Manchester, U.K., pp. 147–151.

[8] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, Cambridge University Press, Cambridge, U.K., 2000.

[9] K. Kanatani, *Statistical Optimization for Geometric Computation: Theory and Practice*, Elsevier, Amsterdam, the Netherlands, 1996.

[10] K. Kanatani, Geometric information criterion for model selection, *Int. J. Comput. Vision*, **26**-3 (1998), 171–189.

[11] K. Kanatani, Statistical optimization and geometric inference in computer vision, *Phil. Trans. Roy. Soc. Lond.*, A-**356** (1998), 1303–1320.

[12] K. Kanatani, Cramer-Rao lower bounds for curve fitting, *Graphical Models Image Process.*, **60**-2 (1988), 93–99.

[13] K. Kanatani, Model selection criteria for geometric inference, in A. Bab-Hadiashar and D. Suter (eds.), *Data Segmentation and Model Selection for Computer Vision: A Statistical Approach*, Springer, 2000, pp. 91–115.

[14] K. Kanatani, Motion segmentation by subspace separation and model selection, *Proc. 8th Int. Conf. Comput. Vision*, July 2001, Vancouver, Canada, Vol. 2, pp. 301–306.

[15] K. Kanatani, Model selection for geometric inference, plenary talk, *Proc. 5th Asian Conf. Comput. Vision*, January 2002, Melbourne, Australia, Vol. 1, pp. xxi–xxxii.

[16] K. Kanatani, Motion segmentation by subspace separation: Model selection and reliability evaluation, *Int. J. Image Graphics/*, **2**-2 (2002), 179–197.

[17] K. Kanatani, Evaluation and selection of models for motion segmentation, *Proc. 7th Euro. Conf. Comput. Vision*, May 2002, Copenhagen, Denmark, Vol. 3, pp. 335–349.

[18] K. Kanatani and C. Matsunaga, Estimating the number of independent motions for multibody motion segmentation, *Proc. 5h Asian Conf. Comput. Vision*, January 2002, Melbourne, Australia, vol. 1, pp. 7–12.

[19] Y. Kanazawa and K. Kanatani, Infinity and planarity test for stereo vision, *IEICE Trans. Inf. & Syst.*, **E80-D**-8 (1997), 774–779.

[20] Y. Kanazawa and K. Kanatani, Stabilizing image mosaicing by model selection, in M. Pollefeys, L. Van Gool, A. Zisserman and A. Fitzgibbon (eds.), *3D Structure from Images–SMILE 2000*, Springer, Berlin, 2001, pp. 35–51.

[21] Y. Kanazawa and K. Kanatani, Do we really have to consider covariance matrices for image features? *Proc. 8th Int. Conf. Comput. Vision*, July 2001, Vancouver, Canada, Vol. 2, pp. 586–591.

[22] Y. G. Leclerc, Constructing simple stable descriptions for image partitioning, *Int. J. Comput. Vision*, **3**-1 (1989), 73–102.

[23] C. Matsunaga and K. Kanatani, Calibration of a moving camera using a planar pattern: Optimal computation, reliability evaluation and stabilization by model selection, in *Proc. 6th Euro. Conf. Comput. Vision*, June–July, 2000, Dublin, Ireland, Vol. 2, pp. 595–609.

[24] B. A. Maxwell, Segmentation and interpretation of multicolored objects with highlights, *Comput. Vision Image Understand.*, **77**-1 (2000), 1–24.

[25] S. J. Maybank and P. F. Sturm, MDL, collineations and the fundamental matrix, *Proc. 10th British Machine Vision Conference*, September 1999, Nottingham, U.K., pp. 53–62.

[26] D. D. Morris, K. Kanatani and T. Kanade, Gauge fixing for accurate 3D estimation, *Proc. IEEE Conf. Comput. Vision Pattern Recog.*, December 2001, Kauai, Hawaii, U.S.A., Vol. 2, pp. 343–350.

[27] N. Ohta and K. Kanatani, Moving object detection from optical flow without empirical thresholds, *IEICE Trans. Inf. & Syst.*, **E81-D**-2 (1998), 243–245.

[28] D. Reisfeld, H. Wolfson and Y. Yeshurun, Context-free attentional operators: The generalized symmetry transform, *Int. J. Comput. Vision*, **14** (1995), 119–130.

[29] J. Rissanen, Universal coding, information, prediction and estimation, *IEEE Trans. Inform. Theory*, **30**-4 (1984), 629–636.

[30] J. Rissanen, *Stochastic Complexity in Statistical Inquiry*, World Scientific, Singapore, 1989.

[31] J. Rissanen, Fisher information and stochastic complexity, IEEE Trans. Inform. Theory, **42**-1 (1996), 40–47.

[32] C. Schmid, R. Mohr and C. Bauckhage, Evaluation of interest point detectors, *Int J. Comput. Vision*, **37**-2 (2000), 151–172.

[33] S. M. Smith and J. M. Brady, SUSAN—A new approach to low level image processing, *Int. J. Comput. Vision*, **23**-1 (1997), 45–78.

[34] Y. Sugaya and K. Kanatani, Outlier removal for feature tracking by subspace separation, *IEICE Trans. Inf. & Syst.*, **E86**-D (2003), 1095–1102.

[35] P. H. S. Torr, An assessment of information criteria for motion model selection, *Proc. IEEE Conf. Comput. Vision Patt. Recog.*, June 1997, Puerto Rico, pp. 47–53.

[36] P. H. S. Torr, Geometric motion segmentation and model selection, *Phil. Trans. Roy. Soc. Lond.*, A-**356** (1998), 1321–1340.

[37] P. H. S. Torr, "Bayesian model estimation and selection for epipolar geometry and generic manifold fitting," *Int. J. Comput. Vision*, **50**-1 (2002), 35–61, 2002.

[38] P. H. S. Torr, A. FitzGibbon and A. Zisserman, Maintaining multiple motion model hypotheses through many views to recover matching and structure, *Proc. 6th Int. Conf. Comput. Vision*, January 1998, Bombay, India, pp. 485–492.

[39] P. H. S. Torr and A. Zisserman, "Concerning Bayesian motion segmentation, model averaging, matching and the trifocal tensor," in *Proc. 6th Euro. Conf. Comput. Vision*, June–July, 2000, Dublin, Ireland, Vol. 1, pp. 511–528.

[40] Iman Triono, N. Ohta and K. Kanatani, Automatic recognition of regular figures by geometric AIC, *IEICE Trans. Inf. & Syst.*, **E81-D**-2 (1998), 246–248.