An Evaluation Measure of Image Segmentation Based on Object Centres

J.J. Charles[1], L.I. Kuncheva[1], B. Wells[2], and I.S. Lim[1]

[1] School of Informatics, University of Wales, Bangor, LL57 1UT, UNITED KINGDOM
[2] Conwy Valley Systems Ltd, UNITED KINGDOM**

Abstract. Classification of organic materials obtained from rock and drill cuttings involves finding multiple objects in the image. This task is usually approached by segmentation. The quality of segmentation is evaluated by matching the whole detected objects to a reference segmentation. We are interested in representing each object by a single reference point called the "centre". This paper proposes an evaluation measure of image segmentation for such representation. We argue that measures based only on distance between obtained centres and a set of predefined centres are insufficient. The proposed measure is based on a list of desirable properties of the segmentation. The three components of the measure evaluate the under/over segmentation of the objects, the proportion of centres placed in the background rather than in objects, and the distance between the guessed and the true centres. The ability of the measure to distinguish between segmentation results of different quality is illustrated on three sets of examples including an image containing microfossils and pieces of inert material.

Keywords: Image segmentation, evaluation measures, discrepancy methods, microfossils, palynomorphs

1 Introduction

Image segmentation is a fundamental process within automatic image analysis. The large variety of practical applications has resulted in a spectrum of generic and specific segmentation algorithms being currently available. The choice of a segmentation algorithm suitable for the problem at hand is not simple. To aid this choice, here we propose a measure of quality of object segmentation using "centres" of the objects.

A survey was conducted by Zhang [1] where evaluation methods for image segmentation were categorised as analytical and empirical. Analytical methods examine the principles and properties of the segmentation algorithms themselves. Empirical methods evaluate the output of the segmentation algorithms on test images. Of the two groups, Zhang recommends the empirical methods.

** *To appear in the Proceedings of ICIAR 2006, Portugal*

The quality of segmentation of an image is judged by the so called "goodness" measure [1]. Examples of goodness measures are the entropy of the partitioned image, intraregion uniformity, region shape, colour uniformity etc. Goodness measures are usually defined by human perception of the ideal segmentation. They are evaluated on the segmented image alone without requiring a reference segmentation, i.e. without a ground truth. Empirical methods that use goodness measures are known as *goodness methods*. Alternatively, the performances of segmentation algorithms can be evaluated relative to a ground truth; these types of methods are called *discrepancy methods*. They measure the inconsistency or some form of distance between the ground truth (ideal segmentation) and the actual segmented image.

There are four types of low level discrepancy approaches presented by Beauchemin and Thompson [2]:- pixel, area, point-pair and boundary. The pixel-based discrepancy approach is the most common one and consists of counting the number of misclassified pixels in the segmentation output relative to a reference partition. Using a similar approach Cardoso and Corte-Real [3] formulate a general measure, an important asset of which is that it is a metric. Area-based methods evaluate area of overlap between corresponding segments [4–6] while boundary-based schemes compare the perimeters of the segments. The point-pair discrepancy approach measures the agreement between two segmented images [7] without explicitly solving the correspondence problem between the regions. Segmentation produces a partition of the pixels and hence can be thought of as a clustering technique. Thus the discrepancy between the obtained segmentation and a ground truth segmentation can be evaluated by any measure of agreement or similarity between two partitions. Along with the Rand index, various other measures of similarity have been proposed in the literature, the most widely used being Jaccard index, adjusted Rand index, correlation, mutual information and entropy [7–9].

In this paper we propose an evaluation measure that belongs to the point-pair group within the discrepancy approach. The segmented image is represented as a set of coordinates of object centres. A reference image is used whereby the centres are marked by hand. The proposed measure consists of three indices evaluating different aspects of the positioning of centres. The rest of the paper is organised as follows. Section 2 describes the real-life problem which prompted this study. The new measure is proposed in Section 3. Toy examples and real images are used to demonstrate the insufficiency of simple distance metrics for evaluating the match between manually placed centres and centres obtained from automatic segmentation. Experiments with microscopy images of palynomorphs are reported in Section 4. Section 5 concludes the study.

2 Segmentation of microscopy images of palynofacies

Microfossil analysis is an essential task in micropaleontology, aiding the interpretation of the history of regional and global climate and of the evolution of the biosphere [10]. Figure 1 shows a typical image of a slide containing microfossils and other organic debris. There are hundreds of types of microfossil and each image may contain many of them. The end task is to locate and subsequently classify the microfossils in the image.

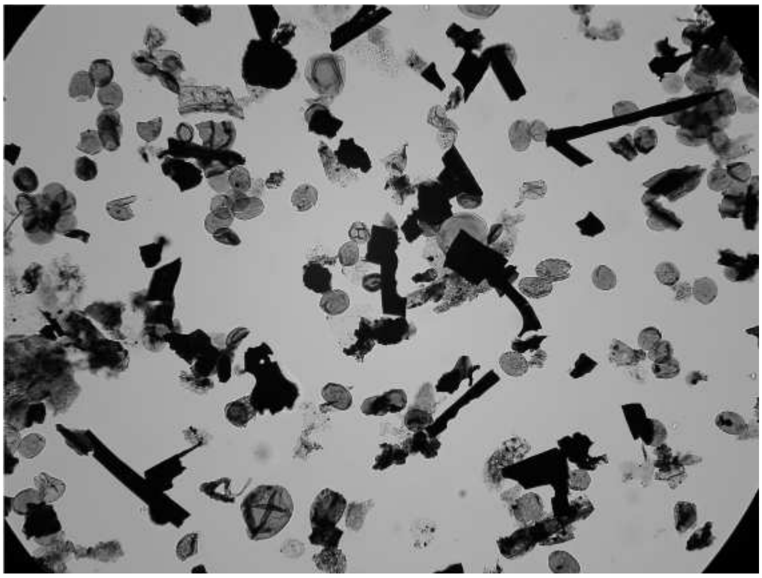


Fig. 1. Thin section of rock containing microfossils

Figure 1 demonstrates the difficulties in locating the objects. First, pieces of inert material (darker objects of irregular shapes and sizes) have to be eliminated so that the image contains only the microfossils. Second, the microfossils appear in different orientation, partly or completely overlapping, clustered together, overshadowed by inert material, distorted by pressure forces in the rock, etc. Third, the colour of some of the microfossils is barely distinguishable from the image background. Fourth, the background is not uniform across the image. Pixel intensities which correspond to background at some places may well be classed as object at a different place in the image.

Although the different types of fossils have different texture and structure, they can be roughly perceived as round or elliptical objects. Hence we are interested in finding "centres" so that we can crop sub-images of microfossils at a later stage applying a higher resolution.

Before object segmentation is attempted, the background must be removed. The difficulty in our case comes from the specific illumination of microscope slides. Usually the centre of the slide is brighter and the intensity fades towards the edges. Also, since the microscopic view is a circle, dark corners might appear. The approach adopted here was to look for and eliminate dark corners, model the background as a function of x and y and remove it from the image. A parabola was fitted to model the intensity of the background for each x and then for each y. The pixels whose true intensity was substantially lower than the scores on both parabolas were marked as non-background (intensity margin of 20 was empirically chosen here).

The measure proposed in the next section is intended for any segmentation method which produces a set of object centres. The standard watershed segmentation was tried as well as a recently developed algorithm called floodfill segmentation [11]. Here we look for a measure to compare object segmentation methods on the basis of their output.

3 Evaluation measure for segmentation methods

Here we construct an empirical discrepancy measure to determine how close two segmented images are.

3.1 Definition of "centre" of an object

Suppose that a distance transform is applied to a black and white image so that each pixel, p, is associated with a function $D(p)$ [12]. The value of $D(p)$ gives the Euclidean distance from p to the nearest white pixel.

Definition 1. A *centre* of an object is the pixel p with the largest distance $D(p)$ within the object. If there is a tie, any of the tied pixels can be chosen to be the centre.

Note that this definition does not imply that the centres will be positioned at the centres of gravity of the objects. Also, one object may have infinite amount of candidate centres with the same highest $D(p)$. This will happen, for example, in a ring-shaped object. The geometric position of points at equal highest $D(p)$ will be a circle with radius equal to the average of the inner and outer radii of the ring. Any point on the circle will be a valid centre according to Definition 1.

The obtained centres are required for extracting the microfossils with a higher resolution for the purposes of subsequent classification. However, more generally, centres of objects may be required for other purposes such as setting the initial position of an active contour [13] or object tracking within moving images [14]. The centres provide a handle for our evaluation method.

3.2 Comparison of sets of centres based on distances

Let $C^* = \{c_1^*, \ldots, c_n^*\}$ be the true centres and $C = \{c_1, \ldots, c_m\}$ be the centres obtained through automatic segmentation. The most intuitive matching measure would be the sum of distances to the nearest centre. Let

$$R(C^* \to C) = \sum_{i=1}^{n} \min_{j=1}^{m}\{\mathrm{dist}(c_i^*, c_j)\} \tag{1}$$

be the representation of C^* by C. In the ideal case where the two sets of centres are identical, $C^* = C$, we have

$$R(C^* \to C) = R(C \to C^*) = 0.$$

If $C \subset C^*$, we have under-segmentation. In this case $R(C^* \to C) > 0$ and $R(C \to C^*) = 0$. If $C^* \subset C$, the image is oversegmented, $R(C^* \to C) = 0$ and $R(C \to C^*) > 0$. To account for both under- and over-segmentation, and also for the discrepancies in the centre location, we can use the following measure

$$M_d(C^*, C) = R(C^* \to C) + R(C \to C^*) \tag{2}$$

$$= \sum_{i=1}^{n} \min_{j=1}^{m}\{\mathrm{dist}(c_i^*, c_j)\} + \sum_{j=1}^{m} \min_{i=1}^{n}\{\mathrm{dist}(c_j, c_i^*)\} \tag{3}$$

Here "dist" can be any distance. In the illustration below we use Euclidean and City-block distances. The smaller the value of M_d, the more similar the two segmentations are. We note that M_d is a metric on the space of sets of centres because $M_d(C^*, C) \geq 0$ with $M_d(C^*, C) = 0$ iff $C^* = C$ (nonnegativity); $M_d(C^*, C) = M_d(C, C^*)$ (symmetry) and it can be proved that $M_d(C_1, C_3) \leq M_d(C_1, C_2) + M_d(C_2, C_3)$ (triangle inequality), where C_1, C_2 and C_3 are sets of centres.

The problem with the distance-based measures is that they do not take into account the specific objectives of segmentation. An example is shown in Figure 2. The true centre of the grey object, according to Definition 1, is situated in the point with the largest $D(p)$ (marked with 'x'). Two guessed centres are displayed in the figure. Clearly Guess 1 is closer to the true centre and any distance measure will favour it over Guess 2. However, Guess 2 sits on the next highest peak of $D(p)$ in the object and is a much better representation of the object than Guess 1. This deficiency of the distance-based measures is addressed by the measure proposed below.

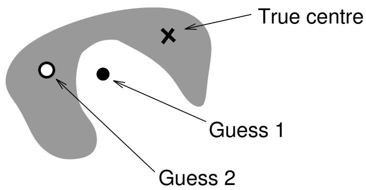


Fig. 2. Examples of a true and two guessed centres.

3.3 Comparison of sets of centres based on segmentation heuristics

The three most desirable properties of a segmented image can be specified as follows

– A perfectly segmented image exhibits no under- or over-segmentation.
– There are no centres of objects which lie outside objects boundaries.
– The centre of each segment should coincide with the relevant object centre.

We shall assume that a reference segmentation is available to represent the ground truth. We also assume that the segmentation process returns a set of m centres, C. We propose to evaluate the quality of segmentation by the following three measures.

Definition 2. Let n_i be the number of centres placed by the automatic segmentation within object i (the object is defined by the ground truth segmentation). The measure of *under- or over-segmentation* of i is

$$r_i = \begin{cases} 1 - 1/n_i, & \text{if } n_i > 0 \\ 1, & \text{if } n_i = 0 \end{cases} \tag{4}$$

If there is no centre in the object or if there are large number of centres there, r_i approaches 1. The most desirable value of r_i is 0 which is achieved if there is only one centre in the object.

Definition 3. The measure of *background segmentation* is

$$v = 1 - \frac{1}{m}\sum_{i=1}^{n} n_i. \tag{5}$$

Note that v is the proportion of automatic centres that are not contained within the boundaries of any objects. Thus $v = 0$ corresponds to the ideal situation where the background is free of centres placed by mistake by the segmentation algorithm.

The values r and v completely represent the under and over-segmentation of the image regardless of the location of the centres within the objects. Hence the third measure evaluates how close the approximations are to the ideal centres within the objects.

Definition 4. Let c_i^* be the true centre of object i and let $c' \in C$ be the nearest centre from the automatic segmentation which lies within object i, i.e.,

$$c' = \arg \min_{c \,\in\, \text{object } i} \{\text{dist}(c_i^*, c)\} \tag{6}$$

The *centre discrepancy* is defined as

$$q_i = 1 - D(c')/D(c_i^*), \tag{7}$$

where $D(c)$ is the distance transform value for point c. By Definition 1, $D(c_i^*)$ is the maximum distance within object i therefore $D(c') \leq D(c_i^*)$, and $q_i \in [0, 1]$. Approximated centres near the boundaries of objects should be assigned high discrepancy value, q_i, while those in the middle of objects should be assigned a low discrepancy value.

To illustrate the rationale for introducing the centre discrepancy, q_i, consider an elongated object as shown in Figure 3. There are infinitely many possible centres, according to Definition 1, situated along the ridge of the distance function for this object. Thus any centre on the ridge should have a lower error value q_i than centres on the edge of the object. Figure 3 shows that Euclidean distance will be misleading in this case as a centre at the periphery will be preferred to one of the true centres with the largest $D(p)$.

(a) Small q_i, large Euclidean distance (preferred)

(b) Large q_i, small Euclidean distance

Fig. 3. Illustration of the advantage of the centre discrepancy measure q_i over Euclidean distance in evaluating an approximated centre (circle) with respect to the true centre (cross).

The three measures r_i, v and q_i can be combined so that the quality of the segmentation is measured by a single value.

Definition 5. The measure of *quality of segmentation* represented by the set of centres C with respect to a ground truth segmentation with a set of centres C^* is

$$S(C, C^*) = \frac{1}{3}\left(v + \frac{1}{n}\sum_{i=1}^{n}(r_i + q_i)\right), \tag{8}$$

where r_i, v and q_i are calculated as in Definitions 2-4, respectively.

4 Experimental results

4.1 A single-object illustration

We start with a simple example showing why distance-based measures $M_d(C, C^*)$ are insufficient for the purposes of segmentation. Figure 4 displays seven copies of an image containing a single elliptical object with different nonempty sets of centres C. In all 7 cases the set C^* consists of one element which is the geometrical centre of the ellipse.

Three measures of quality of the segmentation represented by the centres C (dots) are shown in the table:- $M_d(C, C^*)$ for Euclidean and City-block distance and the quality of segmentation $S(C, C^*)$ as in Definition 5. The rows in the table are sorted with respect to $S(C, C^*)$ starting with the best case (minimum, $S = 0$) and ending with the worst case (maximum, $S = 1$). There are discrepancies in

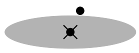
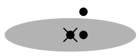
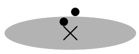
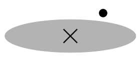
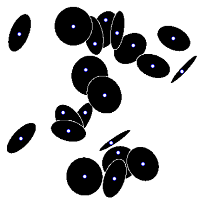
No	Image (one object)	M_d Euclidean	M_d City-block	r	v	q	S
1		0.00	0.00	0.00	0.00	0.00	0.00
2		1.43	1.90	0.00	0.50	0.00	0.17
3		3.70	3.70	0.50	0.00	0.00	0.17
4		7.40	7.40	0.00	0.00	0.70	0.23
5		2.41	3.00	0.50	0.33	0.00	0.28
6		2.95	3.80	0.00	0.50	0.80	0.47
7		4.88	6.80	1.00	1.00	1.00	1.00

Fig. 4. Seven examples of segmentation of one object using centres and the values of the measures of quality of the segmentation. The true centre is marked with '×' and the guessed centres are marked with '•'

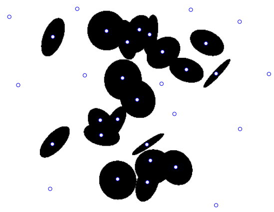
the ranking of the 7 cases according to the three measures. While the trivial case where the single guessed centre coincides with the true centre is the most preferred case for all the measures, $S(C, C^*)$ disagrees with both distance-based measures M_d about the least preferred case. Based on distances, case 4 is the worst because the guessed centre is far from the true centre. However, it is important for our segmentation purposes that the centres lie within the objects. Thus case 7 should be the least preferable one because the object has not been found as the single centre lies in the background. Also, the oversegmented case 5 will be preferred to case 3 by both distance measures. In both cases there is a perfect centre within the set C. Note that there is an extra centre in the background in case 5, while there is no such centre in case 3. The disagreement between the two distance-based measures is minimal. They rank differently only cases 3 and 6. The major flaw of the distance measures is that they do not take into account any object boundaries. Hence the advantage of M_d over S would be speed of calculations. However, the better match with the desiderata for segmentation quality leads us to choose S.

4.2 A multiple-object illustration

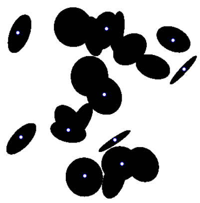
An example involving multiple objects is considered next. Figure 5 (a) presents the objects and the ideal segmentation. Three special cases are shown in plots (b), (c) and (d). The centres are placed manually in all three images.



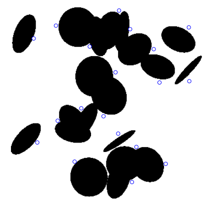
(a) Segmented original (C^*)



(b) Over-segmented background (C)



(c) Under-segmented objects (C)



(d) Completely misplaced centres (C)

Fig. 5. The original segmentation of a multiple object image and three special cases of segmentation (centres have been placed manually in (b), (c) and (d))

For the task of classifying fossils cropped around the centres, missing an object should be penalised stronger than placing an extra centre in the background. In the former case, an important piece of information may be overlooked. In the latter case the analysis time will increase due to the extra centres of nonexistent objects but all the microfossils will be detected in the image. For this reason, the undersegmented image (c) should be ranked worse than (b) and (d) should be ranked worse than (c). Table 1 displays the three measures M_d-Euclidean, M_d-City block and S for the images in plots (b), (c), and (d).

Table 1. Measures of segmentation quality for the images in Figure 5 (b), (c) and (d) with respect to the ground truth (a).

Subplot	M_d Euclidean	M_d City-block	r	v	q	S
(b) Over-segmented background	1180	1513	0.00	0.18	0.38	0.18
(c) Under-segmented objects	585	735	0.50	0.61	0.00	0.37
(d) Completely misplaced centres	1068	1327	1.00	1.00	1.00	1.00

The table shows that the distance-based measures fail to produce the required ranking while S clearly distinguishes between the three cases.

4.3 Results on microfossil images

The results from applying the watershed [15] and floodfill [11] segmentation methods on the image shown in Figure 1 are displayed in Figures 6 and 7, respectively. Table 2 compares the two segmentation approaches. The distance-based measures M_d-Euclidean and M_d-City block agree with S that the floodfill method performs better than the watershed algorithm. There is a considerable difference in the score of the distance-based measures between the two types of segmentations indicating that these two approaches are very different, whereas S shows that floodfill method only slightly improves over the watershed method. This corresponds precisely to how we would interpret the data by visual inspection. The watershed algorithm (Figure 6) has failed to capture 3% of the objects, with most of the captured objects being oversegmented. The floodfill algorithm (Figure 7) has failed to capture 13% of the objects but the detection was completed with nearly no oversegmentation. The watershed method makes up for its oversegmentation by achieving a high capture rate. The small difference between the two values of measure S in Table 2 account for the fact that both methods have assets and flaws and a choice between the two cannot be made with high certainty.

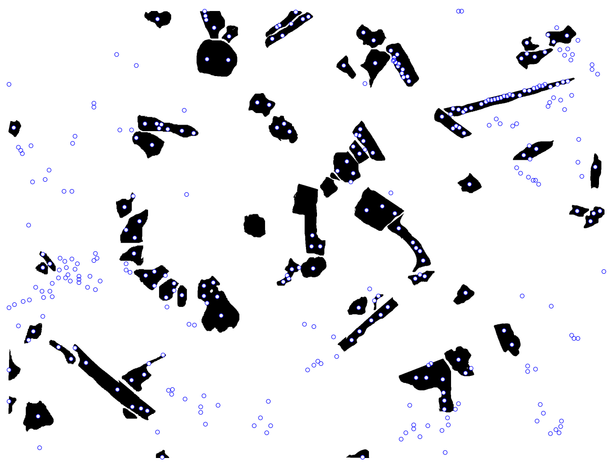


Fig. 6. Manually segmented inert material from Figure 1 overlaid with centres found through the *watershed* segmentation method.

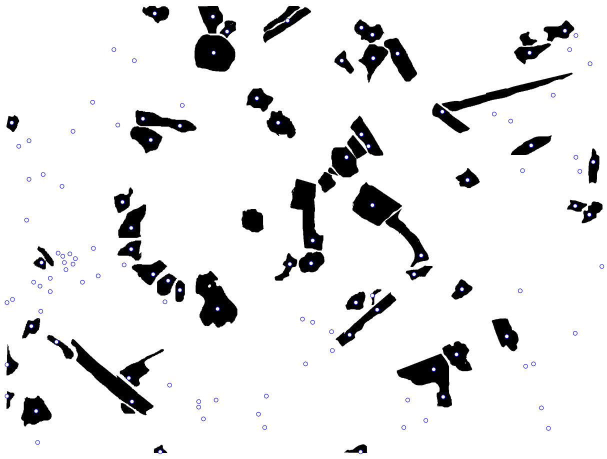


Fig. 7. Manually segmented inert material from Figure 1 overlaid with centres found through the *floodfill* segmentation method.

Table 2. Segmentation quality of watershed and floodfill segmentation of Figure 1.

Method	M_d Euclidean	M_d City-block	r	v	q	S
Watershed	35407	44784	0.56	0.45	0.45	0.49
Floodfill	13254	16783	0.28	0.53	0.39	0.40

5 Conclusions

This paper proposes a new measure of the quality of segmentation of images containing objects. The measure operates on a set of ideal centres of the objects, C^*, assumed to be the ground truth and a set of guessed centres, C, obtained through a segmentation algorithm. Thus the proposed measure falls into the *discrepancy* category of evaluation methods, as detailed by Zhang [1].

Based on a list of desired properties and examples with generated and real images, we argue that the proposed measure of quality, $S(C, C^*)$ is better than measures based on the distances (M_d) between the centres in sets C and C^*. The three components of the proposed measure comply with the intuition for evaluating segmentation results represented by centres. Another advantage of S over M_d is that S has practically useful lower and upper limits while the distance-based measures have only a lower limit. A disadvantage of S is that it is slower to calculate than M_d because it requires knowledge of the objects in the image. We are more concerned with locating the inert material through segmentation rather than extracting the material, so even though two different segmentations can result in the same set of centres the value S will still provide an accurate comparison between segmentations.

The intended application of S is for choosing a segmentation method and tuning it for the specific practical application. In our case, a small number of images will be labelled manually by an expert palynologist, and used as the ground truth. After tuning, the segmentation method will be applied as a standard routine to find the microfossils in other images coming from the same domain.

Acknowledgements

The EPSRC CASE grant Number CASE/CNA/05/18 is acknowledged with gratitude.

References

1. Zhang, Y.J.: A survey on evaluation methods for image segmentation. Pattern Recognition **29**(8) (1996) 1335–1346
2. Beauchemin, M., Thomson, K.P.B.: The evaluation of segmentation results and the overlapping area matrix. Remote Sensing **18**(18) (1997) 3895–3899
3. Cardoso, J.S., Corte-Real, L.: Toward a generic evaluation of image segmentation. IEEE Transactions on Image Processing **14**(10) (2005) 1773–1782
4. Hoover, A., Jean-Baptiste, G., Jiang, X., Flynn, P., Bunke, H., Goldgof, D., Bower, K., Eggert, D., Filtzbiggbon, A., Fisher, R.: An eperimental comparison of range image segmentation algorithms. IEEE Transactions on Pattern Analysis and Machine Intelligence **18** (1996) 673–689
5. Borsotti, M., Campadelli, P., Schettini, R.: Quantitative evaluation of color image segmentation results. Pattern Recognition Letters **19** (1998) 741–747
6. Liu, J., Yang, Y.H.: Multiresolution color image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence **16**(7) (1994) 689–700
7. Rand, W.: Objective criteria for the evaluation of clustering methods. Journal of the American Statistical Association **66** (1971) 846–850
8. Elisseeff, A.B.H.A., Guyon, I.: A stability based method for discovering structure in clustered data. In: Proc. Pacific Symposium on Biocomputing. (2002) 6–17
9. Hubert, L., Arabie, P.: Comparing partitions. Journal of Classification **2** (1985) 193–218
10. Bollmann, J., Quinn, P.S., Vela, M., Brabec, B., Brechner, S., Cortes, M.Y., Hilbrecht, H., Schmidt, D.N., Schiebel, R., Thierstein, H.R.: Automated particle analysis: Calcareous microfossils. Image Analysis, Sediments and Paleoenvironments **7** (2004) 229–252
11. Charles, J.J., Kuncheva, L.I., Wells, B., Lim, I.S.: Object location within microscopic images of palynofacies. (2006) (submitted).
12. Borgefors, G.: Distance transformations in digital images. Computer Vision Graphics and Image Processing **34**(3) (1986) 344–371
13. Wang, Y., Chou, J.: Automatic segmentation of touching rice kernels with an active contour model. Transactions of the ASAE **47**(5) (2004) 1803–1811
14. Paul, G.V., Beach, G.J., Cohen, C.J.: A realtime object tracking system using a color camera. Applied Imagery Pattern Recognition Workshop (2001) 137–142
15. Vincent, L., Soille, P.: Watersheds in digital spaces: an efficient algorithm based on immersion simulations. IEEE Transactions on Pattern Analysis and Machine Intelligence **13**(6) (1991) 583–598