Ultramicroscopy 19 (1986) 337–348
North-Holland, Amsterdam

ODD MEN OUT: A QUANTITATIVE OBJECTIVE PROCEDURE FOR IDENTIFYING ANOMALOUS MEMBERS OF A SET OF NOISY IMAGES OF OSTENSIBLY IDENTICAL SPECIMENS

Michael UNSER

Biomedical Engineering and Instrumentation Branch, Division of Research Services, National Institutes of Health, Bethesda, Maryland 20892, USA

Alasdair C. STEVEN

Laboratory of Cellular and Developmental Biology, National Institute of Diabetes, and Digestive and Kidney Diseases, National Institutes of Health, Bethesda, Maryland 20892, USA

and

Benes L. TRUS

Computer Systems Laboratory, Division of Computer Research and Technology, National Institutes of Health, Bethesda, Maryland 20892, USA

Received 13 March 1986

Achievement of an optimal improvement in signal-to-noise ratio from image averaging techniques depends crucially on the assumption that all members of the set of images to be averaged are fundamentally alike. In HREM of biological macromolecules, this assumption may be invalid for such reasons as variations in viewing geometry, non-uniformity of staining, or structural perturbations caused by specimen preparation procedures or radiation damage. Inclusion of data that are compromized by these or other factors will degrade the information content of the averaged image. Here we present an algorithm which provides an objective quantitative method for the identification and elimination of anomalous members of a set of pre-aligned images. Based on a statistical criterion of mutual consistency, the algorithm forms an ordered list in which the individual images are ranked from most to least reliable. On specification of the noise statistics – in the formulation given here, of stationary white noise – an acceptability threshold in this ordered list is imposed. The derivation and implementation of this algorithm are presented, its properties discussed, and its application illustrated using both real and model electron micrograph data.

1. Introduction

The use of image averaging to improve signal-to-noise ratios of electron micrographs has proliferated greatly in recent years, particularly as applied to structural investigations of biological macro-molecules [1–3]. Initially confined to translational or rotational averaging in real space of crystalline or rotationally symmetric specimens, or to (formally equivalent) Fourier space filtering, the scope of image averaging has been considerably extended by the introduction of correlation alignment techniques to rectify poorly ordered lattices [4–6], or to bring images of free-standing particles into registration [7–9]. The improvement in signal-to-noise to be realized by averaging depends crucially on the premise that the images to be combined are intrinsically the same, apart from their content of random additive noise. In practice, however, this assumption may break down:
(i) there may be intrinsic heterogeneity among the particles analyzed;

0304-3991/86/$03.50

(ii) some particles may have adsorbed in different orientations relative to the plane of the support film;
(iii) the particles may have undergone systematically different modes of negative staining (e.g. depth of stain, degree of infiltration, amount of coincident positive staining);
(iv) some particles may have undergone major distortions upon adsorption to the substrate or when they were dehydrated;
(v) the correlation alignment procedures may have failed, in some cases, to achieve correct registration.

In such circumstances, inclusion of the inconsistent data in the average will not be beneficial, particularly when the goal is to exhibit detail corresponding to the highest resolution at which the data set is potentially meaningful. In applications of correlation averaging, it is common practice for images suspected of being anomalous to be weeded out at the discretion of the investigator, prior to forming the average. This procedure has the drawback of being somewhat arbitrary, and is neither quantitative nor objective. In this paper, we present an algorithm that systematically screens a set of pre-aligned images for anomalous members.

1.1. Algorithm design: general considerations

In the absence of any absolute reference to serve as a basis for discriminating between acceptable and anomalous images, this decision must be based on a criterion of mutual consistency. This approach led us to the definition of the maximum-likelihood ordering principle described in the appendix. Its implementation requires that a parametric form of the statistical distribution of the noise be specified. This design principle has been applied to the case of independent additive Gaussian noise, resulting in the algorithm presented in this paper.

Starting with N pre-aligned images, the algorithm searches for the set of $(N-1)$ images that have the minimum variance; the image lying outside this set is thereby eliminated. This procedure is then reiterated on the subset of $(N-1)$ images to identify the second most deviant member, and so on, ultimately yielding a fully ordered data set. Having produced this ordering, a probabilistic estimate is then computed so as to decide whether an image should be accepted or rejected for averaging purposes. This estimate depends on the statistics of the noise and allows a cut-off point in the ordered set to be defined. As implemented here, the algorithm is based on a statistical model of equal variance and independently distributed Gaussian noise. However, it may be adapted for situations in which other noise distributions may be more appropriate (for example: Poisson noise). A more detailed account of both phases of the algorithm is presented in section 2, and further information concerning their derivation and implementation is given in the appendix. Section 3 presents some examples of its applications.

2. Description of the algorithm

In the following text, the initial images are represented by a set of N vectors, $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$, defining N data points in a space of dimension M, where M is the number of pixels in each image.

2.1. Data ordering

From the initial data set, the algorithm forms an ordered list according to a declining level of consistency: $\boldsymbol{x}_{(1)}, \ldots, \boldsymbol{x}_{(N)}$. This ordering is performed iteratively by removing, at each step, the observation least consistent with the remainder of the set. Let n denote the number of remaining images at any step of iteration, and, for convenience, let us rename them as $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$. Two slightly different procedures may be adopted for choosing the observation to be removed (denoted by $\boldsymbol{x}_{(n)}$), depending on the strategy used for computing the mean. These are called exclusive or inclusive estimation.

(i) *Exclusive center estimation*: When the estimate of the mean does not include the image that is currently considered, the observation by the algorithm is defined by:

$$\boldsymbol{x}_{(n)} = \boldsymbol{x}_r:$$

$$\sum_{\substack{i=1 \\ i \neq r}}^{n} \left\| \boldsymbol{x}_i - \boldsymbol{m}_{n-1,r} \right\|^2 = \min_{k=1,\ldots,n} \left\{ \sum_{\substack{i=1 \\ i \neq k}}^{n} \left\| \boldsymbol{x}_i - \boldsymbol{m}_{n-1,k} \right\|^2 \right\} \tag{1}$$

where $\| \cdots \|^2$ represents the square norm of a vector and where $\boldsymbol{m}_{n-1,k}$ is the mean computed by excluding $\boldsymbol{x}_k$:

$$\boldsymbol{m}_{n-1,k} = \frac{1}{n-1} \sum_{\substack{i=1 \\ i \neq k}}^{n} \boldsymbol{x}_i. \tag{2}$$

The algorithm therefore selects the $(n-1)$ points with minimum variance. The decision is based on the computation of the Euclidean distance between a data point and the current mean estimate. As shown in appendix A.3, these equations define the maximum-likelihood solution in the case of independently distributed and equal variance Gaussian noise. It follows that the $(n-1)$ points selected by the algorithm are those that are "best explained" by the Gaussian model in the sense that their estimated joint likelihood (or probability) is the maximum possible. Moreover, the quantity $\boldsymbol{m}_{n-1,r}$ provides the best current maximum-likelihood estimate of the distribution's mean.

(ii) *Inclusive center estimation*: A simpler procedure is obtained when the mean includes all currently remaining observations. The corresponding rule, obtained by substituting $\boldsymbol{m}_{n-1,k}$ by $\boldsymbol{m}_n$ in eq. (1), is now equivalent to the simpler equation: $\boldsymbol{x}_{(n)} = \boldsymbol{x}_r$:

$$\| \boldsymbol{x}_r - \boldsymbol{m}_n \|^2 = \max_{k=1,\ldots,n} \left\{ \| \boldsymbol{x}_k - m_n \|^2 \right\}, \tag{3}$$

where $\boldsymbol{m}_n$ is current sample mean defined by:

$$\boldsymbol{m}_n = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{x}_i. \tag{4}$$

In this case, selection of the $(n-1)$ points with minimum variance is equivalent to choosing the $(n-1)$ points least distant from the central point $\boldsymbol{m}_n$, which, according to eq. (3), is exactly the same as removing the most distant point.

The exclusive alternative is conceptually more attractive because it excludes the effect of the observation that is currently under consideration as possibly being anomalous. In turn, the inclusive version of the algorithm is computationally simpler (but not necessarily faster, as shown in appendix A.4). For n sufficiently large, however, both methods should give similar results.

2.2. Testing for consistency

When an image is isolated by the algorithm, we might test the hypothesis of whether it belongs to the same population as the $(n-1)$ remaining samples. However, this decision requires some assumptions to be made about the noise distribution.

For the case of additive stationary white Gaussian noise described in appendix A.3, we use the following test statistic:

$$d(\boldsymbol{x}_i, \boldsymbol{m}_n) = a_n \cdot \| \boldsymbol{x}_i - \boldsymbol{m}_n \|^2 / \sigma^2,$$

where

$$a_n = \begin{cases} n/(n+1) & \text{(exclusive option)}, \\ n/(n-1) & \text{(inclusive option)}. \end{cases} \tag{5}$$

Here σ^2 is the (known) variance and $\boldsymbol{m}_n$ the inclusive or exclusive estimate (based on n observations) of the unknown mean vector $\boldsymbol{\mu}$. For randomly selected observations that are identically Gaussian distributed with true distribution parameters $\boldsymbol{\mu}$ and σ^2, $d(\boldsymbol{x}_i, \boldsymbol{m}_n)$ has a χ^2 distribution with M degrees of freedom (the dimensionality of $\boldsymbol{x}_i$). Therefore, when $d(\boldsymbol{x}_i, \boldsymbol{m}_n)$ is above a certain threshold T corresponding to a specified probability of false rejection *,

$$P = \text{Prob}\{ d(\boldsymbol{x}_i, \boldsymbol{m}_n) > T \},$$

the hypothesis that $\boldsymbol{x}_i$ is statistically consistent with the remainder of the data set should be rejected. The choice of the rejection threshold may be facilitated by using the fact that, for M sufficiently large (typically: $M > 30$), the quantity $(d(\boldsymbol{x}_i, \boldsymbol{m}_n) - M)/(2M)^{1/2}$ has a distribution that is very closely approximated by a standardized Gaussian, with zero mean and unit variance.

This test can also be extended for additive non-Gaussian stationary white noise. In such a case, the test statistic $d(\boldsymbol{x}_i, \boldsymbol{m}_n)$, which is obtained from the summation of M independently and identically distributed random variables, will tend

* Note that because of prior ranking of the data, we are only testing for the worst case so that the true level of significance of the test is rather $\alpha = 1 - (1-P)^n$, which represents the probability of the presence of at least one image out of n with a d value greater than T.

to be Gaussian as a consequence of the Central Limit Theorem [10]. Its mean is simply

$$E\{d(\boldsymbol{x}_i, \boldsymbol{m}_n)\} = M,$$

and its variance, provided that n is sufficiently large for $\boldsymbol{m}_n$ to be reasonably close to $\boldsymbol{\mu}$, is closely approximated as follows:

$$\begin{aligned} &E\{d(\boldsymbol{x}_i, \boldsymbol{\mu})^2\} \\ &\quad = E\{\|\boldsymbol{x}_i - \boldsymbol{\mu}\|^4\}/\sigma^4 \\ &\quad = E\left\{\left(\sum_{k=1}^{M}(x_{ik}-\mu_k)^2\right)^2\right\}/\sigma^4 \\ &\quad = \sum_{k=1}^{M} E\{(x_{ik}-\mu_k)^4\}/\sigma^4 \\ &\qquad + 2\sum_{k\neq l}\sum E\{(x_{ik}-\mu_k)^2(x_{il}-\mu_l)^2\}/\sigma^4 \\ &\quad = M\cdot\kappa + M^2 - M, \end{aligned}$$

so that

$$\begin{aligned} &\mathrm{Var}\{d(\boldsymbol{x}_i, \boldsymbol{m}_n)\} \\ &\quad \cong \mathrm{Var}\{d(\boldsymbol{x}_i, \boldsymbol{\mu})\} \\ &\quad = E\{d(\boldsymbol{x}_i, \boldsymbol{\mu})^2\} - E\{d(\boldsymbol{x}_i, \boldsymbol{\mu})\}^2 = M(\kappa - 1), \end{aligned} \tag{6}$$

where κ is the noise distribution's Kurtosis or normalized fourth moment defined as:

$$\kappa = \frac{E\{(x_k-\mu_k)^4\}}{E\{(x_k-\mu_k)^2\}^2} = \frac{m_4}{\sigma^4}, \quad k = 1, \ldots, M. \tag{7}$$

This quantity has the value 3 for Gaussian data. Therefore, for M sufficiently large, the normalized test variable which we evaluate,

$$z(\boldsymbol{x}_i, \boldsymbol{m}_n) = \frac{d(\boldsymbol{x}_i, \boldsymbol{m}_n) - M}{(M(\kappa-1))^{1/2}}, \tag{8}$$

is Gaussian with zero mean and approximately unit variance.

So far, we have considered the distribution parameters σ^2 and κ as known quantities. In practice, however, this is not the case and these quantities have to be replaced by their best available estimates. Nevertheless, when $n \gg 2$ and $M \gg 30$, the limited variability of these quantities will not significantly affect the null distribution of test statistics defined by eqs. (5) and (8), so that the use of the Gaussian approximation is still justified.

3. Experimental results

A set of negatively stained capsomer images of Herpes Simplex Virus (HSV) [11] has been used to illustrate the application of our algorithm. The method was first applied to a set of 30 pre-aligned images, and the inconsistent data identified and rejected to form an optimally averaged image. This image was then used as a test object to further investigate the performance of the algorithm in the presence of various levels of Gaussian noise, and by otherwise simulating different types of artifacts.

In each of these experiments, the images were ranked using both inclusive and exclusive versions of the OMO (Odd Men Out) algorithm. In all cases these produced identical results. The test for statistical consistency was performed using eq. (8), and evaluating its corresponding probability from the right tail of a standardized Gaussian distribution. The estimates of the distribution's second and fourth moments (σ^2 and κ) were computed only once from the full set of N images.

3.1. Application to a set of experimental HSV capsomer images

The set of images was first aligned both rotationally and translationally by correlation alignment techniques [7,8,12]. The ranking assigned by the algorithm is shown in fig. 1 where the image labels correspond to the (arbitrary) order in which the images were entered into the program. The probabilities and cumulative variances computed at each step are plotted in fig. 2. The shape of the probability curve is distinctly sigmoidal (and such – qualitatively at least – has also been the case in the other applications which we have made to date). The first twenty images all have probabilities > 0.75, indicating a high level of mutual consistency. There is a certain latitude in imposing

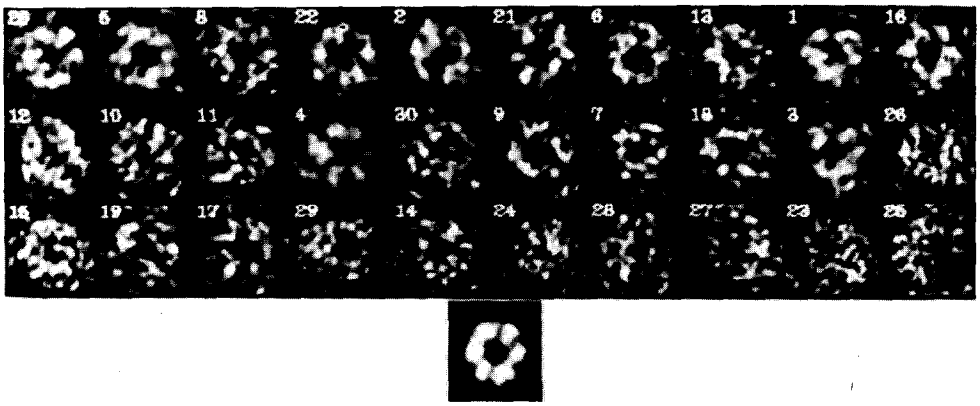

Fig. 1. (a) Set of 30 experimental capsomer images of Herpes Simplex Virus Type II displayed in the order in which they were ranked by the OMO algorithm; (b) average image obtained by excluding the 6 last (ie. least consistent) images.

an acceptability threshold (see section 4), but the last six images have probability values < 0.01 and thus appear to be anomalous. Accordingly these were rejected in forming the average image shown in fig. 1. Note that the presence of outlying data may also be inferred from the variance curve in fig. 2 which starts diverging near the cut-off point ($n = 24$ or 25).

3.2. Comparison with identifying anomalous images as outliers in a two-factor plot obtained by correspondence analysis

In principle, at least, a similar screening procedure could be carried out by assessing a two-factor plot obtained by correspondence analysis of an image set [13,14,16]. The expectation here would be that the consistent data should form a relatively compact cluster and the (randomly) anomalous images could be picked out as outliers. A two-dimensional plot is presented in fig. 3 that maps the same data set according to their two most prominent factors, which account for 9.0% and 6.7% of the total power respectively. The images furthest displaced from the bulk of the data set are Nos. 25, 23, 27, 28, 4, and 29. The identification of the first four as anomalous con-

Fig. 2. Computed probability values and normalized variance estimates obtained at successive iterations of the OMO algorithm ($n = 30,\ldots,1$) for the HSV capsomer images in fig. 1.

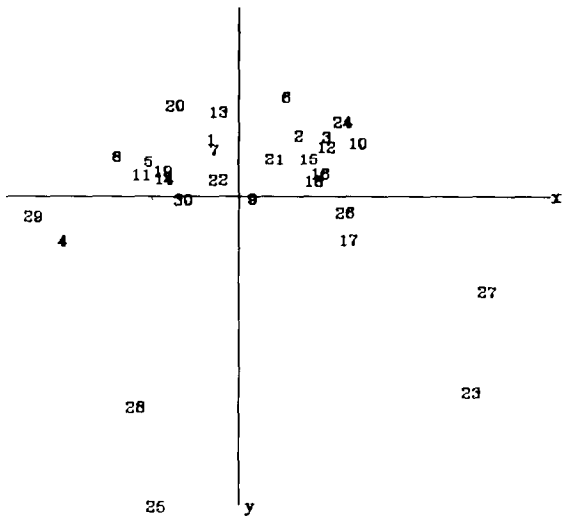

Fig. 3. Factorial map of the HSV capsomer images (fig. 1) according to their two most significant factors obtained using correspondence analysis.

curs with the outcome of the OMO algorithm, but the other two do not. Moreover, images 14 and 24 which were rejected by OMO are not differentiable from the main cluster in fig. 3. Thus we conclude that while grossly inconsistent images may be discriminated in a 2-factor plot, this approach is less reliable than the OMO algorithm, most likely because it is based on only a minor fraction of the information (i.e. a fraction of the total image power).

3.3. Analysis of test images affected by simulated noise

To explore the functioning of the algorithm in terms of a fully defined model system, we used as reference image the averaged HSV capsomer produced in the previous analysis and contaminated it with several kinds of simulated noise.

(a) *Increasing levels of Gaussian noise.* A set of images degraded by the addition of increasing levels of random Gaussian noise was generated. These corresponded to signal-to-noise ratios in the range of 4.0 to 0.1. The algorithm was able to rank these images correctly as shown in fig. 4. For comparison, we estimate the signal-to-noise ratio of HSV capsomer images to be about 0.8 to 1.0.

(b) *Images with uniform signal-to-noise ratios.* To examine the behavior of the algorithm when confronted with a set of mutually consistent images with the same noise level, 20 images were generated each based on the average HSV capsomer contaminated with random Gaussian noise at a S/N of 1.0. The algorithm output is shown in fig. 5. The imposition of probability thresholds for rejection at P-values of 0.5, 0.25, 0.01 and 0.05 leads to omission of 10, 6, 2 and 1 images, respectively, in remarkably good agreement with the expected numbers (10, 5, 2, 1 out of 20, respectively).

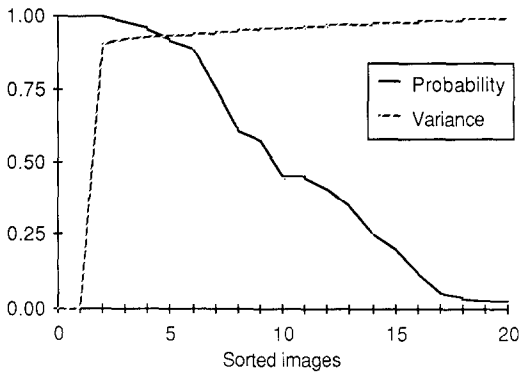

Fig. 5. Computed probability values and normalized variance estimates obtained at successive iterations of the OMO algorithm ($n = 20,\ldots,1$) for a set of 20 generated HSV capsomer images contaminated with Gaussian white noise at a signal-to-noise ratio of 1.

3.4. Detection of particles with adsorbed contaminants

One situation in which anomalous images might arise that is relatively easy to simulate computationally is in the presence of additional stain-excluding features on some particles. These might occur as a result of fortuitous binding of some other protein(s), or through co-adsorption to the electron microscope grid. To model such an event, we added a feature to the capsomer image that was ranked 11th out of 20 in the previous experiment (in order to start with a typical rather than an exceptionally good or bad image). This feature was in each case located in the same part of the

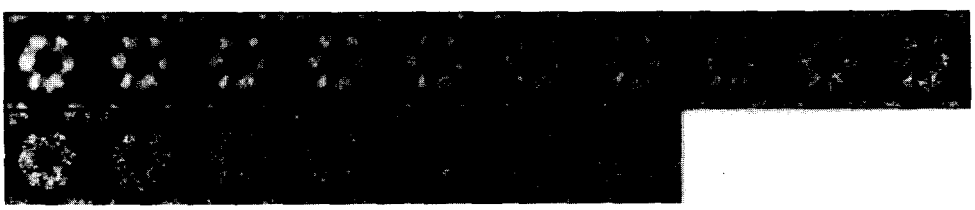

Fig. 4. HSV capsomer image overlaid with various amounts of computer-generated Gaussian additive white noise displayed in the order in which they were ranked by the OMO algorithm. The corresponding signal-to-noise ratios are 4.0, 3.5, 3.0, 2.5, 2.0, 1.5, 1.25, 1.0, 0.9, 0.8, 0.7, 0.6, 0.5, 0.4, 0.3, 0.2 and 0.1.

image, but its size was increased systematically. The initial image was the same as used in the experiment of section 3.3 (b). Upon inclusion of an additional stain-excluding spot of 1 pixel radius (covering 5 pixels, and corresponding to a diameter of 0.7 nm at the specimen), this image moved from 11th to 18th of 20 in the ordered list, but it would not yet be clearly anomalous ($P = 0.183$). Increasing its diameter to 1.0 nm (covering 9 pixels) reduced its ranking to 20th ($P = 0.028$), and a further increase to 1.3 nm reduced the P-value to 0.004. The latter two cases would be rejected at an acceptability threshold of 0.05. The capacity of the algorithm to pick out images of particles impaired by adsorbed contaminants will, in general, depend on such factors as the S/N ratio of a particular image, the shape and staining characteristics of the particle, and the location of the contaminant, so that it is not possible to give a general prescription for detectability by this procedure. Nevertheless this experiment does show that the OMO algorithm provides an effective tool for this purpose.

3.5. Detection of misalignment of images

Another potential source of anomalous images lies in the eventuality that the correlation alignment procedures may fail in some cases to achieve sufficiently precise registration. We have examined the sensitivity of the algorithm to systematic translational and rotational offsets of the same image analyzed in the previous section. For a lateral translation of 0.3 nm (1 pixel), the image moved from 11th to last in ranking and its P-value was reduced from 0.45 to 0.084; an offset of 0.6 nm further reduced this value to $< 10^{-4}$. We conclude that, for statistically similar data, translational offsets of 0.5 nm or worse would be readily detected by the algorithm. When the image was rotated by 5° it moved from 11th to last in ranking and a P-value of 0.033, and 6° further reduced this to 0.006. We conclude that in images of this type, errors in rotational registration of more than 5° would readily be detected by the algorithm. However, the HSV capsomer possesses a high degree of azimuthal symmetry making these images relatively insensitive to rotational misalignments. With non-annular, and in particular, with elongated particles, a much lower level of rotational offset should be detected by the algorithm.

4. Discussion

The algorithm presented here is optimal for screening images affected by stationary white Gaussian noise. In practice the noise content of experimental micrographs can originate from several different sources (cf. section 1) which cumulatively are unlikely to conform precisely to any idealized statistical distribution. However, this consideration should usually not present an obstacle to its applicability. First, the ranking is entirely based on the notion of inter-image proximity and therefore should still be useful in any situation where it is valid to express similarity between images in terms of the Euclidean distance between them. Second, it has been shown in section 2.2 that testing for consistency is still possible for non-Gaussian data provided that the test statistic is suitably normalized using the fourth moment of the noise distribution. It is only when the hypothesis of independent and identically distributed noise is dubious that the probabilistic estimates assigned to the images should be interpreted in relative rather than absolute terms.

Selection of an appropriate probability threshold for acceptance in the ordered list of images depends on striking a judicious compromise between (a) omitting an excessive number of good images, and (b) including an excessive number of sub-standard images. With statistically homogeneous data, the P-value of an imposed acceptability threshold determines the probability of erroneously rejecting an acceptable image. The probability of failing to reject an anomalous image is a less well defined proposition. In practice, we advocate some experimentation in the range of $0.01 > T > 0.005$. It is to be expected that the P-value distribution will fall off rather sharply in this mid-range, and this has indeed been the case for the data sets that we have analyzed so far. Accordingly, the end-point average image should be rather stable against the inclusion or exclusion of the few marginal images.

In the illustrative example given of a set of 30 HSV capsomer images (section 3.1), the algorithm judged that 20% should be omitted (see figs. 1 and 2). The numbers of images involved in this case are so small that omission of the anomalous images did not, in fact, significantly affect the interpretable features of the resulting average image (cf. fig. 3 of ref. [11]). When applied to larger data sets, however, it is to be expected that elimination of a substantial fraction of anomalous images should result in improved statistical definition of the (relatively noise-sensitive) higher spatial frequencies and thus improve the resolution of the asymptotically averaged image.

It should be noted that, because the algorithm is based on a criterion of mutual consistency, it is essential that the initial data set should contain at least a substantial minority of "good" images, in order to ensure stable convergence towards this subset. Moreover, the remainder should be randomly anomalous, i.e. there should be only one coherent subset of images. Thus, it may be useful to first test for the presence of multiple classes by using cluster analysis or other classification methods [14] in a factorial map computed by correspondence analysis [13,16] or some other multivariate statistical mapping technique [15]. However, as shown above (section 3.2), the present algorithm provides a more rigorous test for suspected anomaly than picking outliers from a two-dimensional factorial map. Accordingly, the algorithm should be applied to sets of images that include all possible candidates for membership of specified clusters.

An immediate extension of the algorithm, which might be useful in the presence of large proportions of inconsistent data, is to compute the test statistic z (e.g. (8)) based on current variance update until some cut-off point is reached. The cut-off point, however, should be chosen relatively high (typically $z > 5$) to prevent this quantity from becoming too small (due to the variance decreasing property of the algorithm (A.3), and thus to underestimate the true value of σ^2.

5. Conclusion

A new algorithm (OMO) is presented which ensures the formation of optimal average images from a set of pre-aligned noisy images of ostensibly identical objects. It operates by eliminating those members of the image set that are anomalous and whose inclusion would degrade the average image. The elimination criteria are objective, stable, and quantitative, having a rigorous statistical basis.

Acknowledgements

We thank Drs. C. Roberts, J. Hay and Ms. M. Bisher for their collaboration in obtaining the HSV capsomer images.

Appendix. A maximum-likelihood data ordering principle

This appendix presents the mathematical principles on which the OMO algorithm is based. The first section introduces a general maximum-likelihood-based criterion for the detection of an inconsistent observation. General properties related to the iterative application of this principle are discussed in section A.2. These results are applied in section A.3 to the case of additive stationary white Gaussian noise, leading to the form of the OMO algorithm presented in the paper. Finally, several procedures are described which minimize the computation time.

A.1. General principle

Let us consider n independent observations of a multivariate random variable $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ (e.g. $\boldsymbol{x}_i$ would be the ith image) where all observations except one are known to be drawn from a population that is distributed according to a parametric probability density function $p(\boldsymbol{x} \mid \boldsymbol{\theta})$ in which case this particular point is referred to as extraneous. The problem is to define a procedure that selects $n - 1$ points so as to isolate the point that is most likely to be extraneous.

If the distribution parameter θ were known, the usual statistical approach would be to exclude the least likely observation or, equivalently, to choose the set of $(n-1)$ points with maximum likelihood (or probability) relative to the basic probability density function $p(x|\theta)$. However, this strategy is not directly applicable because θ is unknown and has to estimated. An intuitively appealing approach is to incorporate the estimation of θ in the maximization process, so that a new rule may be formally expressed as:

Remove x_r:

$$\prod_{\substack{i=1 \\ i\neq r}}^{n} p(x_i|\theta_{n-1})\} = \max_{k=1,\ldots,n}\left\{\max_{\theta}\left\{\prod_{\substack{i=1 \\ i\neq k}}^{n} p(x_i|\theta)\right\}\right\}. \tag{A.1}$$

This means that we will consider all n possible sets of $(n-1)$ vectors and, for each of them, evaluate the parameter value that maximizes their joint probability density function. The set of points with overall maximum probability then determines, by elimination, the observation to be removed. This approach therefore selects the $(n-1)$ observations with maximum possible likelihood, which are those that can be the "best explained" by the probability model $p(x|\theta)$. In a later step, the estimated value $p(x_r|\theta_{n-1})$, where $\theta_{n-1} = \theta_{n-1}(x_1,\ldots,x_{n-1}, x_{r+1},\ldots,x_n)$ is defined by the left side of eq. (A.1), may be used to test the hypothesis that x_r is drawn from the same population as the other observations. Note that the parameter value that maximizes the joint probability density function of a given set of points defines the maximum likelihood estimate of θ. This general estimation principle is commonly used in statistics because of its interesting properties and mathematical tractability [7].

A.2 Iterative ordering and properties

With an initial set of N observations, the maximum likelihood data selection procedure may be applied iteratively, and will produce an ordered list of vectors that we label as $x_{(1)}\ldots x_{(N)}$ ($x_{(N)}$ is the observation removed at iteration step No. 1 using θ_{N-1} and so forth...). We employ the following notation for the log-likelihood function:

$$L(x_{(1)},\ldots,x_{(n)}|\theta) = \log\left\{\prod_{i=1}^{n} p(x_{(i)}|\theta)\right\} = \sum_{i=1}^{n} L(x_{(i)}|\theta), \tag{A.2}$$

and define $\theta_n = \theta_n(x_{(1)},\ldots,x_{(n)})$ as the maximum likelihood estimate of θ given $x_{(1)},\ldots,x_{(n)}$:

$$L(x_{(1)},\ldots,x_{(n)}|\theta_n) = \max_{\theta}\{L(x_{(1)},\ldots,x_{(n)}|\theta)\}, \quad n=1,\ldots,N. \tag{A.3}$$

An immediate consequence of the maximum likelihood property of the estimators θ_n $(i=1,\ldots,N)$ is that:

$$L(x_{(1)},\ldots,x_{(m)}|\theta_m) \geq L(x_{(1)},\ldots,x_{(m)}|\theta_n), \quad m,\ n=1,\ldots,N, \tag{A.4}$$

which also implies that

$$L(x_{(n+1)},\ldots,x_{(m)}|\theta_m) \geq L(x_{(n+1)},\ldots,x_{(m)}|\theta_n), \quad m\geq n=1,\ldots,N, \tag{A.5}$$

indicating that the estimated likelihood of the points that are removed between two stages m and n of the algorithm can only decrease. Using eq. (A.1), it is possible to show that the average likelihood of the samples that are considered by the algorithm can only increase as the algorithm proceeds:

$$L(x_{(1)},\ldots,x_{(n)}|\theta_n)/n \geq L(x_{(1)},\ldots,x_{(m)}|\theta_m)/m, \quad m\geq n=1,\ldots,N. \tag{A.6}$$

Despite there being no guarantee that the algorithm is absolutely optimum, that is, for any n, the quantity defined by eq. (A.2) is the maximum likelihood that can be obtained from any combination of n points in N, eqs. (A.5) and (A.6) show a generally favorable tendency. There is a reinforcement effect in that the current estimate of the average likelihood of the remaining points increases, whereas the likelihood of the observations that have already been removed decreases monotonically.

Note that these properties are also valid for an inclusive maximum likelihood parameter estima-

tion where the quantities defined in eq. (A.1) would be computed on the basis of a unique parameter estimate $\theta_n'(x_{(1)}, \ldots, x_{(n)})$. In such a case, however, the estimated likelihoods would always be smaller than – or at best equal to – their corresponding values obtained by exclusive estimation.

A.3. Application to independent Gaussian random variables

If we now consider $p(x \mid \theta)$ as being an M-dimensional multivariate Gaussian density function with parameters μ (mean vector) and a diagonal covariance matrix, given by σ^2 times the identity matrix I: $p(x \mid \theta) \sim N(\mu, \sigma^2 I)$, the maximum likelihood procedure discussed here is equivalent to the geometrical algorithm described in section 2. This model in which the M components of the random vector x are statistically independent and identically distributed about the mean vector μ is representative of a situation in which the measurements (images) are corrupted by additive stationary white Gaussian noise with variance σ^2.

The derivation of the optimal ordering procedure for the Gaussian model is based on the following results on maximum likelihood parameter estimation for this particular case. These formulae may be established by setting the partial derivatives of the log-likelihood function with respect to the distribution parameters to zero and solving these equations for μ and σ^2.

– In the case of an unknown mean vector and a known variance, the maximum likelihood estimation of the mean, given n independent observations $x_1, \ldots, x_n$, is the sample mean defined by eq. (4), and corresponds to the maximum log-likelihood function:

$$\begin{aligned}&\max\{L(x_1, \ldots, x_n; \mu)\}\\ &\quad = L(x_1, \ldots, x_n; m_n)\\ &\quad = \frac{1}{2\sigma^2}\sum_{i=1}^{n}\|x_i - m_n\|^2 - \frac{Mn}{2}\log\{2\pi\sigma\},\end{aligned} \tag{A.7}$$

where $\|x_i - m_n\|^2$ is the squared Euclidian distance between M-dimensional vectors x_i and m_n.

– In the case of unknown mean and variance, the maximum likelihood estimate of the mean is still given by eq. (4) and the maximum likelihood estimate of the variance is:

$$s_n^2 = \frac{1}{n}\sum_{i=1}^{n}\|x_i - m_n\|^2, \tag{A.8}$$

and corresponds to the maximum log-likelihood function:

$$\begin{aligned}&\max\{L(x_1, \ldots, x_n; \mu, \sigma)\}\\ &\quad = L(x_1, \ldots, x_n; m_n, s_n)\\ &\quad = -n(M\log\{2\pi s_n\} + 1)/2.\end{aligned} \tag{A.9}$$

For both cases (known and unknown variance), it is then straightforward to show, by substitution of the maximized Gaussian likelihood functions in eq. (A.1), that the proposed maximum likelihood procedure is equivalent to the geometrical rule (1) based on an exclusive estimation of the mean. For this particular set of equations, the meaning of properties (A.5) and (A.6) is the following. At any step of iteration, the observations that have already been eliminated by the algorithm tend to be located further and further from the current mean estimate while the variance of the remaining points decreases with n.

Note that these results are also indirectly applicable in the case of independent Poisson noise, using the non-linear transformation $y = (x + 1/4)^{1/2}$ which converts a Poisson variable x with parameter μ to a random variable y whose distribution is approximately Gaussian with mean $\sqrt{\mu}$ and variance 1/4, provided that μ is not too small (typically greater than 4).

A.4. Efficient implementation of the OMO algorithm

Within each iteration of the algorithm, the computation of the exclusive variance and mean estimates in eqs. (1) and (2) may be simplified by using the following expressions:

$$m_{n-1,k} = (nm_n - x_k)/(n-1), \tag{A.10}$$

$$\begin{aligned}&\sum_{\substack{i=1\\ i\neq k}}^{n}\|x_i - m_{n-1,k}\|^2\\ &\quad = \left(\sum_{i=1}^{n}\|x_i\|^2\right) - \|x_k\|^2 - (n-1)\|m_{n-1,k}\|^2,\end{aligned} \tag{A.11}$$

where $\boldsymbol{m}_n$ is the inclusive mean estimate based on n data points. Thus, provided that the $\| \boldsymbol{x}_i \|^2$'s have been evaluated in advance and that all sums are updated recursively, the exclusive and inclusive versions of the algorithm can be implemented with approximately the same number of operations. For an exclusive mean estimation, the number of operations required for an iteration is now proportional to nM and not n^2M as it would be if the means and the variances were computed using eqs. (2) and (1), respectively. Following the same principle, the sample mean may be updated recursively as the algorithm proceeds,

$$\boldsymbol{m}_n = \left[(n+1)\boldsymbol{m}_{n+1} - \boldsymbol{x}_{(n+1)} \right] / n,$$
$$n = N-1, \ldots, 1 \qquad \text{(A.12)}$$

Starting with mean $\boldsymbol{m}_N$ computed from all data points, this allows an overall saving of a factor $N/2$ over a direct evaluation using eq. (4). In this last expression, $\boldsymbol{x}_{(n+1)}$ represents the data point that has been removed at iteration step $(n+1)$. Note that these computational simplifications are applicable independently of the system of axes in which the data points are represented.

References

[1] D.L. Misell, Image Analysis, Enhancement, and Interpretation, in: Practical Methods in Electron Microscopy, Vol. 7, Ed. A.M. Glauert (North-Holland, Amsterdam, 1978).

[2] T.S. Baker, in: Electron Microscopy in Biology, Vol. 1, Ed. J.D. Griffith (Wiley, New York, 1981) pp. 189–290.

[3] W. Baumeister and W. Vogell, Eds., Electron Microscopy at Molecular Dimensions (Springer, Berlin, 1979).

[4] J. Frank, in: Computer Processing of Electron Microscopic Images, Ed. P.W. Hawkes (Springer, Berlin, 1980).

[5] R.H. Crepeau and E.K. Fram, Ultramicroscopy 6 (1981) 7.

[6] W.O. Saxton and W. Baumeister, J. Microscopy 127 (1982) 127.

[7] J. Frank, Ultramicroscopy 1 (1975) 159.

[8] J. Frank, A. Verschoor and M. Boublik, Science 214 (1981) 1353.

[9] M. Kessel, in: Proc. 8th European Congr. on Electron Microscopy, Budapest, 1984, Vol. 2, Eds. A. Csanády, P. Röhlich and D. Szábo, pp. 1326–1329.

[10] J.R. Blum and J.I. Rosenblatt, Probability and Statistics (Saunders, Philadelphia, 1972).

[11] A.C. Steven, C.R. Roberts, J. Hay, M.E. Bisher, T. Pun and B.L. Trus, J. Virol. 57 (1986) 578.

[12] B.L. Trus et al., to be published.

[13] M. Van Heel and J. Frank, Ultramicroscopy 6 (1981) 187.

[14] M. Van Heel, Ultramicroscopy 13 (1984) 165.

[15] M. Radermacher and J. Frank, Ultramicroscopy 17 (1985) 117.

[16] L. Lebart, A. Morineau and J.P. Fenelon, Traitement des Données Statistiques (Dunod, Paris, 1973).