

## Automated grain boundary detection by CASRG

Kingshuk Roy Choudhury <sup>a,\*</sup>, Patrick A. Meere <sup>b</sup>, Kieran F. Mulchrone <sup>c</sup>

<sup>a</sup> Department of Statistics, National University of Ireland, Cork, Ireland

<sup>b</sup> Department of Geology, National University of Ireland, Cork, Ireland

<sup>c</sup> Department of Applied Mathematics, National University of Ireland, Cork, Ireland

Received 12 December 2005; accepted 14 December 2005

Available online 7 February 2006

### Abstract

Constrained automated seeded region growing (CASRG) is an algorithm for automated grain boundary detection. It uses as input a single digitised microphotograph, such as ones obtained from a polarising microscope with an attached digital camera. In addition to this, it requires the user to click on the clasts within the microphotograph that the user wishes to obtain boundaries for. The algorithm requires no subsequent human input. The algorithm is based on the seeded region growing (SRG) algorithm of Adams and Bischof [Adams, R., Bischof, L., 1994. Seeded region growing. *IEEE Transactions on Pattern Analysis Machine Intelligence* 16, 641–647]. We have modified this algorithm to be guided by constraints and to adapt to the heterogeneity of colour information in the image. Imposition of these pre-determined additional conditions enables automated grain boundary detection without human intervention. The accuracy of CASRG has been validated through two benchmarking comparisons; one lithology with low tectonic strain and a second with high strain are used. The CASRG measurements are compared with those from hand drawn boundaries, which are used as a gold standard. Comparison is made using (a) a non-overlap statistic, (b) shape features, (c) strain estimates. In each case, the CASRG method compares very favourably with the gold standard.

© 2006 Elsevier Ltd. All rights reserved.

**Keywords:** Grain boundaries; Seeded region growing; Adaptive threshold; Automated; Constraints; Non-overlap; Shape features; Strain analysis

### 1. Introduction

Strain analysis and the study of regional strain patterns are invaluable tools in interpreting the tectonic history of a region. However, in considering 12 recently published studies, a wide variability in sampling density for finite strain characterisation is demonstrated. The data in Table 1 indicates that, independent of the size of the study area, there is an upper limit of around 30 to the number of samples used. This results in very low sampling densities when the study area becomes large. There might be a number of reasons for this observation, e.g. homogeneity of deformation in an area, availability of suitable exposure, etc. However, we believe that the primary reason is the laborious and time consuming methods available for obtaining the raw data required for strain analysis.

There is one notable exception to the 30 sample limit provided by the study of Mukul and Mitra (1998). They analysed 119 samples of quartzite from an area of 200 km<sup>2</sup>

around the Sheeprock Thrust Sheet, Sevier Fold-and-Thrust belt, Utah, USA. However, they employed a semi-automatic procedure for obtaining the data for strain analysis as described by Mukul (1998). In this paper we develop CASRG, a semi-automatic algorithm for strain analysis that enables rapid and accurate extraction of data for strain analysis. Automation of this process will allow strain analysis studies to break the 30 sample limit and introduce the possibility of statistical analysis of spatial strain variation, as in Mukul (1998). This paper concentrates on the problem of extracting data for strain analysis from sandstones and looks at deformed examples from the Variscides of southwest Ireland (Meere, 1995) and the Moine of northwest Scotland. The CASRG algorithm will yield data that is applicable to strain analysis methods based on marker shape (e.g. the mean radial length method of Mulchrone et al. (2003)) and to methods based on the relative position of markers such as those by Fry (1979) and Mulchrone (2003).

Traditional methods of measurement required sustained use of the polarising microscope with skilled manipulation of the rotating stage and knowledge of the use of various graticules (Ramsay, 1967, section 5.2). With the proliferation of digital cameras, it is now common place to obtain digital images of a field of view as seen through the microscope (Fig. 1a). Digital images are easily manipulated by computer graphic software

\* Corresponding author. Tel.: +353 21 4902546.

E-mail address: K.RoyChoudhury@ucc.ie (K. Roy Choudhury).

Table 1

Area, number of samples and sampling density for a selection of recent studies which included strain analysis, at least in part

Author(s)	Study area (km <sup>2</sup> )	No. Samples	Samples per km <sup>2</sup>
Meere (1995)	70	23	0.32
Srivastava et al. (1995)	$2.5 \times 10^{-9}$	4	$1.6 \times 10^9$
Yin and Oertel (1995)	12	8	0.66
Mukul and Mitra (1998)	200	119	0.59
Roig et al. (1998)	500	8	0.02
Bresser and Walter (1999)	180	15	0.08
Althoff et al. (2000)	800	6	0.01
Hippert and Davis (2000)	12	9	0.75
Hippert and Davis (2000)	18	3	0.16
Hippert and Davis (2000)	4	6	1.50
Simancas et al. (2000)	60	28	0.46
González-Casado and García-Cuevas (2002)	2000	37	0.02
Mulchrone (2002)	55	18	0.33
Bailey and Eyster (2003)	14	8	0.57

packages that allow the outline of clast boundaries to be easily traced. Alternatively the boundaries may be manually traced from a printout and then scanned into a digital image (i.e. as suggested by Mukul (1998) and Mulchrone et al. (in review)). Given a set of such boundaries, it is possible to make the measurements required for strain analysis either manually or using automated methods (Mulchrone et al., in review).

Although methods that require manual identification of boundaries represent a striking improvement on totally manual methods for data extraction, there is room for further improvement. The primary aim of automatic clast boundary detection is to remove the manual step of marking boundaries. The task is especially onerous if one is engaged in a strain mapping study, where thousands of clasts have to be marked for reliable measurements of strain. Concomitantly, one is also seeking increased speed and accuracy in making strain measurements (Meere and Mulchrone, 2003). Speed is guaranteed not so much by the efficiency of the algorithm itself, as by the increase in processing power of computers. The issue of accuracy is of course paramount in any scientific endeavour. Given the nature of the current problem, clasts will always exist where manual marking of boundaries will be better than any automatic identification algorithm. In fact, given time, patience and

practice, manual marking will be at least as good as the best automatic clast boundary detection algorithm. In practice, however, lack of dexterity with the mouse or pen can cause manually identified boundaries to deviate from the true boundary of the clast. In most cases, these errors will be negligible in terms of the accuracy of measurements made on the clast. Therefore, the aim is to develop a method that will deliver parameter estimates which are close (in an average sense) to those obtained by careful manual marking.

Previous work on automated clast boundary detection (e.g. Heilbronner, 2000; Ailleres et al., 1995; Bartozzi et al., 2000), demonstrates the difficulty of the problem. These papers address the difficulty by introducing extra information about the grain boundaries: for example Heilbronner (2000) has utilised multiple images of the same field of view and Bartozzi et al. (2000) use SEM images. Automatic clast boundary identification from a single image is an even harder problem. Clasts will be adjacent to each other and appear to be the same colour (e.g. in Fig. 1b, clasts 10 and 11). It will be very difficult to tell these apart. Fortunately, for strain measurement, we do not have to measure *all* clasts, but only enough for an accurate strain analysis (around 150 according to Meere and Mulchrone (2003)). In any given thin section image, there will be some clasts that appear well defined due to a sharp colour contrast with their immediate neighbourhood. It is the boundaries of these clasts that we will seek to identify.

## 2. Region based identification

Previous work on automatic clast boundary identification (e.g. Heilbronner, 2000; Bartozzi et al., 2000) utilise *edge-detection* based methods to identify the boundary of the clasts. Typically the initial boundaries produced by edge detection have many imperfections (some edges occur *within* the clasts as well as on the actual boundary and sometimes edges are absent on the real boundary). These initial boundaries are then post processed to obtain more realistic boundaries. While this approach appears to work satisfactorily for measurements such as clast count and clast size distribution, the subjectivity introduced by the post processing methods make them unsuitable for strain analysis, where crucial parameters are commonly the *physical* orientation of the clast (as opposed to

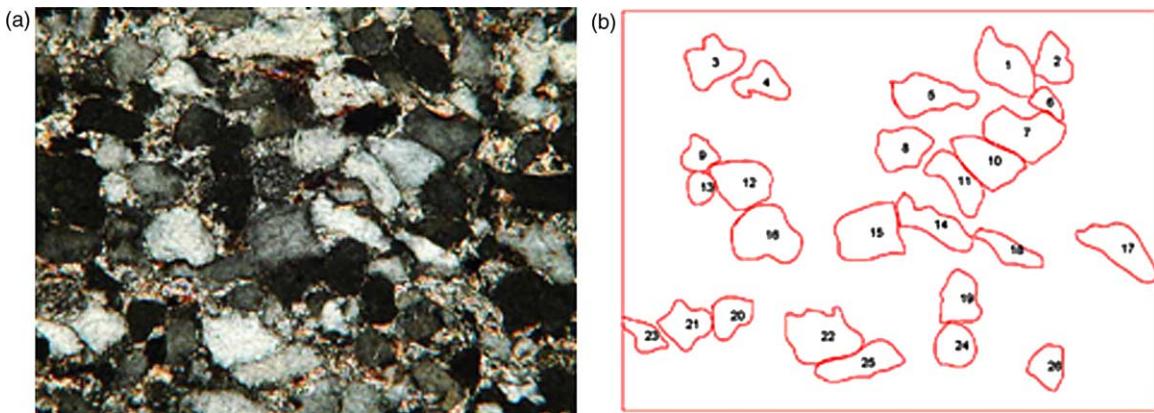


Fig. 1. (a) Original microphotograph. (b) Hand drawn boundaries.

the *c-axis* orientation), aspect ratio or relative location. An alternative approach is taken here, where instead of directly identifying the boundary, we identify a ‘region’, i.e. a set of points that purportedly belong to the clast. From an image processing perspective, region and edge based methods are equivalent. Instead of imperfections in the clast boundary by edge detection, we have holes in the identified clast (see Fig. 5). But for extracting features like major and minor axes, orientation, centroid, etc., the presence of these holes has a far diminished role to play than incorrectly identified edges. Consequently, *no* post processing is required in this approach. In other words, the region based method of identifying clasts is potentially more convenient for an automated feature extraction algorithm.

### 2.1. Seeded region growing algorithm

In the previous section, we have already mentioned that we are interested in identifying a few well-defined grains per field of view. The most convenient method in image processing for identifying a few well-defined regions in an image is the seeded region growing algorithm (SRG) (Gonzalez and Wintz, 1987; Adams and Bischof, 1994). The basic idea in SRG is to start with a point (or *seed*) which we know belongs to the region of interest. The region is then *grown* by adding points that are

similar to the seed. In our variation we deal with one grain at a time, as follows: Basic SRG algorithm

1. Select a seed point. It is added to a queue called a sequential search list (SSL).
2. Remove a point from the SSL. Look at the neighbouring points of this chosen point. If they are *similar* to the region already grown, add them to the region (and the SSL). If they are not, add them to the boundary of the region.
3. Repeat step 3 till the SSL is exhausted. At that point, we have our region identified.

### 2.2. Selecting seeds

It remains to describe how the seed points are selected and how we judge if a point is *similar* to a region. The seed is selected by the user. The user looks at the picture of the thin section and selects the grains they want to identify by clicking at a point somewhere inside the grain of interest. Although this procedure requires human intervention, the clicking itself takes very little time to accomplish. There is very little subjectivity involved, since the algorithm produces more or less the same result no matter where in the region you click, provided the region is of reasonably uniform colour. The effort of the user (and also the subjectivity) is in identifying suitable grains for

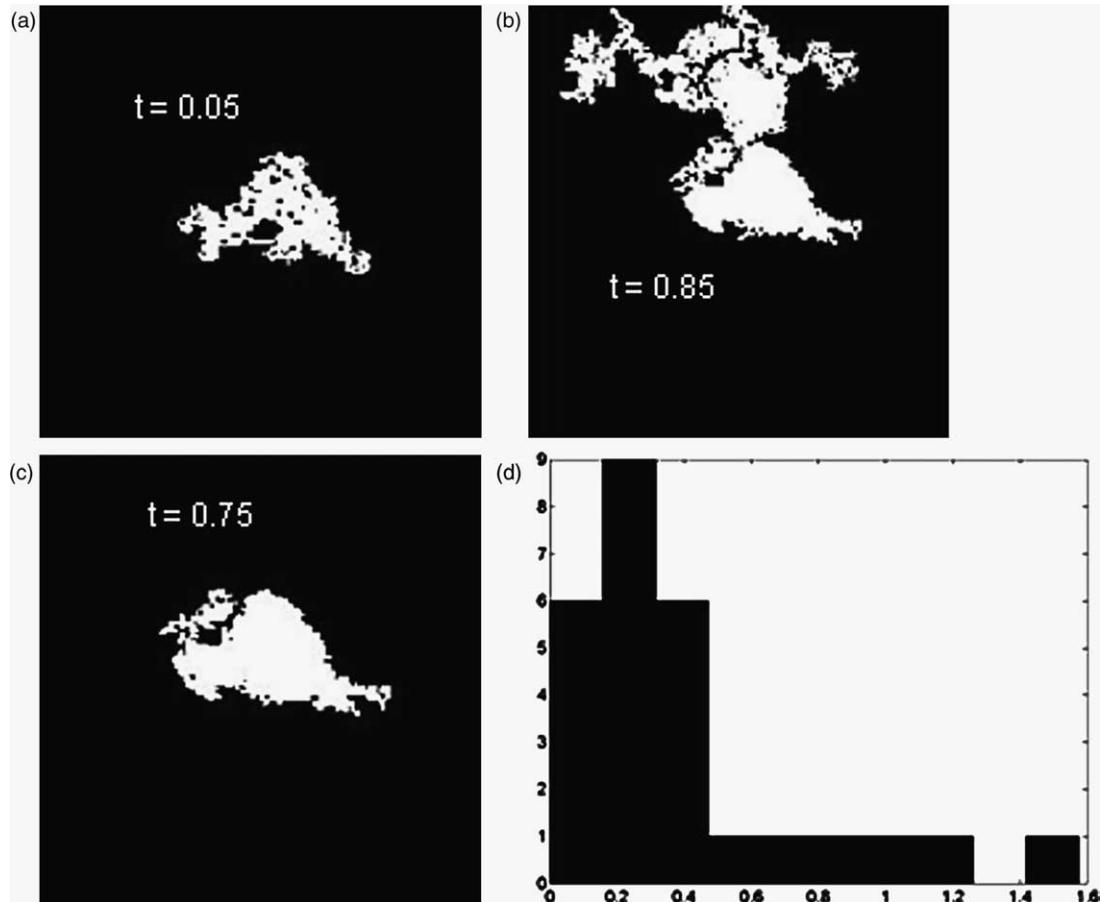


Fig. 2. (a) Threshold too low. (b) Threshold too high. (c) Optimal threshold. (d) Distribution of optimal thresholds across grains.

clicking. Apart from utilising the excellent perception of colour contrast in humans, this step also allows the user to incorporate their expert knowledge of the fabric. For instance they may want to avoid certain grains that appear polycrystalline or belong to a different mineral than the one currently of interest.

### 2.3. Testing pixels for similarity

The basic idea behind ‘testing’ a pixel is to judge whether it is similar, in a sense to be defined, to pixels already identified to be within the region. There are a variety of criteria available for judging whether a point is similar in colour to a region (Haralick and Shapiro, 1985). The criterion of similarity used in this study is now described.

Colour information can be recorded in many different formats. Here it is viewed as a three-dimensional vector  $\mathbf{c} = (c_R, c_G, c_B)$  with each component giving a level (between 0 and 1) of Red, Green or Blue. Thus (0,0,0) would be Black (absence of all colour), (1,0,0) would be Red, (1,1,1) would be White, etc. A simple way to compare whether a point  $p_1$  and a region  $R$  are of similar colour is by measuring the Euclidean distance between their colour vectors  $\mathbf{c}_1$  and  $\bar{\mathbf{c}}$ :

$$d(c_1, \bar{\mathbf{c}}) = \sqrt{(c_{1R} - \bar{c}_R)^2 + (c_{1G} - \bar{c}_G)^2 + (c_{1B} - \bar{c}_B)^2}$$

Here the colour of a region is represented by its average colour value  $\bar{\mathbf{c}}$ . This is just the average of all the colour vectors belonging to points in the region.

In step 2 of the SRG algorithm, we judge whether a new point belongs to the current region  $R$ , we need to check whether this distance is above or below a certain threshold  $t$ , i.e.

$$p_i \in R \text{ if } d(c_i, \bar{\mathbf{c}}) < t$$

The choice of threshold is delicate: it should be such that it will be large enough to allow for natural variation within the region (even relatively uniform regions will have some variability in colour information), but small enough to be able to detect a change from points outside the region. Thus, if  $t$  is too small, the identified region will be too small and vice versa (see Fig. 2a and b). The choice of optimal  $t$  can be made

by trial and error (by a visual comparison of the identified region to the original picture). Because of a variety of colours and brightness levels present in the same image, different grains may have substantially different optimal thresholds. Fig. 2d presents the distribution of optimal thresholds for the grains identified in Fig. 1b. Classical SRG (Gonzalez and Wintz, 1987) uses a single threshold value for all grains. Given the big range of thresholds and the relative sensitivity of grain identification to the choice of threshold (as demonstrated in Fig. 2a–d), we need to choose the optimal threshold separately for each grain. However, to set the optimal threshold manually using the trial and error method may be quite time consuming. Thus the need arises for an automated method of choosing the optimal threshold.

### 2.4. Automated choice of threshold

It is possible to choose the optimal threshold without human intervention. The choice of optimal threshold is guided by the level of variability present in the interior of the clast. The less variable a region, the lower the threshold required. Variability inside a region can be measured by the generalised coefficient of variation, which is defined as follows:

$$\text{CV}(R) = \frac{\sqrt{\frac{1}{n} \sum_i (c_{Ri} - \bar{c}_R)^2}}{\bar{c}_R} + \frac{\sqrt{\frac{1}{n} \sum_i (c_{Gi} - \bar{c}_G)^2}}{\bar{c}_G} + \frac{\sqrt{\frac{1}{n} \sum_i (c_{Bi} - \bar{c}_B)^2}}{\bar{c}_B}$$

Here the sum is overall  $n$  pixels  $i$  in the region  $R$ . Colour data typically have skewed distributions, where brighter regions will have correspondingly higher variances. In order to adjust for varying levels of brightness across clasts, we need to scale the variance by the average brightness level of the region. It may seem natural to calculate the CV for the region that has been grown, but, in practice, this will always give a low value. This is because the region growing algorithm will always select pixels of similar colour information, which means that the region grown by SRG will always have low

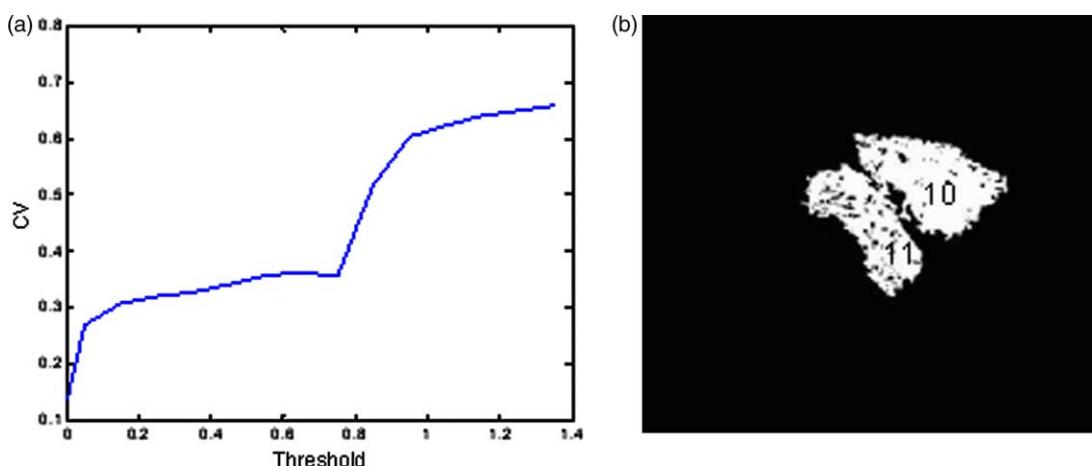


Fig. 3. (a) Generalised coefficient of variation (CV) as a function of threshold. (b) Grains 10 and 11 of Fig. 1b at optimal threshold before application of constraints.

variability. For instance, imperfections in the grain are excluded from this calculation. Instead, we have found it more useful to calculate the CV for an area which *encloses* the grown region. The principle is as follows: if the grown region includes primarily all or part of a grain, the enclosing region should have a low CV overall; if, on the other hand, the region spans two or more grains and/or inter-granular matter, which have different colour schemes, the enclosing region would have a high CV. For our convenience, we have used a fitted ellipse as the enclosing area, but one could choose any other shape, such as a rectangle or a convex hull, to the same effect. The important requirement is that the enclosing shape be as *tight* as possible to the identified region.

As an illustration of this idea, clast 4 in Fig. 1b has an approximate generalised CV of 0.35, whereas the overall CV for the whole image is 0.75. This low variability inside a grain can be utilised for optimal threshold selection. For instance, a low threshold will yield a small region inside a grain, as in Fig. 3. In this case the CV is 0.27. A larger threshold will yield a larger region, but as long as this region is within the grain, as in Fig. 2c (CV=0.35) the CV will not increase substantially. Once the region exceeds the grain, as in Fig. 2d the CV rises sharply to 0.55. This behaviour is shown in Fig. 3a. Based on this property of the CV statistic, a grain can be characterised as the *maximal* region with a CV less than a given cutoff CV value. The choice of this cutoff will depend on the variability in the image. For the image in Fig. 1a (and other similar images), we have empirically established that a cut-off CV value of around 0.4 works satisfactorily.

## 2.5. Imposition of additional constraints

The automated region growing algorithm described above should work well when we have sharp contrast between the grain and its immediate neighbourhood. However, when we

have two grains of similar colour side by side, as for instance in the case for grains 10 and 11 in Fig. 1b, it will be hard to separate them based on variability alone. In this case the CV of the region combining the two grains (shown in Fig. 3b) is around 0.33. So how can one separate them? To do so, we utilise the information provided by the user in the seeding process. By seeding both grains, the user is instructing the program to grow them separate from each other. Therefore any threshold that generates a region containing two or more seed points is too large for the job.

As a general principle, when we have situations where the colour contrast alone is not sufficient to delineate a clast, we must make use of additional constraints, such as user input as described above, or other constraints such as size or shape. The downside of this is that imposition of such constraints often needs additional user input. However, in this case, we are fortunate that the information required is already supplied during the seeding process. This may not always be the case. In this case, we may need to supply additional constraints in the form of ‘spoiling seeds’, i.e. seeds supplied purely for the purposes of limiting the size of the automatically identified region.

The discussion in Section 2 can be summarised in the following algorithm: Constrained automated seeded region growing (CASRG) algorithm: First, the user selects seeds (numbered 1– $k$ ) and a set of spoiling seeds (1– $j$ ). For each seed in this list, the following (automated) loop is carried out:

1. Fix threshold  $t$
2. Grow region by SRG at threshold  $t$
3. Compute CV of grown region
4. if ( $CV < \text{cutoff}$ ) increase threshold;  
else decrease threshold
5. check for constraints (and modify threshold appropriately)
6. Iterate steps 1–5 to convergence of threshold.



Fig. 4. Regions identified by CASRG with hand-drawn boundaries from Fig. 1b superimposed.

The algorithm has been coded for the MATLAB® environment ([www.mathworks.com](http://www.mathworks.com)). This code can be downloaded from <http://euclid.ucc.ie/pages/staff/kingshuk/casrg.zip>. Users should be aware that to run this code they need to have the MATLAB Image processing toolbox installed.

### 3. Methodology for the validation of CASRG

The result of applying the CASRG is shown in Fig. 4. With a few exceptions, the identified regions are broadly similar to the clast boundaries outlined in Fig. 1b. However, there are significant differences. The CASRG identified regions are full of holes that represent areas of non-uniform colour within the targeted clasts. While small holes can be dealt with easily by employing simple morphological operations such as ‘closing’ (see e.g. Matheron, 1975), larger holes are not as tractable. However, such holes do not seriously affect the value of many quantitative features obtained from these regions. Another difference is that many identified regions have appendages that lie outside the actual clast boundary. This is primarily due to the presence of areas of similar colour adjacent to the clast.

While visual inspection of the results gives us a rough idea, ‘objective’ measures of the quality of region fit are essential to make accurate comparisons. Of course, to measure the quality of fit, we need to know the ‘true’ regions. As with most real problems there are no ‘absolutely true’ regions available to us. However, in the introduction, we have argued that the manually drawn boundaries, if carefully drawn by an expert, are potentially the closest approximation to the grains in the field of view. Therefore, we will use the hand drawn boundaries, for example those in Fig. 1b as the ‘gold standard’ against which to compare the CASRG regions.

We will evaluate three different aspects of the quality of fit:

- The ‘non-overlap’ of the CASRG regions with the ‘gold standard’
- A comparison of extracted features between CASRG regions and the ‘gold standard’
- Comparison of strain estimates from CASRG regions and the ‘gold standard’

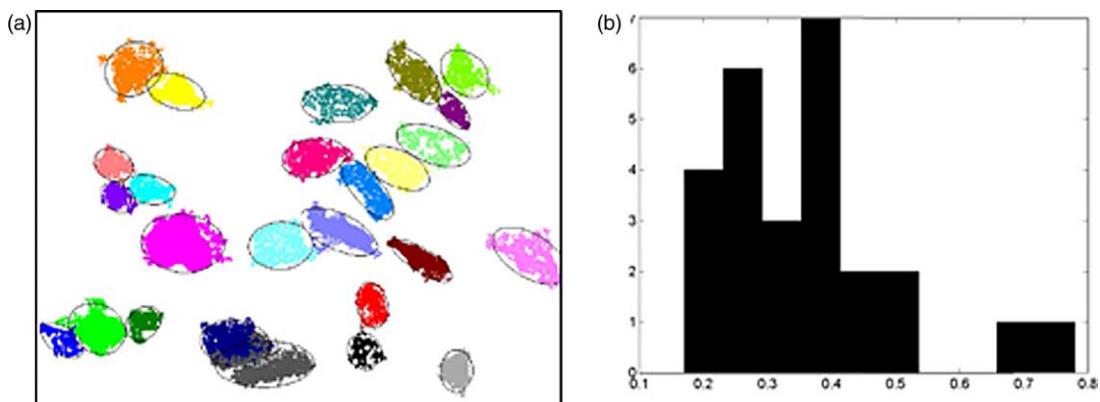


Fig. 5. (a) Ellipses fitted by moment method to regions identified by CASRG. (b) Distribution of non-overlap statistic for CASRG regions with hand drawn boundaries as reference.

### 3.1. Non-overlap statistic

Given two regions A and B, the non-overlap statistic  $\text{no}(A, B)$  (Mulchrone et al., in review) measures the area where the two regions do not overlap. The non-overlap area is divided by the area of B (the base area) to express the non-overlap as a scale-invariant proportion. Mathematically, this statistic can be expressed using the symmetric difference of two sets:

$$\text{no}(A, B) = \frac{\text{area}((A - B) \cup (B - A))}{\text{area}(B)}$$

The non-overlap statistic is a unit-less quantity that can be used to measure how similar (in shape, size and location) two regions are to each other. If two regions are identical, their non-overlap should be zero. The bigger the non-overlap, the more they are dissimilar. It is possible for the non-overlap to be greater than 1 (for instance, if A and B do not intersect with each other), but that would indicate a pretty bad fit. Fig. 5b shows us the distribution of the non-overlap values for regions drawn using CASRG (Fig. 4) versus those drawn by hand (Fig. 1b), with the hand drawn regions used as the base area B. The majority of the regions have a non-overlap statistic between 15 and 55%. Two values are between 65 and 80%. The average (median) value is 36%. While these values do not have an immediate interpretation, these values can be used as a benchmark for future algorithms to improve on.

### 3.2. Feature based comparison

The primary application of CASRG in structural geology is for the purposes of mensuration of clasts. Common features requiring mensuration are the area, centroid, major and minor axes and the orientation of each clast. Apart from their relevance to procedures such as strain analysis, these features are intrinsic geometric properties of the objects, which describe their shape, size and location. Therefore comparisons of such features are suitable for evaluating the quality of fit. One advantage of region based processing is that such features can be quickly computed from the moments of pixel co-ordinates of points belonging to each region. For instance, the area is obtained as the ‘zero’-th moment and the centroid as the first

moment. The other parameters mentioned can be obtained using an eigen decomposition of the ‘moments of inertia’, or second moment matrix, of these points (Mulchrone and Roy Choudhury, 2004). A graphical representation of these features, in the form of fitted ellipses, is shown in Fig. 5a.

### 3.3. Strain analysis

The features described in Section 3.2 are often used for strain analysis in structural geology. Therefore an application specific validation of CASRG is by comparison of strain estimates with the gold standard. A number of methods of finite strain calculation appear in the literature (e.g. Robin, 1977; Yu and Zheng, 1984; Robin and Torrance, 1987; Mulchrone and Meere, 2001; Mulchrone et al., 2003). For reliable strain calculation by the  $R_p/\phi$  method, we need at least 150 clasts (Meere and Mulchrone, 2003). For this purpose, we require larger scale examples, which are considered in the next section.

## 4. Large scale examples

The photomicrograph in Fig. 1a is of sufficiently high magnification so that problems associated with automated clast boundary detection could be appreciated by visual inspection. However, for reliable strain analysis, an example with far greater number of clasts is required. For this purpose, we have chosen two examples, one with low strain (Fig. 6a) and one with high strain (Fig. 6d). The low strain example is an Upper Devonian quartz arenite from the Variscides of southwest Ireland, while the higher strain example is a deformed Cambro-Ordovician quartzite from the Moine of northwest Scotland. Although it is difficult to examine individual clasts in any detail, it appears from visual inspection of Fig. 6c–f that CASRG does a reasonable job of identifying the majority of the clasts, although it is unsuccessful in a few. It should be noted that the choice of clasts was first made during the process of hand drawing the boundaries in Fig. 6c and e. CASRG was

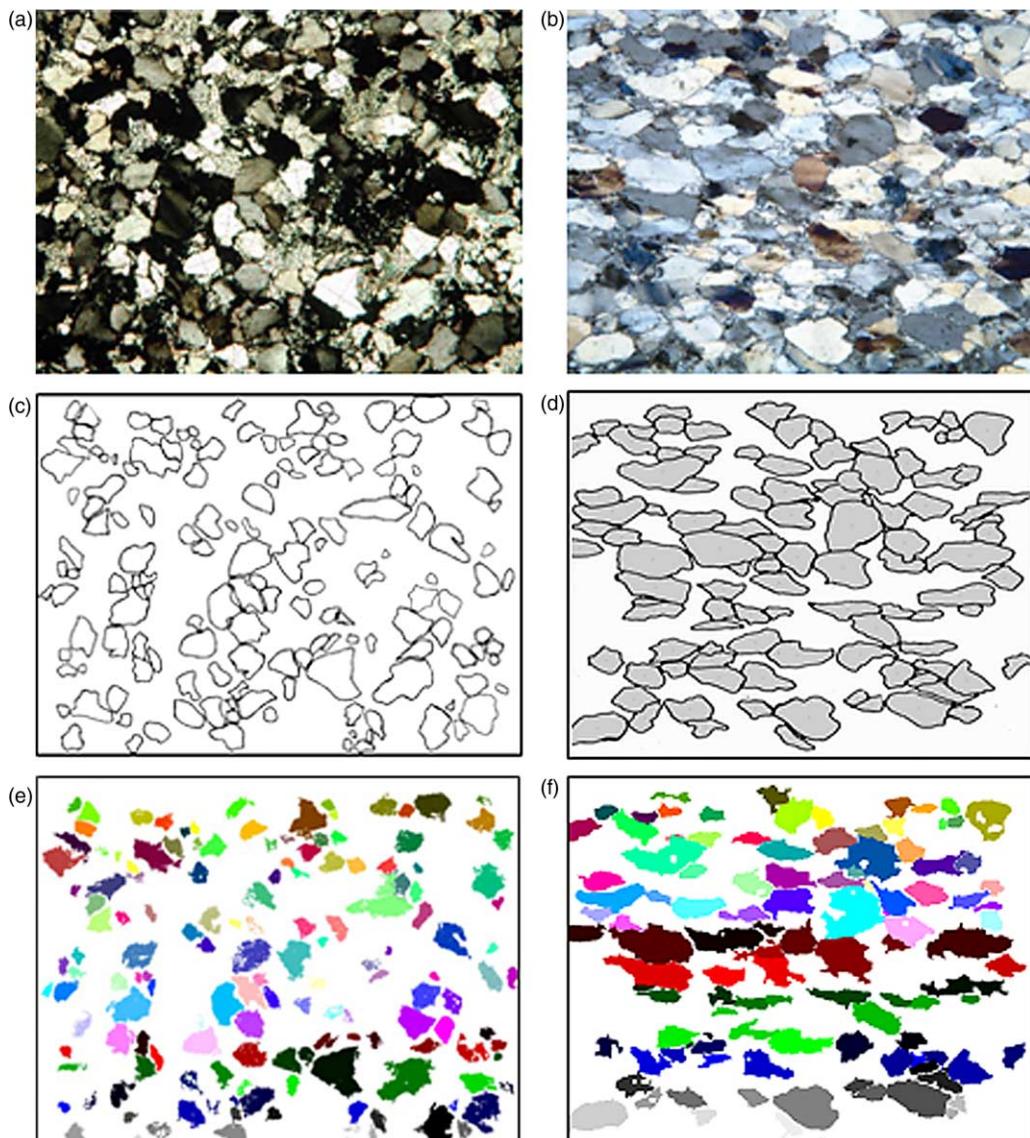


Fig. 6. (a) Low strain example. (b) Low strain boundaries. (c) Low strain regions. (d) High strain example. (e) High strain boundaries. (f) High strain regions.

subsequently applied on all the clasts selected in this process. Given the large number of clasts in this example, it is not feasible to discuss each of these clasts individually. Therefore statistical methods are needed to check the correspondence between the various methods.

#### 4.1. Non-overlap comparisons

As described in Section 3.1, the fidelity of the regions identified by the CASRG method (Fig. 6a and d) can be measured by comparing them with the hand drawn boundaries (Fig. 6b and e) using the non-overlap statistic. The distribution of the resulting non-overlap values is given in Fig. 7. As can be seen from the histograms (Fig. 7a and b), both distributions have a heavily skewed shape, with the majority of values less than 0.5, but a reasonable number greater than 0.5. The average (median) non-overlap value for low strain is about 0.31, while that for high strain is 0.37.

To explore the pattern of non-overlap values, we look at a plot of the non-overlap value against the clast area (Fig. 7c and d). The results are somewhat contradictory. In Fig. 7c, the plot clearly shows that the non-overlap generally decreases with the size of the clast. This is not unexpected, given that the non-overlap statistic is a measure of relative error. It can be argued that if the magnitude of error remains the same, the relative error will decrease with increased size. A more mathematical

discussion of this result is given by Mulchrone et al. (in review). Even accounting for this decreasing trend, one set of points in the extreme left of Fig. 7c stand out. These points correspond to clasts of a very small size, which also have a very high non-overlap value. This cluster of points contains all clasts with a non-overlap value of 1 or more, with the exception of two clasts of intermediate size. Clearly CASRG does not identify such small clasts very well. It is a moot point whether measurements made on these clasts by other methods also suffer similarly. It can be argued (with some justification) that any measurements made (by CASRG method) on these clasts be treated with a pinch of salt. One option is to ignore such clasts from all further calculations. In practice, the judgement of which clasts to omit cannot be made on non-overlap values (which cannot be computed as hand drawn boundaries will not be available). An obvious alternative is to use the size of the clast. Based on this observation, we suggest omitting clasts of size 500 or lower. The size threshold for exclusion of clasts is somewhat arbitrary, given that there are errors in all automatically identified regions. This line of reasoning would lead to calculations where observations are given weight roughly proportional to the clast size, but for reasons of conciseness, will not be pursued here. Another factor that can cause poor identification is if the clasts are in extinction. In this case, they are often hard to distinguish from the background, which also appears dark (as it is a cross polar image).

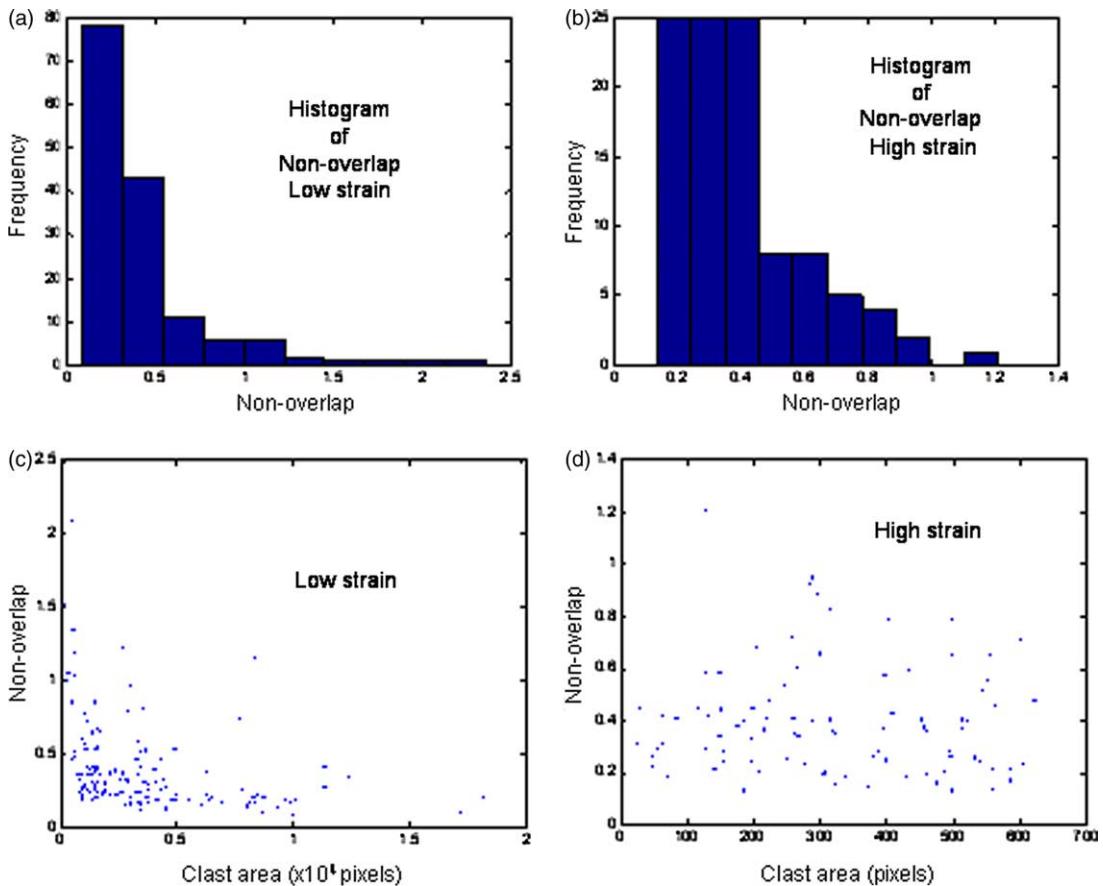


Fig. 7. (a) Histogram of non-overlap for low strain example. (b) Histogram of non-overlap for high strain example. (c) Variation of non-overlap against clast size for low strain example. (d) Variation of non-overlap against clast size for high strain example.

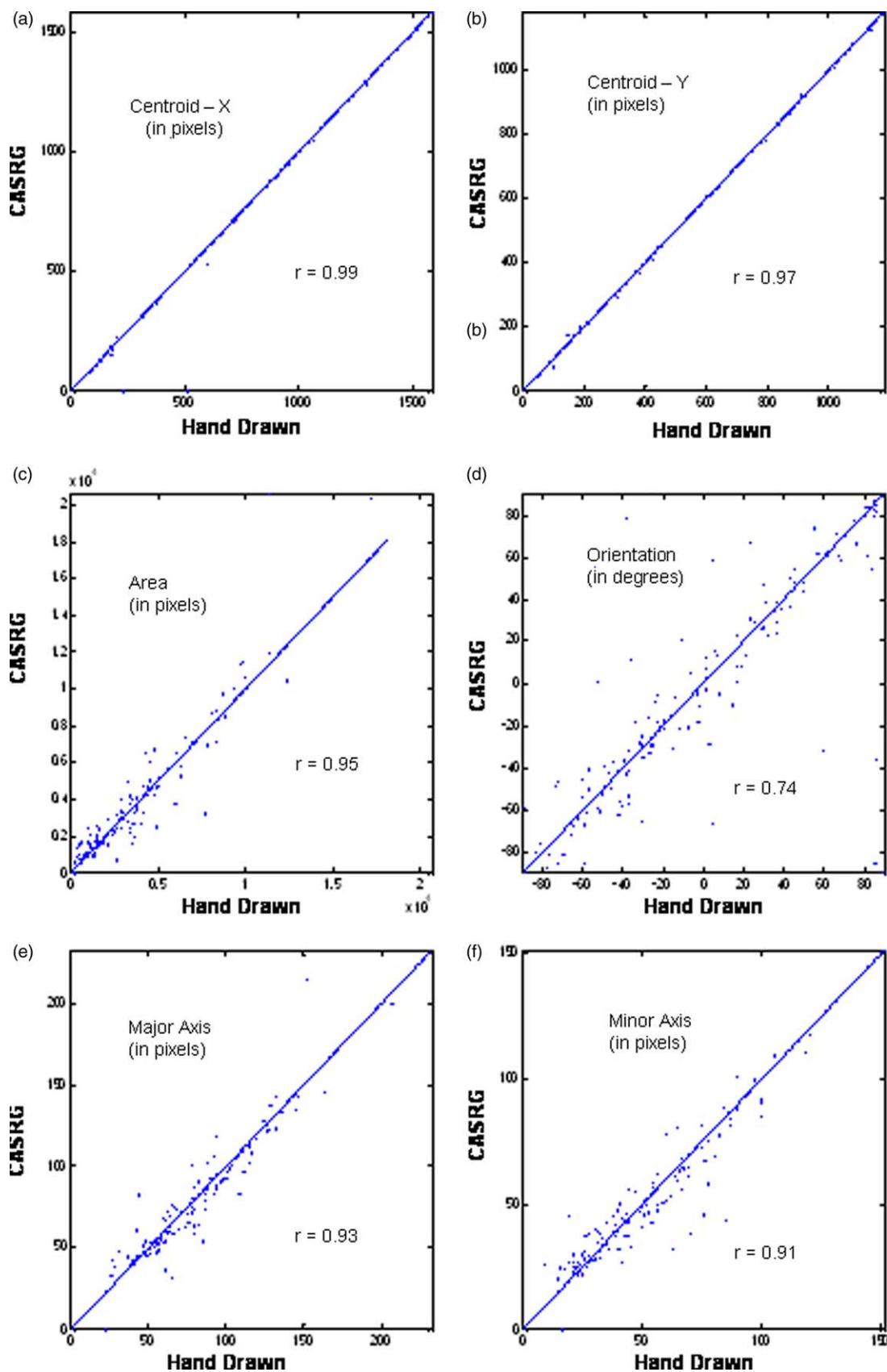


Fig. 8. Comparison of features of CASRG vs. hand drawn regions for low strain example.

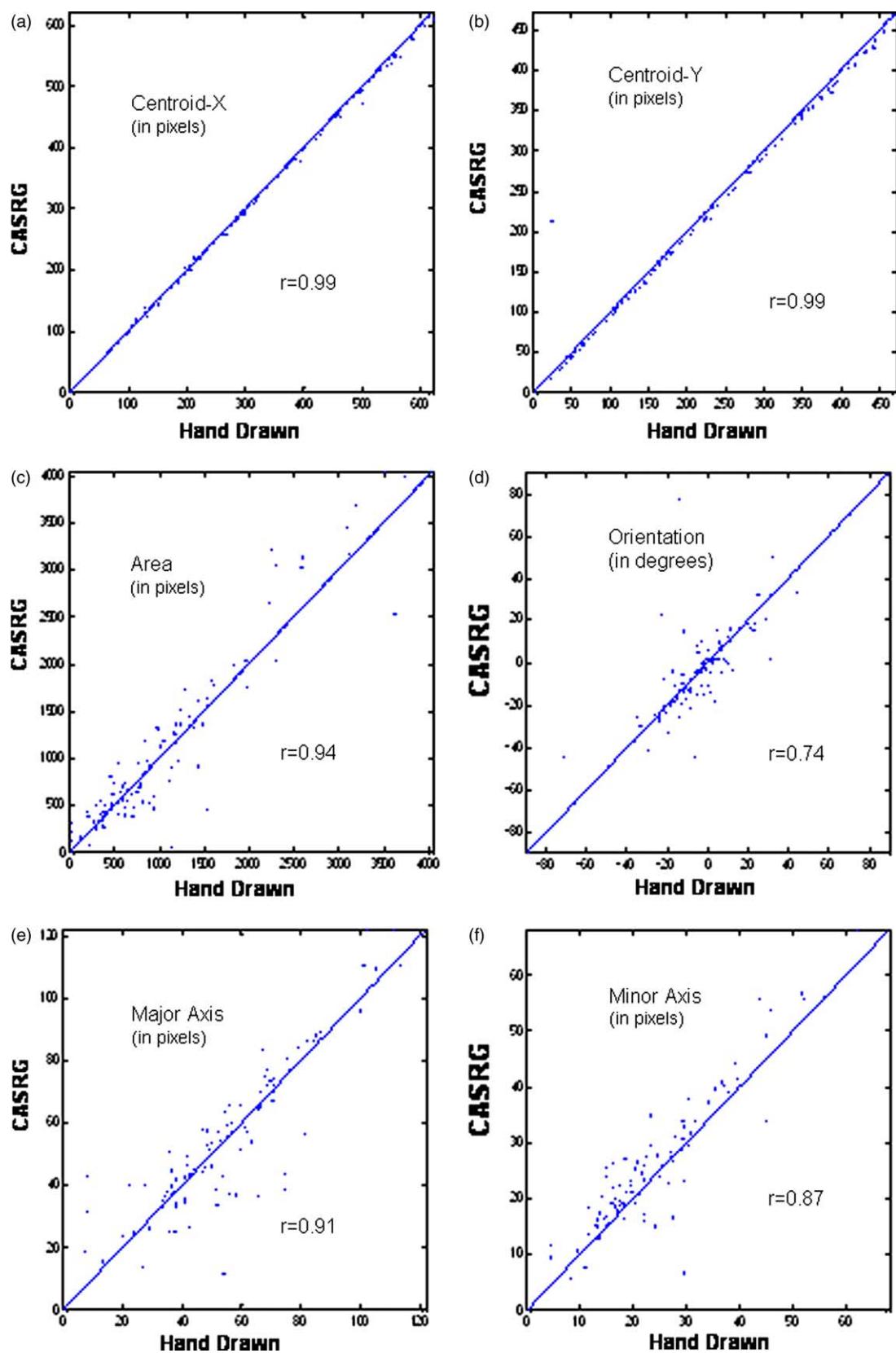


Fig. 9. Comparison of features of CASRG vs. hand drawn regions for high strain example.

The average non-overlap value for such clasts is 0.53, which is substantially higher than the overall average.

We now turn our attention to the plot for high strain (Fig. 7d). The pattern is not quite the same, in that there are a fair few large grains with larger non-overlap values and not that many small grains with large overlap values. It should be noted that the scale of these images is different: the low strain image is  $1200 \times 1600$  pixels in size while the high strain image is  $480 \times 640$ . Therefore the areas in Fig. 7d need to be multiplied by a factor of  $2.5 \times 2.5 = 6.25$  when comparing with Fig. 7c. But that alone does not explain the relatively larger values of non-overlap for the larger clasts. On examination of the image in Fig. 6d, we see that there is a lot more undulose extinction in this field of view compared with Fig. 6a, no doubt due to the presence of greater strain. Undulose extinction means that there is systematic variation in colour from one part of the clast to the other. This typically causes CASRG to leave out parts of the clast that are different in colour from the area where the seed lies. This leads to high non-overlap with the hand drawn boundaries, because experts can in most cases differentiate between undulose extinction and grain boundaries. The reason this affects large clasts more than small ones is because large clasts are more likely to suffer from undulose extinction than smaller ones. Despite these problems for a number of clasts, it is reassuring to note that the average non-overlap for the high strain example (0.37) is only slightly higher than that of the low strain example (0.31). One mitigating factor in this comparison is that the high strain example has far fewer grains in extinction, which caused the average performance in the low strain case to suffer. Other factors that can cause poor identification are polycrystalline clasts and neo-grain formation. Clasts with a high concentration of inclusions would also cause difficulties.

#### 4.2. Comparison of features

Feature based comparisons are important from the viewpoint of applications. These comparisons between features, namely centroid, area, major and minor axis and orientation, extracted from CASRG and hand drawn boundaries are shown in Figs. 8 and 9. As explained in Section 4.1, we have excluded very small clasts (clasts with area less than 500 and 75, respectively, for the low and high strain case) from these comparisons. The diagonal line through these figures is the line  $y=x$ , which represents a perfect match between two sets of measurements.

If we first look at Fig. 8, the low strain example, we can see that there are roughly equal numbers of points on either side of the line in all six plots. This indicates that one method does not systematically over- or underestimate the parameters with respect to the other method. Moreover, most points tend to be very close to this line. This pattern gives rise to a high linear correlation between the two sets of measurements. In particular, the centroids obtained by both methods appear virtually identical in most cases. The correlation between the two measurements is less marked for the orientation parameter. This is caused by some clasts for which the orientation measurement differs widely, although most tend to agree.

Some of this disagreement is spurious, caused by the artificial separation of  $+\pi/2$  and  $-\pi/2$ . In actual terms, these two orientations are the same. It is just that they cannot be reconciled on a linear plot. Adjusting for these points will make the correlation substantially higher (0.85). The other features, namely size, major and minor axes, are all estimated very well. In relation to the data itself, two observations that can be made are: (a) the orientation plot appears to have an approximately uniform spread of values between  $+\pi/2$  and  $-\pi/2$ ; (b) the range of values in the major and minor axes are close to each other, indicating low axial ratio, as one would expect in a low strain situation.

In Fig. 9 (high strain example), too, the correlations are all very high except for the orientation, indicating good agreement between CASRG and the hand drawn boundaries. Correlations for the major and minor axes are slightly lower than those for the low strain case. There appears to be a slight downward bias in the CASRG y centroid measurements (most points in this plot are below the line  $y=x$ ). This could be an artefact caused by resizing of the hand drawn boundary image, whose size is enhanced to facilitate drawing. Despite this, the agreement in the two sets of centroid measurements is still remarkable. In relation to the data from the high strain example, we can note that: (a) the range of orientations appears to be restricted between  $+\pi/4$  and  $-\pi/4$ , indicating the presence of a strong fabric. (b) The range of values in the major axes is roughly double that of the minor axes indicating an axial ratio of two or more.

#### 4.3. Comparison by strain analysis

A number of methods of finite strain calculation appear in the literature based on the shape characteristics of strain markers (e.g. Robin, 1977; Yu and Zheng, 1984; Robin and Torrance, 1987; Mulchrone and Meere, 2001; Mulchrone et al., 2003) or based on the position of nearest neighbours (e.g. Fry, 1979; Mulchrone, 2003). Results of strain calculations done using the algorithms given in Mulchrone et al. (2003) (mean radial length) and Mulchrone (2003) (nearest neighbour) are shown in Table 2. Calculations done using the other methods

Table 2

A comparison of finite strain estimates ( $R_s$ ,  $\phi$ ) using the method of Mulchrone et al. (2003) from data acquired manually and using CASARG

	$R_s$ (manual)	$R_s$ (CASARG)	$\phi$ (manual)	$\phi$ (CASARG)
Low strain sample				
Mean radial length	(1.03) <b>1.08</b>	(1.03) <b>1.09</b>	(−44) <b>131</b>	(−43) <b>−33</b>
Centre to centre	(1.01) <b>1.06</b>	(0.01) <b>1.01</b>	(−42) <b>−15</b>	(−42) <b>38</b>
High strain sample				
Mean radial length	(1.86) <b>1.98</b>	(1.75) <b>1.85</b>	(−2) <b>−1</b> (2)	(−5) <b>−4</b>
Centre to centre	(1.28) <b>1.70</b>	(1.25) <b>1.66</b>	(−19) <b>3</b> (25)	(−25) <b>−2</b>

mentioned gave similar results and are omitted for brevity. Table 2 shows us that the  $R_s$  value estimated using manual and CASRG derived data are virtually identical. The wide confidence interval for the  $\phi$  parameter in the low strain case is expected (Meere and Mulchrone, 2003). The remarkable similarity of the strain results can be attributed to the fact that strain calculation is, broadly speaking, an averaging procedure. In the previous section, it was demonstrated by plots of the measurements that there were no relative biases between the methods. This combined with the averaging possibly leads to a cancellation of relative errors. A formal error analysis is not presented here.

## 5. Discussion and conclusions

The CASRG algorithm is an attempt to address what from an image processing perspective is a difficult problem. As an image, a sandstone microphotograph has quite complex structure, both within and outside the clast components. Use of colour information alone fails to deliver satisfactory grain boundary detection. Other authors have addressed this problem by injecting additional information, such as multiple images of the same field of view (Heilbronner and Pauli, 1993; Heilbronner, 2000) or orientation contrast images using SEM (Bartozzi et al., 2000). Acquisition of such images is often a tedious and time-consuming task. Moreover, it may require the use of special equipment (such as specialised electron microscopes) to which many geologists may not have access. In the absence of such information, i.e. limiting ourselves to a single microphotograph, we have attempted to derive additional information by user input. The nature of input required (seeding) is such that from the user's perspective it is quick, simple to perform and reasonably objective (i.e. user independent). From an image processing point of view, however, the seeding is very informative. It provides both a starting point and a set of constraints that the resulting regions must follow. Another strong point of this approach is that it does not require the resulting clasts to be internally homogenous (i.e. the region can have 'holes').

From an automation viewpoint, the strength of the CASRG algorithm is that it does not require human intervention subsequent to the seeding process. This is achieved by an automatic choice of threshold, which in turn is derived from a measure of homogeneity of the grain identified. In contrast, other algorithms, such as the Lazy Grain Boundary method proposed in Heilbronner (2000) or that in Bartozzi et al. (2000), involve human intervention in a series of steps leading to grain boundary detection. Ideally, one would like to eliminate the seeding procedure as well, i.e. to have a completely automated procedure. However, any such algorithm will not possess the information inherent in seeding, thus making the problem that much harder.

Measurements made from CASRG regions appear to be in close agreement with those made by careful manual measurement. However, it must be remembered that manual measurements are themselves subject to error (Mulchrone et al., in review). These observations are only empirical:

a theoretical analysis of the approximation qualities of CASRG does not exist at present. There are obvious limitations to the applicability of CASRG. It works in a scenario where the clasts have (roughly) the same colour internally, i.e. for clasts such as quartz grains. Thus it will fail in situations where clasts have a more complicated structure, such as twinned feldspars. Even in this limited setting, CASRG appears not to perform well in certain situations. In particular, three situations that we have noted are (a) when clast sizes are small, (b) when clasts are in extinction and (c) when there is significant undulose extinction. These problems can be significantly overcome by adopting a seeding selection protocol that selects larger clasts that are optically not in extinction and exhibit low internal contrast due to undulose extinction. When one considers that the optimum sample size for most stain analysis methods is less than 200 data points (Meere and Mulchrone, 2003), there should be no difficulty in obtaining sufficiently large data sets with such a protocol.

Like many automation processes, there is a trade-off here between speed and accuracy. What the validation in Section 4 demonstrates is that within the limitations discussed above, CASRG has the potential of saving time, with reasonably accurate answers. This is particularly true in the context of the features examined in Section 4.2. Although the CASRG algorithm itself is quite computationally intensive (requiring several minutes on a fast workstation), this time should not be a consideration because (a) it requires no subsequent intervention after the seeding and (b) this time will decrease with more efficient algorithms and faster computers.

There are undoubtedly aspects of this algorithm that can be improved by subsequent research. One possibility that suggests itself is the incorporation of additional, case specific, constraints. Also the nature of the algorithm is such that it should prove relatively straightforward to extend it to situations where we have more information, such as multiple polarising angles.

## References

- Adams, R., Bischof, L., 1994. Seeded region growing. *IEEE Transactions on Pattern Analysis Machine Intelligence* 16, 641–647.
- Ailleres, L., Champenois, M., Macaudiere, J., Bertrand, J.M., 1995. Use of image analysis in the measurement of finite strain by the normalized Fry method: geological implications for the 'Zone Houillère' (Briançonnais zone, French Alps). *Mineralogical Magazine* 59, 179–187.
- Althoff, F., Barbey, P., Boullier, A., 2000. 2.8–3.0 Ga plutonism and deformation in the SE Amazonian craton: the Archaean granitoids of Marajoara (Carajás Mineral Province, Brazil). *Precambrian Research* 104, 187–206.
- Bailey, C.M., Eyster, E.L., 2003. General shear deformation in the Pinaleño Mountains metamorphic core complex, Arizona. *Journal of Structural Geology* 25, 1883–1892.
- Bartozzi, M., Boyle, A.P., Prior, D.J., 2000. Automated grain boundary detection and classification in orientation contrast images. *Journal of Structural Geology* 22, 1569–1579.
- Bresser, G., Walter, R., 1999. A new structural model for the SW Irish Variscides: the Variscan front of the NW European Rhenohercynian. *Tectonophysics* 309, 197–209.
- Fry, N., 1979. Random point distributions and strain measurement in rocks. *Tectonophysics* 60, 806–807.

Gonzalez, R.C., Wintz, P., 1987. Digital Image Processing. Addison-Wesley, Reading, Massachusetts.

González-Casado, J.M., García-Cuevas, C., 2002. Strain analysis from calcite e-twins in the Cameros basin, NW Iberian Chain, Spain. *Journal of Structural Geology* 24, 1777–1788.

Haralick, R.H., Shapiro, L.G., 1985. Image segmentation techniques. *Computer Vision, Graphics, and Image Processing* 29, 100–132.

Hippert, J., Davis, B., 2000. Dome emplacement and formation of kilometre-scale synclines in a granite–greenstone terrain (Quadrilátero Ferrífero, southeastern Brazil). *Precambrian Research* 102, 99–121.

Heilbronner, R., 2000. Automatic grain boundary detection and grain size analysis using polarization micrographs or orientation images. *Journal of Structural Geology* 22, 969–981.

Heilbronner, R., Pauli, C., 1993. Integrated spatial and orientation analysis of quartz c-axes by computer-aided microscopy. *Journal of Structural Geology* 15, 369–382.

Matheron, G., 1975. Random Sets and Integral Geometry. John Wiley & Sons, New York.

Meere, P.A., 1995. The structural evolution of the western Irish Variscides: an example of obstacle tectonics? *Tectonophysics* 246 (1–3), 97–112.

Meere, P.A., Mulchrone, K.F., 2003. The effect of sample size on geological strain estimation from passively deformed clastic sedimentary rocks. *Journal of Structural Geology* 25, 1587–1595.

Mukul, M., 1998. A spatial statistics approach to the quantification of finite strain variation in penetratively deformed thrust sheets: an example from the Sheeprock Thrust Sheet, Sevier Fold-and-Thrust belt, Utah. *Journal of Structural Geology* 20, 371–384.

Mukul, M., Mitra, G., 1998. Finite strain and strain variation analysis in the Sheeprock Thrust Sheet: an internal thrust sheet in the Provo salient of the Sevier Fold-and-Thrust belt, Central Utah. *Journal of Structural Geology* 20, 385–405.

Mulchrone, K.F., 2002. A statistic for estimating strain with confidence intervals from deformed line distributions with an application to schists and gneisses of the Western Gneiss Region, west central Norway. *Journal of Structural Geology* 24, 545–556.

Mulchrone, K.F., 2003. Application of Delaunay triangulation to the nearest neighbour method of strain analysis. *Journal of Structural Geology* 25, 689–702.

Mulchrone, K.F., Meere, P.A., 2001. A Windows program for the analysis of tectonic strain using deformed elliptical markers. *Computers and Geosciences* 27, 1253–1257.

Mulchrone, K.F., Roy Choudhury, K., 2004. Fitting an ellipse to an arbitrary shape: implications for strain analysis. *Journal of Structural Geology* 26, 143–153.

Mulchrone, K.F., O'Sullivan, F., Meere, P.A., 2003. Finite strain estimation using the mean radial length of elliptical objects with bootstrap confidence intervals. *Journal of Structural Geology* 25, 529–539.

Mulchrone, K.F., Meere, P.A., Roy Choudhury, K., 2005. SAPE: a program for semiautomatic parameter extraction for strain analysis. *Journal of Structural Geology* 27, 2084–2098.

Ramsay, J.G., 1967. Folding and Fracturing of Rocks. McGraw Hill, New York.

Robin, P.F., 1977. Determination of geologic strain using randomly oriented strain markers of any shape. *Tectonophysics* 42, T7–T16.

Robin, P.F., Torrance, J.G., 1987. Statistical analysis of the effect of sample size on paleostain calculation. I. Single face measurements. *Tectonophysics* 138, 311–317.

Roig, J., Faure, M., Truffert, C., 1998. Folding and granite emplacement inferred from structural, strain, TEM and gravimetric analyses: the case study of the Tulle antiform, SW French Massif Central. *Journal of Structural Geology* 20, 1169–1190.

Simancas, J.F., Galindo-Zaldívar, J., Azor, A., 2000. Three-dimensional shape and emplacement of the Cardenchosa deformed pluton (Variscan Orogen, southwestern Iberian Massif). *Journal of Structural Geology* 22, 489–503.

Srivastava, H.B., Hudleston, P., Earley III, D., 1995. Strain and possible volume loss in a high-grade ductile shear zone. *Journal of Structural Geology* 17, 1217–1231.

Yin, A., Oertel, G., 1995. Strain analysis of the Ninemile fault zone, western Montana: insights into multiply deformed regions. *Tectonophysics* 247, 133–143.

Yu, H., Zheng, Y., 1984. A statistical analysis applied to the  $R_p/\phi$  method. *Tectonophysics* 110, 151–155.