Pattern Recognition Letters 12 (1991) 13-23
North-Holland

January 1991

Transition region determination based thresholding

Y.J. Zhang and J.J. Gerbrands

Department of Electrical Engineering, Delft University of Technology, Delft, Netherlands

Received 25 April 1990
Revised 28 August 1990

Abstract

Zhang, Y.J. and J.J. Gerbrands, Transition region determination based thresholding, Pattern Recognition Letters 12 (1991) 13-23.

We present a newly developed thresholding technique which is not based on the image's gray-level histogram. This technique is fully automatic and quite robust in the presence of noise and unexpected structures. Moreover, no empirical parameters are used, and no limitations on shape and size of objects are imposed. A comparison with histogram based threshold selection is also discussed.

Keywords. Automatic threshold selection, transition region, effective average gradient.

1. Introduction

Thresholding is a popular tool used in image segmentation. A wide range of thresholding techniques have been developed, a survey of them can be found in Sahoo et al. (1988).

Standard approaches to threshold selection are often based on locating valleys or peaks in the image's gray-level histogram. Two classes can be distinguished: (1) direct histogram analysis (e.g., Prewitt and Mendelsohn (1966), Rosenfeld and Torre (1983)); (2) histogram transformation (e.g., Weszka and Rosenfeld (1979)).

Here we introduce a new threshold selection method which is not based on the gray-level histogram, but on the determination of the transition region between objects and background.

The existence of the transition region in the discrete image has been discussed by Gerbrands (1988). The conclusion is that "even if the continuous image contains ideal step edges, the discrete image, which results from sensing the image and sampling according to the Shannon theorem, will contain transition regions which are quite distinct" (at least one pixel wide). This region is geometrically located between objects and background, and is composed of pixels which have intermediate gray-levels between those of objects and of the background (Zhang (1989)).

2. Development of the method

In the following, we first introduce a new parameter that represents some interesting characteristics of the image. In addition we introduce a technique to reduce the effect of noise, which makes use of this parameter. We then relate

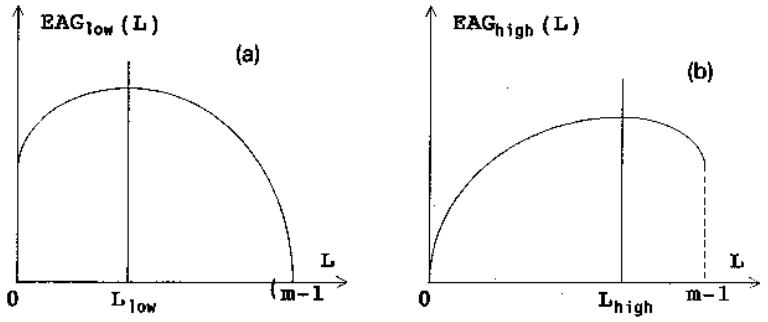

Figure 1. Typical curves of $EAG(L)$ versus L. (a) Typical $EAG_{low}(L)$ curve. (b) Typical $EAG_{high}(L)$ curve.

this parameter to the definition and the determination of the transition region. Once the transition region has been determined, threshold selection can take place.

2.1. Effective Average Gradient

First, we introduce the concept of EAG (Effective Average Gradient). Let $f(i,j)$ be an image function:

$$f(i,j), \quad f \in G, \ i,j \in S \tag{1}$$

where $S=[1,2,\ldots]$ is a set of integers representing spatial coordinates of the pixels, and $G=[0,1,\ldots]$ is a set of integers representing the gray-levels of pixels. The EAG of a gradient image $g(i,j)$ obtained from an original image $f(i,j)$ by using gradient operators (such as discussed by Pratt (1978)), is defined by the following equation:

$$EAG = \frac{TG}{TP} \tag{2}$$

where

$$TG = \sum_{i,j \in S} g(i,j) \tag{3}$$

is the total magnitude value of the gradient image, and

$$TP = \sum_{i,j \in S} p(i,j) \tag{4}$$

is the total number of pixels with non-zero gradient values, as $p(i,j)$ is defined by:

$$p(i,j) = \begin{cases} 1 & \text{if } g(i,j) \neq 0, \\ 0 & \text{if } g(i,j) = 0. \end{cases} \tag{5}$$

From the above definitions, we can see that only the pixels which have non-zero gradient value are involved in the calculation of EAG, which is the meaning of 'effective'. The EAG represents a selected statistic of the image.

To reduce the influence of noise, a type of image transformation, called clip transformation, is introduced. Given an original image $f(i,j)$ and a gray-level L ($L \in G$), a transformed image $f_L(i,j)$ can be obtained by using the following clip transformation:

$$f_L(i,j) = \begin{cases} f(i,j) & \text{if } f(i,j) > L, \\ L & \text{if } f(i,j) \leqslant L. \end{cases} \tag{6}$$

From the transformed image $f_L(i,j)$, a gradient image $g_L(i,j)$ can be obtained. The EAG of image $g_L(i,j)$ is a function of L (for a given gradient operator). It should be written as $EAG(L)$. Moreover, $EAG(L)$ is also dependent upon which part of the image is clipped out by the transformation. In equation (6), the low gray-level part of the original image has been clipped (all gray-levels not exceeding L are changed to L). If we clip the high gray-level part of the image by using the transformation as follows:

$$f^L(i,j) = \begin{cases} L & \text{if } f(i,j) \geqslant L, \\ f(i,j) & \text{if } f(i,j) < L \end{cases} \tag{7}$$

a different $EAG(L)$ will be obtained.

To distinguish, we call the effective average gradient based on the transformation of (6) the low part clipped $EAG_{low}(L)$ while the effective average gradient based on the transformation of (7) is called high part clipped $EAG_{high}(L)$.

Two drawings to show the typical curves of $EAG(L)$ versus L are given in Figure 1(a) for $EAG_{low}(L)$ and in Figure 1(b) for $EAG_{high}(L)$. Common properties of both curves are that without clipping, the EAGs do not equal zero; for increasing L, the EAGs increase in the beginning and decrease after one maximum; and finally, both EAGs become zero. These points will be made evident in the following discussions.

2.2. Maximum points of Effective Average Gradient

From Figure 1 we can see that both $EAG_{low}(L)$ and $EAG_{high}(L)$ reach a maximum at two given gray-levels L_{low} and L_{high}, respectively:

$$L_{low} = \text{Arg}\left\{\underset{L\in G}{\text{Max}}\,[EAG_{low}(L)]\right\}, \tag{8}$$

$$L_{high} = \text{Arg}\left\{\underset{L\in G}{\text{Max}}\,[EAG_{high}(L)]\right\}. \tag{9}$$

Those two maximum points have the following important properties regardless of the relative gray-levels of the objects and background:

First property. *There exist one and only one* L_{low} *and* L_{high}, *respectively.*

Second property. *Both* L_{low} *and* L_{high} *have significant discrimination meaning* (see proof).

Third property. L_{high} *is never smaller than* L_{low}.

The physical interpretations of the maximum points and their properties can be found in Zhang (1989). We will give the analytical proofs of these properties in Appendix A.

2.3. Definition of transition region

A transition region in an image is a 2-D region whose gray-level range is limited by two borders in

Figure 2. The transition region of a real image.

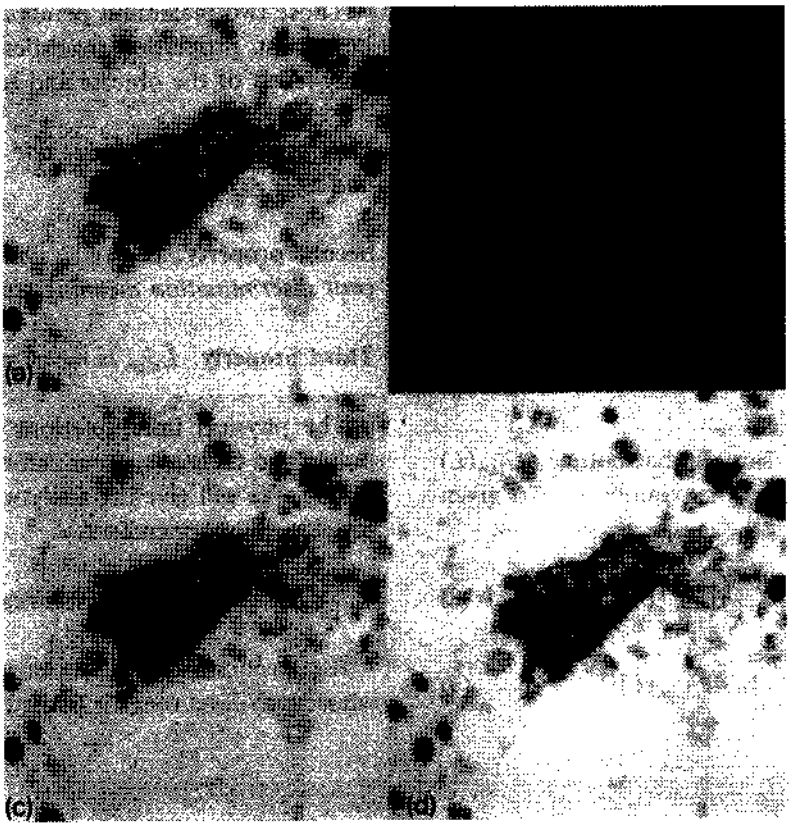

Figure 3. Comparison example. (a) One real world image. (b) Histogram of image (a). (c) Segmentation result using threshold determined by histogram modification method. (d) Segmentation result using threshold determined by our method.

1-D gray-level space. These two borders are L_{low} and L_{high}, which can be determined by equations (8) and (9), respectively. The transition region can be written as:

$$t(i,j), \quad i,j \in S, \ t \in R \tag{10}$$

where $R = [L_{low}, L_{low}+1, \ldots, L_{high}]$ is a subset of G.

A real example of the transition region for a cell image is shown in Figure 2, where the transition region is geometrically limited by two contours. It is easy to see that the real border separating the object from the background is located somewhere between these two contours, i.e., inside the transition region. The uniqueness of the transition region for a given object is guaranteed by the first property of the maximum point of $EAG(L)$. From the second property, we know that the pixels in the transition region generally have a high contrast compared to their neighborhoods. The third property is illustrated by the fact that the two contours do not cross.

2.4. *Transition region determination and threshold selection*

The determination of the transition region is related to its definition. To detect a transition region, it is necessary to find its two borders L_{low} and L_{high}, which can be calculated according to equations (8) and (9), respectively. Intuitively, we should determine $EAG_{low}(L)$ and $EAG_{high}(L)$ for all $L \in G$ to discover their maximum points. However, according to equation (2), EAG is determined by the pixels in a range of gray-levels. For a slight change of L, only a small fraction of these pixels are affected, so that the EAG value will not change radically. In fact, the curve of EAG versus L is a rather smooth curve (see examples in the next

section). To find the maximum value from such a curve, one fast iterative procedure to short-cut the calculations has been designed, as is discussed in Appendix B. Using this procedure, the computation time for images with 256 gray-levels can be reduced by a factor of about 25.

Delineation of the transition region makes the selection of a threshold straightforward. As the boundary pixels are contained in the transition region, a threshold can be determined on the basis of the pixels in this region. For example, the threshold value can be the mean gray-level of all pixels belonging to the transition region, or the mode of the histogram of the transition region.

3. Comparison and discussion

Liedtke et al. (1987), after discussing several segmentation techniques, have pointed out that "the amount of relevant *a priori* knowledge that can be incorporated into the segmentation algorithm is decisive for the reliability of the method." Below, we discuss and compare our method with histogram based methods in terms of what *a priori* information of the image is used. We compare also the performance of different methods in the presence of noise as real images are always contaminated by noise.

In direct histogram analysis techniques, only 1-D gray-level information—the pixel population—is used. All pixels of the image have equal contribution to the threshold selection. The influence of noise corrupted pixels increases as the noise increases, since the relative difference between the gray-levels of the object and background will decrease in such cases.

In histogram transformation techniques, local contrast information is employed. The contribution of each pixel to the threshold selection is weighted according to its 'edge value' (the rate of change of gray-level at this pixel). However, no distinction between real edges and noise edges is made. Since noise corrupted pixels generally have higher edge values, the influence of noise on the threshold selection becomes even more important with the introduction of local contrast information.

In the transition region determination based thresholding technique, the contribution of each pixel to the threshold selection is weighted according to its gradient value and its gray-level related to the whole gray-level distribution of the image. The gradient value of noise corrupted pixels belonging to the object and/or background will be diminished or even suppressed by the clip transformation. The 1-D gray-level information, such as the expected absolute gray-level of objects and background, the 2-D local contrast information, such as the gradient value of pixels, and the 2-D global information, such as the pixel arrangement, are combined to improve the performance of threshold selection.

Except the above difference about the incorporation of a priori knowledge, another important point is how such knowledge is used in the threshold selection. The technique of histogram modification for threshold selection, discussed by Weszka and Rosenfeld (1979) and many others, has several variants. The most general one consists of constructing a transformed histogram by calculating the average of the 'edge values' for each gray-level of the image, and then choosing the

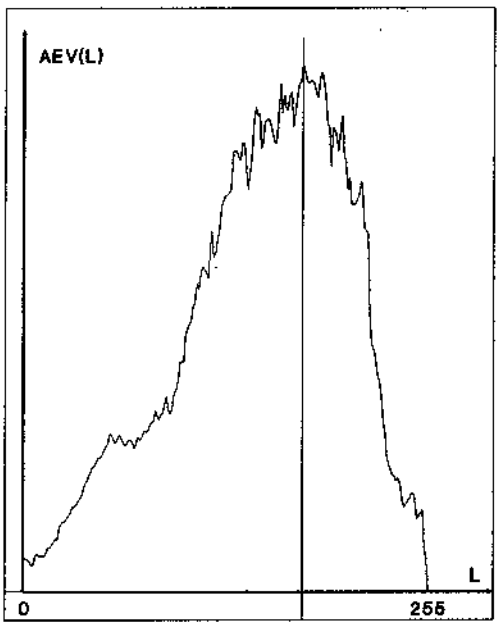

Figure 4. The curve of average edge value (Weszka and Rosenfeld (1979)) for Figure 3(a).

peak from the transformed histogram. The authors claim that this average should certainly be higher for the border pixel gray-levels than that for the interior pixel gray-levels. However, as each time the calculation is based on the pixels with only one gray-level, a possible difficulty can arise. Unwanted structures and/or noise having high edge value will cause the peak of the transformed histogram to be located at a gray-level corresponding to those structures or noise. The threshold value thus obtained will not be appropriate. On the contrary, in our method, the calculation of EAG is based on all pixels which have non-zero gradient value. Since the pixel number involved is rather high, the disturbance of undesirable structures and/or noise is limited.

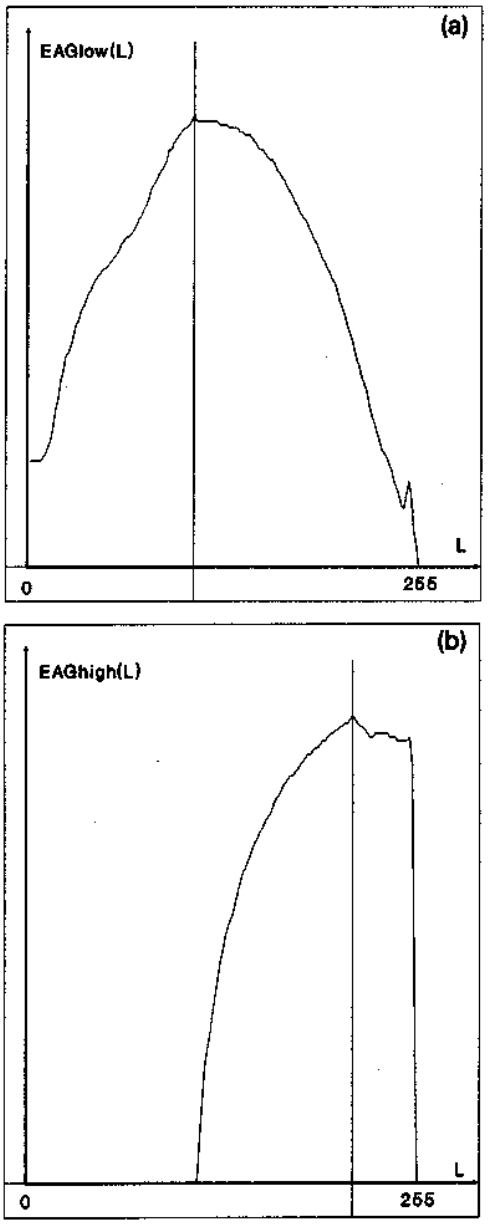

Figure 5. $EAG(L)$ curves for the image of Figure 3(a). (a) $EAG_{low}(L)$ curve. (b) $EAG_{high}(L)$ curve.

The above theoretical comparison can be further illustrated by the following real world example. In Figure 3(a), a cell image which involves some noise and undesirable structures is shown. Its histogram is given in Figure 3(b). The difficulty of using direct histogram analysis techniques arises as the valley is not visible. The results of segmentation using the above mentioned histogram transformation method and our method are shown in Figures 3(c) and 3(d), respectively. By comparing Figure 3(c) with Figure 3(d), the advantage of our method is quite noticeable.

To see the difference more clearly, let us compare Figures 4 and 5. In Figure 4, the transformed histogram curve (i.e., the average 'edge value' curves) for the image of Figure 3(a) is presented. The EAG_{low} and EAG_{high} curves for the same image are shown in Figures 5(a) and 5(b), separately. When we compare Figure 4 with Figure 5, we notice the irregularity of the transformed histogram curve as well as the smoothness of the EAG curves. The vulnerability of the histogram modification technique and the robustness of our method are obvious. We should note that although we only show one real example here, such situations are generally expected, as we have discussed before showing this example.

4. Conclusions

We have introduced a new threshold selection method. Since it is based on the determination of the transition region, no histogram calculation is needed. A comparison with histogram based threshold selection methods is also presented.

This technique is possible to be extended to the case of multi-thresholding. In fact, we have used it in the quantitative analysis of megakaryocyte in

bone marrow tissue (see Zhang (1990)). In that analysis task, the nucleus and cytoplasm of megakaryocytes (they have different mean gray levels) need to be separated from other structures. A two-step procedure has been designed. In the first step, the whole cell is extracted from the image. This result is then used in the second step where the cytoplasm is separated from the nucleus. Finally, the images are split into three types of regions: nucleus, cytoplasm and other structures. Our technique has been used in both steps to determine two thresholds (one for each step). The results are quite satisfactory. In one experiment with about 200 images, more than 95% of nucleus and cytoplasm are appropriately segmented.

We conclude by indicating some principal characteristics of the transition region determination based thresholding technique:

(1) *Fully automatic and quite general*:

As described earlier, the transition region between objects and background is located and delimited automatically, without any human intervention. The threshold value can then be calculated automatically on the basis of this region. Additionally, no subjective experimental parameters and no limitation on shape and/or size of objects are introduced, so that the calculations can be deterministically carried out. Moreover, as no connectivity constraints have been introduced during the determination of the transition region, the method is not limited to the single object case.

(2) *Robust and accurate*:

The clip transformation has the property to limit the influence of disturbing structures. The final threshold value can hardly be affected by the presence of such structures. Moreover, there is an averaging process inherent to the calculation of the *EAG*. The *EAG* values are determined on the basis of a large number of pixels from a range of gray-levels, so that the effect of noise to the threshold selection is decreased and more accurate results can be expected.

(3) *Computational efficiency*:

Theoretically, the clip transformation as well as the calculations of the gradient, *TG* and *TP* are all operations which may be implemented in parallel. Practically, these operations are standard ones which are included in most commercially available image processing systems. Finally, the short-cut processes discussed above permit fast computation of the transition region.

Acknowledgments

This work has been possible thanks to the generous help and invaluable advice of Professor G. Cantraine and Dr. J.M. Paulus, to whom we extend heartfelt gratitude. We are thankful to the members of the electronic group and the hematology group of the Liège University, Belgium, for their help. The support of the "Netherlands' Project Team for Computer Application (SPIN), Three-dimensional Image Analysis Project" is also acknowledged. We are very grateful to the referee for the comments and suggestions to improve the presentation of this paper.

Appendix A

Proof for three properties of maximum points of EAG

For the sake of clarification, we assume, without loss of generality, that the given image has one object on the background, and the object is surrounded by a noisy border ramp. Moreover, since the calculation is performed for the entire image, we will only treat an edge ramp which can be considered as the average of all profiles of the border ramp. As a result, such an edge ramp will be rather smooth and from a practical point of view can be considered as monotonous.

Below, the proof is based on continuous functions. Digital images are only approximations of continuous functions, but the extension to real images is straightforward (the continuous case can be considered as the sets G and S going to infinity).

First property. *There exist one and only one* L_{low} *and* L_{high}, *respectively.*

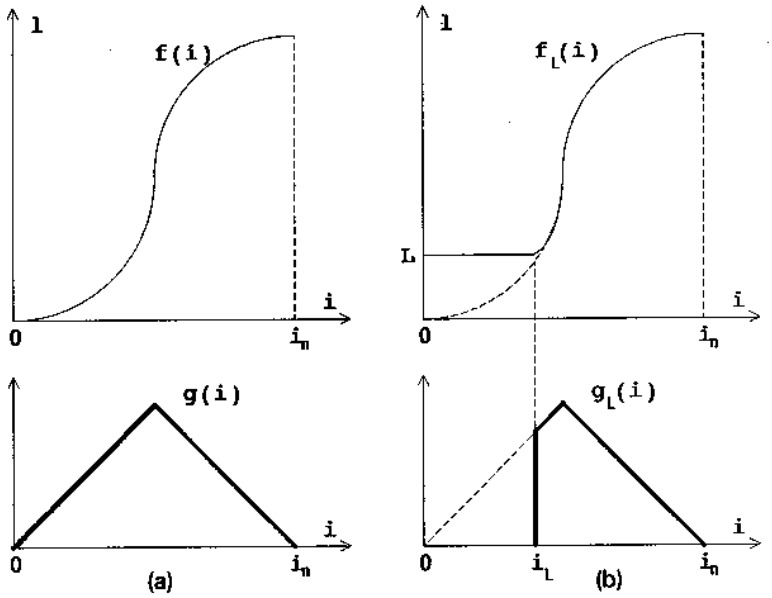

Figure 6. Profile representation of the edge ramp. (a) An edge ramp and its gradient. (b) Clipped edge ramp and its gradient.

The proof is only given for L_{low}; a similar proof can be derived for L_{high}.

An edge ramp and its gradient are illustrated in Figure 6(a), where:

$$l = f(i) \tag{A.1}$$

is a function of gray-level versus spatial coordinate.

For every l, a corresponding i exists. The inverse function can be defined as:

$$i = f^{-1}(l). \tag{A.2}$$

For each given L, if we use it to clip $f(i)$, one clipped edge ramp $f_L(i)$ and one clipped edge ramp gradient $g_L(i)$ are produced. Figure 6(b) shows an example.

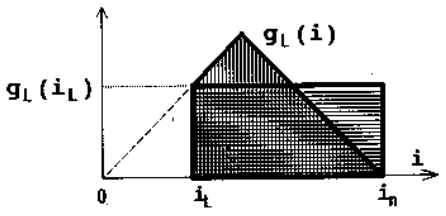

Figure 7. Geometric interpretation of the numerator in the right side of equation (A.7).

As L can be variable, according to (A.2), we have:

$$i_L = f^{-1}(L). \tag{A.3}$$

$EAG_{low}(L)$ can be, according to Figure 7, written as:

$$EAG_{low}(L) = \frac{\int_{i_L}^{i_n} g_L(i)\,di}{i_n - i_L}. \tag{A.4}$$

Now, we calculate the derivative of $EAG_{low}(L)$ with respect to L. Keeping in mind that i_L is a function of L, and according to the Leibniz's theorem for differentiation of an integral (see, e.g., Abramowitz and Stegun (1964)), we have:

$$EAG'_{low}(L) = \frac{i'_L\left(\int_{i_L}^{i_n} g_L(i)\,di - (i_n - i_L)g_L(i_L)\right)}{(i_n - i_L)^2}. \tag{A.5}$$

The derivative of i_L with respect to L can be written successively:

$$i'_L = \{f^{-1}(L)\}' = 1/f'(i_L) = 1/g_L(i_L). \tag{A.6}$$

Taking (A.6) into (A.5), we finally obtain:

$$EAG'_{low}(L) = \frac{\int_{i_L}^{i_n} g_L(i)\,di - (i_n - i_L)g_L(i_L)}{(i_n - i_L)^2 g_L(i_L)}. \tag{A.7}$$

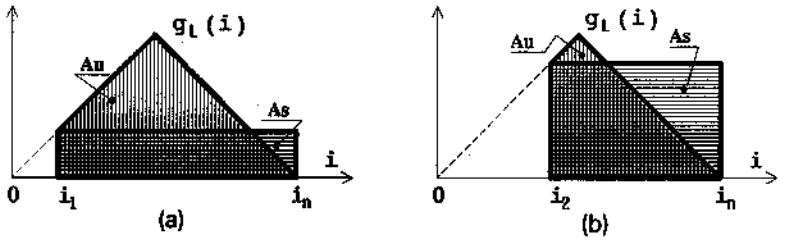

Figure 8. Drawings to demonstrate the existence of the stationary point of $EAG_{low}(L)$. (a) The clip gray level L corresponds to a rather small value i_1. (b) The clip gray level L corresponds to a rather big value i_2.

The denominator in the right side of equation (A.7) is positive for $0<i_L<i_n$. The sign of $EAG'_{low}(L)$ is then determined by the numerator. The two terms of the numerator in (A.7) have simple geometric interpretations (see Figure 7):

The first term equals the area under the curve of $g_L(i)$ from $i=i_L$ to $i=i_n$, while the second one equals the area of the rectangle with height $g_L(i_L)$ and width (i_n-i_L). At the stationary point, these two areas are equivalent.

Let us now consider the two drawings of Figure 8 to see the existence of the stationary point. It is evident that in Figure 8(a), we will have $Au>As$, i.e.:

$$\int_{i_1}^{i_n} g_L(i)\,\mathrm{d}i-(i_n-i_1)g_L(i_1)>0, \tag{A.8}$$

while from Figure 8(b), we can get $Au<As$, i.e.:

$$\int_{i_2}^{i_n} g_L(i)\,\mathrm{d}i-(i_n-i_2)g_L(i_2)<0. \tag{A.9}$$

As $EAG'_{low}(L)$ is a continuous function, according to the mean value theorem, there exists an $i=i_{low}$ (and also a corresponding $L=L_{low}$), where $i_1<i_{low}<i_2$, which makes:

$$EAG'_{low}(L_{low})=0. \tag{A.10}$$

L_{low} is thus a stationary point of $EAG_{low}(L)$. It

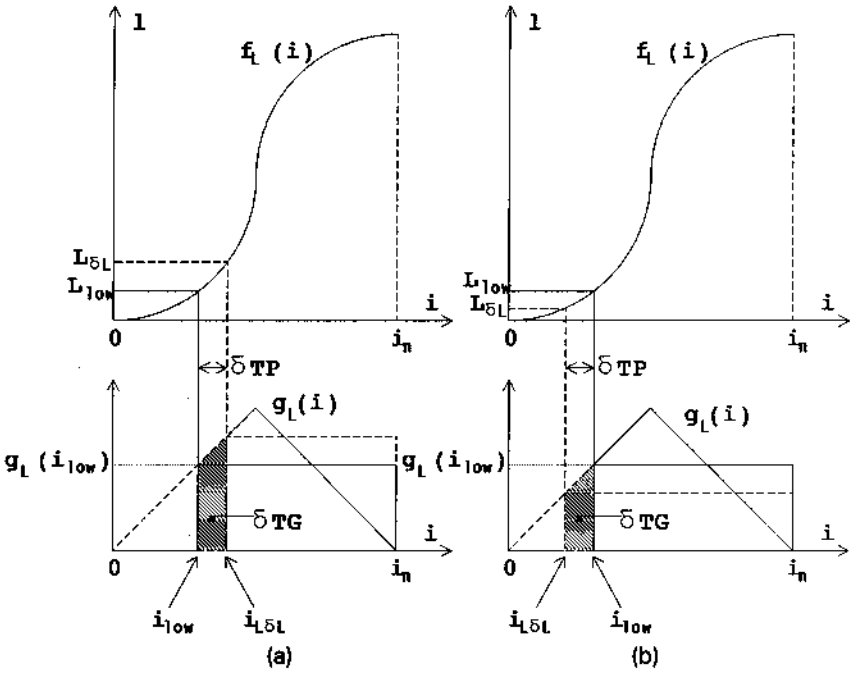

Figure 9. The changes of $EAG_{low}(L)$ with respect to a small increment δL of L_{low}. (a) If $\delta L>0$. (b) If $\delta L<0$.

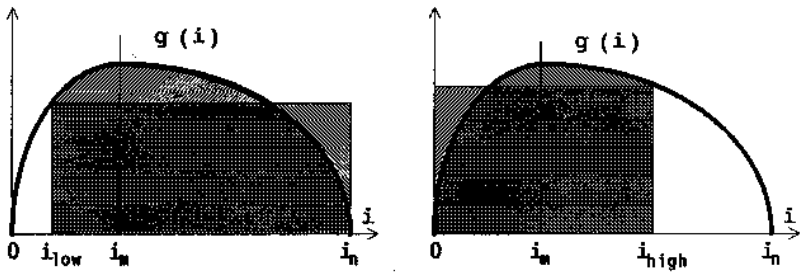

Figure 10. Drawing to show the relation of L_{high} and L_{low}.

is also the only stationary point in the interior region, since Au decreases monotonically whereas As increases monotonically before Au attains zero. Moreover, by taking equations (A.8) and (A.9) into (A.7), respectively, we will get $EAG'_{low}(L)>0$ for $L<L_{low}$ and $EAG'_{low}(L)<0$ for $L>L_{low}$. We conclude that $EAG_{low}(L)$ attains the maximum at $L=L_{low}$ and there is one and only one L_{low}.

Second property. *Both L_{low} and L_{high} have significant discrimination meaning.*

Combining (A.8) and (A.9), we know that at the maximum point (L_{low}) there is:

$$\int_{i_{low}}^{i_n} g_L(i)\,\mathrm{d}i = (i_n - i_{low})g_L(i_{low}). \qquad \text{(A.11)}$$

Comparing (A.11) with equation (A.4), we get:

$$EAG_{low}(L_{low}) = g_L(i_{low}). \qquad \text{(A.12)}$$

Now let us see what happens if we give L_{low} a small increment δL (Figure 9):

(1) For $\delta L>0$, i.e., $i_{L\delta L}>i_{low}$, Figure 9(a) shows:

$$\delta TG/\delta TP > EAG_{low}(L_{low}) \quad (\delta L>0). \qquad \text{(A.13)}$$

The inequality of (A.13) means that a positive increment δL will cause some pixels with high gradient values to be discarded from $EAG_{low}(L)$.

(2) For $\delta L<0$, i.e., $i_{L\delta L}<i_{low}$, Figure 9(b) shows:

$$\delta TG/\delta TP < EAG_{low}(L_{low}) \quad (\delta L<0). \qquad \text{(A.14)}$$

The inequality of (A.14) means that a negative increment δL causes some pixels with low gradient values to be introduced into $EAG_{low}(L)$. Combining (A.13) and (A.14), we know that L_{low} is the gray-level at which the image preserves maximum number of pixels with high gradient value and minimum number of pixels with low gradient value in the low gray-level range of the image. L_{high} has a similar discriminaton meaning as L_{low}, but in the high gray-level range.

Third property. *L_{high} is never smaller than L_{low}.*

Suppose that $g(i)$ attains its maximum value at $i=i_m$. According to the geometric representation of Figure 7, i_{low} must be at the left side of i_m, i.e., $i_{low}<i_m$, and on the other hand, i_{high} must be at the right side of i_m, i.e., $i_{high}>i_m$. Combining these two inequalities, we have $i_{high}>i_{low}$. These relations are depicted in Figure 10. As I is a monotonic function of i, we can conclude that L_{high} will always be greater than L_{low} in real images.

Appendix B

Fast procedure to calculate maximum points of EAG

This procedure is adapted from an optimization technique, i.e., the fraction method (Wang (1979)). This method permits to find the maximum of a continuous function with minimum trials.

As we treat the digital image, we make use of the series of Fibonacci defined as follows:

$$\begin{aligned} &F(1)=1, \quad F(2)=1, \\ &F(k)=F(k-2)+F(k-1), \quad k=3,4,\ldots \end{aligned} \qquad \text{(B.1)}$$

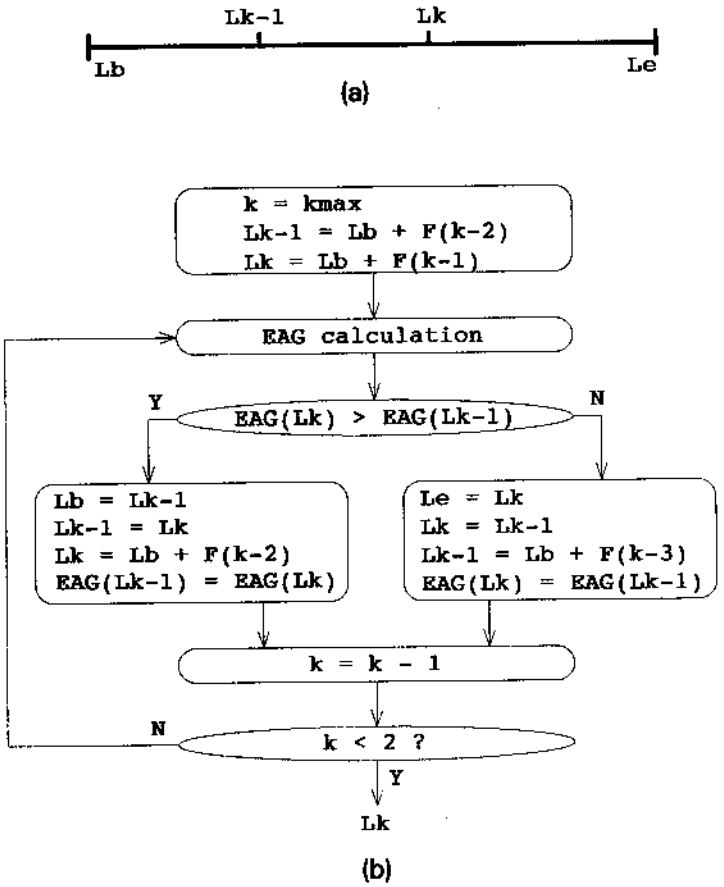

Figure 11. Fast procedure to calculate the maximum points of $EAG(L)$. (a) Choice of clip level L. (b) Flowchart of procedure.

to calculate EAG only at some discrete points.

Figure 11 shows the diagram of the adapted procedure which makes it possible to choose L from the Fibonacci series for fast determination of maximum points of EAG, where L_b and L_e are the lowest and highest gray-levels of the original image, respectively. $k_{\max}$ can then be determined from:

$$F(k_{\max}+1) > L_e - L_b + 1 > F(k_{\max}). \qquad \text{(B.2)}$$

References

Abramowitz, M. and I.A. Stegun (1964). *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables.* U.S. Department of Commerce.

Gerbrands, J.J. (1988). Segmentation of Noisy Images. Ph. D. dissertation, Delft University, The Netherlands.

Liedtke, C.E., T. Gahm, F. Kappei and B. Aeikens (1987). Segmentation of microscopic cell scenes. *Analytical & Quantitative Cytology* 9 (3), 197–211.

Pratt, W.K. (1978). *Digital Image Processing.* Wiley, New York.

Prewitt, J.M.S. and M.L. Mendelsohn (1966). The analysis of cell images. *Ann. New York Acad. Sci.* 128, 1035–1053.

Rosenfeld, A. and P. De la Torre (1983). Histogram concavity analysis as an aid in threshold selection. *IEEE Trans. Syst. Man Cybernet.* 13, 231–235.

Sahoo, P.K., S. Saltani, A.K. Wong and Y.C. Chen (1988). A survey of thresholding techniques. *Computer Vision, Graphs, and Image Processing* 41, 233–260.

Wang, L.X. et al. (1979). *Handbook of Mathematics.* Beijing.

Weszka, J.S. and A. Rosenfeld (1979). Histogram modification for threshold selection. *IEEE Trans. Syst. Man Cybernet.* 9, 38–52.

Zhang, Y.J. (1989). Development of a 3-D Computer Image Analysis System: Application to Megakaryocyte Quantitation. Ph.D. dissertation, Liège University, Belgium.

Zhang, Y.J. (1990). A 3-D image analysis system: quantitation of megakaryocyte. Submitted to *Cytometry.*