Appendix D “Sieving” outliers in image registration data

D L B Jupp
November/December 2001.

Introduction – Basic assumptions and framework

This note explains the basis for the program SIEVER that was part of the former microBRIAN image processing system. SIEVER aimed to help users identify and remove image control points that were in error and (in particular) gross errors or “outliers”.

The context of SIEVER is that there exists an image or a number of images (perhaps overlapping frames or the same area at different times) and a map base or coordinate system within which they will be referenced. Points in the images (called here “GCP”s or Ground Control Points) are selected either pairwise between images or from the map base that identify the image or map coordinates of spatial features as accurately as possible. This may be done by a person or by a program (such as a correlation program). In the case of the identification of positions from the map base it is generally necessary for a person to identify the GCPs.

For the activity being described here, it is necessary that each identified feature have a unique identifier so that a given feature can be located in all of the images or the map base frame if it has been located or identified in any one. In microBRIAN this was accomplished simply by means of associating a separate list of GCPs with each image or map base and associating each record in the file with a unique spatial feature. In this way, if a feature is not present in an image of map base its location in the associated file is empty. If this or similar method of identifying features across

multiple images is available then the GCPs that are common to all of a group of images can be identified quickly.

The second aspect to the best use of SIEVER is that it is assumed that major geometric distortions have been “nominally” removed. That is, panoramic distortion or various artefacts created by the geometry of the image view and data collection have been taken into account in the coordinates of the GCPs in the images. This is assumed to be done to a point where remaining geometric differences between any images or a map base can be accounted for (or well described) by an “affine” transformation.

An affine transformation is linear such that if one image has coordinates (x,y) and the other (x’,y’) then the relationship between the coordinates has the form:

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} + \begin{bmatrix} a_{01} \\ a_{02} \end{bmatrix}$$

This six parameter transformation can be thought of as some combination of shift of origin, rotation, skew and possibly a separate “stretch” in the x and y directions.

Fitting Transformation Models

For any pair of images there can be transformations from each coordinate system to the other. If all of the transformations have been determined then any feature in one image can be located in each of the others. In the case where all of the transformations can be represented by a simple affine transformation then if the transformation is exactly known in one direction it can be inverted to obtain the transformation in the other direction.

However, transformations are never known exactly. Normally, the GCPs are used to estimate the coefficients by least squares or some other measure of goodness of fit between the image coordinates and those predicted by the transformation. The GCPs are normally subject to error or “noise” and in this process any outliers present will play a disturbing role.

There will always be some error or noise level that limits the accuracy of location of the GCPs and therefore also limit the accuracy of the estimated coefficients of a transformation. However, “outliers” are points that are inaccurate well beyond this base of error or “noise”.

When the transformation is being estimated it is usual to suppose that errors in GCPs are only in those of the “TO” side of the transformation. That is in the affine transformation above we are calling the (x,y) data the FROM side of the transformation (or the coordinates the transformation acts on) and the (x’,y’) data the TO side of the transformation (or the coordinates produced by the action of the transformation). In this case, assuming the errors are uncorrelated, it is possible to separate the fit to the x and y components of the TO side as:

$$x_i' = a_{10} + a_{11}x_i + a_{12}y_i + \varepsilon_i$$

and

$$y_i' = a_{20} + a_{21}x_i + a_{22}y_i + \eta_i$$

where ε_i and η_i and noise variances and the coefficients can be efficiently found by two separate leastsquares solutions.

However, the errors usually occur in both the FROM and the TO coordinate systems – a fact that is utilised when the inverse transformation is estimated – where the errors are then assumed to be only in the FROM side of the transformation.

In the case of an affine transformation it is possible to solve for the coefficients assuming errors in both variables – but only with a simplified error model. This solution is a by-product of the work described here and will be described in passing.

However, the more significant issue relates to outliers. In this case, if the outlier is in a GCP that is on the FROM side of the system then it is difficult if not impossible to actually identify the outlier. Its presence is often only indicated by a very poor fit to the data. Moreover, if the outlier is on the TO side of the system then because least squares balances all errors it is likely the fit to the outlier will be better than to other data.

The situation where the outliers are strongly fitted by the model – and hence very difficult to identify as outliers – is closely related to a concept called “predictive error”. The predictive error is the error at a point between the data and the model prediction based on all data *except* the point in question. Conversely, it measures the sensitivity of the model to the data at the point or the degree of control that the data point exerts on the model. The sensitivity can be expressed as a “predictive error multiplier” which will be large if the model is very sensitive to the presence of a data point.

High predictive error multipliers are both good and bad. A high value indicates a very important point in the modelling – or one that exerts a lot of influence over the model. However, such points are also the ones that introduce the greatest effects of error and when they are an outlier the effect can be very great but the actual error at the point is usually small – since the model fits the outlier in preference to other points!

The action of SIEVER therefore does three things. Firstly, it solves for the affine model without assuming the errors are all on one of the FROM and TO sides of the transformation model. Then it estimates the predictive errors of the residuals as one means of identifying outliers. Then it plots the squares of the errors (which would have a Chi-square distribution if they were from a normal population or errors) against the Gamma distribution in what is called a Q-Q (or Quantile-Quantile) plot. This also allows errors that are too large to be explained as simply large but still possible error values to be identified.

The SIEVER model

The SIEVER model assumes there are a set of N_v images and/or map base with common GCPs as identified from the individual GCP files of the images.

The matrix with M rows (where M is the number of common GCPs) and $2N_v$ columns and form:

$$A = \left[\underset{\sim}{x}_1, \underset{\sim}{y}_1, \underset{\sim}{x}_2, \underset{\sim}{y}_2, \ldots \underset{\sim}{x}_N, \underset{\sim}{y}_N \right]$$

(where the curled underscore indicates a column vector) is the starting point for the analysis.

Because we have assumed that nominal geometric transformations have been carried out on the data to remove the major image distortions and that the relationships between the images can (in the absense of noise) be modelled as affine transformations, the situation is one of the Generalised Linear Model (GLM) or Factor Analysis.

That is, we are assuming that there is an underlying coordinate system (μ, ν) such that for each of the image coordinate systems we have:

$$x_j = a_{j0} + a_{j1}\mu + a_{j2}\nu + \varepsilon$$

and

$$y_j = b_{j0} + b_{j1}\mu + b_{j2}\nu + \eta$$

where (ε, η) are the errors in the coordinates.

We are going to assume for convenience that the images coordinates have had the means extracted from them so that the data have “zero mean”. The underlying coordinates will also be assumed to have “zero mean”. In this case it is possible to express the form of the underlying transformation as:

$$\begin{bmatrix} x_j \\ y_j \end{bmatrix} = \begin{bmatrix} a_{j1} & a_{j2} \\ b_{j1} & b_{j2} \end{bmatrix} \begin{bmatrix} \mu \\ \nu \end{bmatrix} + \begin{bmatrix} \varepsilon \\ \eta \end{bmatrix}$$
$$= B_j \begin{bmatrix} \mu \\ \nu \end{bmatrix} + \begin{bmatrix} \varepsilon \\ \eta \end{bmatrix}$$

It follows that we could express the matrix A in the form:

$$A = \begin{bmatrix} \mu_1 & \nu_1 \\ \mu_2 & \nu_2 \\ \ldots & \ldots \\ \mu_M & \nu_M \end{bmatrix} \left[B_1^T \quad B_2^T \quad \ldots \quad B_N^T \right] + E$$

where E is the matrix of random errors or “noise”. That is, The underlying model is one where A has basic Rank 2 or A is the product of a matrix of size Mx2 and one of size 2x2N_v.

Solving for the underlying model and errors

The separate the underlying model from the errors we will use the theorem of Eckert and Young () which states that the best approximation to a matrix by one of lower rank can be conveniently found using the Singular Value Transformation (SVD) of the matrix.

The SVD (introduced by Lancos, 1958) provides an eigenvalue decomposition for an arbitrary matrix. That is, if A is any MxN matrix then there exist matrices U (MxN), S (NxN and diagonal) and V (NxN) that satisfy the conditions:

$$\begin{aligned} S &= diag\left[s_1, s_2, ..., s_N\right] \\ s_1 &\geq s_2 \geq ... \geq s_N \geq 0 \\ U^T U &= I_N \\ V^T V &= V V^T = I_M \\ A &= USV^T \end{aligned}$$

Since, using these rules we have that:

$$A^T A = VS^2 V^T$$

it follows that the squares of the singular values are the eigenvalues of $A^T A$ and the matrix V is the matrix of its eigenvectors.

An alternative way to write the SVD is in terms of the column vectors of U and V in the form:

$$A = \sum_{j=1}^{N} s_j \underset{\sim}{u}_j \underset{\sim}{v}_j^T$$

The Eckert-Young theorem can be stated in the way used here in the following form. If B is the best approximation to A by a matrix of rank p<N where “best” means in the sense of minimising the sums of squares of differences between all elements of the matrices then:

$$A = B + E$$

$$B = \sum_{j=1}^{p} s_j \underset{\sim}{u}_j \underset{\sim}{v}_j^T$$

$$E = \sum_{j=p+1}^{N} s_j \underset{\sim}{u}_j \underset{\sim}{v}_j^T$$

$$\|A - B\|^2 = \sum_{i=1}^{M}\sum_{j=1}^{N} (a_{ij} - b_{ij})^2$$

$$= \|E\|^2 = \sum_{j=p+1}^{N} s_j^2$$

As with many cases of least squares there is an "Analysis of Variance" (AOV) for this generalised system in the form:

$$TSS = MSS + RSS$$

$$\sum_{i=1}^{M}\sum_{j=1}^{N} a_{ij}^2 = \sum_{i=1}^{M}\sum_{j=1}^{N} b_{ij}^2 + \sum_{i=1}^{M}\sum_{j=1}^{N} e_{ij}^2$$

$$\|A\|^2 = \|B\|^2 + \|E\|^2$$

or

$$\sum_{k=1}^{N} s_k^2 = \sum_{k=1}^{p} s_k^2 + \sum_{k=p+1}^{N} s_k^2$$

where "TSS" stands for the Total Sum of Squares, "MSS" the Model Sum of Squares and "RSS" the Residual Sum of Squares.

Of special interest to us is the AOV for a single row of each matrix. Let the i'th row of a matrix be denoted by $\underset{\sim}{r}_i(A)$. It is a row vector that will, in this case, have N components. Then:

$$TSS_i = MSS_i + RSS_i$$

$$\|\underset{\sim}{r}_i(A)\|^2 = \|\underset{\sim}{r}_i(B)\|^2 + \|\underset{\sim}{r}_i(E)\|^2$$

or

$$\sum_{k=1}^{N} s_k^2 u_{ik}^2 = \sum_{k=1}^{p} s_k^2 u_{ik}^2 + \sum_{k=p+1}^{N} s_k^2 u_{ik}^2$$

From this (or directly) it can be seen that $\|\underset{\sim}{r}_i(A)\|^2 = \|\underset{\sim}{r}_i(US)\|^2$.

It is easy to see that in our specific case, where p=2 and N as used above is actually $2N_v$ then the corresponding solution to the SIEVER model is:

$$\begin{bmatrix} \mu_1 & \nu_1 \\ \mu_2 & \nu_2 \\ \dots & \dots \\ \mu_M & \nu_M \end{bmatrix} = \begin{bmatrix} s_1 \underset{\sim}{u}_1 & s_2 \underset{\sim}{u}_2 \end{bmatrix}$$

$$\begin{bmatrix} B_1^T & B_2^T & \dots & B_N^T \end{bmatrix} = \begin{bmatrix} \underset{\sim}{v}_1^T \\ \underset{\sim}{v}_2^T \end{bmatrix}$$

Because of the ortho-normal properties of V it therefore follows that the AOV above on US is a direct analysis of the errors between the underying 2 factor model of the GCPs and the selected GCPs. Moreover, the row AOV above is an analysis of each row or identifiable spatial feature and our objective is to isolate the outliers to the features and images in which they occur.

Note that these steps have overcome the problem of the very serious problems that can occur due to assuming the errors are all in one "side" of the equations. However, we still have a difficulty in that least squares "spreads" error across all the points and can dominated by a few points with strong "control". If these points are also outliers then their identification can be very difficult. In many cases the best tool to use is the human eye and so in the following there will be a number of suggests plots that are sometimes much more decisive than the accompanying statistics!

Analysing the SIEVER solution

The first action in SIEVER is to normalise the matrix A so that the columns have zero mean and the sum of squares of the columns is 1.0. In this case, the matrix $A^T A$ is in "correlation" matrix form.

The value of the normalisations is twofold. First, as noted above, the formulation is more convenient when the columns have zero mean but secondly the normalisation frees the data of scale changes that different coordinate systems may have.

The SVD of the matrix is then formed to obtain the components discussed above.

The possible actions following this are:

Basic "PCA" Summary

The SVD of the normalised matrix A can be interpreted as a Principal Component Analysis or PCA of A in correlation form (see Harrison & Jupp, 1990; 1995).

A simple summary set of statistics comprises (for each of the N components) the Principal Variance (P_{var}), the percent variance ($Var\%$), the Total Variance ($T_{var}\%$) and the Noise to Signal Ratio (%) (NSR%) defined by:

$$P_{\text{var},k} = s_k^2$$

$$Var_k(\%) = 100 \times \frac{s_k^2}{\sum_{j=1}^{N} s_j^2}$$

$$T_{\text{var},k}(\%) = 100 \times \frac{\sum_{j=1}^{k} s_j^2}{\sum_{j=1}^{N} s_j^2}$$

$$NSR_k(\%) = 100 \times \frac{\sum_{j=k+1}^{N} s_j^2/(N-k)}{\sum_{j=1}^{k} s_j^2/k}$$

These can be calculated for any PCA or SVD exercise and in our case, if the rank of the system is 2 we should have the Total variance in the first two components very high and the contribution of the remaining components as very low and NSR should reach its minimum at k=2.

It may be shown that these statistics are not very reliable for only two images and it is better in every way if SIEVER is used on a number of images in combinations – such as up to 5 in an initial investigation and down to studies of pairs of images in final outlier searches. This will be discussed later.

Row or Feature based Error Analysis

The issue, of course, is how to measure what is "high" and "low" with regard to the model fit. To help with this we can also generate some statistics for the rows of A – that is for the spatial features that are being mapped as GCPs in the images being registered.

For this we will work assuming that the assumption of the rank of *A* (*p*) being 2 is confirmed by the PCA analysis. The consequences of its not being true are discussed later. The assumption we have from this is that the estimated variance of the residual can be written:

$$\tilde{\sigma}^2 = \frac{1}{N-p} \sum_{j=p+1}^{N} s_j^2$$

The row or feature statistics we can generate to help to go on from here are, for Feature *i*, the Normalised Error (e_i^2), the feature Chi-square (X_i^2) and the Predictive Error Weight ($PE_i(\%)$) as defined by:

$$e_i^2 = \frac{1}{N-p}\sum_{k=p+1}^{N} s_k^2 u_{ik}^2$$

$$X_i^2 = M\frac{e_i^2}{\tilde{\sigma}^2}$$

$$PE_i = 100\times\left(\frac{1}{1-\sum_{k=1}^{p} u_{ik}^2} - 1\right)$$

The derivation of the Predictive Error formula and its relationship with “jacknifing” is to be described fully in a separate document since it is also used in other modules – such a MODEL to characterise the sensitivity of the models to presence or absence of image features.

Using the Q-Q Plot

To investigate the distribution of the errors remaining after fitting the general affine model, SIEVER uses a technique that is very useful from exploratory data analysis called “Q-Q Plotting”.

The Quantile-Quantile Plot

A Q-Q (or Quantile-Quantile) plot is constructed o test the statistical distribution of data against a known model. A “quantile” is a fraction (which when expressed as a percentage is called a “percentile”) between zero and one.

Consider that you have M data values (y_j) which are sorted into increasing order such that:

$$y_1 \le y_2 \le \ldots \le y_M$$

Associated with each y_i is a estimate of the fraction (or quantile) of data values less than or equal to y_i. This is simply:

$$q_i = \frac{i-1/2}{M}$$

Suppose *F(x)* is the cumulative distribution function for a probability distribution function *P(x)* such that:

$$F(x) = \int_{-\infty}^{x} P(x')dx'$$

The set of *x* values that satisfy:

$$F(x_i) = q_i \quad i = 1, M$$

are the set of values for which the expected fraction of values from a sample of M "drawings" which are less than or equal to x_i is the quantile q_i.

It follows that if the data yi are drawings from the distribution function *P(x)* then the plot of y_i against x_i for i=1,M will be close to a straight line with slope one and intercept zero.

This plot is the Q-Q plot and tests whether the data values are samples from the distribution *P(x)*.

Gamma Probability Plotting

To test the distribution of the residual errors from the SIEVER model, SIEVER uses a general method called Gamma probability plotting. This allows the distribution to be tested against a family of possible distributions, one of which is χ^2.

The hypothesis is that if the residuals are normal or near normal then the statistic X_i^2 defined in the previous section will be distributed as Chi-square with (*N-p*) degrees of freedom – or χ^2_{N-p}.

The cumulative distribution for the incomplete Gamma distribution can be defined as:

$$F(x;\alpha,\lambda,\eta) = \int_\alpha^x P(x';\alpha,\lambda,\eta)dx'$$

$$P(x;\alpha,\lambda,\eta) = \frac{\lambda^\eta}{\Gamma(\eta)}(x-\alpha)^{\eta-1}e^{-\lambda(x-\alpha)}$$

If *x* is χ_r^2 then $\alpha = 0$, $\lambda = 1/2$ and $\eta = r/2$. It follows that if the y_i values above are the sorted X_i^2 defined in the previous section and you obtain the values x_i by solving:

$$F(x_i; 0, 1/2, (N-p)/2) = q_i$$

then the Q-Q Plot of the x_i values against the y_i values should be a straight line with slope one and intercept zero. In SIEVER this is tested by a linear regression and associated statistics.

Interpreting the Q-Q Plot

The plot can depart from the ideal for a number of reasons. One is that the errors are not normal so the X_i^2 are not distributed as χ^2. It is known that for the general incomplete Gamma distribution the fitted linear regression will have slope $1/2\lambda$ and intercept α. Also, if the value (η) for which the data become well approximated by a straight line is different from the theoretical value it may mean the effective degrees of freedom are different from (*N-p*).

However, any of these reasons simply relates to the distributions and not the extreme events that represent outliers. As discussed further later, outliers are often indicated by very high values of the y_i data for the high quantiles. In this case, the slope is usually much greater than one and the intercept large and negative.

Once outliers are detected they should be traced to the images where the features have been located and checked very closely. Removing points suspected of being outliers is not always justified as in any well defined distribution there can be large values. SIEVER actually provides a good indication of when there are too few large errors (which might happen if points with large errors are arbitrarily deleted) since in this case the Q-Q plot will often "flatten" at the higher quantiles indicating there has been truncation of larger but still statistically possible errors.

Data Plots to indicate Outliers

In addition to the Q-Q Plot, SIEVER provides for some other plots that can give some assurance that the analysis is working as expected or for the location of outliers. These are listings and plotting of columns of the matrix US and or the Error Matrix.

Listing and Plotting the Matrix US

As noted above, the matrix first two columns US constructed from the SVD can be identified with the modelled "underlying" set of feature coordinates:

$$\begin{bmatrix} \mu_1 & \nu_1 \\ \mu_2 & \nu_2 \\ \cdots & \cdots \\ \mu_M & \nu_M \end{bmatrix} = \begin{bmatrix} s_1 \underset{\sim}{u}_1 & s_2 \underset{\sim}{u}_2 \end{bmatrix}$$

The last N-2 columns can be identified as orthogonal transformations of the errors between the affine model and the data.

If only two images are being considered there will be two columns of the underlying coordinates and two columns of orthogonalised errors. If more than two images have been considered then there will be more possible error plots.

An XY Plot of the first two columns will indicate if the transformation is working well as the coordinates should be recognisable as linear transformations (such as rotations and scaling) of the coordinates in any one of the data sets.

An XY Plot of pairs of the columns of orthogonalised errors can locate outliers or large errors and confirm any findings from the Q-Q Plot. These alternative analyses of the data are important as in many cases outliers so affect the SIEVER model that the actual location is hidden apart from the general lack of fit or improbability of the model result.

Listing and Plotting the Error Matrix

To provide another way to search for errors in the location of features and to identify the image from which they are likely to have come, SIEVER allows you to list and plot the Error Matrix.

The Error Matrix (*E*) is obtained by "rotating the errors back" into the original system. That is, it is a re-construction of the data matrix but with the model components left out:

$$E = \sum_{k=p+1}^{N} s_k \underset{\sim}{u}_k \underset{\sim}{v}_k^T$$

It is the estimated error matrix in the model for the data matrix A into model and error as:

$$A = B + E$$

XY Plots of the columns of E are sometimes very helpful but are also often either hard to interpret of not much more informative than the Q-Q Plot and plots of the columns of US.

Using SIEVER to locate outliers

SIEVER provides a set of tools but the effective location of outlier locations of features needs so skill in its use. The main problem is that outliers, especially when they occur at features with a high Predictive Error weight can affect the model to a degree where the outlier or outliers may well be fitted better than the majority of the features.

Guarding against this is important and is best done by ensuring as much as possible that the features SIEVER identifies as have a large predictive error (PE) weight are as accurate and well placed as possible. A high PE weight indicates that a point has a lot of control over the model. This can be good and often some of the most important features are those with high PE weight. However, if a point with high PE weight is in error it can be very bad for the model.

The existence of a number of outliers to the point where there is not a clear distinction between the fitted errors and the outliers is usually indicated by a very poor fit of the model and a lack of clear support for the rank of the A matrix as 2. All of these situations should lead you to look carefully among the located features and images for the problem points.

What is more common is for the outliers to be clearly indicated as large errors after the model is fitted. These unusually large errors will show up in the Q-Q Plot and in the XY Plots of the columns of US and/or the columns of the estimated error matrix E.

Because of the possibility that eliminating points with large but still statistically feasible errors from the data set will create a truncated distribution it is important to improve the locations of features rather than deleting them. The Q-Q Plot allows you to see if this has occurred. It is important not to over-SIEVE data sets before models are fitted. The objective is to find outliers and not remove points with large but statistically feasible error values.

The objective of SIEVER is to identify and advise. You should always go back the feature identifications in images before action is taken with a specific point. If this is kept in mind SIEVER provides a very powerful means for screening collections of GCPs and preparing the corresponding data sets for MODEL and MOSMOD.

References

Harrison, BA and Jupp, DLB (1990). Introduction to Image Processing. Part TWO of the microBRIAN Resource Manual. CSIRO Australia.

Harrison, BA and Jupp, DLB (1995). Image Classification & Analysis. Part THREE of the microBRIAN Resource Manual. CSIRO & MPA.

Lanczos, C. (1958). Linear systems in self-adjoint form. *Am. Math. Monthly*, **65**, 665-679.