COMPUTERS AND BIOMEDICAL RESEARCH 12, 615–630 (1979)

Preprocessing of Electron Micrographs of Nucleic Acid Molecules for Automatic Analysis by Computer. II. Noise Removal and Gap Filling

P. LEMKIN*, B. SHAPIRO*, L. LIPKIN*, J. MAIZEL†, J. SKLANSKY‡, AND M. SCHULTZ*

*Image Processing Unit, Division of Cancer Biology and Diagnosis, National Cancer Institute, †National Institute of Child Health and Human Development, National Institutes of Health, Bethesda, Maryland 20014; and ‡School of Engineering, University of California, Irvine, Irvine, California 92717

Received February 16, 1979

A technique is proposed for computer preprocessing of digitized electron micrographs of nucleic acid strands to facilitate their automatic segmentation and subsequent analysis. This technique, applied after high pass notch filtering the image, removes almost all spurious artifactual objects in the background. This enables the effective use of segmentation and gap filling operations that otherwise could not previously have been applied due to the combinatorics of the computations.

1. INTRODUCTION

We have previously reported (*1*) on the use of a high pass notch filter as a preprocessor for electron micrographic (EM) images of metalized nucleic acid molecules. The preprocessing was a necessary prerequisite to image segmentation, which in turn was required before image-based data could be available for the automatic computation of a secondary structure map (*2–6*). For a wide variety of images and electron microscopic conditions, notch filtering is not a sufficient preprocessor. In particular, the process of metal deposition may produce large numbers of small (with respect to nucleic acid molecular size) granular artifacts in the background. This results from the metal vapor condensing on the supporting substrate, and these deposits may and do produce image density values which approach those of metal on nucleic acid. This granularity is different (larger by at least two orders of magnitude in size) from the granularity inherent in the developed emulsion.

Removal of the background granular artifacts is still insufficient for subsequent automatic processing to the level of secondary structure maps. The nucleic acid strand images do not necessarily present a uniform gray-level density along their entire length. One frequently encounters local regions within the strand which are lighter than adjacent portions. Indeed, there may be an actual gap between portions.

0010-4809/79/060615-16$02.00/0

The algorithm in this paper removes most of the background artifacts. This makes it practical to apply gap filling tests and procedures to the relatively few remaining objects which are much more likely to be molecular strands. Essentially the procedure relatively "enhances" the image by preserving potential nucleic acid objects and by removing small background noise objects. The gap filling procedures are computationally quite expensive (in time and memory). It is therefore necessary that most of the artifacts in the background be removed, so that a combinatorially and computationally excessive burden may be avoided.

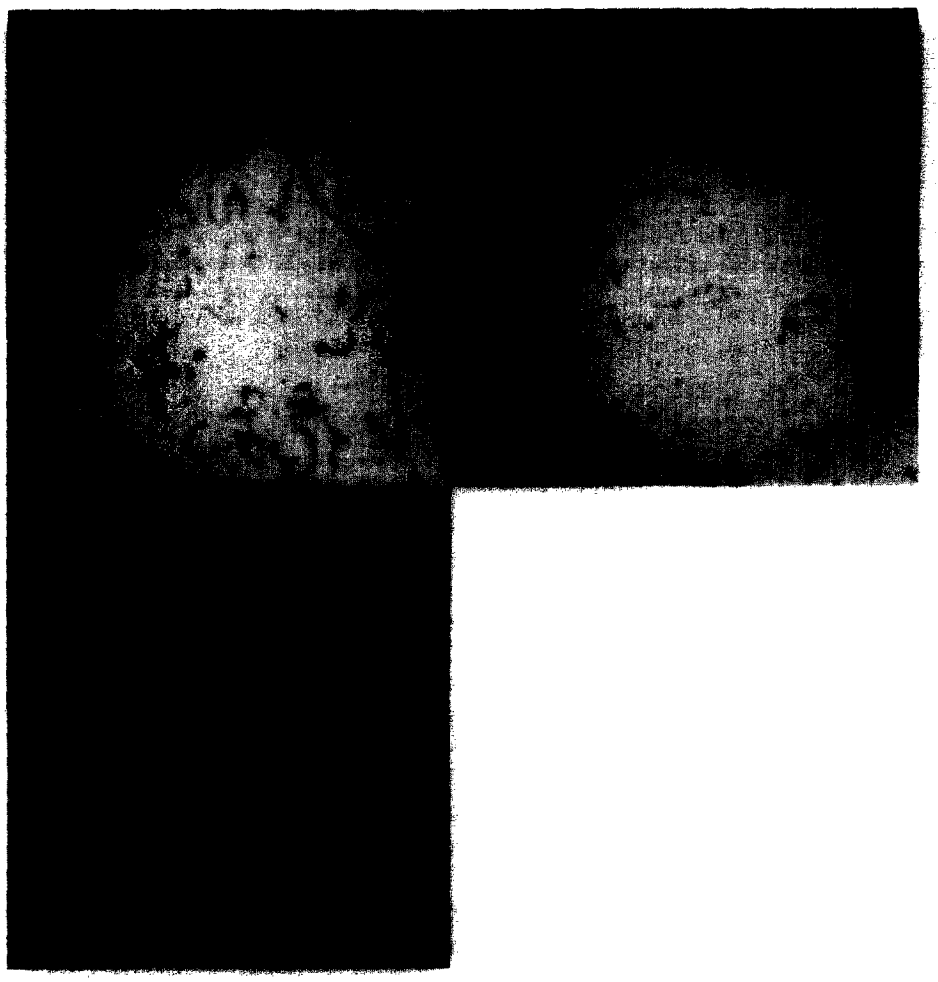

FIG. 1. Original images. Adenovirus type 2 (a) ribosomal RNA image from EM micrograph negative 006323, (b) m-RNA image from EM micrograph negative 006361. (c) SV40 virus DNA molecule from EM micrograph. Notch filtered images of (d) ribosomal RNA, (e) m-RNA, (f) SV40 viral DNA.

The notch filter technique was demonstrated in (*1*) on both ribosomal and m-RNA of adenovirus type 2 RNA. These images as well as a sample of SV40 viral DNA are used to illustrate the background noise removal algorithms in this paper.

2. Materials and Methods *012*

Specimen preparation

The EM micrograph negatives of the RNA molecules were obtained using a Philips 300 electron microscope at 10 000×, with Kodak 4489 film, developed with D76 for 3 min. Adenovirus type 2 m-RNA prepared as previously described (*7*) was

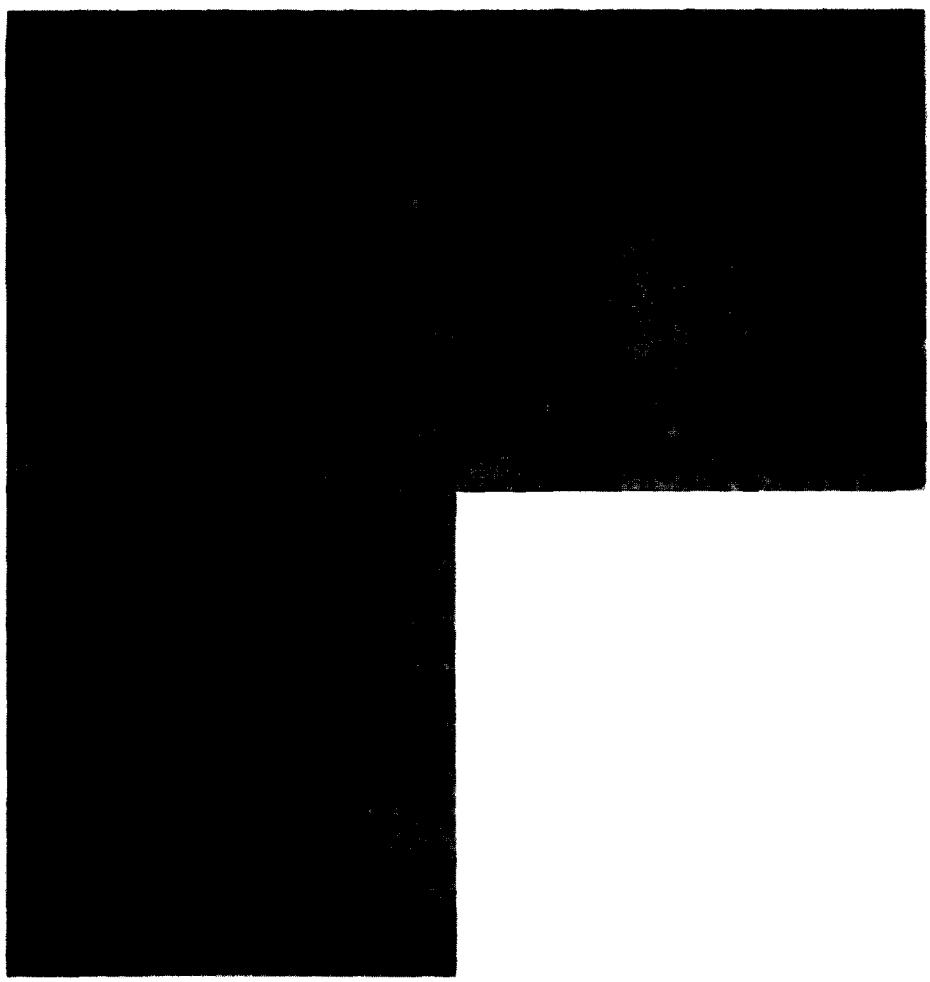

FIG. 1—*continued*

a gift from Dr. H. Westphal. Ribosomal RNA from rabbit reticulocytes was prepared as previously described for HeLa ribosonal RNA (*8*) was a gift from M. L. Stewart. The RNA samples were prepared with 70% formamide solution, 0.01 *M* TRIS (hydroxymethlaminomethane), and 0.001 *M* EDTA onto distilled water. SV40 viral DNA prepared and cleaved with restriction enzyme ECO as described (*9*) was a gift from Dr. N. P. Salzman in (*10*). The DNA was denatured (*4*), spread for EM and photographed at 7290× magnification on 5302 fine grain release positive 35mm film, developed with D76 for 3 min. Three such EM micrographs (positives) are shown in Figs. 1a–c. The ribosonal RNA image in Fig. 1a has a lower overall shading error than the m-RNA image in Fig. 1b or the SV40 in Fig. 1c. The initial data acquisition and gray-scale reversal is identical to that described in (*1*) using the BMON2 image processing system (*11*, *12*) on the real time picture processor, RTPP, (*13–15*).

The digital notch filter, discussed in (*1*, *16*) is a linear transformation that removes low spatial frequencies from the image. Briefly, the procedure is as follows: an $n \times n$ pixel sampling window is moved through the image and its average is subtracted from the center point of the window for each point in the image. For EM micrographs in the present study, the strands are narrow—only a few pixels wide. Consequently, we chose $n = 32$. When applied to digitized nucleic acid images, the notch filter removed most of the low spatial frequencies corresponding to image shading. The notch filtered images are shown in Figs. 1d–f.

Background noise object removal

Under the conditions of metal deposition, and at the magnifications employed, most of the small artifacts in the background are on the order of 3 to 10 pixels in diameter. The molecules of interest, on the other hand, are on the order of 30 pixels or greater in extent. The molecules are elongated while the coarse granules are more or less round and considerably smaller in area.

The tactic adopted here for removing these granules from the EM image depends on labeling each background noise object. Subsequent subtraction of each granule from the image levels—in the best case—an image containing only nucleic acid strands against a blank background. The worse case is an image which is considerably richer in nucleic acid strands but still contains a few artifacts in the background. As will be discussed below, procedures dependent on sequential analysis of each object and labeling background noise on the basis of size and shape or combinations of image properties are too slow and memory consuming to be used for any except the most exploratory processing.

The globally applied shrink procedure developed as part of the background noise removal algorithm, combines in a sense the identification, decision, and labeling processes into one. In principle, if an object can be reduced to a single isolated point after a small number (about 10 or less) shrink passes, it is a background object, since it would take many more such passes to reduce a molecular strand to a single point.

This is the case even through background noise objects and strands may be of the same "thickness" and density. Various algorithms are available for thinning binary images (*17, 18*).

Before blob removal, some preliminary thresholding is necessary, i.e., the notch filtered image is sliced at a gray-scale threshold such that gaps in molecules are minimized. This lower bound on the threshold is determined by the maximum allowable proportion of background noise fragments expressed as a percentage of total image area. If the threshold is too high, it results in fragmentation of the nucleic acid strands. If the threshold is too low it increases the amount and size of the

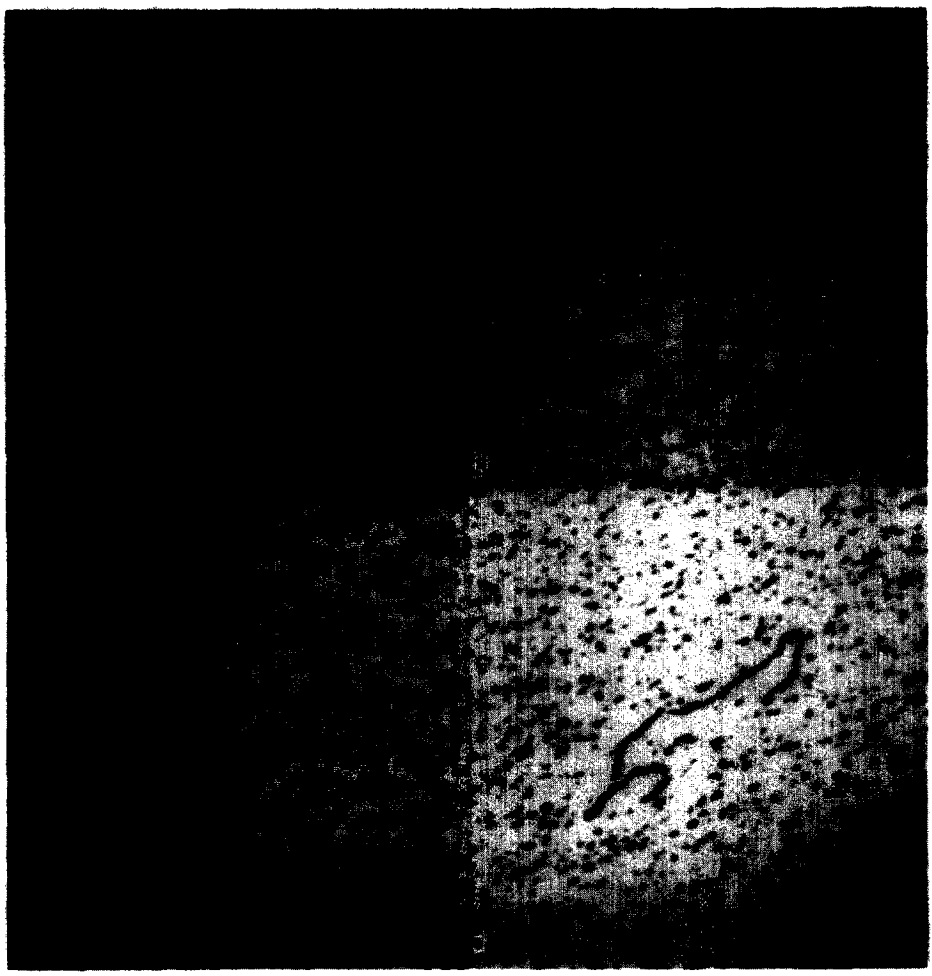

FIG. 2. Notch filtered SV40 viral DNA image. (a) Threshold sliced at [T_{max}: 255], (b) threshold sliced at [75th percentile: 255], (c) threshold slices at [80th percentile: 255], (d) threshold sliced at [90th percentile: 255].

background objects as is seen in Fig. 2. Figure 2b shows thresholding at the 75th percentile of gray-level distribution. Although this threshold preserves the integrity of the DNA molecule (i.e., does not generate a spurious gap), the background fragments are too large. Thresholding at the 90th percentile (Fig. 2d) reduces the background fragments to a reasonable size, but now has resulted in many gaps. Figure 2c thresholding at the 80th percentile gives a reasonable compromise between too many background objects and too many gaps. The actual percentile value to be used for a set of images will depend on the type and amount of noise as well as magnification.

The background noise blob removal algorithm is given below.

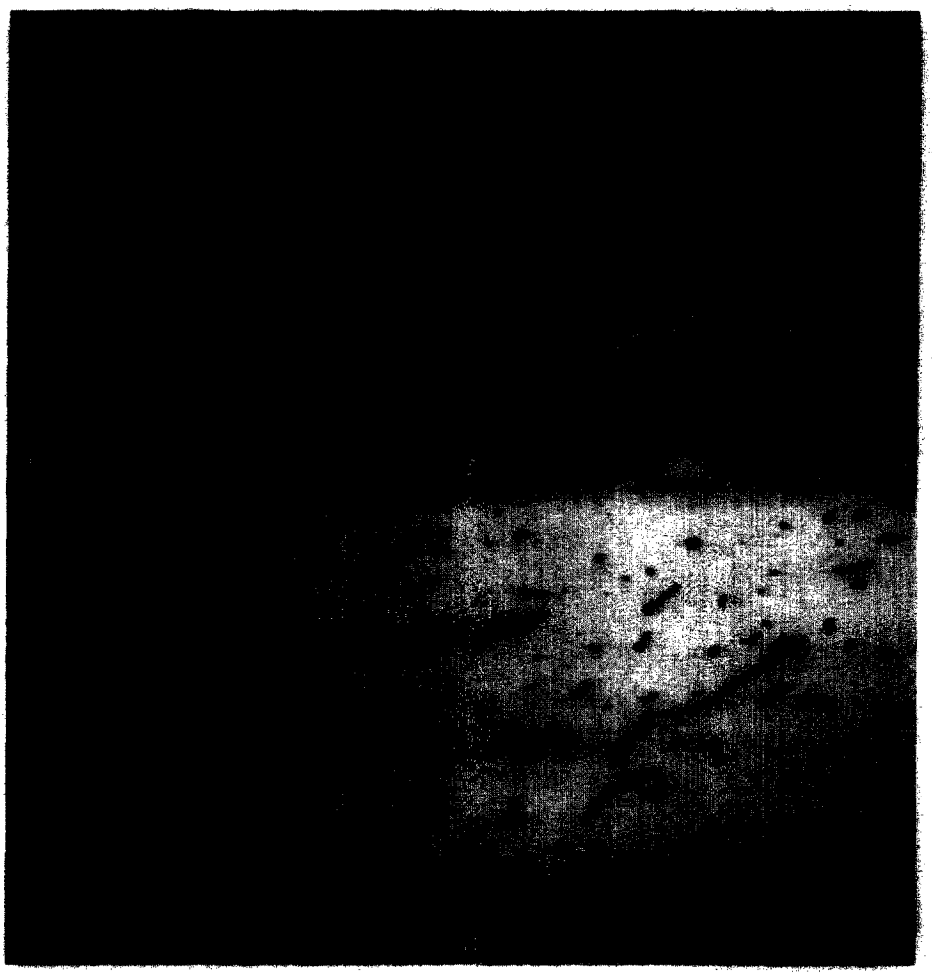

FIG. 3. Binary image of SV40 viral DNA slices at 80th percentile of gray scale. (a) Before shrinking, (b) after shrinking 1 iteration, (c) after shrinking 4 iterations, (d) after shrinking 10 iterations.

Algorithm

[1] Notch filter (see (*1*)) the input image into image *Ii*.

[2] Generate a binary image *Ij* and its copy *Ij'* from the notch filtered image *Ii*. This is done by thresholding *Ii* at the 80th percentile of its gray-value histogram.

[3] Shrink *Ij* in *r* passes, preserving connectivity and remembering isolated points. This is carried out as follows. For each pass test each (x,y) neighborhood to determine whether:
(a) It contains an isolated pixel in which case set $Ij'[x,y]$ to 0 and its (x,y) position added to the list of isolated points.

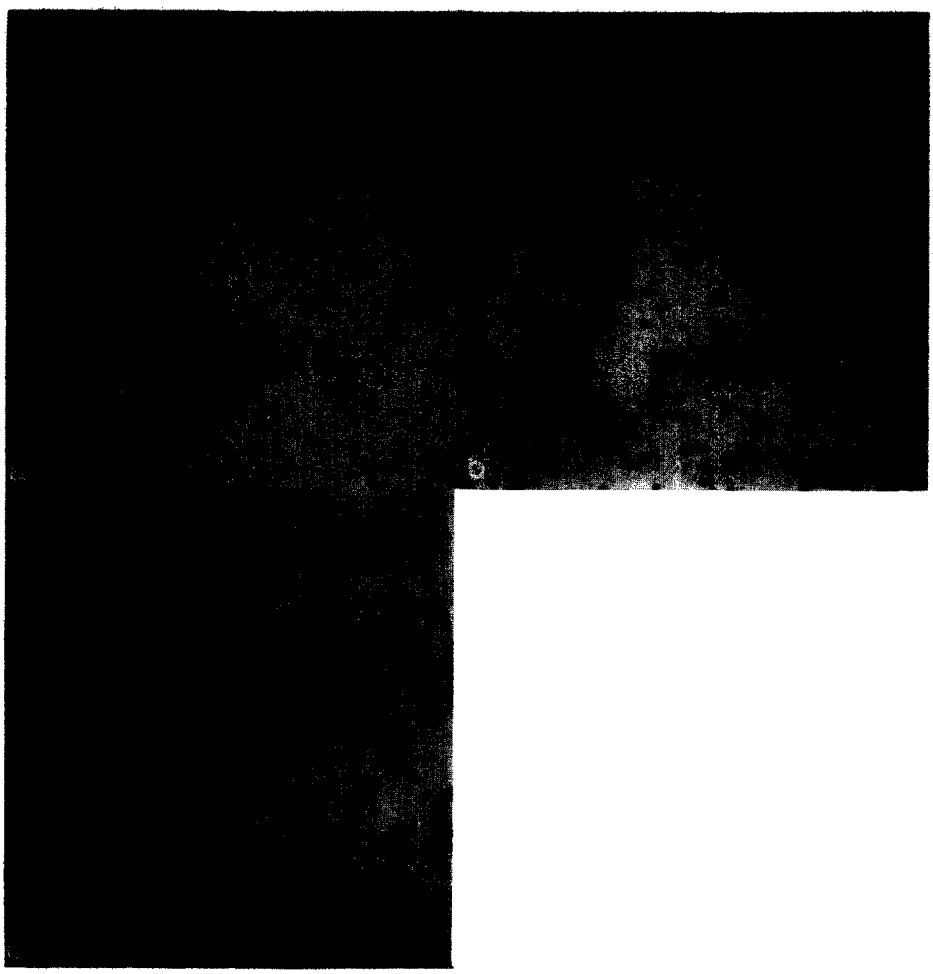

FIG. 4. Noise cleaning for ribosomal RNA (using 10 shrink iterations). (a) Pixels of objects shrunk to isolated points, (b) background noise mask, (c) notch filtered image less the background noise mask.

and

(b) any of the 28 finite state acceptors ($c = [1:28]$) ACCEPT (c,x,y) (see below) holds, in which case $Ij'[x,y]$ is set to 0.

At the end of each pass copy Ij' into Ij.

[4] Copy image Ii into image Ij, then label image Ij' with the value 255 at all isolated points in the list previously generated.

[5] Propagate the image Ij 255 values which are 8-neighbor connected to 1's until all are changed to 255's.

[6] Finally, extract a mask of pixels with 255 values from Ij. These regions

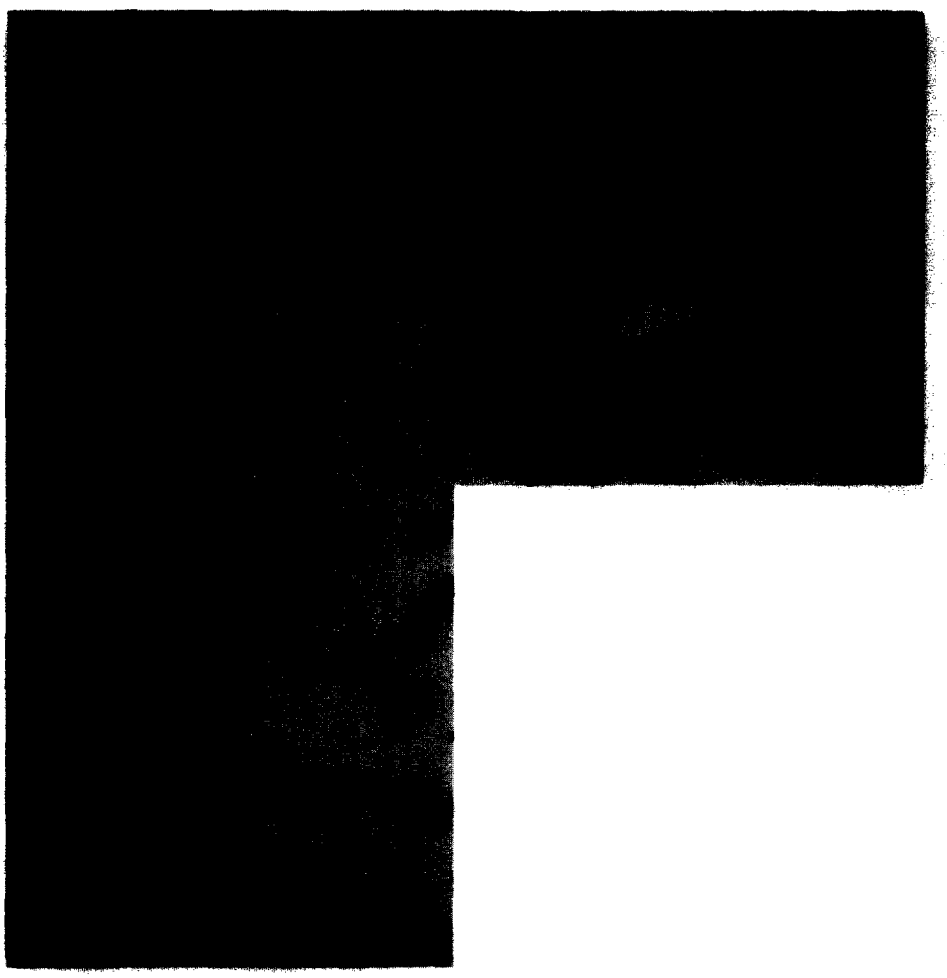

FIG. 5. Noise cleaning for m-RNA (using 10 shrink iterations). (a) Pixels of objects shrunk to isolated points, (b) background noise mask, (c) notch filtered image less the background noise mask.

correspond to the noise objects deleted. It may then be used to remove noise blobs from *Ii*.

This shrink algorithm interatively removes boundary points of objects, while preserving connectivity in the sense of 8-neighborhood adjacency. On encountering or producing an isolated point, it places its coordinates on a list of objects for subsequent deletion. In order to arrive at such a list, the entire image must be analyzed in terms of 3×3 neighbors in raster mode. It is most convenient to construct all possible patterns of 3×3 neighborhoods for template matching in this analysis.

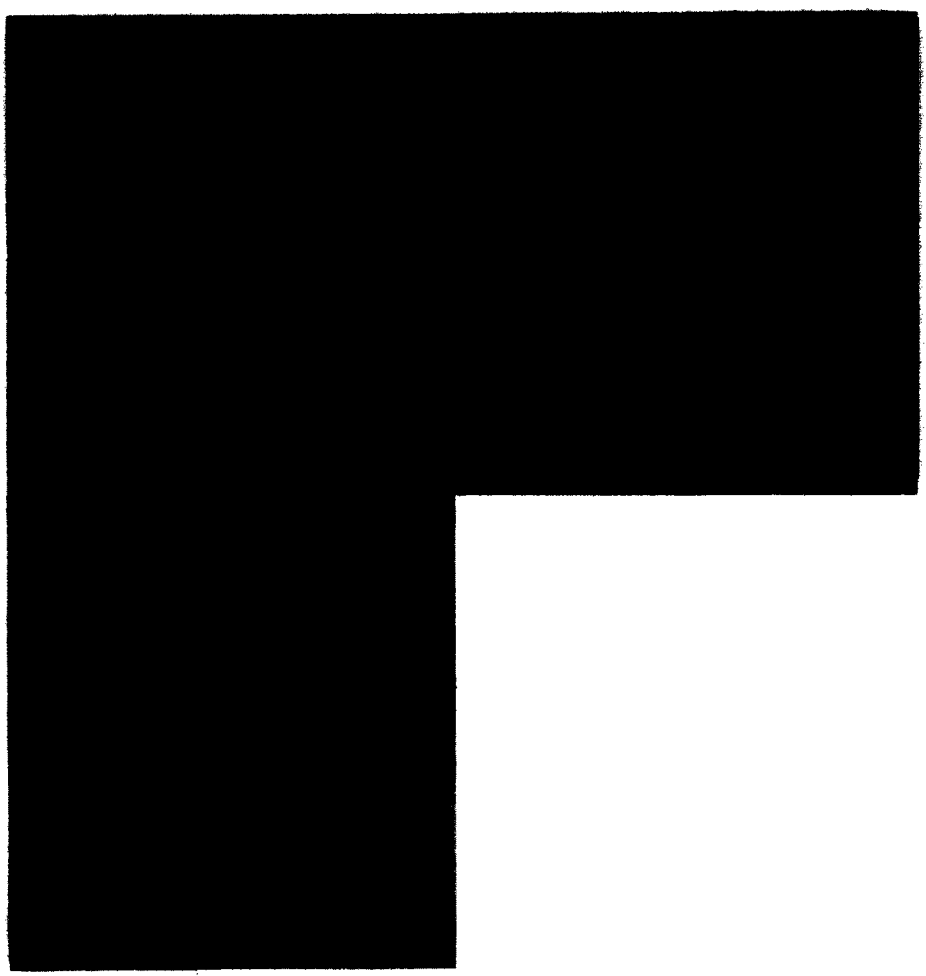

FIG. 6. Noise cleaning for SV40 viral DNA (using 10 shrink iterations). (a) Pixels of objects shrunk to isolated points, (b) background noise mask, (c) notch filtered image less the background noise mask.

There are 28 such patterns for which the central pixel is a 1, and which will inhibit breaks in connectivity. The patterns along with the procedures form what may be termed finite state acceptors (FSA) each of which constitutes a state of a finite state machine. If one of the FSAs accepts the neighborhood, the central pixel is changed to a 0. The following 28 finite state acceptors, {FSA(1), FSA(2), . . ., FSA(28)}, are given in Table I. One notes that the first FSA group (a) has only four permutations compared to eight permutations for the other three (groups b, c, d). This is due to the fact that if all eight permutations of group a are used, then the connectivity of neighborhoods is not preserved.

TABLE I

FINITE STATE ACCEPTORS

a. Face erode

Code:	707	761	175	437
	011	111	110	000
	011	111	110	111
	011	000	110	111

b. Corner erode

Code:	407	603	701	341	161	071	035	017
	000	001	011	111	110	100	000	000
	011	011	011	010	110	110	110	010
	011	001	000	000	000	100	110	111

c, Triangle erode

Code:	007	403	601	301	141	161	031	015
	000	000	001	011	110	100	000	000
	010	011	011	010	010	110	110	010
	011	001	000	000	000	000	100	110

d. Hair erode

Code:	401	201	101	041	021	011	005	003
	000	001	010	100	000	000	000	000
	011	010	010	010	110	010	010	010
	000	000	000	000	000	100	010	001

[a] Twenty-eight finite state acceptors, {FSA(1), FSA(2), . . ., FSA(28)}, for shrinking. If the FSA is true, then the central 1 is changed to a 0. Only central 1's may be changed to 0's. The set of 28 binary 3 × 3 neighborhood templates is divided into four groups a, b, c, and d. Group a has four 90° permutations while groups b, c, and d have all eight 45° permutations.

The finite state acceptor procedure uses an efficient encoding scheme of a neighborhood configuration c into a number. Thus, testing to see if a state is accepted is done by simply matching its code against that of the code of the input data. The data of a 3 × 3 neighborhood is encoded into a 9-bit binary number using

the following neighborhood nomenclature:

$$\begin{matrix} i3 & i2 & i1 \\ i4 & i8 & i0 \\ i5 & i6 & 77 \end{matrix}$$

Thus the resulting binary number is the concatenation of the 9 bits of the neighborhood in the following order and is called the CODE of a given binary neighborhood

$$i0\ i1\ i2\ i3\ i4\ i5\ i6\ i7\ i8.$$

For example, the neighborhood

000
011
011

is encoded as

100000111 binary,

and has the code

407 octal.

The procedure ACCEPT(c,x,y) accepts a neighborhood at postition (x,y) for an acceptor c if the following Boolean expression is true:

$$(\text{CODE}(Ij,x,y) = \text{FSA}(c)) \text{ And } (\text{Not ISOLATED}(Ij',x,y)).$$

ISOLATED(.) is a Boolean procedure to test for isolated central points in a 3×3 neighborhood which is true if the test is met. The restriction (Not ISOLATED(Ij',x,y)) prevents missing objects greater than one pixel (in image Ij) which disappear in a single pass (from the resultant image computed to that point Ij').

3. Results and Discussion

Figure 3a shows the binary image produced from step [2] of the algorithm. Figure 3b shows the results of applying step [3] shrinking one iteration; Fig. 3c—four interations; Fig. 3d—ten iterations. The background noise starts to disappear immediately with isolated pixels being removed after one iteration shown in Fig. 3b. Table II lists the number of isolated pixels found during each of the 10 shrink passes for each of the three images. Both visually and from this table it can be inferred that much of the noise is removed by the end of 4 iterations and most by 10 iterations.

The subsequent illustrations show the results of using 10 iterations. Figures 4a, 5a, and 6a show all of the isolated points produced from step [4] of the algorithm.

TABLE II

ISOLATED PIXELS FORMED DURING SHRINKING[a]

Pass number	Ribosomal RNA isolated pixels	m-RNA isolated pixels	SV40 viral DNA isolated pixels
1	437	678	258
2	247	442	279
3	129	186	253
4	47	102	130
5	24	77	95
6	10	33	63
7	10	37	54
8	6	19	40
9	4	10	25
10	4	7	13

[a] The number of isolated pixels formed during each of the 10 successive shrink iterations (passes) of the blob removal algorithm is given for each of the three nucleic acid images.

Figures 4b, 5b, and 6b show the results of propagating Ij in the algorithm step [5] (i.e., the background noise mask); Figs 4c, 5c, and 6c show the final cleaned up image resulting from algorithm step [6] which masks out the background noise blobs from the notch filtered image. As can be seen from Figs. 4c, 5c, and 6c, several gaps are present.

Size distribution of background noise

Occasionally, the size distribution of the background noise objects show a distinct multimodal distribution. The local peaks are in the neighborhoods of 1–10 and 10–20 pixels extent (greatest dimension). When the peaks appear, they are quite separate. Their possible relationship to metal–substrate interaction is discussed below.

Gap filling

Given a preprocessed image, it is now feasible to attempt the isolation and processing of nucleic acid strands. It is necessary however to repair gaps in the strand, both those originally present and those introduced as a result of the thresholding operation.

There exist several classes of tactics for gap repair. The procedures are all more or less boundary driven and some consider all objects two at a time. They do differ markedly in computational complexity depending on the kinds and weight of context dependency. Some possibilities follow:

a. Gradient tracking to cross the gaps during boundary following.

b. Joining nearest molecules if a (distance/density) heuristic is not too great after all molecule fragments have been segmented.

c. Template matching gap filling filter using models of how gaps appear in the scene prior to segmentation.

d. Heuristic "smart" boundary follower (similar to that proposed in (*19*)).

e. Expanding cleaned up image by 1 pixel, then shrinking it 1 pixel (*20*).

This last method (e) was used here, although methods a–d hold more promise for a universal solution. Figures 7a, 8a, and 9a show the noise cleaned images while 7b,

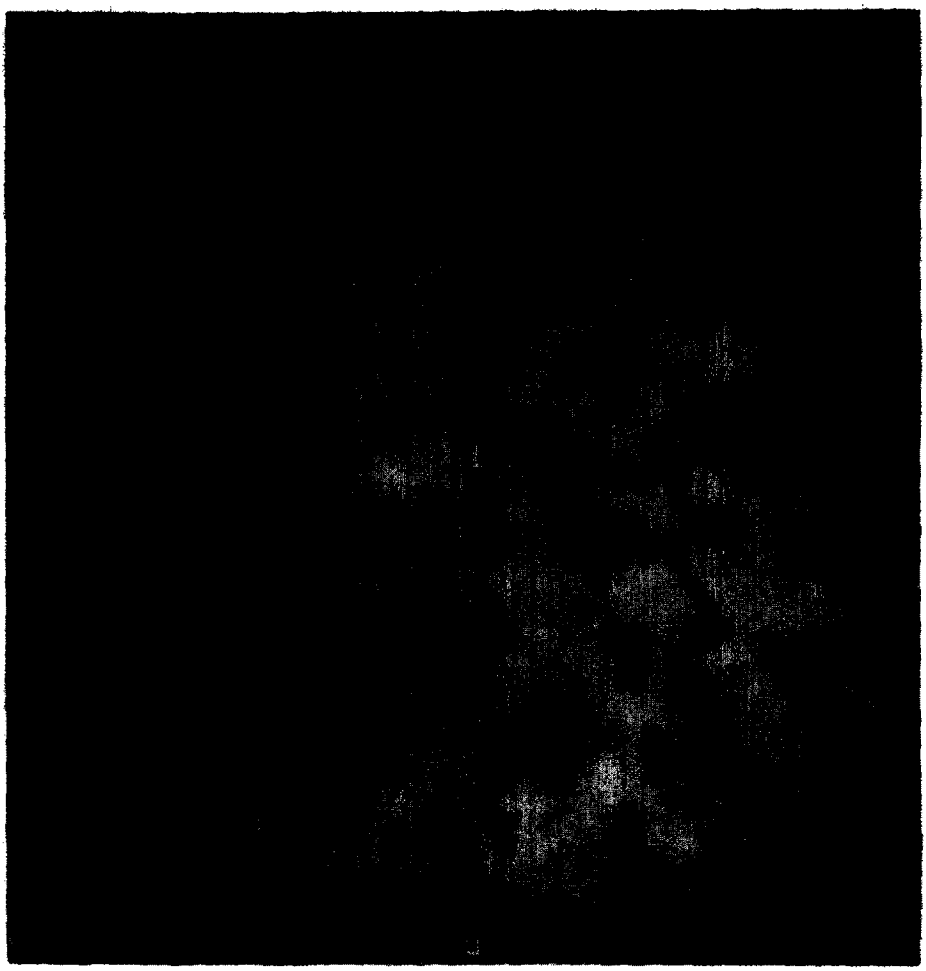

FIG. 7. Gap filling for ribosomal RNA. (a) Cleaned image, (b) cleaned image expanded 1 pixel, (c) expanded cleaned image now shrunk 1 pixel, (d) boundary trace segmentation applied to (c) using minimum perimeter sizing to eliminate fragments and remaining noise objects.

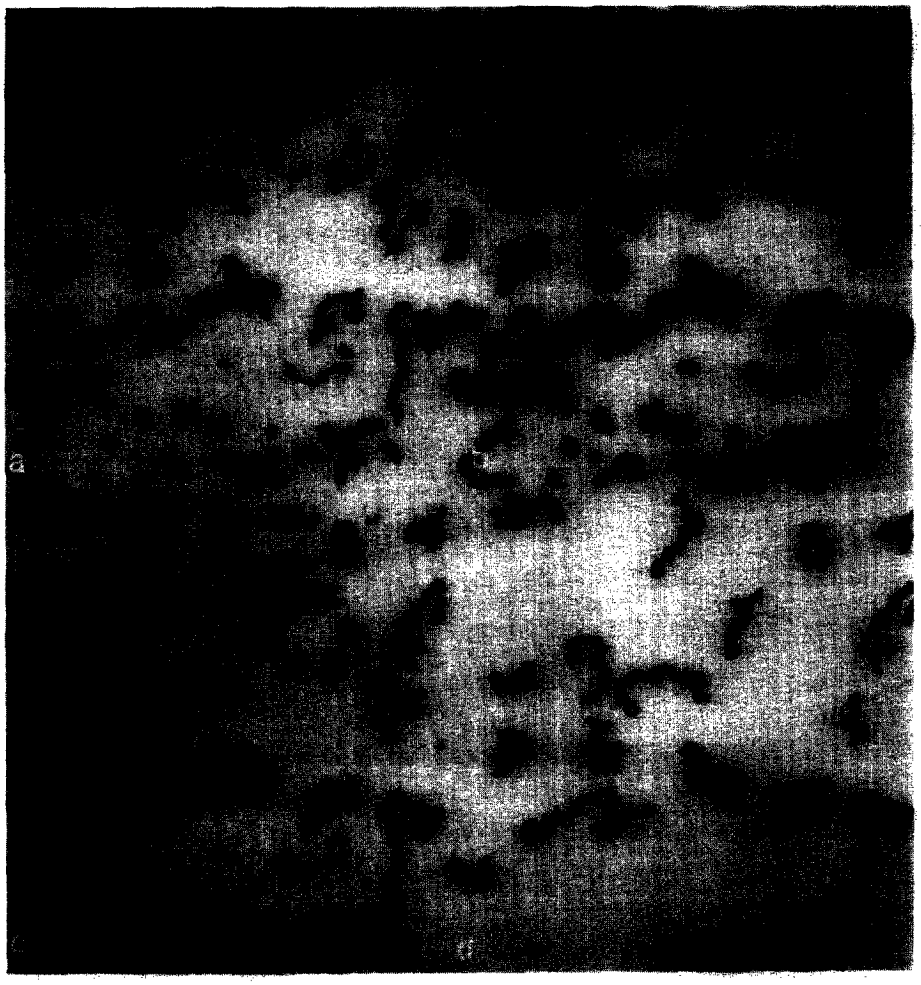

FIG. 8. Gap filling for m-RNA. (a) Cleaned image. (b) cleaned image expanded 1 pixel. (c) expanded cleaned image now shrunk 1 pixel. (d) boundary trace segmentation applied to (c) using minimum perimeter sizing to eliminate fragments and remaining noise objects.

8b, and 9b show the cleaned images expanded by 1 pixel. Figures 7c, 8c, and 9c show them subsequently shrunk by 1 pixel. The gaps are acceptably filled on the high-resolution image Fig. 9c but the lower-resolution images had parts of the molecules merge together. Figures 7d, 8d, and 9d show an automatic boundary follower segmentation applied to the gap-filled cleaned images from Figs. 7c, 8c, and 9c.

A solution to the problem of removing background noise from nucleic acid electron micrographs in order to facilitate their subsequent automatic segmentation

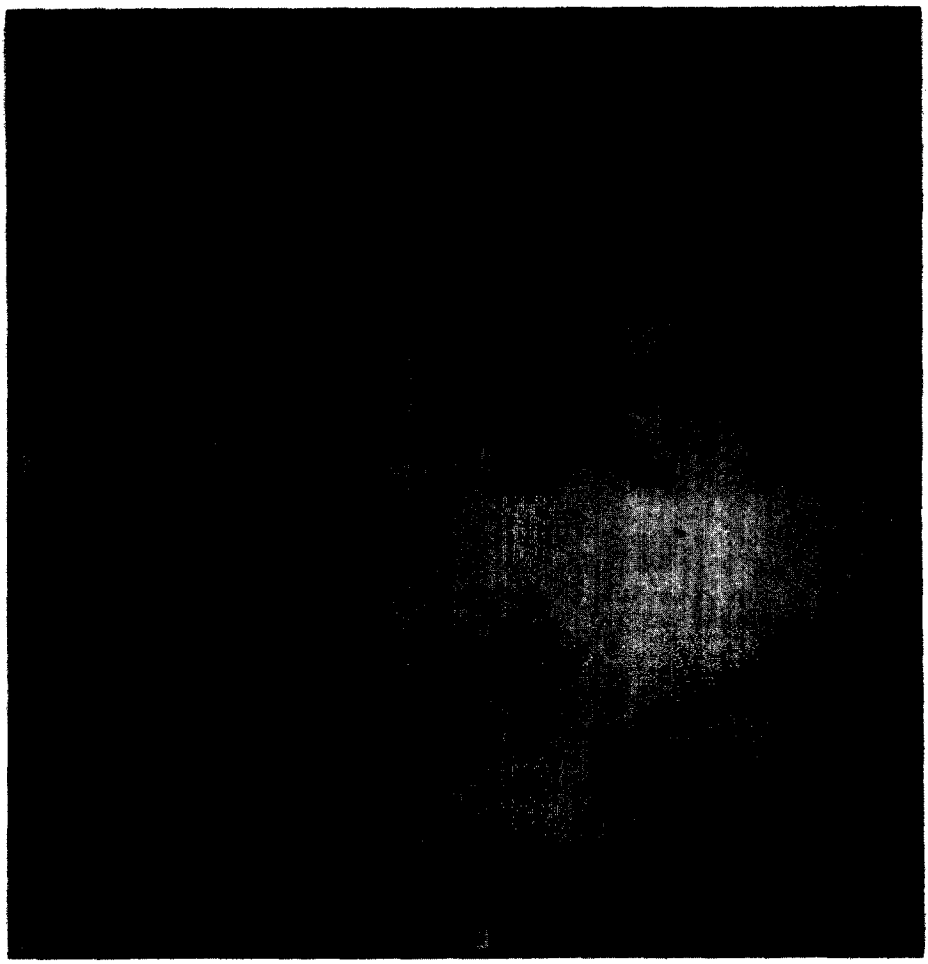

FIG. 9. Gap filling for SV40 viral DNA. (a) Cleaned image, (b) cleaned image expanded 1 pixel, (c) expanded cleaned image now shrunk 1 pixel, (d) boundary trace segmentation applied to (c) using minimum perimeter sizing to eliminate fragments and remaining noise objects.

(using sophisticated gap filling algorithms) has been proposed. The noise removal algorithm relies on the fact that background noise artifacts are disjoint and of a marked smaller size than the nucleic acid molecules of interest. The algorithm seems to work fairly well over a range of EM nucleic acid materials and magnifications.

REFERENCES

1. LIPKIN, L., LEMKIN, P., SHAPIRO, B., AND SLANSKY, J. Preprocessing of electron micrographs of nucleic acid molecules for automatic analysis by computer. *Comput. Biomed. Res.* **12**, 279 (1979).

2. WELLAUER, P. K. AND DAVID, I. B. Secondary structure maps of ribosomal RNA and DNA. I. Processing of *Xenopus laevis* ribosomal RNA and structure of single-stranded ribosomal DNA. *J. Molec. Biol.* **89,** 379 (1974).
3. WELLAUER, P. K., DAVID, I. B., KELLEY, D. E., AND PERRY, R. P. Secondary structure maps of ribosomal RNA and DNA. II. Processing of L-cell ribosomal RNA and variations in the processing pathway. *J. Molec. Biol.* **89,** 397 (1974).
4. SHEN, C-K. J. AND HEARST, J. E. Mapping of sequences of 2-fold symmetry on the simian virus 40 genome: A photochemical crosslinking approach. *Proc. Nat. Acad. Sci. USA* **74,** 1363 (1977).
5. HSU, A. AND JELINEK, W. R. Mapping of inverted repeated DNA sequences within the genome of simian virus 40. *Proc. Nat. Acad. Sci. USA* **74,** 1631 (1977).
6. SHAPIRO, B. "Shape Description Using Boundary Sequences," Ph.D. dissertation, University of Maryland, 1978.
7. ERON, L. AND WESTPHAL, H. Cell-free translation of highly purified adenovirus messenger RNA. *Proc. Nat. Acad. Sci. USA* **71,** 3385 (1974).
8. WELLAUER, P. K. AND DAVID, I. B. Secondary structure maps of RNA: Processing of Hela ribosomal RNA. *Proc. Nat. Acad. Sci. USA* **70,** 2827 (1974).
9. FARRED, G. C., GARON, C. F., AND SALZMAN, N. P. Origin and direction of simian virus 40 dexoyribonucleic acid replication. *J. Virol.* **10,** 484 (1972).
10. LEVOWITZ, J., GARON, G. G., CHEN, M. C. T., AND SALZMAN, N. P. Chemical modification of simian virus 40 DNA by reaction with a water soluble carbodiimide. *J. Virol.* **18,** 205 (1976).
11. LEMKIN, P. "Buffer Memory Monitor System for Interactive Image Processing." NCI/IP Technical Report 21b, Nat. Tech. Info. Serv. PB278789 (listing PB278790), April, 1978.
12. LEMKIN, P. AND LIPKIN, L. BMON2—A distributed monitor system for biological image processing. Submitted.
13. CARMAN, G., LEMKIN, P., LIPKIN, L., SHAPIRO, B., SCHULTZ, M., AND KAISER, P. A real time picture processor for use in biological cell identification. II. Hardware implementation. *J. Histochem. Cytochem.* **22,** 732 (1974).
14. LEMKIN, P., CARMAN, G., LIPKIN, L., SHAPIRO, B., SCHULTZ, M., AND KAISER, P. A real time picture processor for use in biological cell identification. I. System design. *J. Histochem. Cytochem.* **22,** 725 (1974).
15. LEMKIN, P., CARMAN, G., LIPKIN, L., SHAPIRO, B., AND SCHULTZ, M. "Real Time Picture Processor—Description and Specification." NCI/IP Technical Report 7a, Nat. Tech. Info. Serv. PB269600/AS, June 1977.
16. SCHWARTZ, A. A. AND SOHA, J. M. Variable threshold zonal filtering. *Appl. Opt.* **16,** 1779 (1977).
17. STEFANELLI, R. AND ROSENFELD, A. Some parallel thinning algorithms for digital pictures. *J. Assoc. Comput. Mach.* **18,** 255 (1971).
18. ROSENFELD, A. AND KAK, A. "Digital Picture Processing." Academic Press, New York, 1977.
19. MARTELLI, A. An application of heuristic search methods to edge and contour detection. *Comm. ACM* **19,** 73 (1976).
20. NAKAGAWA, Y. AND ROSENFELD, A. "A Note on the Use of Local Min and Max Operations in Digital Picture Processing." TR-590, Univ. Maryland Computer Science Center, 1977.