technique which excels in certain feature extraction tasks because of ease of programming, high speed, and versatility.

REFERENCES


Computer Analysis of Chromosome Patterns: Feature Encoding for Flexible Decision Making

ALLEN KLINGER, MEMBER, IEEE, ARNOLD KOCHMAN, AND NIKITAS ALEXANDRIDIS, STUDENT MEMBER, IEEE

Abstract—Experimental pattern recognition techniques for processing chromosome slides with a computer are described. The purpose of the computer program is twofold: to illuminate the basic mechanisms by which a human recognizes an object, such as a chromosome, and distinguishes it from other entities; and the employment of these mechanisms is an automatic and precise extraction of chromosome features.

Index Terms—Feature extraction, figure properties, hierarchical decisions, integral projections, pattern recognition, picture processing.

INTRODUCTION

The purpose of this study was to find methods of abstracting the characteristics of a chromosome and to develop a dynamic method of encoding this data for feature extraction. The computer program must locate a single entity—chromosome or nonchromosome—then identify as many of its characteristics as possible and employ them to make a decision. If the characteristics noted are sufficient to satisfy the observer that his decision will be correct, he will make a decision. Otherwise, he will investigate alternative ways of looking at the object. On a computer, this takes the form of a decision tree (Fig. 1).

The chromosome is defined and recognized by the presence or absence of "arms" and of their shape and size. For the purpose of the computer program described here, it is necessary to formulate this definition more explicitly. The chromosome recognition program developed by Frey [1] was unable to process many cases without human intervention because it used an overly strict operational definition of "chromosome." The program developed by Neurath et al. [2], has some similarities to our program but adopts a fixed-processing format for each chromosome. The techniques reported here permit the recognition of distinct situations, including overlapping chromosomes; hence the best criterion can be applied to each situation.

The structure of the program permits several relatively complex functions. For example, once a chromosome has
been detected, not only the modified data, but also the sequence of modifications that have occurred, is considered. In this way, when the program makes a decision it can take into account the extent and type of modifications to the data as well as possible alternatives.


Our work presents new methods for data reduction and isolation of visual entities (chromosomes, blobs, touching and overlapping chromosomes) as well as boundary encoding via a scaled polar plot which enables analysis for overlap resolution. In essence we describe an experimental program which exemplifies feature encoding and decision making in hierarchical fashion. Many of the papers cited describe more extensive analysis and computation on actual chromosome data.

**DATA REDUCTION AND ISOLATION OF SEPARATE ENTITIES**

The first phase of the computer program modifies the data representing the picture so that it becomes as small a data base as possible. The scanned picture is input in the form of numbers representing gray levels, or relative darknesses, corresponding to a particular small area of the original photograph. A picture of 1 000 000 points is converted into a different form and reduced to some 10 000 pairs of coordinates. Each set of coordinates in the reduced picture is on the left or right facing edge of a dark area of the picture. In a picture of the type under consideration, most locations are not associated with any information about the objects to be studied. It is the coordinates of the intersections of the boundary of the dark chromosome with each horizontal line slicing it that are of particular interest.

The data are input line by line, each line corresponding to one sweep of the flying spot scanner digitizer across the photograph. Thus, after the first phase of the program, the picture is reduced to a list of points known to be on the boundaries of objects in the picture. Entries in this list start at the upper left corner of the picture; further entries occur as these points of the picture are examined from left to right and from top to bottom, as English is read. This ordering is crucial to the operation of the second phase.

The second phase addresses the problem of isolating particular objects from other objects. A good system for accomplishing this must associate all parts of an object with each other, but it must not join together two or more distinct objects. Two basic approaches were originally considered, the first being that all points found in the first phase would be encoded according to vertical continuity. This involves finding the tops and bottoms of all vertically oriented curves. Objects would then be detected by finding points where these curves are joined together. However, this method was rejected since it requires excessive bookkeeping in order to obtain reliable results.

The method selected utilizes boundary tracing; this tracing is done symbolically using the reduced table. The criterion used for finding the next point on the boundary is that the last point and the next point must be on the same (left or right) sides of vertically adjacent overlapping cross-sectional line segments. Fig. 2 illustrates schematically the gray levels in a hypothetical picture; the solid black is intended to represent the images of objects. In Fig. 3 which is derived from Fig. 2, the ends of the lines are the points whose coordinates would be retained in the list. Table I indicates the corresponding contents of the computer program's internal table; this table contains all the required information about the picture in a compact form. (A minus sign associated with a point of the table indicates that the point was at the left of an object.)

A pointer starts at location 1, symbolically moving down the left side of the object, jumping down the table by two until the first overlap is detected. The first overlap detected while going downward is the left-most on that line. The pointer is set at this new position and the process of searching for overlaps is continued until no overlap is found on the next line of the picture. At this time the lowest point on the leftmost edge has been reached. The pointer is then increased by one and moved up the right side as it was previously moved down the left side.

When the boundary has been traced as completely as possible, the process is repeated, this time moving clockwise instead of counterclockwise. Locations at which the pointer...
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16

Fig. 2. A sample picture with two objects.

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16

Fig. 3. Tracing sequence for boundary of the object in Fig. 1.

### TABLE I

**Contents of the Computer Program's Internal Table Corresponding to the Picture in Fig. 1**

<table>
<thead>
<tr>
<th>Before Boundary Tracing</th>
<th>After Boundary Tracing</th>
<th>Before Boundary Tracing</th>
<th>After Boundary Tracing</th>
</tr>
</thead>
<tbody>
<tr>
<td>( x, y )</td>
<td>( x', y' )</td>
<td>( x, y )</td>
<td>( x', y' )</td>
</tr>
<tr>
<td>(-5, 1)</td>
<td>(-5, 1011)</td>
<td>(-14, 7)</td>
<td>(-14, 2027)</td>
</tr>
<tr>
<td>(-5, 2)</td>
<td>(-5, 1012)</td>
<td>(-5, 8)</td>
<td>(-5, 1018)</td>
</tr>
<tr>
<td>(-5, 4)</td>
<td>(-5, 1012)</td>
<td>(-5, 8)</td>
<td>(-5, 1018)</td>
</tr>
<tr>
<td>(-8, 2)</td>
<td>(-8, 1012)</td>
<td>(-9, 8)</td>
<td>(-9, 1018)</td>
</tr>
<tr>
<td>(-9, 2)</td>
<td>(-9, 1012)</td>
<td>(-10, 8)</td>
<td>(-10, 1018)</td>
</tr>
<tr>
<td>(-5, 3)</td>
<td>(-5, 1013)</td>
<td>(-5, 9)</td>
<td>(-5, 1019)</td>
</tr>
<tr>
<td>(-4, 4)</td>
<td>(-4, 1014)</td>
<td>(-9, 9)</td>
<td>(-9, 1019)</td>
</tr>
<tr>
<td>(-10, 4)</td>
<td>(-10, 1014)</td>
<td>(-10, 10)</td>
<td>(-10, 1019)</td>
</tr>
<tr>
<td>(-6, 5)</td>
<td>(-6, 1015)</td>
<td>(-4, 10)</td>
<td>(-4, 1020)</td>
</tr>
<tr>
<td>(-10, 5)</td>
<td>(-10, 1015)</td>
<td>(-6, 10)</td>
<td>(-6, 1020)</td>
</tr>
<tr>
<td>(-14, 5)</td>
<td>(-14, 2025)</td>
<td>(-9, 10)</td>
<td>(-9, 1020)</td>
</tr>
<tr>
<td>(-15, 5)</td>
<td>(-15, 2025)</td>
<td>(-10, 10)</td>
<td>(-10, 1020)</td>
</tr>
<tr>
<td>(-6, 6)</td>
<td>(-10)</td>
<td>(-4, 11)</td>
<td>(-4, 1021)</td>
</tr>
<tr>
<td>(-10, 6)</td>
<td>(-14)</td>
<td>(-5, 11)</td>
<td>(-5, 1021)</td>
</tr>
<tr>
<td>(-14, 6)</td>
<td>(-14, 2026)</td>
<td>(-9, 11)</td>
<td>(-9, 1021)</td>
</tr>
<tr>
<td>(-15, 6)</td>
<td>(-15, 2026)</td>
<td>(-10, 11)</td>
<td>(-10, 1021)</td>
</tr>
<tr>
<td>(-6, 7)</td>
<td>(-19)</td>
<td>(-4)</td>
<td>(-39)</td>
</tr>
<tr>
<td>(-10, 7)</td>
<td>(-20)</td>
<td>(-10)</td>
<td>(-40)</td>
</tr>
</tbody>
</table>

Some corresponding Fig. 3 tracing sequence numbers are: *1, 63; 23, 61, and 24, 60.

Fig. 4. Points are assigned to an object as the boundary is traced.

---

Some points are signaled by adding a multiple of 1010 to the \( y \) coordinate at this location. The point has now been "assigned" to an object. Points associated with each other by tracing a boundary are signaled by adding the same multiple of 1010 to the \( y \) coordinate. Since the value of the \( Y \) coordinate has not changed modulo 1010, the data has not been disturbed by the procedure which isolates objects. Thus an error or omission can be corrected at a later point in processing. Throughout the procedure the spatial relationships of all points with each other are preserved, even though the representation of the data changes. Therefore, once a point has been found to be part of an object (not necessarily complete), that point can be used in further attempts to recognize complete objects.

When associating a pair of points, they must belong to overlapping segments in two consecutive lines. If the difference in \( y \) coordinates of two segments is greater than some small integer, overlapping is not considered sufficient to connect them. Since both ends of segments are adjacent in the table, overlapping can be easily determined: if one end of a segment can be recognized as lying on the boundary of an object, the other end of the segment must also be part of a boundary of the same object. Recognizing both ends of a line segment as belonging to the object causes internal boundaries to be recognized at the same time that the external boundary is traced (Fig. 4).

Fig. 4 shows how the boundary of an object is traced first in one direction of rotation and then in the other. The only part of the object that is not discovered is the small bump at the top C since neither the right nor the left sides of this part have been traversed. To have done so would have

1 The number chosen was 5. A larger number might cause separate objects to appear joined; a smaller number would be more likely to cause spurious changes in vertical direction because of a noise line.
required the computer program to detect the possibility at $B$ of changing directions as at $A$. A change from downward travel to upward travel must be made at $A$ since it is the lowest point of the object. A subroutine to detect that the direction could be changed at $B$ could be written, but it would be entered at every stage of the boundary tracing. After completing the tracing of the boundary as shown, an attempt could be made to trace a path starting at $C$, retracing the main body. When this is done, only the identification of the object need be changed in order to connect the missing part to the main object.

**Dynamic Encoding of Information for Feature Extraction**

The actual encoding of information is the major task of the computer program. Two techniques are incorporated in the program. These are integral projections [9], [10], or chord profiles [11], and the production of a polar plot of the suspected chromosome. These two methods are not parallel. The integral projection technique is used as a pretest to determine which branches of the tree should be followed. At a later time the polar plot is used for the more explicit decision making.

**Integral Projections**

The third part in row 1 of Fig. 4 shows how the first integral projection from the left of a chromosome is obtained. It will be a plot of the lengths of the leftmost segments of horizontal chords which intersect the chromosome as a function of the vertical position $y$. If the part of the object represented by the first integral projection is removed, the first integral project of what is left is the second integral projection of the original object. Some nonchromosomes—convex blobs—may be eliminated by the presence of only a first integral projection on both the $x$ and $y$ axes. In many cases, the possibility of an object's being a chromosome is eliminated immediately because the number of points which constitute its boundary, henceforth called its size, are either much too small or much too large. (The size is obtained by the computer program during boundary tracing; the examples given here required $60 \leq \text{size} \leq 2000$.)

One of the drawbacks in the use of integral projections is that they change with the orientation of an object. Two identical objects each rotated one from the other would have different integral projections. Integral projections are also very sensitive to error, particularly when an attempt is made to obtain them in the direction perpendicular to the direction of scanning. The data produced by scanning a photograph is susceptible to the least error in the direction of the scan. Any errors produced in scanning have unpredictable effects on integral projections. Isolated points representing noise could radically alter the functional form of the integral projection data. (In Fig. 5, the $E'$s indicate places where problems occurred with the integral projec-
tions.) There are two further reasons, in addition to this noise sensitivity, for ruling out integral projection storage of the chromosome data for quantitative processing.

1) Such encoding makes it unclear to which chromosome arm a given chord is to be associated.

2) Overlapping chromosomes cannot be resolved from integral projection data.

Alternatively, the availability of the scanned picture as a data base larger than the computer core memory, the processing of the picture by reading one line into memory at a time, and the ease with which one projection can be calculated from the boundary data, lead us to employ this mode of picture encoding for the "first test" or qualitative processing.

Polar Plots

For each object thus far identified, the polar plot is made as follows. All boundary points are transformed into polar coordinates with respect to an origin set to the average coordinates of the object's boundary points. A table with 360 spaces (10 rings, 36 sectors) stores entries which indicate whether the object's boundary passes through the corresponding region of the plane. Note that the table $r(\theta)$, is a functional representation of the boundary curve of the object, and that this is not in general single valued. Let the single-valued function $R(\theta)$, the radial extremum, be derived from the table by:

$$R(\theta) = 9 \cdot \max \left( \frac{r(\theta)}{r_{\text{max}}} \right)$$

where $r_{\text{max}}$ is the radius of the point furthest from the center. The derived function $v(\theta)$ in (2) is used for actual decision making and is called the polar plot. (Brackets indicate the greatest integer function.)

$$v(\theta) = \sum_{i=0}^{[R(\theta)]} (i + 1).$$

Gross-shape considerations regarding chromosomes indicate that a point relatively far from the center contains more information than a point near the center, since the closer a point is to the center, the less likely it is to be at the end of a chromosome arm. The effect of the function $v(\theta)$ is to magnify the amplitude of peaks in $R(\theta)$. However, peaks in $R(\theta)$ of almost equal amplitudes will usually have exactly the same amplitude in $v(\theta)$ because of the quantization $v(\theta)$ introduces. Peaks that have a somewhat greater difference in amplitude will have their relative difference in amplitude magnified by the mapping from $R$ into $v$.

Some of the problems that limit the usefulness of the integral projections do not affect the utility of a polar plot. Recognizing a chromosome in this way does not necessarily require that the orientation of the object be taken into account. A rotation is merely a translation of the angular coordinate. Furthermore, use of integral projections (and possibly of other schemes such as chain encoding) would require one to deal with the problem of pattern recognition at a lower level of logic than is absolutely necessary. Integral projection implicitly requires that the results be compared with some possible prototypes, searching for an approximate match. By depending on the very properties that make the image appear to be the image of a chromosome, and by recognizing the obvious property of chromosomes that they tend to consist of arms extending generally outward from a center, it is possible to make correct detections of chromosomes even under adverse circumstances.

Fig. 5 illustrates an object which should be recognizable as an $X$ chromosome, and at the same time it should be possible to recognize it as a less than perfect example. The integral projections of an object such as that in Fig. 4 are radically altered by the presence of a hole which could be ignored (as the result of uneven staining, i.e., optical noise) and by the outcropping on the upper right arm which cannot be ignored (abnormal chromosome possibility). Hence they do not facilitate distinguishing between nonchromosomes and abnormal chromosomes. Too many assumptions about the data are required before the analysis can be made.

However, the polar plots in Figs. 6, 8, and 10 can be utilized in a different way. The number of chromosome arms and the number of peaks in $v(\theta)$ should be the same; furthermore, distinguishing a chromosome from a blob is essentially a problem of recognizing arms. Thus, the polar plot is a device for extracting those features by which chromosomes can be identified.

Figs. 6, 8, and 10 are examples of plots of $v^i(\theta)$. A smoothed version of $v(\theta) = v^0(\theta)$ obtained from

$$v^{i+1}(\theta) = \frac{1}{2} \left[ v^0(\theta_{n-1}) + \frac{v^0(\theta_{n+1}) + v^0(\theta_n)}{2} \right],$$

$$v^0(\theta_n) \leq \max \left[ v^0(\theta_{n-1}), v^0(\theta_{n+1}) \right]$$

otherwise (i.e., if $v^0(\theta_n)$ is found to be the apex of a peak), $v^{i+1}(\theta_n) = v^0(\theta_n)$. This smoothing process tends to make the peaks triangular.

Figs. 6, 8, and 10 are plots of $v^i(\theta)$ for the objects shown in Figs. 5, 7, and 9, respectively. These data have undergone
a transformation which is of a specific form, but which still depends largely on the data itself. The transformation taking \( r(\theta) \) into \( u^0(\theta) \) is defined such that a plot of \( u^0(\theta) \) consists only of straight lines. If, as in Fig. 10, the plot is not commensurable with a pattern of straight lines, the transformation is not correct. Under such circumstances, the procedure must climb to the next node of the encoding tree. In the case of the pattern depicted in Fig. 9 and plotted in Fig. 10, a successful choice of this transformation will occur when the origin of the polar coordinates is translated in the direction indicated by the arrows. Once a suitable \( u(\theta) \) is found, the essence of the plot can be represented by a vector of the relative maxima and minima in the polar plot. This vector, along with a record of the sequence of operations that have been performed in order to obtain \( u^0(\theta) \) constitute all of the encoded information about the object. The nature of the sequence of data modifications indicates the decision criterion, while the vector of the extrema provides the input to the decision apparatus.

**Analysis and Decision Making**

Objects encountered by the decision making apparatus are being evaluated for possible membership in four categories: normal \( X \) chromosome, abnormal \( X \) chromosome, \( Y \) chromosome, and nonchromosome. Abnormal \( Y \) chromosome is not included since it is not yet obvious what will be necessary and sufficient to distinguish an abnormal \( Y \) chromosome from a member of one of the classes already mentioned. In particular, some imperfect \( X \) chromosomes look very much like slightly misshapen \( Y \) chromosomes.

**Classification**

Fig. 7 is an example of what is obviously a complete and probably normal chromosome, although it is not an excellent image. It was selected as an example to demonstrate that the conclusion that the data represents a good \( X \) chromosome can be reached by analysis of the radial plot shown in Fig. 8. Four major peaks can be seen in Fig. 7. The smaller peaks are small enough in comparison to the four large peaks that they can probably be ignored. The significance of these smaller peaks is that, if one were to draw the largest and smoothest outline of an ideal \( X \) or \( Y \) figure (Fig. 7) superimposed on the chromosome that does not enclose the chromosome, the fraction of the total area in the plot constituted by the minor peaks and the closeness with which the major peaks can be matched according to size, can be used as a measure of the level of confidence that may be attributed to the decision. It should be noted that the peaks are usually composed of two distinct areas that can be compared separately. Thus, there are three ways to compare the sizes of these peaks: by their total areas and by the areas of the top and bottom parts.

Fig. 5 shows what is clearly an \( X \) chromosome and it is clearly imperfect. The first and fourth peak areas in Fig. 6
match very well and the second peak is fairly close to the same size. The second peak does not match well with any of the others, except for the bottom part of the third peak. There are many indicators of an $X$ chromosome but not of an ideal $X$ chromosome. Therefore, it is concluded that the object is indeed probably an $X$ chromosome, but probably abnormal.

Fig. 9 is the picture of a nonchromosome which is relatively complex in shape. It might be a chromosome but is eliminated by consideration of the polar plot. Fig. 10 illustrates an instance in which it is clear that none of the peaks compares in size with any of the others. Furthermore, the peaks are not all triangular. Therefore, the object is tentatively found to be a nonchromosome.
Overlap Resolution

The question naturally arises, what does it mean if the radial plot indicates more than four peaks? Such a result could indicate an object of the type shown in Fig. 11. This figure could be interpreted as two \( X \) chromosomes, one on top of the other. The polar plot does not make it immediately obvious that this is the case, but it does indicate that the possibility exists. The object can be considered further, using two different points as new origins of polar axes. This is an attempt to shift attention to one chromosome only, rather than concentrating on both at the same time. The direction from the old origin to a new origin will be in the direction “pointed to” by the largest peaks or pairs of adjacent peaks. This assumes that each of the largest peaks appear so large because a chromosome is displaced from the origin in the direction of that peak. The distance that the origin must be moved in order to place it at the centroid of the chromosome can be approximated from the heights of the peaks, possibly as the difference between the height of the tallest peak and the average height of all peaks.

Once the coordinates have been translated as shown in Fig. 12, the polar plot can be reexamined in a new light. It is assumed that the largest peaks in this profile are not part of the chromosome that we have tried to center. In Fig. 13(a) these peaks have been deleted. Another criterion for deleting these peaks from the polar plot is the fact that they are heavily skewed to the right. A strongly skewed peak indicates an arm emanating from some point other than the center. Size, however, is a first test of peaks to be deleted since a skewed peak can result when an arm of a chromosome is merely bent. (Two examples of such skewed peaks are shown in Fig. 13.) With this first simplification, we now have six peaks left from the original eight.

Of these six remaining peaks, one can be eliminated because it is much smaller than the others. (It should be remembered that the decision is really made on the basis of the function \( v^4(\theta) \), in which the differences in sizes would be greater.) When peak \( 4 \) is deleted, the polar plot has arrived at the state depicted in Fig. 13(b). None of the remaining peaks can be removed because of its size or shape.
If any of the remaining peaks should be removed to find an X chromosome, it must be peak 6, because only in this way could the object have axial symmetry. Once the deletion-generated sharp corners are removed by the smoothing procedure, the problem has been reduced to a situation with which we already know how to deal. If, however, either peaks 5 or 7 were deleted, this would result in the finding of a possibly misshapen chromosome. Thus, of the five possible interpretations, only the one resulting from the deletion of peak 6 reflects a complete and well-defined chromosome image.

**Conclusion**

The experimental pattern recognition program described here employs several feature extraction techniques to enable computer analysis of chromosome patterns. The program performs data reduction to isolate objects from the scene in an efficient manner based on the line at a time input of scanned picture data from magnetic tape, and retaining the useful boundary information in this process. One test—a convexity indication eliminating blobs—is based on the easily computed integral projections/chord profiles used elsewhere in pattern recognition by Pavlidis and Ball. Boundary data is encoded and features obtained (peaks in a polar plot of the boundary indicating chromosome arms) as the program continues past this test through its decision tree. Features are summarized in a pattern vector which consists of the relative maxima and minima in the final polar plot of the chromosome, while the record of translations and deformation which led to this plot is retained.

Methods for using the data to classify the chromosomes (normal X chromosome, abnormal X chromosome, Y chromosome, and nonchromosome) are described which could easily be incorporated in the existing program, to make it an operational program. Finally, a method for resolving cases of overlapping chromosomes by selecting subsets of the polar plot peaks was presented, as was a method for assigning a confidence level to each decision.

The key concept presented here is a two-level logical structure of the pattern recognition program which allows it to select the procedures and criteria most likely to lead to a correct decision. In particular, the program uses a changing organization of the features of the object discovered as the decision-making task changes. The resulting pattern recognition concept we term dynamic encoding for feature extraction, while the overall task this program is intended to perform is that of flexible decision making needed for picture processing.

**References**


