A Combinatorial Algorithm for Detecting Encircling Regions

Christos Liambas and Leonidas Pitsoulis

Department of Mathematical, Physical, and Computational Sciences, Faculty of Engineering, Aristotle University of Thessaloniki
54006 Thessaloniki, Greece
chrili@itc.auth.gr
pitsoulis@gen.auth.gr

Abstract: In this study a combinatorial algorithm is developed for the detection of regions that could be approximated by circular disks. The figure of a region is a section where the boundary is inordinately asymmetrical and jagged. Three circular disks are computed for each region. The indisk which is the largest disk contained in the region, the outdisk which is the smallest disk containing the region and has the same center as the indisk, and the approximation disk with area equal to the area of the region and has the maximum intersection with the region. Reasonably there are some problems described as a considerable remark for efficient use of the proposed algorithm: in meteorology, where hail pads have to be analyzed in order to determine the number and characteristics of the hailstones that collided with the pad, in biology especially in cytology where cells have to be identified for the reason that contains useful information that should be extracted. Computational results on a set of benchmark images from actual data are presented.

Key words: Circular Disk, Combinatorial Algorithm, Image Processing, Pattern Recognition

INTRODUCTION

The main object is to investigate the relevance between the proposed algorithmic technique in various existing applications that arises in several areas. These applications may serve as a test bed for inspecting the benefits of the recommended method. This document also describes and clarifies the development of a combinatorial algorithm in order to detect encircling regions from a noisy image. Encircling region is a set of black pixels that under specific circumstances produce an irregular circular disk with abnormal border. Also, region could be called shape.

Only few comparative studies on the effectiveness of combinatorial optimization techniques to solve combinatorial problems in image analysis are currently available [Alexander Toet 01]. This report is oriented for researchers specialized in combinatorial approach.

The organization of the rest of this report is as follows. First some known scientific problems are introduced from image analysis. Then the main algorithm implementation is supervised. Afterwards a required optimization of the above method is needed because decreases significantly the processing time. Some useful test images are examined for checking the algorithm’s efficiency. Finally, some concluding remarks and suggestions are obtained.

1. Problem Statement

The Identification of Encircling Region Problem (IER) is defined as follows:

Definition 1 (IER) Given a noisy black and white image where the black areas define irregular shapes unknown in number and size, which overlap each other and can be approximated by circular disks, identify these areas. For any irregular shape S an indisk which is the largest disk contained in S and an outdisk which is the smallest disk containing S, say D_I and D_O respectively, are associated to S. Shape S can be approximated by a circular disk D_A if the centers of D_I and D_O coincide, say at c (Figure 1). The area of the approximation disk D_A is equal to the area of the shape S therefore its radius \( r_{D_A} \) is fixed. Its center should lie in the area defined by the disk with center c and radius \( r_{D_O} - r_{D_I} \), since it must fully contain the indisk. Its exact location would be that
point in this area where the intersection of the approximation disk and the shape, with respect to the black pixels, is maximum. The approximation disk defines an associated shape in the sense that its area and location are known.

The following difficulties were acknowledged with the IER problem, which the proposed algorithm resolves:

- Overlapping between the shapes in the image
- The number and size of the shapes is unknown
- Noise in the image

Depending on the application, limited computational power may also be an issue. For instance, the images have a resolution of 300 dots per inch.

1.1. Hail pad analysis in Meteorology

The IER problem has a direct application in meteorology where images from hail pads have to be analyzed automatically, in order to determine the number and characteristics of the hail stones that collided with the pad. This is a very important issue, in the specific science, because the automated statistical evaluation increases the research progress dramatically, since it does not require any subjective human intervention.

A hail pad is a pad with a coated surface of area approximately $27 \times 27\text{cm}^2$, where upon impact, a hailstone leaves a permanent mark, from which some conclusions about the size, kinetic energy and other information about the stone, are derived. These primitive parameters are considered necessary in order to adequately be used to any statistical evaluation procedure. A typical image from a hail pad is shown in Figure 2 with an added technical uniform noise 10%.

The main concept is to use a network of hail pads which are placed in an area where hailstone storms occur, in order to measure the characteristics of the storm [T.S. Karacostas 02].

1.2. Cytological image analysis in Biology

The IER problem also occurs in Biology where images with cells should be examined and studied. -A known respectable issue and a very good application target-An important step in the diagnosis of a cervical cytology specimen is estimating the proportions of various present cell types. This is usually done with a cell classifier.[Benjamin S. White, Kenneth R. Castleman 06]

1.3. Other Applications

The IER problem could also be applied to other areas as well. For example, in oncology images from X-rays have to be analyzed to identify tumours which is well known that they grow in a spherical manner and may appear overlapping in the image. In astronomy, images that contain a large number of stars could be analyzed in order to enumerate individual stars. In dermatology, where warts appear in the same way as encircling regions. This is very useful for automatic locating, even for further study or for laser therapy.
2. The Proposed Algorithm

The employment of the proposed algorithm is a combinatorial approach for solving the IER problem, where the image will be recursively scanned until all the approximation disks are located thereby identifying all the regions.

The algorithm is composed of three main phases, which will be described in detail in the sections below. The following parameters are defined:

- \( r_{\text{max}}, r_{\text{min}} \): the maximum and minimum radius of any possible indisk respectively in the image \( r_{\text{max}} \geq r_{\text{min}} \)
- \( \tau \): the overlapping factor
- \( \omega \): the irregularity factor
- \( \theta \): the noise

All of the above mentioned parameters are application specific. Given a set of test images for a specific application, a calibration procedure could be employed to evaluate these parameters. For example \( r_{\text{max}} \) and \( r_{\text{min}} \) essentially define the range of possible sizes of regions that are going to be scanned. The overlapping factor \( \tau \) takes values in \([0, 1)\), and is directly proportional to the degree of overlapping that can occur between the shapes in the image. Overlapping means the relative size of the areas in the intersections between the shapes. For instance \( \tau = 0 \) implies that our application commands for no overlapping at all between the objects, while \( \tau = 0.99 \) means that some objects may completely covered by others. The irregularity factor \( \omega \) limits the maximum difference that can occur between the radiuses of the indisk and the outdisk for a given shape. Given an indisk with radius \( r_{\text{D}} \), the radius of the outdisk is bounded by

\[
r_{\text{D}} (1+\omega) \geq r_{\text{D}} \geq r_{\text{D}}
\]

This parameter essentially describes the degree of irregularity in the boundary region of the shape. A very low irregularity factor value means that our shapes are almost circular. Finally, a noise parameter \( \theta \in [0, 1) \) is defined. When evaluating whether or not a set \( T \) of pixels is black in color, if \( |T| \times (1 - \theta) \) pixels are black then the whole set considered as being black.

2.1. The indisks procedure

In the first phase takes place the computation of the indisks for all shapes in the image. This is the most computationally expensive part of the algorithm, since every pixel of the image has to be scanned multiple times, depending on the range of possible radiuses for a given application. The procedure is illustrated in pseudo code in Figure 4. Let \( I \) denote the set of pixels from the original image, \( I \) the set of pixels that are possible candidates for centers of indisks, and \( D(c, r) \) be the set of pixels as defined by the disk with center \( c = (x, y) \) and radius \( r \). The set IDisks contains all the indisks computed at every iteration of the procedure.

The procedure iterates for every possible radius \( r_{\text{max}}, r_{\text{max}} - 1, \ldots, r_{\text{min}} \), and for each iteration it examines every pixel from the set \( I \) whether it is a center of an indisk or not. This is done in line 5 by the procedure IsBlack, where given a pixel \( c \) and a radius \( r \) it examines whether the area defined by the disk \( D(c, r) \) is black in color, taking into consideration the noise parameter \( \theta \). Since it is computationally expensive to check every pixel in the area of the disk, a sampling technique is employed which can be considered as a combination of the annular ring and wedge shape sampling techniques [Sing-Tze Bow 03].
c and radiuses \( r_{\text{min}}, 2r_{\text{min}}, 4r_{\text{min}}, \ldots \) are black in color. The sampling technique is illustrated in Figure 5. Given that an indisk \( D(c, r) \) is found, it is added to the set \( \text{IDisks} \) (line 6).

\[
\begin{align*}
\tau \geq 0.5 & \quad \tau \leq 0.5
\end{align*}
\]

\( \begin{array}{c}
\tau \geq 0.5 \\
\tau \leq 0.5
\end{array} \)

\( \begin{array}{c}
\tau \geq 0.5 \\
\tau \leq 0.5
\end{array} \)

**Figure 6. Overlapping**

Line 7 together with lines 10-13 deal with the overlapping problem. Consider that at some point in the procedure InDisks an indisk \( D_i \) is found with radius \( r_{D_i} \) (Figure 6). If the overlapping factor \( \tau \) is less than or equal to 0.5 (The right diagram in Figure 6), then the area defined by the disk with center that of \( D_i \) and radius

\[
r = r_{D_i} + (r_D - 2\tau r_{D})
\]

which is illustrated as a shaded area in the figure, cannot contain a center of a future indisk \( D \) with radius \( r_{D} \leq r_{D_i} \), since that would violate the allowed degree of overlapping. As the considered radius of possible indisks \( r_D \) decreases, \( r \) decreases and approaches its lower bound \( r_{D_i} \) as \( r_D \) decreases. If the overlapping factor \( \tau \) is greater than or equal to 0.5 (the left diagram in Figure 6) a similar situation exists, where \( r \) is again defined as in (2), but \( r \) increases and approaches its upper bound which is \( r_{D_i} \). Note that when \( \tau = 0.5 \) the indisk itself is the area of forbidden centers of future indisks, and it remains constant throughout the iterations. Since the procedure InDisks searches for the center of an indisk with a given radius \( r \) at every iteration, in lines 7 and 10-13 are computed the areas of forbidden centers with respect to the set of currently computed indisks.

2.2. The outrisks procedure

In the second phase of the algorithm an outrisk is computed for every indisk found in the first phase. This is done by increasing the radius of the corresponding indisk incrementally, and terminating if a fraction, which is a function of \( \theta \), of the pixels in the circumference of the resulting circle is white in color or the radius reaches its upper bound as defined in (1). The procedure is illustrated in Figure 7. The set \( \text{ODisks} \) denotes the set of outrisks computed, while the routine \( \text{count}(c, p) \) returns the fraction of the white pixels to the black pixels in the circumference of the circle with center \( c \) and radius \( p \). In the case where overlapping occurs only portions of the circumference will be examined, and specifically those portions

```
procedure OutDisks(IDisks, \( \theta, \omega \))
1 \( O\text{Disks} := \emptyset; \)
2 do \( D(c, r) \in \text{IDisks} \to \)
3 \( p := r; \)
4 while \( p \leq r(1 + \omega) \text{ or count}(c, p) \leq (1 - \theta) \) do
5 \( p = p + 1; \)
6 endwhile;
7 \( O\text{Disks} := O\text{Disks} \cup D(c, p); \)
8 od;
9 \text{return}(O\text{Disks})
end OutDisks;
```

**Figure 7. The OutDisks procedure**

which do not intersect with the areas defined by the overlapping indisks. The number of black pixels in each computed outrisk can be considered as an approximation to the area of the corresponding shape.

2.3. The approximation disk procedure

In the third and final stage of the algorithm the approximation disks are computed. For each outrisk an approximation disk is drawn for the corresponding shape. Its radius \( r_{D_2} \) is defined by the number of black pixels \( B \) between the indisk and the outrisk, i.e.

\[
r_{D_2} = \sqrt{\frac{\pi r_D^2 + B}{\pi}}
\]

since the indisk is fully contained within the shape. Its center should be in the disk with radius \( r_{D_0} = r_{D_1} \) and the same center as the outrisk, in such a position such that the intersection of the approximation disk and the shape is maximum.

It should be noted that the proposed algorithm could be modified such that the initial parameters \( r_{\text{max}}, r_{\text{min}}, \tau, \theta \), and \( \omega \) could be dynamically computed depending on the input image.

3. Optimization

In the description of the computation of the indisks, it was assumed that every pixel in the image was scanned multiple times. This assumption was made for expository purposes, but in practice this is infeasible considering the large number of pixels involved in images with a high resolution [R.O. Duda, P.E. Hart, and D.G. Stork 04]. Therefore there is a need for further discretization of the digital image which results from an analog image. Consider that there is a total of \( X \times Y \) pixels, where \( X \) is the width and \( Y \) the height of the image respectively, in pixels. Establishing a step size \( \sigma \), such that the original digitized image will be represented, by only those
pixels from the total $X \times Y$ which are at least $\sigma$ pixels apart from each other, horizontally or vertically.

In the discussion that follows, the maximum error that can occur in the radius $R$ of a computed indisk by the procedure InDisks, using a step size $\sigma$ is computed. Consider for instance Figure 8, in which the small black dots represent the original pixels and the larger red dots the pixels considered for a step size $\sigma = 3$. The circle with center $x$ and radius $R$ represents the indisk $D$ of a shape in the original image, while the circle with center $x_c$ and radius $R_c$ represents the computed indisk $D_c$. The disk $D_c$ would be the largest disk contained in $D$, with center one of the pixels contained in $D$. It is easy to see that $R_c$ is bounded by

$$R - \frac{\sqrt{\sigma^2}}{2} \leq R_c \leq R$$

where $\frac{\sqrt{\sigma^2}}{2}$ is the maximum distance that can occur between $x$ and $x_c$. Following the discussion in section 1, the center of the indisk determines the center of the outdisk, while the center of the approximation disk which represents the shape, should be at most $\omega R$ away from $x$. Considering that using a step size $\sigma$ would induce a maximum error of $\frac{\sqrt{\sigma^2}}{2}$ between the centers of the computed and actual indisks, we can conclude that the step size should be bounded by

$$0 \leq \sigma \leq \frac{2\omega R}{\sqrt{2}}$$

Equation (4) essentially states that the maximum step size is directly proportional to the degree of irregularity in the boundary of the shape and the size, in pixels, of the shape. This is also true in practice, since given a constant irregularity factor the step size could be increased to identify larger shapes in contrast with smaller in area shapes, where higher accuracy is required thereby a smaller step size. Consequently the proposed algorithm could start with a high step size as bounded by (4) and decrease at every iteration depending on the radius of indisks that it examines, thereby reducing the total number of iterations considerably.

4. Computational Results

The proposed algorithm has been implemented in Delphi, and the computational experiments have been conducted on an Intel Pentium IV 3.2GHz Prescott processor with 1GB of memory. The main data structure used is a doubly linked list for quick access to the set of permissible indisk centers in each iteration.

4.1. Hail pad image analysis

The test images from the hail pads are taken from the results of the study in [R. Rudolph, C. Giannaris, C. Boufidis, and J. Flueck 05]. These are black and white images with a resolution of 300 dpi. The algorithm executed on these images with the following parameter settings:

- $r_{\text{max}} = 300$ pixels, $r_{\text{min}} = 30$ pixels
- $\tau = 0.2$
- $\theta = 0.1$
- $\omega = 0.2$
- $\sigma = 5$

The maximum and minimum radius are confined as indicated above, since the maximum and minimum diameter hail stone that could occur is 5 cm and 0.5 cm respectively. The overlapping factor $\tau$ was set to 0.2 empirically, since that is the maximum overlapping observed. Similarly for the $\omega$ parameter. The noise parameter is set $\theta = 0.1$ since there was added 7% of artificial uniform noise to the original image. The step size $\sigma$ was set to a constant value of 5, although it could have been set to take values in the range

Figure 8. Discretization

Figure 9. Solution procedure
between \( r_{\text{max}} \) to 8 for \( r_{\text{min}} \). This was done to achieve higher accuracy, since the computational time for a constant step size of 5, was less than a minute for all test instances. In Figure 9 are presented the results of the proposed algorithm for the hail pad image shown in Figure 2. The disks shown in solid light blue color are the indisks computed by the algorithm for the shapes present in the image, while for each indisk the associated large circle is the outdisk and the smaller circle denotes the approximation disk. The indisks for the four shapes present in the image are found in order of decreasing radius, that is A, B, C and D. Overlapping occurs only between the shapes associated with the disks A and B where the algorithm correctly discriminates the two shapes.

Figure 11. Computed Shapes

4.2. Biology image analysis

In order to verify the algorithm’s effectiveness in biology images, some representative instances are obtained. The goal is to detect cells of a noisy image and it’s easier than hail pad analysis for two obvious reasons:

- Cells have almost the same size
- Cell region is almost circular

Consequently, for the first reason, the difference between \( r_{\text{max}} \) and \( r_{\text{min}} \) is small and for the second reason, the omega parameter varies in an area near to zero because the border’s form is mainly normal.

Figure 12. Cells treated with griseofulvin courtesy of National Academy Science
*Figure 12* is an instance with low complexity for the reason that only two cells appear with a small overlap. As well it should be noticed that in this case cells appear with symmetrical and uniform jagged boundary. Supplying the transformed black and white image to the proposed algorithm, regions successfully are calculated (*Figure 13*).

*Figure 13. Calculated regions*

Effectively, the black area in the green rectangle is not recognized as a possible cell since the desired indisk has radius smaller than the \( r_{\text{min}} \). This method is based upon the principal guidelines discussed in section 2.

*Figure 14. Red cells characterized by a central hemoglobin zed area surrounded by pallor giving the cell the classic "target", Mexican hat or "bulls eye" appearance*

Furthermore, the reflection of a much more remarkable image as a test bed is displayed in *Figure 14*. Similarly, as mentioned above, the required preprocessing step produces a black and white bitmap. The notable consideration point is the enclosed white area in the cells. This area is transformed into a black area in the preprocessing step. According to the sampling technique the algorithm is able to identify cells with enclosed white areas (kernels) and analyze deeper the structure of these kind of shapes as a future ability. For example, the algorithm is capable of separating the kernel and the cytoplasm as well as measuring the area of each one.

Observing *Figure 15* with 31 computed indisks (same as cell population) the main impact is that the algorithm establishes the indisks in the best position. The shape with a yellow indisk needs further discussion due to its oval form. The initialization of some significant preliminary parameters is \( \omega = 0.1 \) and \( \tau = 0.1 \). Dramatically increasing the \( \tau \) parameter (more than 50%) leads the algorithm to identify two indisks in this shape. But like human intervention recognizes that this shape contains only one object (one indisk), correctly the algorithm falls into the above consideration. As mentioned below (section 5), as a feature study, a different way for approaching shapes with ellipsis instead of circles is required for cases like this one.

*Figure 15. Computed indisks*

5. Conclusions and Further Research

In this paper an algorithm is presented to identify overlapping shapes of varying size in a digitized image which can be approximated by a disk. The main difficulties encountered were the overlapping requirement and the highly irregular boundary of the regions. The proposed algorithm was tested on a set of test instances from real hail pad and biology images, and it performed well in identifying correctly the shapes. By correct identification it is meant that a human would have probably performed in an equivalent manner.

The following could be investigated furthermore to improve the algorithm.

- Following the discussion in section 3, a variable step size could be employed to decrease the computational effort dramatically. For example for an indisk radius of 200 and \( \omega = 0.1 \) the maximum allowed \( \sigma \) according to (4) is 28, which in comparison with \( \sigma = 5 \) translates to examining 10000 pixels instead of 313600, in
an image with a total of 2800 × 2800 pixels.

- The proposed algorithm may examine a pixel several times but does not utilize this information. A procedure could be developed to use this information and know the color of a pixel in future iterations.
- A generalization of the proposed algorithm could be examined, where the shapes would be approximated using ellipses instead of circles.
- According to the sampling technique it is possible to examine further the cells. For example the discrimination between the kernel and the cytoplasm of a cell.

The IER problem could also be applied to other areas as well. For example in oncology images from X-rays have to be analyzed to identify tumors which are known to be growing in a spherical manner and may appear overlapping in the image. Furthermore, in astronomy, images that contain a large number of stars could be analyzed in order to enumerate individual stars

ACKNOWLEDGMENT

The authors would like to thank Prof. Karakostas from the Department Meteorology and Climatology of the Aristotle University of Thessaloniki, for introducing the hail pad image identification problem as well as for providing the test instance images.

REFERENCES

[03] Sing-Tze Bow, Pattern Recognition and Image Preprocessing, Marcel Dekker, 2002