Reprinted from

SIGNAL
PROCESSING

Signal Processing 39 (1994) 43–54

Objective and quantitative segmentation evaluation and comparison[†]

Y.J. Zhang[a,*], J.J. Gerbrands[b]

[a]Image Processing Section, Department of Electronic Engineering, Tsinghua University, Beijing, 100084, People's Republic of China
[b]Information Theory Group, Department of Electrical Engineering, Delft University of Technology, 2600 GA Delft, The Netherlands

Received 6 August 1992; revised 3 December 1992, 3 May 1993 and 13 January 1994

ELSEVIER

ELSEVIER

Signal Processing 39 (1994) 43–54

SIGNAL PROCESSING

Objective and quantitative segmentation evaluation and comparison[†]

Y.J. Zhang[a,]*, J.J. Gerbrands[b]

[a]*Image Processing Section, Department of Electronic Engineering, Tsinghua University, Beijing, 100084, People's Republic of China*
[b]*Information Theory Group, Department of Electrical Engineering, Delft University of Technology, 2600 GA Delft, The Netherlands*

Received 6 August 1992; revised 3 December 1992, 3 May 1993 and 13 January 1994

Abstract

A general framework for segmentation evaluation is introduced after a brief review of previous work. The accuracy of object feature measurement is proposed as a criterion for judging the quality of segmentation results and assessing the performance of applied algorithms. This goal-oriented approach has been shown useful for an objective and quantitative study of segmentation techniques.

Zusammenfassung

Nach einer kurzen Übersicht über bekannte Arbeiten wird die Bewertung von Segmentierungen in einen allgemeinen Rahmen gestellt. Als Kriterium zur Qualitätsbeurteilungen von Segmentierungsergebnissen und zur Bewertung der Leistungsfähigkeit angewandter Algorithmen wird die Genauigkeit der Bestimmung von Objektkenngrößen vorgeschlagen. Dieser zielgerichtete Ansatz hat sich als nützlich erwiesen für die objektive und quantitative Untersuchung von Segmentierungsresultaten.

Résumé

Un cadre général pour l'évaluation de la segmentation est introduit après un bref passage en revue des travaux existants. La précision de la mesure de l'attribut d'objet est proposée comme critère de jugement de la qualité des résultats de segmentation et d'évaluation des performances des algorithmes appliqués. Cette approche orientée objectif est montrée utile pour une étude quantitative objective des techniques de segmentation.

Key words: Segmentation evaluation; Quantitative assessment; Performance comparison; Image synthesis; Quality criteria; Object feature measurement

*Corresponding author.

[†]This work was partially supported by The Netherlands' Project Team for Computer Science Research (SPIN), Three-dimensional Image Analysis Project.

SSDI 0165-1684(94)00037-Z

1. Introduction

Image segmentation is the process that subdivides an image into its constituent parts. It is one of the most critical tasks in automatic image analysis because the segmentation results will affect all the following tasks, such as feature extraction and object classification. Due to its importance, much effort has been devoted to the segmentation process and technique development in the last decades. This has already resulted in quite many (more than thousands) different algorithms, and the number is still increasing. Several survey papers have been published (e.g., [3, 4, 12, 14]), but they only partially cover the large number of techniques developed.

One important fact in the development of segmentation techniques is that no general theory exists [4], so this has traditionally been an ad hoc and problem-oriented process [16]. Although some initial attempts in the direction of a unified theory were reported (e.g., [13]), this problem is far from being solved. Note that the models discussed in [13] for certain classes of images are rather informal (as indicated by the authors) and relatively limited. As a result, none of the developed techniques is generally applicable. Given a particular application, finding the appropriate segmentation algorithm is still quite a problem [12]. The performance evaluation and comparison of competing techniques become indispensable in this context. We will present a general framework for segmentation evaluation and comparison in this paper. It is also useful for providing guidelines in refining existing techniques.

This paper is organised as follows. In the next section a brief review of previous work for segmentation evaluation is given. The limitations of these proposed methods and some open problems are pointed out. Our new approach for an objective and quantitative study will be presented in Section 3, where a general evaluation framework and new criteria for segmentation performance assessment are described. In Section 4 several commonly used object features are studied by comparing their behaviours in different evaluation situations. The results show that the measurement accuracy of object features is related to the quality of segmented images and different features can have various technique evaluation abilities. Finally, in Section 5 the advantages of our approach are summarised.

2. Previous work and open questions

2.1. Early proposed methods

In contrast to the abundant number of existing segmentation techniques, only few methods for their evaluation and comparison have been proposed. One important reason for this is the lack of appropriate measures for judging the quality of segmentation results and then the performance of applied algorithms. Below, these methods are briefly reviewed in the sequence of their appearance, with a special attention to the performance criteria used by these methods.

Yasnoff et al. [16] proposed to take the number of mis-classified pixels and their positions into account for computing two measures: the percentage of area mis-classified and the pixel distance error. The first measure is self-explaining. The sum of the distances between pixels that have been assigned to a wrong class and the nearest pixels that actually belong to the correct class is taken as the second criterion. Similar ideas appear in Abdou et al.'s Figure Of Merit (FOM) for edge detection evaluation [1], so the second criterion suffers from the same problem as the FOM, namely that various configurations can be found for which the same pixel distance error is obtained. Some examples are depicted in Fig. 1, where the pixel distance errors for the situations in (A), (B) and (C) are equal (the numbers of mis-classified pixels are also equal). Without further processing the three mis-classified pixels in Fig. 1(A) could be counted as object pixels, whereas the three mis-classified pixels in Fig. 1(B) would be counted as background pixels. The consequences of these two cases are different, for example, both for the influence on the size and the shape of objects. Moreover, the pixel distance error cannot distinguish several isolated mis-classified pixels (Fig. 1(C)) from a cluster of mis-classified pixels (Figs. 1(A) and (B)), although the error produced by the latter is more serious than that by the former.

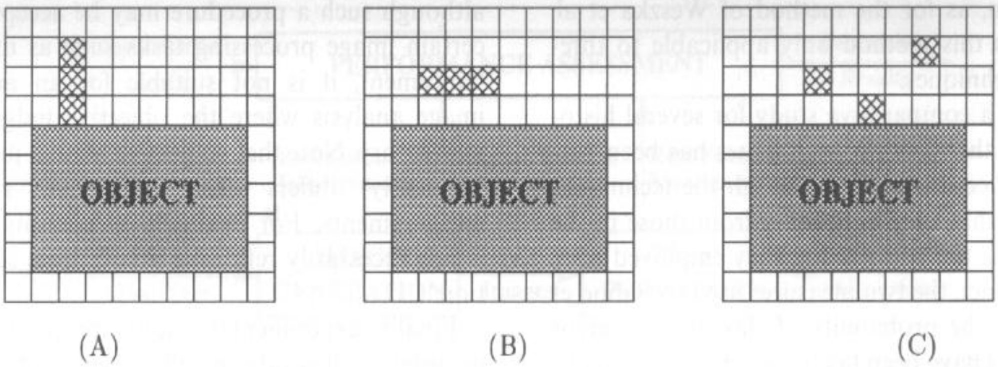


Fig. 1. Various configurations that have equal pixel distance measures and also the same number of misclassified pixels.

Weszka et al. proposed two perceptive 'goodness' measures in their method for selecting and evaluating the threshold [15]. One measure is called the discrepancy between the original image and the segmented image, which is actually measured by the classification error and therefore equal to Yasnoff et al.'s method [16]. The other one is the so-called busyness of the segmented image. Its calculation is based on the gray level co-occurrence matrix whose entries are estimates of the joint probabilities of gray levels for adjacent pixels. A common point is that no shape property of objects is considered in both of them, so that various segmentation results can yield the same 'goodness' values. Besides, other techniques than thresholding ones cannot be evaluated by using this method as the threshold values are indispensable for computing these measures.

In the context of developing a rule-based system, Levine et al. [8] proposed to take the intra-region uniformity (which is similar to Weszka et al.'s busyness measure [15]), and the inter-region contrast into account. These criteria are also based on human intuition of 'goodness' so that they are again the subjective criteria in principle. In addition, these measures consider neither the shape nor the size of segmented regions in the evaluation, so that they can be used only as some complementary measures for the quality of segmentation.

Instead of examining experimentally segmented images, Liedtke et al. [9] proposed to study the segmentation algorithms themselves. By using the amount of relative 'a priori knowledge' incorporated they qualitatively compared six different algorithms. They found that this amount is decisive for the reliability of an algorithm. However, the amount of 'a priori knowledge' has not been precisely defined and cannot be measured quantitatively. Moreover, the way in which how such knowledge can be incorporated into an algorithm can also be decisive [19], which they have not considered.

MacAulay et al. [10] tested four simple thresholding algorithms by visually comparing segmented images with the original images. A segmentation is accepted when the area of segmented objects matches within a margin of 5% to the area of visually detected objects. In contrast to the methods mentioned above, no precise objective measure is calculated and only a subjective estimate of the pixel classification error is applied (see also [16]).

Sahoo et al. [14] also employed two quality measures for the evaluation of thresholding techniques. The one is the uniformity criterion adapted from Levine et al. [8], the other is what they called a shape measure. They calculated this measure by the summation of the gradient values over pixels that have a gray level higher/lower than the average gray level of their neighbours and also higher/lower than a selected threshold value. From this summation, the gradient values over pixels that have a gray level higher/lower than the average gray level of their neighbours but lower/higher than the selected threshold value are subtracted. Apart from the vague relation of this measure to the form of objects (for instance, such a measure can hardly distinguish a square from a circle), this measure is also very sensitive to isolated noise pixels as these will be heavily weighted in the computation. In addition, the role of the threshold value in the computation, as for the method of Weszka et al. [15], makes this method only applicable to thresholding techniques.

Recently, a comparative study for several histogram-based thresholding techniques has been performed by Lee et al. [7]. Although the techniques and images that they used differ from those in the other studies, the criteria that they employed were not new. In fact, the two measures used by Sahoo et al. [14] and the probability of classification error (see [15, 16]) have been taken. So, this study shares the same problems as the others.

2.2. *Some general problems*

The above-mentioned methods have also some common limitations. First, we should stress that there is no unique or standard judgment of segmentation results due to the lack of a general segmentation theory. The various 'goodness' criteria proposed can, at best, only provide a rough condition for a subjectively 'nice' segmentation. Moreover, different criteria may not be compatible. The classification based criteria just exploit pixel properties. When region information is to be taken into account, they may become less important. In general, they evaluate algorithms regardless of the segmentation goal: the results do not necessarily correspond to what one expects from the analysis.

Second, certain evaluation criteria are also used inside some particular segmentation algorithms. That is, the criteria for obtaining a segmentation and for its evaluation are identical. This will introduce some bias when comparing algorithms of different criteria. For example, some algorithms are uniformity oriented in nature as indicated by Sahoo et al. [14]. If the region uniformity criterion is used in evaluation and comparison, these algorithms would have more chances of showing a 'better' performance than others.

The third limitation is related to the involvement of visual perception, because many methods need human inspection to provide a reference. In fact, the quality measurements are defined in terms of their correlation with human performance. Two problems arise here. First, the results heavily rely on the judgement of a particular observer. Second, although such a procedure may be acceptable for certain image processing tasks such as image enhancement, it is not suitable for an automatic image analysis where the objective judgement is mandatory. Note that subjective human judgement frequently differs from objective computer measurements. For example, a 'pleasing' picture is not necessarily reflecting an accurate segmentation.

Finally, experimental results are often supplied in order to illustrate the effectiveness of proposed evaluation methods. In all studies cited above only real images acquired from particular domains are used. The 'random' nature of real images makes the evaluation results not appropriate for different applications. Moreover, many indeterminate characteristics of real images make an analytic evaluation and an accurate comparison complicated because various properties are mixed, and it is difficult to study the influence of these individually.

3. New approach

We now present our approach for the objective and quantitative evaluation and comparison of segmentation algorithms. First we indicate some primary conditions to be satisfied and introduce a general framework. Then we describe new criteria for performance judgement and discuss their advantages with respect to existing criteria. Lastly, we explain how synthetic test images can be generated, how segmentation algorithms should be tested, and how their performance will be assessed.

3.1. *Primary conditions and general framework*

As is made evident from the discussions presented above, the following conditions should be satisfied for an evaluation and comparison procedure:

(1) It should be general, i.e., be suitable for all segmentation techniques and various applications. This implies that no parameters nor properties of particular segmentation algorithms must be involved so that no bias with respect to special techniques will be introduced.

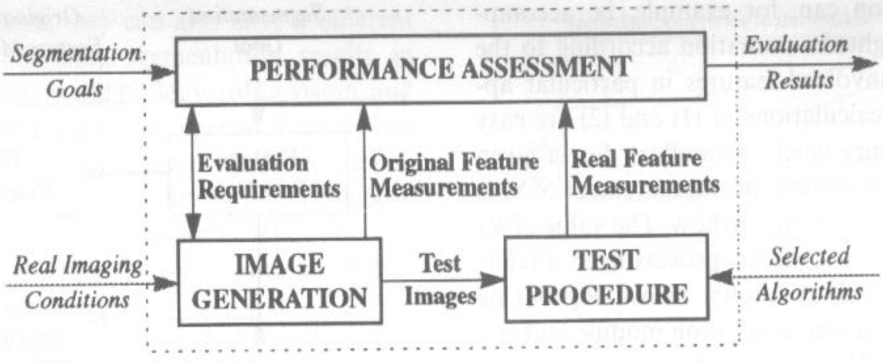


Fig. 2. General framework for segmentation evaluation.

(2) It should use quantitative and objective criteria for performance assessment. Quantitative means here exactness, objective refers to the segmentation goal and reality.

(3) It should use images that can be reproduced by all users for the purpose of algorithm testing. Moreover, these images should reflect common characteristics of real applications.

Taking these conditions into consideration, we have designed a general evaluation framework. It consists of three related modules as shown in Fig. 2. The information concerning segmentation goals, image contents and imaging conditions can be selectively incorporated into this framework to generate test images, which makes it suitable for various applications. This framework is also general in the sense that no internal characteristics of particular segmentation techniques are involved as only segmentation results are taken into account in the test. Other particularities will be discussed in the following sections.

3.2. Criteria for performance assessment

In image analysis, the ultimate goal of image segmentation and other processes is to obtain measurements of object features. These measures are heavily dependent on the results of these earlier processes. It is therefore obvious that the accuracy of these measurements obtained from segmented images would be an efficient index revealing the performance of applied segmentation algorithms. Taking into account this ultimate goal, we propose to use the accuracy of these measurements as the criteria to judge and rank the performance of segmentation techniques. We call them ultimate measurement accuracy (UMA). Referring to the second condition in Section 3.1, such criteria are certainly objective and quantitative.

In different analysis applications, different object features can be important. So, the UMA can represent a group of feature-dependent and so goal-oriented criteria. It can be denoted as UMA_f, where f corresponds to the considered feature. The comparative nature of UMA implies that some references should be available. Depending on how the comparison is made two UMA types can be applied, i.e., absolute UMA (AUMA_f) and relative UMA (RUMA_f) with the following definitions:

$$\mathrm{AUMA}_f = |R_f - S_f| \qquad (1)$$

$$\mathrm{RUMA}_f = (|R_f - S_f|/R_f) \times 100\% \qquad (2)$$

where R_f denotes the feature value obtained from a reference image and S_f denotes the feature value measured from a segmented image. The values of AUMA_f and RUMA_f are inversely proportional to the segmentation quality: the smaller the values, the better the results. It seems from Eq. (2) that the RUMA_f may be greater than 100% in certain cases. In practice, we can limit it to 100% for it does not make sense to distinguish between very inaccurate results.

Features used in (1) and (2) can be selected according to the segmentation goal. Various object features can be employed (some examples are shown in the next section) so that different situations can be covered. In addition, the combination of UMA of different object features is also possible.

This combination can, for example, be accomplished by a weighted summation according to the importance of involved features in particular applications. The calculations of (1) and (2) are easy and do not require much extra effort. For a given image analysis procedure, the measurement of S_f is a task we need to perform anyhow. The value of R_f can be obtained by a similar process from a reference image. As will be shown below, R_f will be produced by the image generation module and can then be used in all evaluation processes.

3.3. Detailed framework

Image generation

In order to satisfy the third condition in Section 3.1, synthetic test images are used. They can be easily manipulated and reproduced. Besides, image synthesis can easily provide reference feature values as being needed in (1) and (2). The image generation module consists of four related components [20]. The first one is to compose the basic image, which can be done on the basis of a simple model of an application. The basic image is the starting point for generating test images. In 'object variation', the object in the basic image is modified to provide images with a variety of object numbers, sizes, shapes, etc., in order to be able to approximate the contents of real images. Different corruption and/or distortion factors, such as various noise and blurring, can be simulated. They will then be combined with a variety of objects to form sets of test image. Before this combination, some features are computed from corruption-free images; they will serve as a reference later on. Such a procedure permits us to incorporate the application domain knowledge, the image content information and the image forming conditions into synthetic images, so these images can represent the reality quite well.

Test procedure

This is a typical image analysis procedure. It consists of two consecutive steps: segmentation and measurement. In segmentation, the algorithm is treated as a 'black box'. We give it a test image as input, and we take the segmented image as output. From the segmented image the real feature values can be obtained from the extracted objects, and these values will be used in the module 'performance assessment'. It is easy to be seen that the first condition in Section 3.1 is satisfied by using this procedure.

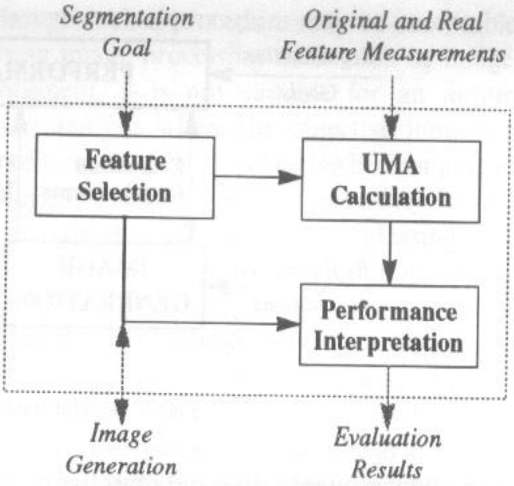


Fig. 3. Detailed scheme of performance assessment module.

Performance assessment

The performance assessment module can be divided into three components as shown in Fig. 3. Note the inter-relation among these components. According to the segmentation goal, different object features should be selected for a particular evaluation. Appropriate synthetic images are then generated in accordance with the selected object features. In other words, various images can be designed and generated to meet the evaluation conditions. With the feature values obtained both from reference images and segmented images, the UMA calculation can be carried out. Combining the results and the image generation conditions, the algorithm's performance can be quantitatively assessed as shown above.

4. Object feature study

4.1. Features selected

Different object features may be important in different applications to describe object properties.

On the other hand, they can also have a different effectiveness in judging segmentation results in various situations. In cell biology, robot vision, and many other applications, geometric features are commonly employed (see, e.g., [2], [11]). In the following, the evaluation ability of four geometric features will be studied as examples. They are area (A), perimeter (P), form factor (F) and circularity (C) of objects.

F is derived from the A and P of objects, which describe the form of objects:

$$F = (P^2)/(4\pi A) \tag{3}$$

C is measured for two circles, whose centers are defined to be at the objects center-of-mass:

$$C = (\text{radius of inscribed circle})/(\text{radius of circumscribed circle}) \tag{4}$$

These are only four of many possible features that can be used. All of them can be directly measured from the extracted objects in the segmented images. Other features can be taken instead.

4.2. Methods

For this study, one set of synthetic test images (16 images) is generated. We use images of size 256 × 256, with 256 gray levels. The basic image is composed of a circular disc (diameter 128) with gray level 144 on the middle of a homogeneous background of gray level 112. This image can be seen, for instance, as a simple model of cytologic image [5]. To simulate objects of different shape, a set of elliptical objects are produced in 'object variation' by simultaneously adjusting the length of the long and short axes of these objects. These objects are numbered from # 1 to # 4 whose eccentricities are 1.0, 1.5, 2.0 and 2.5, respectively. Here we keep the areas of the objects constant so as to avoid the influence of object size, because segmentation techniques generally perform better for images containing relatively bigger objects [21].

The reference values of the four features with these four objects are different. To see more clearly their relative magnitude, which are more important than the absolute magnitudes, these values are normalised to the range [0,1] and shown in Fig. 4.

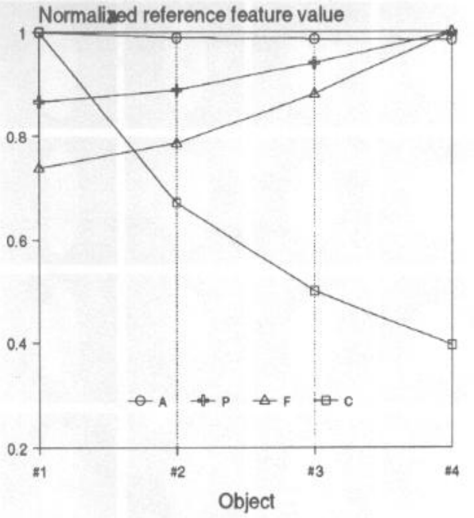


Fig. 4. Normalised reference feature values of different objects in generated images.

Note that except for area, all other features have a monotonic and significant change along the object axis.

In image formation, noise is one of the most important and common corruption factors. The noisy effect has been simulated by adding uncorrelated zero-mean Gaussian noise as examples. The standard deviations of the Gaussian noise used were 4, 8, 16, 32, respectively. The Signal-to-Noise Ratio (SNR) of images is more meaningful than the absolute noise level in images, and is defined as (see [1,6]):

$$\text{SNR} = (h/\sigma)^2 \tag{5}$$

where h is the grey level difference between the object and background regions, and σ is the standard deviation of the noise. According to (5), the SNR of the generated images are 64, 16, 4, and 1, respectively. These values cover the range of most applications (see, e.g., [6]). Another common distortion in image formation is the production of a transition region [18]. In this study the transition region is simulated by filtering the noise-free images with a 3 × 3 uniform filter. The images thus generated are shown in Fig. 5. In Fig. 5, the eccentricity of objects increases along the horizontal axis and the SNR of images increases along the vertical axis. Since different object variation and image

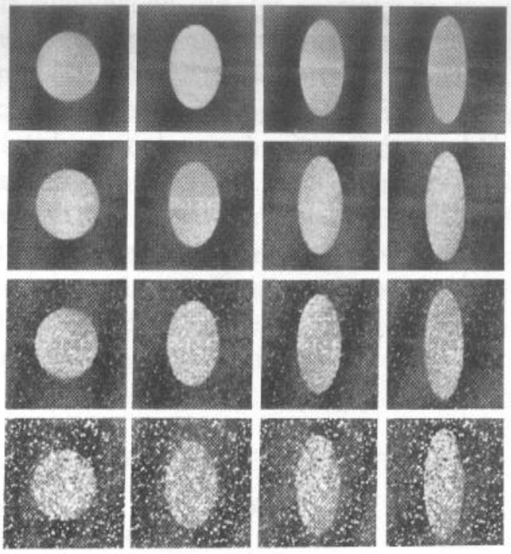


Fig. 5. Examples of synthetically generated images.

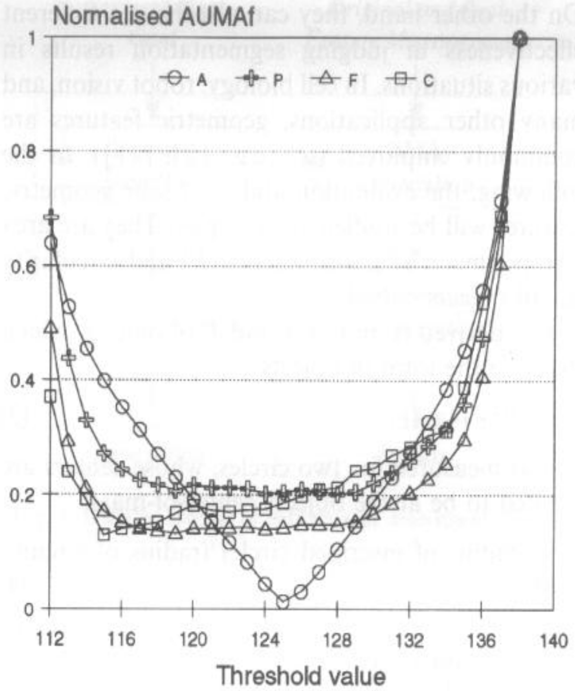


Fig. 6. Results for four object features for the image containing object #3 with SNR = 16.

corruption information are independently incorporated into this set of images, several experiments can be performed to individually study them.

The generated images have been segmented separately by thresholding them with all possible threshold values between the grey levels of object and background. After each thresholding, one opening operation is applied. This process consists of a binary erosion followed by a binary dilation, both with 8-connectivity. The morphological operations help to eliminate incidentally touching noise pixels and make the feature measurement more reasonable. Then the largest object is selected; the inside holes are filled. To cope with the random nature of noise, ten noisy images are generated for each SNR level. The feature measurement is performed after segmenting each of these ten images, and the average of the ten values is taken as the final feature value. Note that this simple procedure based on the thresholding is quite representative. In fact, all segmentation techniques based on global thresholding can be covered as they are mainly different in the way to select a threshold value. Two factors are taken into account for choosing this procedure. One is its simplicity, so that the description and presentation of this study will be clear and so that the emphasis will be focused on evaluation. Another is that this procedure does provide a set of gradually segmented images. As our objective here is to study the behaviour of different features in UMA calculations, this procedure is quite effective.

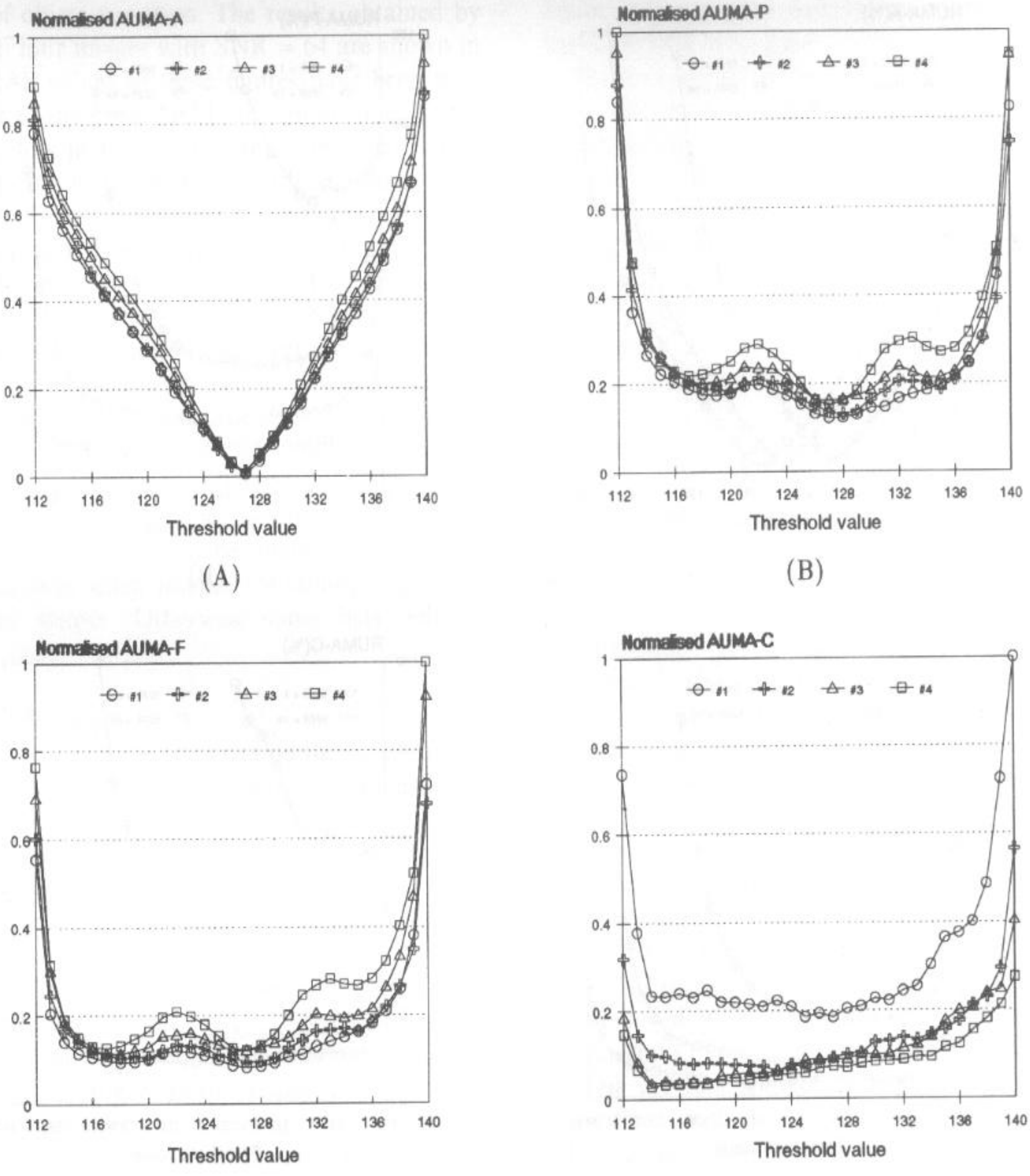
4.3. *Experimental results*

The experimental results are shown by plots of different UMA values (obtained from (1) or (2)) against the threshold values which can be considered as the control parameters of the above segmented images.

First experiment

In this experiment we study the behaviour of these four features as a function of the threshold value. As an example, we show the results obtained by (1) for the image containing object #3 with SNR = 16 in Fig. 6. In order to easily compare the magnitudes of four features, these values have been normalised to the range [0,1].

From Fig. 6 we see that all feature curves have a local minimum located at some threshold value. Along with the change of threshold values, the



(A)

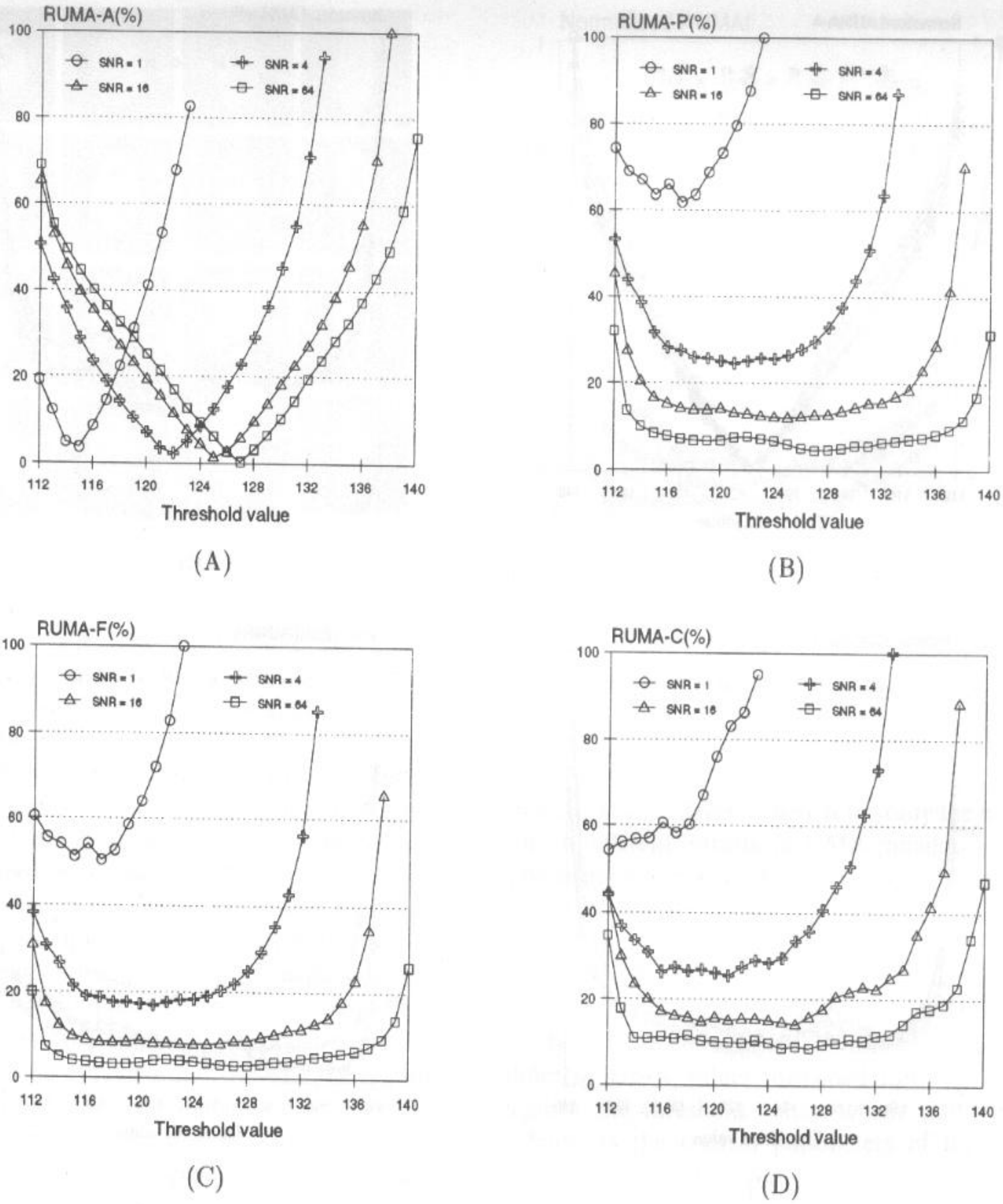
(B)

(C)

(D)

Fig. 7. Results for four object features for four SNR = 64 images.

AUMA$_f$ values are also varied. The AUMA$_f$ value is seen as an index of the quality of segmented images. In other words, all four object features can be used in the AUMA$_f$ calculation for the purpose of judging segmented results. Two important aspects can also be derived from Fig. 6. The first is that the local minima are not located at the same place. This implies that when different features were used different optimal threshold values are expected. The second is that these curves have different forms. The difference among these curves shows various abilities of these features to test a small change of the parameter values (the threshold values) in segmentation techniques (the thresholding techniques). The *A* curve has a pronounced local minimum and this curve steadily follows the change of the threshold value. The other features are less sensitive to this change, which is



(A)

(B)

(C)

(D)

Fig. 8. Results for four object features for four images containing object # 1.

shown by the relative flatness of their curves, especially in the middle range of the threshold values. From one side, it shows that the accurate measurement of object area is highly related to the performance of segmentation algorithms. In other words, a smaller variation of techniques can cause a larger difference of the measurement results. From the other side, it also means that the *A* feature would be more suitable than the other ones to appraise the quality of differently segmented images because even a smaller alteration of the segmented images can be detected.

Second experiment

In this experiment we take a look at the behaviour of these four features and the relation with the object shape. The comparison is made with images having the same SNR level in order to isolate the effect of object variation. The results obtained by (1) from four images with SNR = 64 are shown in Fig. 7. All values in these figures have been normalised to the range [0,1] for comparison.

The plots in Fig. 7 show that the sensitivity of these feature values with respect to the shape variation of objects is different. As an example, we can compare the F and C curves shown in Fig. 7(C) and (D). For more elongated elliptical objects, the accuracy for measuring circularity becomes better whereas that for measuring form factor becomes worse. On the other side, the P curves and C curves shown in Fig. 7(A) and (B) are also not alike. Note that all these results are obtained from the same segmented images. These differences show that the dependence of feature measurements on the object shape must be taken into account in evaluation when images containing objects of different shapes. Otherwise some bias will be introduced.

Third experiment

In this experiment, we focus on the behaviour of the UMA values of these four features and the relation with different SNR levels in images. The results obtained by (2) for images containing the same object #1 but with different SNR (images taken along the vertical axis of Fig. 5) are collected in Fig. 8.

From Fig. 8 we see that the influence of noise upon the measurement of object features is quite different. The A curves are again distinct from the other ones. In Fig. 8(A), the A curves are shifted (it is mainly caused by the opening operation) to the left with the decrease of SNR level, but they still keep the same form. Since the feature value variation is caused by the change of algorithm parameters, the ability of A to test the influence of parameter modification is relatively independent of the SNR level of images. It is thus possible by choosing the appropriate parameter values of an algorithm to obtain comparable results from images of different SNR level. The curves of other features are monotonically moving up with the decrease of SNR level. Besides, the forms of these curves are also modified, which ranged from relative flat for the higher SNR cases to somehow sharp for the lower SNR cases. Both effects mean that the UMA values of these features are more related to the SNR level of images.

The above experiments are only examples that show the behaviour difference among the features in evaluation, with different image contents and imaging conditions. According to the generality of the evaluation framework and performance criteria, other features can be studied and used in evaluation according to the segmentation goal. More complicated algorithms and images can also be tested.

5. Concluding remarks

In this paper, we first presented an overview of existing segmentation evaluation methods, with an emphasis on their limitations and weakness. This survey shows that the appropriate criteria for judging algorithms are critical. In an attempt to study segmentation techniques objectively and quantitatively, we proposed new performance criteria and introduced a general framework. These criteria are based on the measurement accuracy of object features from segmented images. Several object features are studied in the context of their ability to assess segmentation performance within the general framework.

The advantages of our new approach are multifold. First is the generality. This approach is appropriate for a wide range of techniques and real applications. This is because that the evaluation criteria are independent of applied segmentation algorithms and the general framework can be adapted according to the requirement of a special segmentation task. Second is the effectiveness. In image analysis, what we are most interested in is to obtain an accurate measurement of object properties and the performance criteria that we used are suitable for this task. Thirdly, no subjective bias has been introduced in the evaluation process. The results obtained are objectively accurate. Fourthly, our approach is based on the comparison with exact references. As we know the 'best' achievable result, all algorithms can be easily ranked against the 'ground-truth'.

The proposed approach has been applied to study different segmentation techniques with satisfactory results. One study is the comparison of five distinct thresholding techniques with images containing objects of different sizes and distorted by noise [21]. Another comparison of thresholding techniques with other types of segmentation techniques also provides useful information regarding to different imaging and pre-processing conditions [17]. In addition, a complete evaluation and comparison study has also been carried out with valuable results [18]. In this study different types of representative segmentation procedures that include both boundary-based and region-based techniques as well as the algorithms using parallel and sequential strategies are treated. Besides, images of various contents and differently corrupted are used. All those studies show that our approach for objective and quantitative segmentation evaluation and comparison is valid and useful.

Acknowledgments

We are very grateful to the anonymous referees for their helpful comments and valuable suggestions to improve the presentation of this paper.

References

[1] I.E. Abdou and W. Pratt, "Quantitative design and evaluation of enhancement/thresholding edge detectors", *Proc. IEEE*, Vol. 67, No. 5, 1979, pp. 753–763.

[2] S. Bradbury, "Quantitative image analysis", in: G.A. Meek and H.Y. Elder, eds., *Analytical and Quantitative Methods in Microscopy*, 1977, pp. 91–116.

[3] K.S. Fu and J.K. Mui, "A survey on image segmentation", *Pattern Recognition*, Vol. 13, 1981, pp. 3–16.

[4] R.M. Haralick and L.G. Shapiro, "Image segmentation techniques", *Comput. Vision, Graphics and Image Processing*, Vol. 29, 1985, pp. 100–132.

[5] H. Harms and H.M. Aus, "Computer color vision: Automated segmentation of biological tissue sections", in: E.S. Gelsema and L.N. Kanal, eds., *Pattern Recognition in Practice II*, 1986, pp. 323–330.

[6] L. Kitchen and A. Rosenfeld, "Edge evaluation using local Edge Coherence", *IEEE Trans. Systems. Man and Cybernetics*, Vol. SMC-11, No. 9, 1981, pp. 597–605.

[7] S.U. Lee *et al.*, "A comparative performance study of several global thresholding techniques for segmentation", *Comput. Vision, Graphics Image Processing*, Vol. 52, 1990, pp. 171–190.

[8] M.D. Levine and A. Nazif, "Dynamic measurement of computer generated image segmentations", *IEEE Trans. Pattern Anal. Mach. Intell.*, Vol. PAMI-7, No. 2, 1985, pp. 155–164.

[9] C.E. Liedtke *et al.*, "Segmentation of microscopic cell scenes", *Analytical and Quantitative Cytology and Histology*, Vol. 9, No. 3, 1987, pp. 197–211.

[10] C. MacAulay and B. Palcic, "A comparison of some quick and simple threshold selection methods for stained cells", *Analytical and Quantitative Cytology and Histology*, Vol. 10, No. 2, 1988, pp. 134–138.

[11] J.K. Mui and K.S. Fu, "Automated classification of nucleated blood cells using a binary tree classifier", *IEEE Trans. Pattern Anal. Mach. Intell.*, Vol. PAMI-2, No. 5, 1980, pp. 429–443.

[12] N.R. Pal and S.K. Pal, "A review on image segmentation techniques", *Pattern Recognition*, Vol. 26, 1993, pp. 1277–1294.

[13] A. Rosenfeld and L.S. Davis, "Image segmentation and image models", *Proc. IEEE*, Vol. 67, No. 5, 1979, pp. 764–772.

[14] P.K. Sahoo *et al.*, "A survey of thresholding techniques", *Comput. Vision. Graphics and Image Processing*, Vol. 41, 1988, pp. 233–260.

[15] J.S. Weszka and A. Rosenfeld, "Threshold evaluation techniques", *IEEE Trans. System, Man and Cybernetic*, Vol. SMC-8, No. 8, 1978, pp. 622–629.

[16] W.A. Yasnoff *et al.*, "Error measures for scene segmentation", *Pattern Recognition*, Vol. 9, 1977, pp. 217–231.

[17] Y.J. Zhang, "Image synthesis and segmentation comparison", *Proc. 3rd Internat. Conf. for Young Computer Scientists*, Beijing, China, 15–17 July, 1993, pp. 8.21–8.24.

[18] Y.J. Zhang, "Segmentation evaluation and comparison: A study of various algorithms", *Proc. Visual Communications and Image Processing '93*, Cambridge, Massachusetts, USA, Nov. 7–12, 1993, SPIE Vol. 2094, pp. 801–812.

[19] Y.J. Zhang and J.J. Gerbrands, "Transition region determination based thresholding", *Pattern Recognition Lett.*, Vol. 12, 1991, pp. 13–23.

[20] Y.J. Zhang and J.J. Gerbrands, "On the design of test images for segmentation evaluation", *Proc. 6th European Signal Processing Conf.*, Brussels, Belgium, 24–27 August, 1992, Vol. 1, pp. 551–554.

[21] Y.J. Zhang and J.J. Gerbrands, "Comparison of thresholding techniques using synthetic images and ultimate measurement accuracy", *Proc. 11th Internat. Conf. on Pattern Recognition*, The Hague, The Netherlands, 30 August–3 September, 1992, Vol. 3, pp. 209–213.