Document Restoration by Digital Image Processing

W. H. SPUCK, R. J. BLACKWELL, *and* J. M. SOHA

PUBLIC AND PRIVATE ARCHIVES are replete with important, interesting, or sentimental documents that have become nearly illegible through the course of years, repeated handling, or accidental or purposeful damage. Archivists, both professional and amateur, have investigated techniques to preserve these documents and to restore them. Recently, the pertinent journals have recorded contributions, at a rate of perhaps two dozen per year, to document restoration.[1] This paper is to describe another restoration investigation, which applies technology derived from the U.S. space program. The technology is known as digital image processing and may be described most simply as an amalgamation of the camera and the computer.[2]

A document is commonly understood to be a vehicle (paper, book, medal, or similar surface) that carries an inscription such as writing, printing, or drawings. Since the term usually refers to items of historical or legal value, it must include photographic materials, but this would apply only to relatively recent documents.

Digital image processing emerged from the milieu of photography, and restoration techniques for continuous tone images have been studied extensively.[3] Our interest here, however, is in the adaptation of the technology to documents bearing inscriptions of the ilk of writing or line drawings. We will call these graphical documents. Half-tone images, which are on the boundary between graphical and pictorial records, will be considered pictorial and therefore outside our domain of interest.

Image processing experience has shown that rarely does extensive processing reveal information that is not visible in original images to a careful (perhaps trained) observer. This is not to denigrate the value of such processing (even to the

William H. Spuck writes as the manager, Science Data Analysis Section of the California Institute of Technology's Jet Propulsion Laboratory. Blackwell and Soha are supervisors of two different groups within the Science Data Analysis Section.

[1]For additional references to pertinent monographs and to articles in professional journals, the reader should see *Library Literature: An Index to Library and Information Science*, published bimonthly, with periodic cumulations, by the H. W. Wilson Co., New York (under the following headings: Archives—Care and Restoration; Books—Care and Restoration; and Documents—Care and Restoration) and in *Library and Information Science Abstracts*, published bimonthly by the Library Association, London (under the heading Technical Processes and Services, Code SE).

[2]F. C. Billingsley, "Review of Digital Image Processing," *Proceedings*, European Computer Conference on Interactive Systems & CAD, September 1975, ONLINE, Brunel University, Uxbridge, Middlesex, England.

[3]William K. Pratt, *Bibliography on Digital Image Processing and Related Topics* (Los Angeles: University of Southern California, Image Processing Institute, Electronic Sciences Laboratory [USCEE Report 453], 1973); Azriel Rosenfeld, "Picture Processing: 1974," *Computer Graphics and Image Processing*, 4 (1975): 133-55; and Azriel Rosenfeld, "Picture Processing: 1973," *Computer Graphics and Image Processing* 3 (1974):178-94.

most sophisticated of observers), but it does place a lower bound on the quality of image that may be restored. The principal objective of the present investigation is not to enable reading or interpreting of documents, but to restore their quality. Then, not only do they become more legible to the casual reader, but also they possess the best possible qualities for resisting inevitable subsequent deterioration.

We characterize good graphical documents by sharp, clear imprints on an otherwise clean background. This is not to rule out variations in the markings, but to emphasize their deliberateness. In general, graphical documents include color variations of both the markings and the background. We will not discuss these, however. Instead, we will consider only black and white material, and treat color only insofar as it represents a means of emphasizing the contrast between the markings and the background. (For example, printing with a particular hue on an off-white background may yield more contrast when viewed through spectrally selected filters and color controlled illumination than when these are not controlled.)

Document deterioration yields many artifacts, by which is generally meant markings or discoloration not intended by the creator of the document. Mold, water, and dirt stain are typical artifacts in the deterioration of ancient archival documents. These artifacts affect both the markings and the background, or, more significantly, the relationship between the two. Ultimately, we desire to regain (or perhaps gain for the first time) good graphics. That is, we seek deliberate markings on a distractionless background.

An important constraint to any realization of the restoration goal is that the process must be practical. Whether the process is to be made available as a system or a service, some implementation of the process must be feasible technically and economically. In the end, an essential criterion must be the cost per page of restoration. Ease of operation and similar operational factors also affect cost and process acceptability and must be considered.

Image Restoration Specimens

For the current investigation, we treated four pages of graphical material. Three are handwritten and old. The fourth is typewritten. These documents were obtained from technology advisers attached to state and local governmental offices. The advisers asked several archives for examples of deteriorated or damaged documents. Both worst-case and typical specimens were selected, and four pages from these were chosen for processing. Photographs of the four pages appear as Figures 1 through 4.

Figure 1 is an interesting historical document. It is an extract from the will of President Zachary Taylor. It evidences a common problem of inked documents; namely, that ink from the reverse side of the paper has bled through to the front in varying amounts. This makes legibility difficult, but certainly not impossible. Other minor artifacts are also present, the most notable being the effects of transparent tape covering part of the left-hand margin.

Figure 2 is a letter from the Archives of the State of Mississippi. This document shows several problems, but the most significant is the obvious loss of contrast between the writing and the background from the top to the bottom of the page.

Figure 3 is a deed, known to us as one of Mrs. Brally's documents. This deed poses the problems of the Mississippi document, but, in addition, the results of mold that has grown over the central area of the document are quite evident. It is probably true

that better (spectrally selective) photography would decrease the impact of the mold, but, once a black and white photograph has been exposed and developed, the most that can be done is to retrieve the maximum amount of information from the photograph. For a restoration experiment, this is equivalent to starting with a more difficult example.

Figure 4 is a typewritten memorandum from the Archives of the Tacoma, Washington, Police Department. The document evidences considerable deterioration caused by repeated duplication. The net effect is to turn most of the printing into disconnected dots. This artifact is compounded by contrast degradation and by the superposition of random dot patterns in background areas.

Restoration Process

Documents or photographs of documents must be converted to a numerical form (a process called digitization) before they can be restored by a computer. Similarly, computer-restored images are numerical and must be displayed on some medium (paper, film, television monitor) before they can be viewed. This process is illustrated in Figure 5.

The process of digitizing documents for restoration must be performed carefully. The computer industry uses the slogan GIGO (garbage in, garbage out) as a reminder of the precept that good computing results start with good, valid data. This is as true for document restoration as for any other computer application.

The process used to digitize the documents of Figures 1-4 involved two steps. First, the original documents were photographed. Then the photographs were converted to numerical form.

Special attention was given during the photographic step to the safety of the original document. Such a precaution is essential in the milieu of important archival documents, but, having recognized this, we will assume its accomplishment and discuss it no further.

The photographs in Figures 1-4 were, in fact, prepared with specific attention to their suitability for computer processing. Figures 2 and 3 were photographed using standard darkroom techniques and Kodak Contrast Process Pan film. Kodalith/Ortho-Type 3 was used for the Tacoma Police document. No special lighting, filtering, or lens techniques were used. Figure 1 was photographed with somewhat more care. A photographer from the Johnson Spacecraft Center, Houston, carried out the photography under the guidance of one of the authors. Several lighting and filter combinations were explored, and Tri-X Pan film was used. The combination found most suitable for this document was white light (tungsten source) at F16 and 0.5 sec. exposure with a Wratten #11 filter.

While further discussion of this photographic process is beyond the scope of this introductory paper, we note that the following parameters have been observed to be significant to the process:

a) spectral content of illumination source
b) illumination geometry (including combination of epi- and trans-illumination)
c) photographic filter selection
d) lens optical characteristics
e) film spectral response

f) film resolution capability (grain characteristic)
g) film photometric response (fog level, gamma, maximum density)

Subsequently, the photographic film was scanned by an automatic precision microdensitometer. This yielded a digital magnetic tape that recorded the transmissivity (equivalently, the density) of the film at every point on a very fine (10 micron spacing) rectangular network of points covering the entire film surface. Parameters similar to those pertinent for the photographic process affect the scanning step.

Digitization need not be a two-step process. In fact, almost any realization of the methods reported here that was aimed at a production environment would bypass the photographic intermediate. Means exist today (such as television cameras, facsimile scanners, and reflection microdensitometric scanners) to scan the original document directly, and there are some parameters, such as photometric sensitivity, that would be improved in this way. One-step digitization was not used in this investigation simply because the two-step process was more expeditious.

Display of the restored document is merely the reverse of digitization. Numerical images are placed on some medium that can be visually observed. Photographic film is a very convenient medium and was used for this study. The film was exposed by a high precision cathode ray tube (television monitor). Other media (especially direct viewing of cathode ray tubes) might be preferable for other applications.

Processing Results

President Taylor's Will. Following digitization, President Taylor's will was displayed with only a minor digital contrast adjustment (Figure 6). Minor contrast adjustments represent no novel technology since they are equivalent to paper selection, exposure, and chemistry adjustments in photographic reproduction. The adjustments here were made digitally to allow improved quality consistency in the computer-to-film conversion process by fixing the photographic and electronic parameters of the display system.

A histogram (graph of the frequency of occurrence) of the gray levels contained on the document was prepared. This appears in Figure 7. Gray levels had been recorded in 256 steps from 0 (pure black) to 255 (pure white).

Analysis of the histogram shows what can be observed in the original document: to wit, that the gray levels concentrate in two regions (the histogram is bi-modal). These correspond to the background (the whitest or left-most concentration), and the writing and bleedthrough (the darkest or right-most concentration). The bleedthrough is located on the left (lighter) side of the darker concentration. Considerable improvement is expected, then, from saturating to white all points whose gray level is deduced to belong to either the background or the bleedthrough and to black all remaining points. The result of applying this process appears as Figure 8. Because the gray levels of the bleedthrough and the writing overlap, we would not expect perfect results from this process.

Two other artifacts were also removed in Figure 8. First, the margins of the document were located and saturated to background white. They contain no information. Second, the area covered by transparent tape was subjected to an independent histogram analysis and restoration, since the net effect of the tape was to change the contrast relationship.

Mrs. Brally's and Mississippi Documents. Mrs. Brally's document and the Mississippi document were processed together because, except for the mold, they evidenced similar problems. Histogram analysis suggested that extensive contrast manipulation would be helpful. Figures 9 and 10 are the result. Here, the zero gray level remained at zero (black), gray levels above 60 were set to 255 (white), and those in between were uniformly (linearly) adjusted between 0 and 255. Note that this contrast adjustment could be accomplished by modifying the digitizing parameters to utilize fully the available dynamic range. This is often difficult to accomplish in practice, however, so conservative digitizing parameters are usually chosen to avoid the possibility of data loss due to saturation. (This would be particularly true in a production environment.)

It appeared that Mrs. Brally's document would benefit from differing contrast adjustments from top to bottom, and this was performed by subdividing the image into five horizontal swaths and separately adjusting each. The result is Figure 11.

At this point, it appeared desirable to try to sharpen the edges of the letters. Edge sharpening is a standard image-processing technique, although it is somewhat sophisticated mathematically. An edge-sharpening method was applied to the two pages, with the results shown in Figures 12 and 13. The technique used is known as *two-dimensional convolutional high pass filtering.* The edge-sharpening technique encompasses another contrast adjustment, which was again selected to maximize discrimination.

Now some attention was necessary to the distortion caused by the mold. The edge-sharpening technique had a salutary benefit on the mold effects in that it converted them from something like a haze to a random pattern of black dots and lines. In image processing language, this undesirable component of a picture is called noise. A technique was now applied that measured the perimeter of all black objects on the page and eliminated the object if its perimeter was smaller than a specified threshold. This noise-cleaning process, applied to Mrs. Brally's document, resulted in the considerably cleaner image of Figure 14—a result which we judged to be a satisfactory final product. Of course, this process resulted in some loss of inscription data, as well as the artifacts, but this tends to be equally true of each of the previous processes. This fact emphasizes the need for judicious application of each process. Such judgment has not yet been automated.

When applied to the Mississippi document, the perimeter measurement process was much less satisfactory. The problem here is that the noise cannot be separated from the inscription solely on the basis of perimeter. A large perimeter value (Figure 15) removes most noise, but it also results in an unacceptable loss of inscription data. A smaller perimeter value (Figure 16) preserves more of the inscription, but also admits more noise. What is needed is a technique to take advantage of the relative spatial correlation of the writing as opposed to the more random character of the noise. Therefore, an alternative process was applied to the Mississippi document, with its goal to join isolated fragments of letters. The technique used was a crude directional filter that emphasized points which, when aligned with their neighbors, fell on straight lines. After this filtering, the previous perimeter-based noise cleaning procedure was applied to reduce the background artifacts. Finally, the resulting image was combined with the edge-enhanced version of Figure 13, using a computerized version of a double exposure. This produced a picture in which the relative strength of the inscription is increased. The result, in Figure 17, is not totally satisfactory but does, in our opinion, represent some improvement.

Tacoma Police Document. The processing applied to the Tacoma Police document was considerably different from that previously described. The concept was to broaden the isolated dots that formed the letters by emphasizing individual digitizing points that had several neighbors falling within the dots. This was accomplished by means of a five-by-five point low-pass (averaging) filter. The result of this operation appears in Figure 18.

This result was discouraging, and careful thought suggested no simple techniques to improve performance even in the less deteriorated regions. The approach that seemed most promising for this document was one which combined optical character recognition (pattern matching against idealized character templates) coupled with contextual grammatical analysis. Such an approach was beyond the scope of this investigation.

Conclusion

The most significant conclusions to be drawn from this study would probably be those that answer the following three questions:

1. Were the restorations successful?
2. Can restoration be commercially practical?
3. What should be the next investigative steps?

These questions are obviously sequential, each depending for significance on an affirmative answer to its predecessor.

The first question, about the successfulness of the restoration, is perhaps the most difficult question of the three to answer. First, the success is one of degree and the measure is subjective. Second, the authors are not the appropriate experts to measure the success. And, third, the authors have not exhausted their ideas on what can be done (hence, question 3) and, therefore, tend to see the results as not good enough. Nevertheless, a reasonably unbiased comparison of Figures 1-3 with Figures 8, 14, and 17 indicates that the processing yielded improved graphics in the handwritten documents. The blacks are black; the whites, white. The most significant artifacts have been removed, and nearly everything that can be read through careful observation of the original can be read more easily on the product. These restorations, then, possess the desired characteristics of master copies for subsequent document reproduction. Success was not achieved for the Tacoma Police document.

The question of commercialization has two principal components, financial and technical. Although some review of the financial facet has been conducted, this is better the domain of industrial analysts than research and development specialists. Any commercialization analysis must consider the two principal markets; namely, the market for restoration systems and the market for restoration services.

Technically, the feasibility of commercial packaging of the restoration methods used for this investigation is a certainty. Several alternative digitizers and display devices are already commercially available, and the processor can be any of a number of commercially available minicomputers. It is more likely, however, that if the processor were to be packaged today it would use the new microprocessor technology. The selection of system components would depend somewhat on the intended market. For example, drum-type (facsimile) scanners are probably less expensive, but flat-bed scanners might be more appropriate for fragile documents.

Similarly, facsimile display devices are inexpensive, but the desired product may be microfilm. In addition, system design undoubtedly would include further investigation of methodology along the lines discussed below.

Several topics for future investigation emerged from this effort. Several are novel applications of existing continuous tone image methods, but most are adaptations of these to the graphic restoration problem.

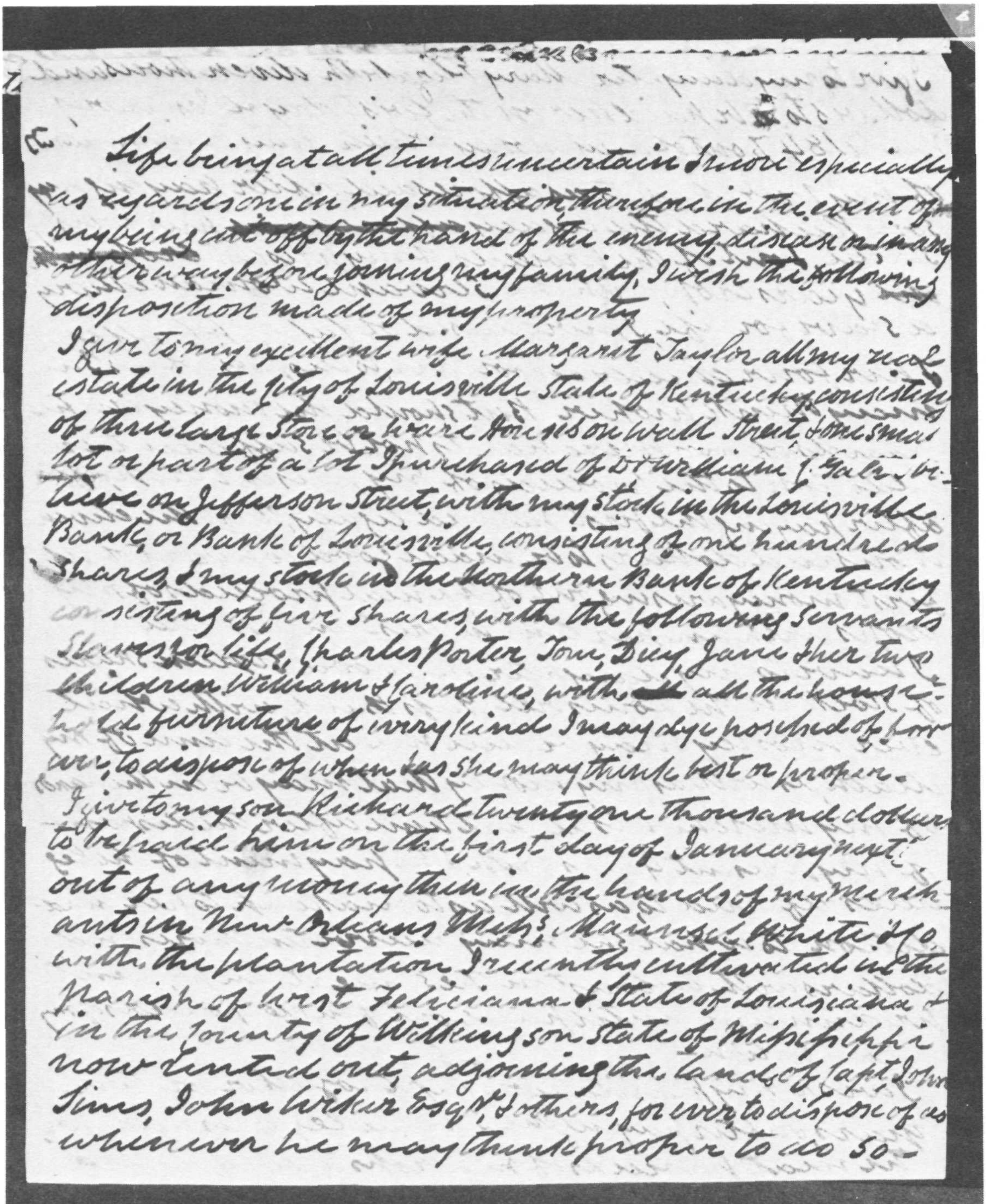
One fruitful area for exploration would be the generation of histograms for small regions of the page to exploit local properties more fully. Methods which measure and adapt their parameters to these properties would emerge.

Another interesting feature to exploit is the inherent linearity of the graphical material. Techniques which follow lines and ones that span short gaps are of particular interest.

The topic of good data collection must not be ignored. As discussed above, restorations are re-creations, not creations. They must start with the best possible measurement of the available qualities of the original document.

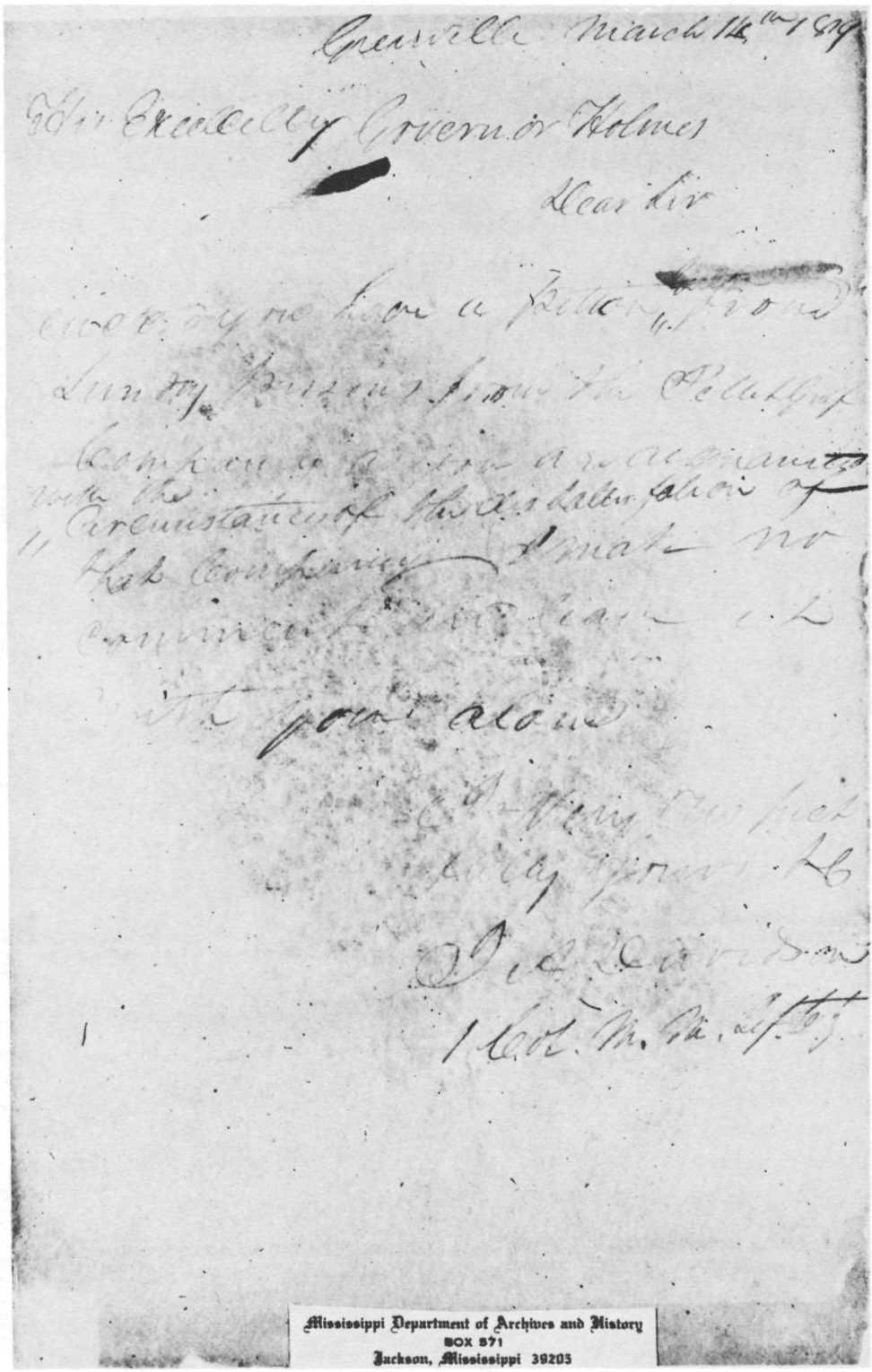
The most significant constraint imposed on the development of any restoration method is the commercialization goal. Today's computing hardware constrains the complexity of the processes that may be employed. The constraint lies mainly in the area of the amount of data that may be retained within the processor at any one time. Current limitations are in the vicinity of a few scan lines. A typical graphical document might have a total of several thousand such lines. Many potential processes require the entire page to be processed simultaneously (accessed randomly). The present trend in computer hardware technology may obviate this complexity constraint within only a few years, but the methods applied in this investigation all satisfied the constraint.[4]

[4]The authors gratefully acknowledge the assistance of L. N. Mogavero and R. L. Gilbert at the headquarters of the National Aeronautics and Space Administration (NASA), and the team they manage of state and local government technology transfer specialists, who contributed to the selection of documents for restoration and the evaluation of the results. John Brinkmann of NASA/Johnson Spacecraft Center photographed President Taylor's will. William Smollen, NASA technology representative to the New Orleans Regional Planning Commission, made the arrangements for this photographic session. John D. Addington, of the JPL Image Processing Laboratory, performed much of the processing of the will, and several inquisitive members of the Image Processing Laboratory staff suggested innovations to improve the restoration methods applied to all of the documents.

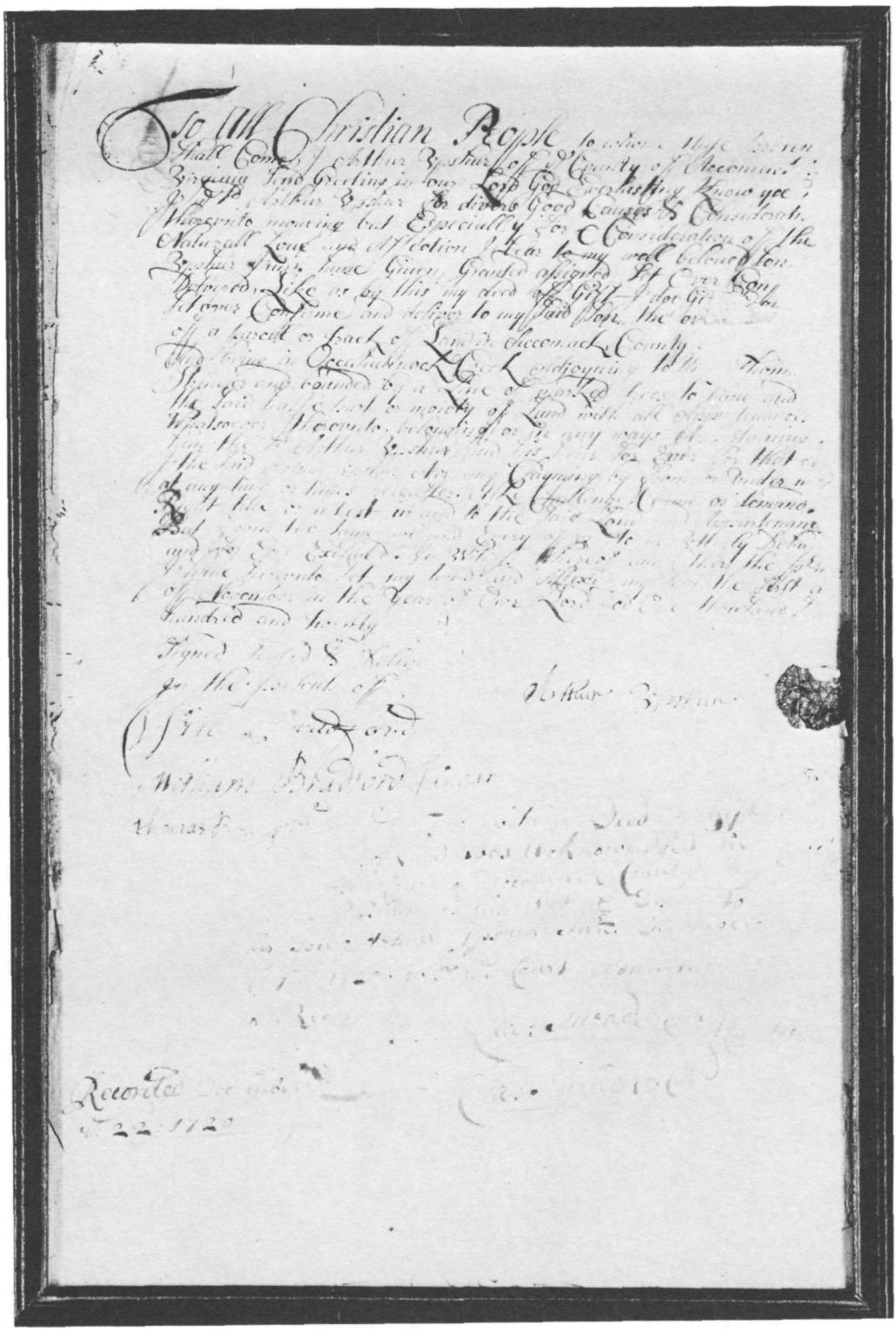


Life being at all times uncertain & more especially as regards one in my situation, therefore in the event of my being cut off by the hand of the enemy, disease or in any other way before joining my family, I wish the following disposition made of my property
I give to my excellent wife Margaret Taylor all my real estate in the City of Louisville State of Kentucky consisting of three large Store or Ware Houses on Wall Street, & one small lot or part of a lot I purchased of Dr William J. Galt lying on Jefferson Street, with my stock in the Louisville Bank, or Bank of Louisville, consisting of one hundred shares & my stock in the Northern Bank of Kentucky consisting of five shares, with the following servants slaves for life, Charles Porter, Tom, Dicy, Jane & her two children William & Caroline, with all the household furniture of every kind I may dye possessed of, forever, to dispose of when & as she may think best or proper.
I give to my son Richard twenty one thousand dollars to be paid him on the first day of January next, out of any money then in the hands of my merchants in New Orleans Messrs. Maunsel White & Co with the plantation I recently cultivated in the Parish of West Feliciana & State of Louisiana & in the County of Wilkinson State of Mississippi now rented out, adjoining the lands of Capt. John Sims, John Walker Esqr. & others, for ever to dispose of as whenever he may think proper to do so -

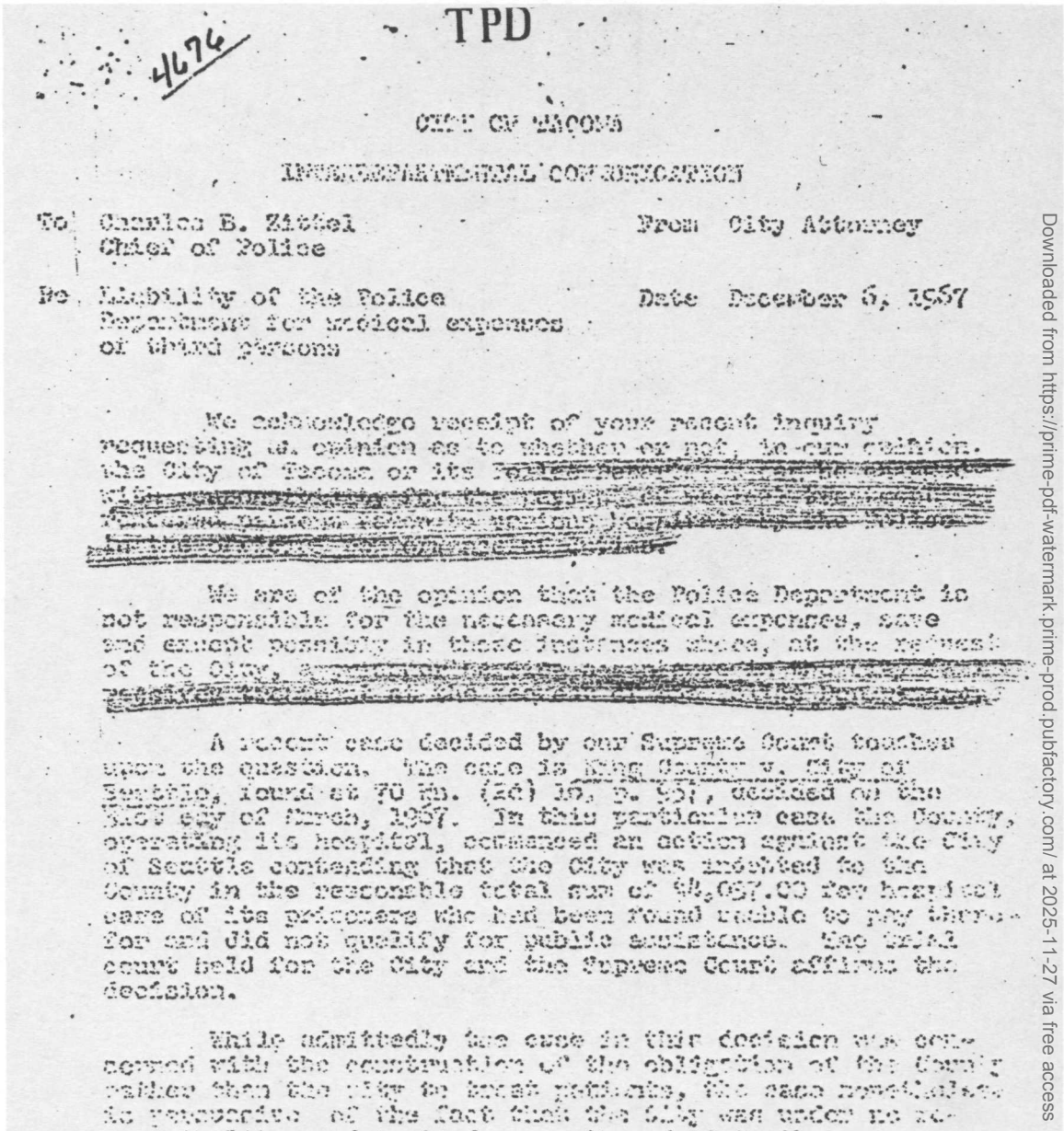
1. A portion of the will of Zachary Taylor. Bleedthrough from the reverse side of this page makes legibility difficult.



2. A letter from the Archives of the State of Mississippi, with mold in center.



3. Mrs. Brally's document.



TPD

4676

CITY OF TACOMA

INTERDEPARTMENTAL CORRESPONDENCE

To: Charles B. Ziebel
Chief of Police

From: City Attorney

Re: Liability of the Police Department for medical expenses of third persons

Date: December 6, 1957

We acknowledge receipt of your recent inquiry requesting an opinion as to whether or not, in our opinion, the City of Tacoma or its Police Department [redacted]

We are of the opinion that the Police Department is not responsible for the necessary medical expenses, save and except possibly in those instances where, at the request of the City, [redacted]

A recent case decided by our Supreme Court touches upon the question. The case is King County v. City of Seattle, found at 70 Wn. (2d) 10, P. 421, decided on the 21st day of March, 1967. In this particular case the County, operating its hospital, commenced an action against the City of Seattle contending that the City was indebted to the County in the reasonable total sum of $4,057.00 for hospital care of its prisoners who had been found unable to pay therefor and did not qualify for public assistance. The trial court held for the City and the Supreme Court affirms the decision.

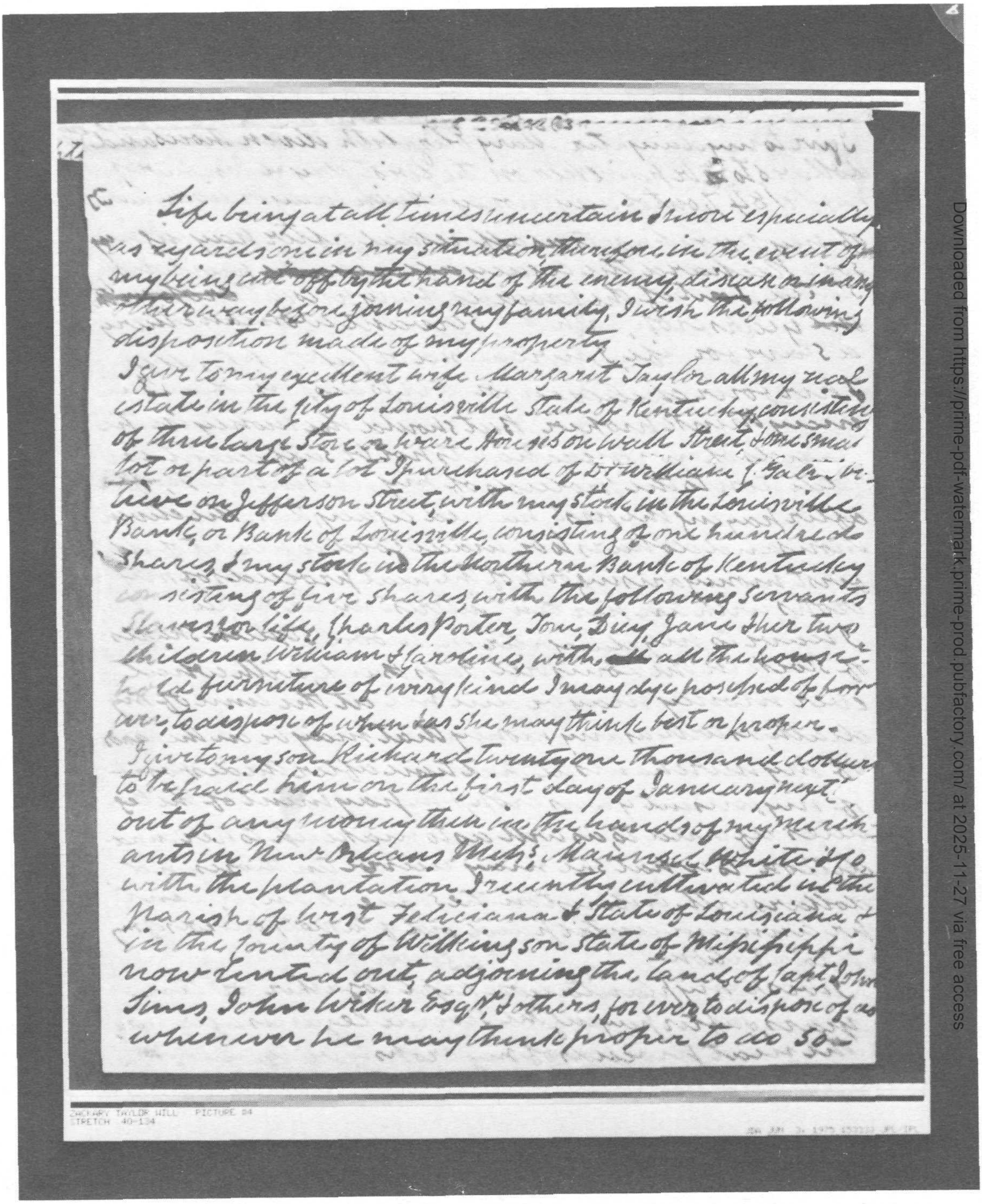
While admittedly the case in this decision was concerned with the construction of the obligation of the County rather than the City to bear patients, the case nevertheless is recursive of the fact that the City was under no responsibility or duty to do more than deliver the prisoner to the institution for appropriate medical attention, and having delivered the patient to the County Hospital (since duty under the statute was to provide free medical service), the City had fulfilled its duty.

What we believe, under the provision of this case, [redacted]

4. A typewritten document from the Tacoma Police Department. The document has suffered from repeated duplication.

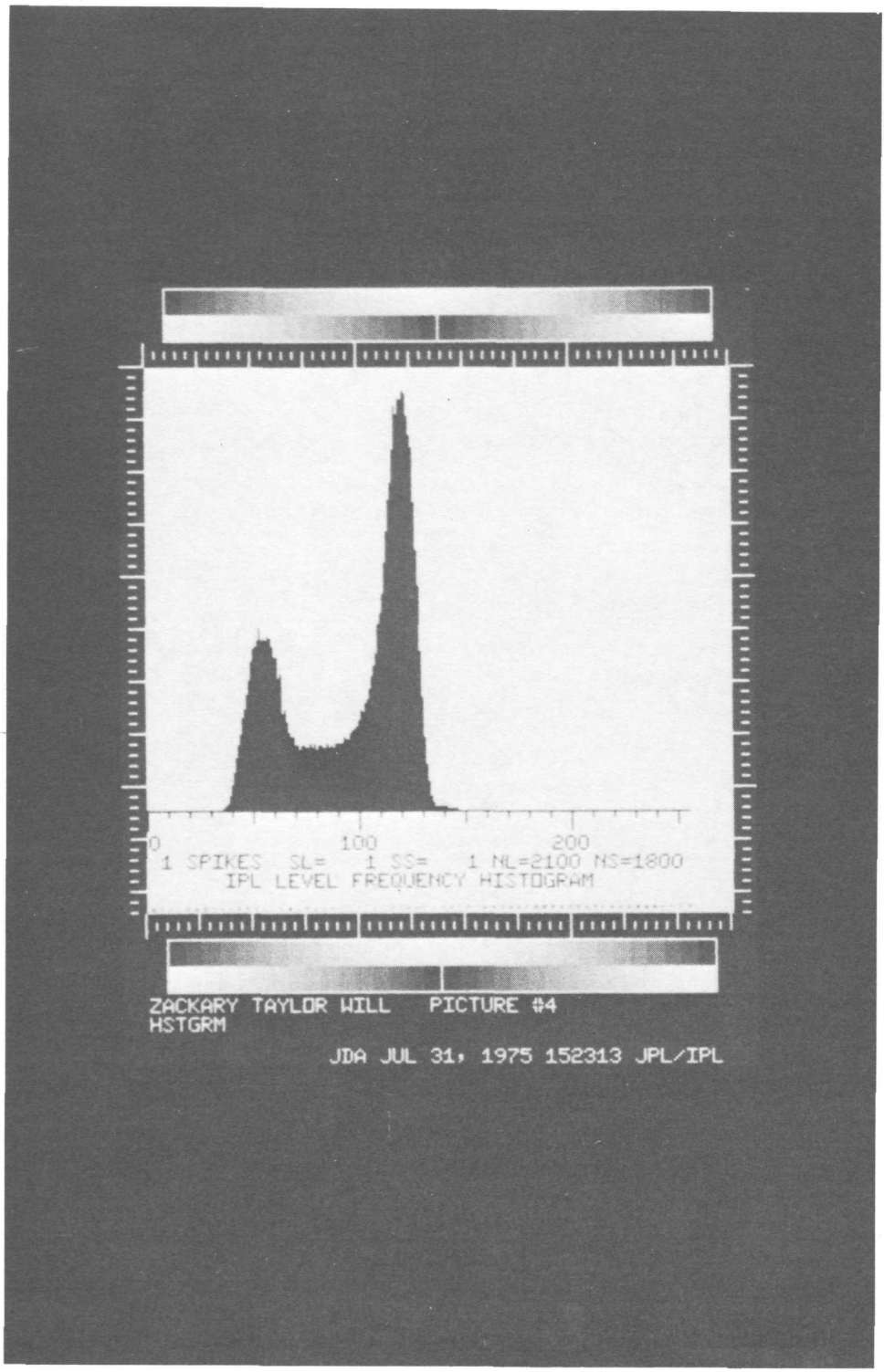
5. Overview of the digital document restoration process.

Figure 5. Overview of the Digital Document Restoration Process

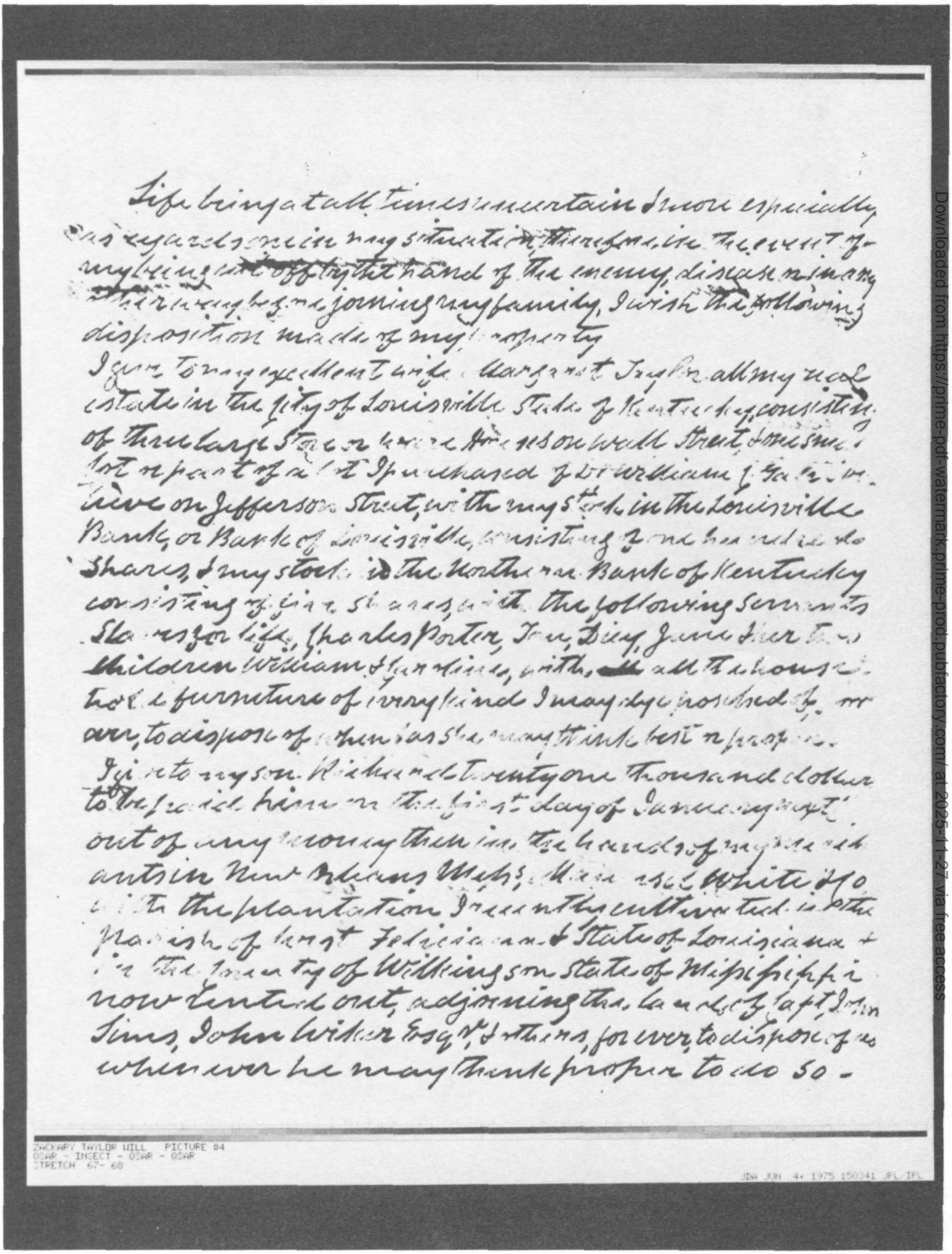


Life being at all times uncertain & more especially
as regards one in my situation, therefore in the event of
my being cut off by the hand of the enemy, disease or in any
other way before joining my family, I wish the following
disposition made of my property
I give to my excellent wife Margaret Taylor all my real
estate in the City of Louisville State of Kentucky consisting
of three large Store or Ware Houses on Wall Street, some small
lot or part of a lot I purchased of Dr William C. Galt [illegible]
on Jefferson Street, with my stock in the Louisville
Bank, or Bank of Louisville, consisting of one hundred
shares & my stock in the Northern Bank of Kentucky
consisting of five shares, with the following servants
slaves for life, Charles Porter, Tom, Biey, Jane & her two
children William & Caroline, with all the house-
hold furniture of every kind I may dye possessed of, for
ever, to dispose of when & as she may think best or proper.
I give to my son Richard twenty one thousand dollars
to be paid him on the first day of January next
out of any money then in the hands of my merch-
ants in New Orleans Messrs Maunsel White & Co
with the plantation I recently cultivated in the
parish of West Feliciana & State of Louisiana &
in the County of Wilkinson State of Mississippi
now rented out, adjoining the lands of Capt John
Sims, John Walker Esqr & others, for ever to dispose of as
whenever he may think proper to do so

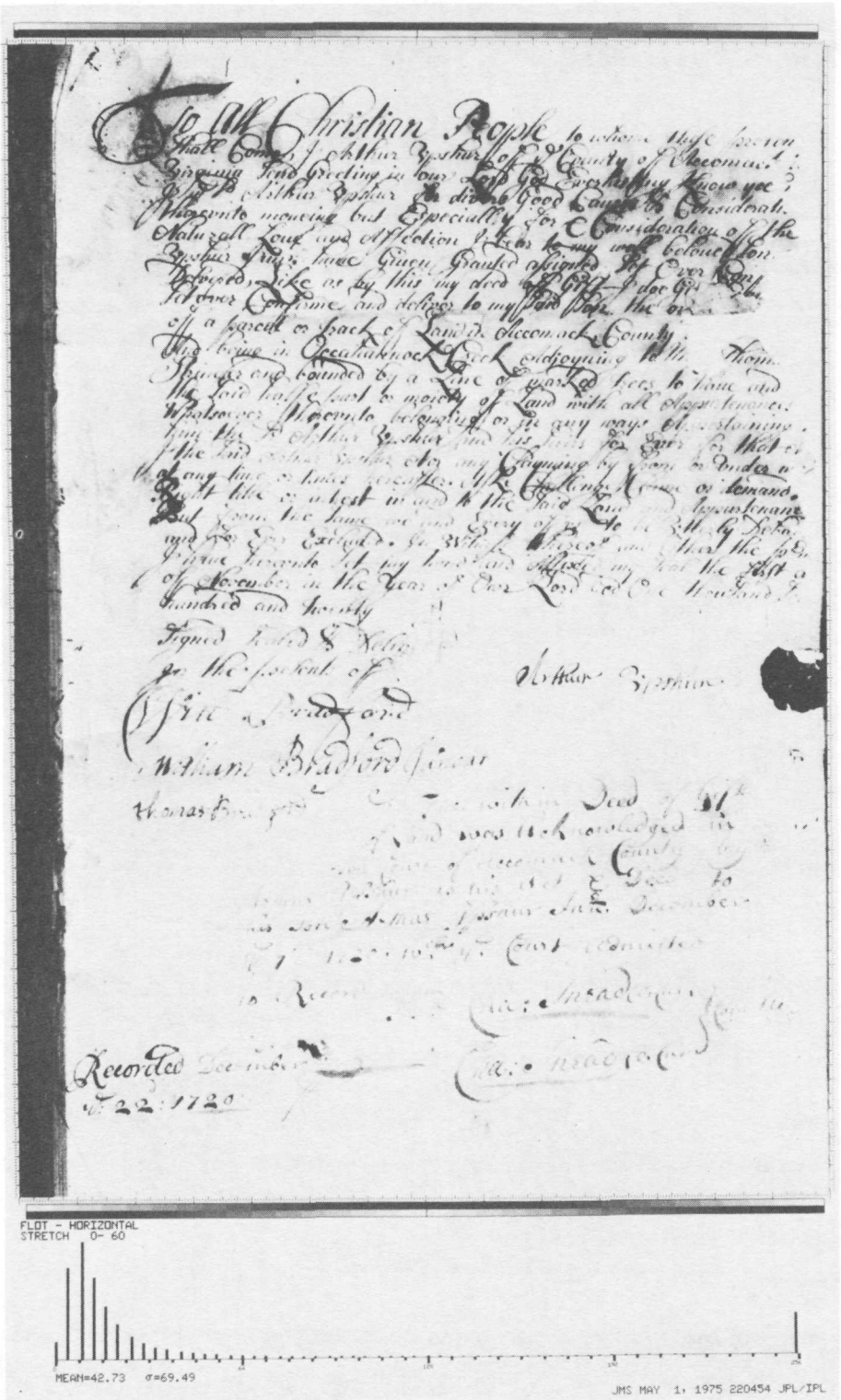
6. Zachary Taylor will after digitization and contrast adjustment.



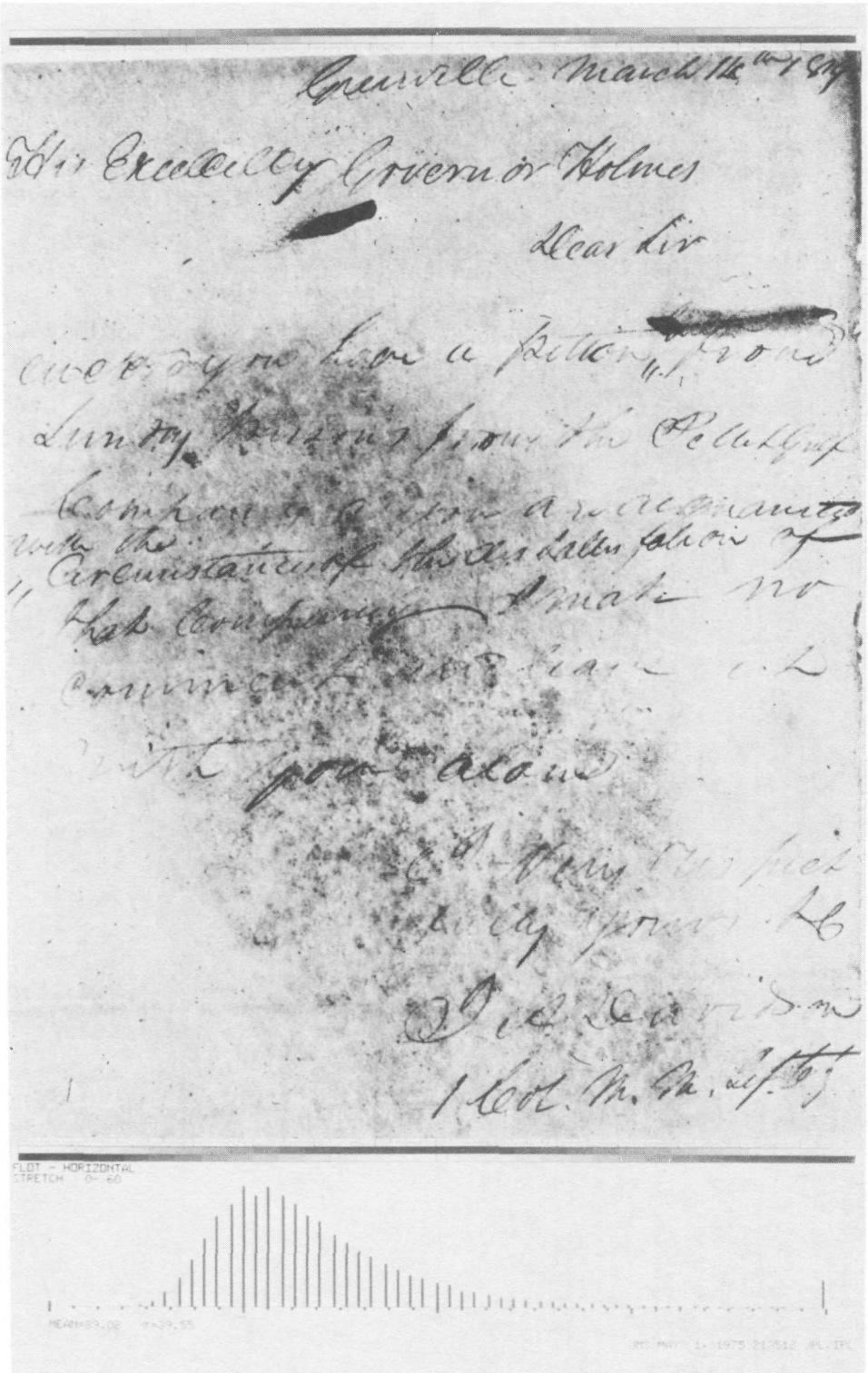
7. Gray level frequency histogram illustrates the separation of the writing from the bleed-through and the background.



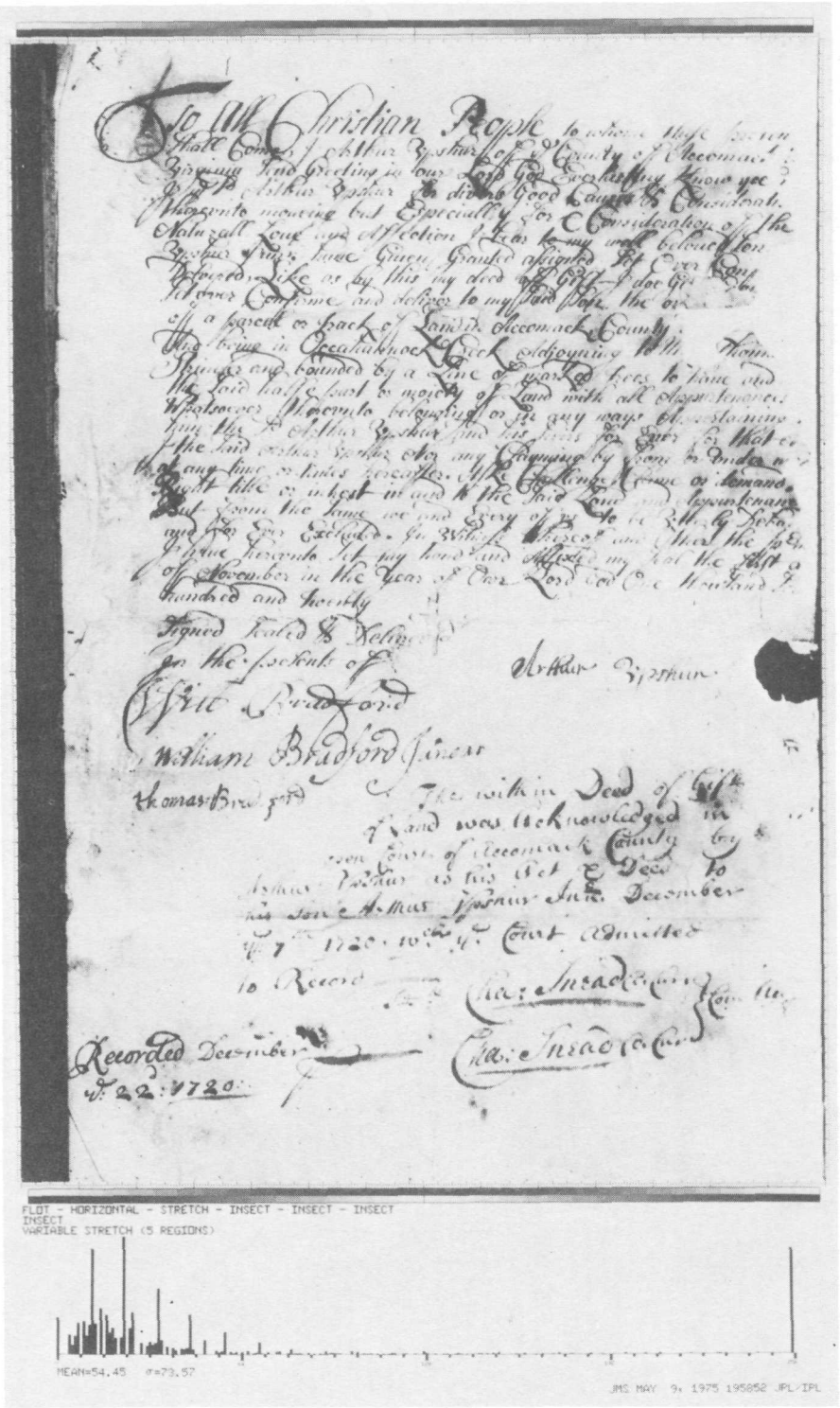
8. Page of the Taylor will after performing binary contrast adjustment to separate writing.



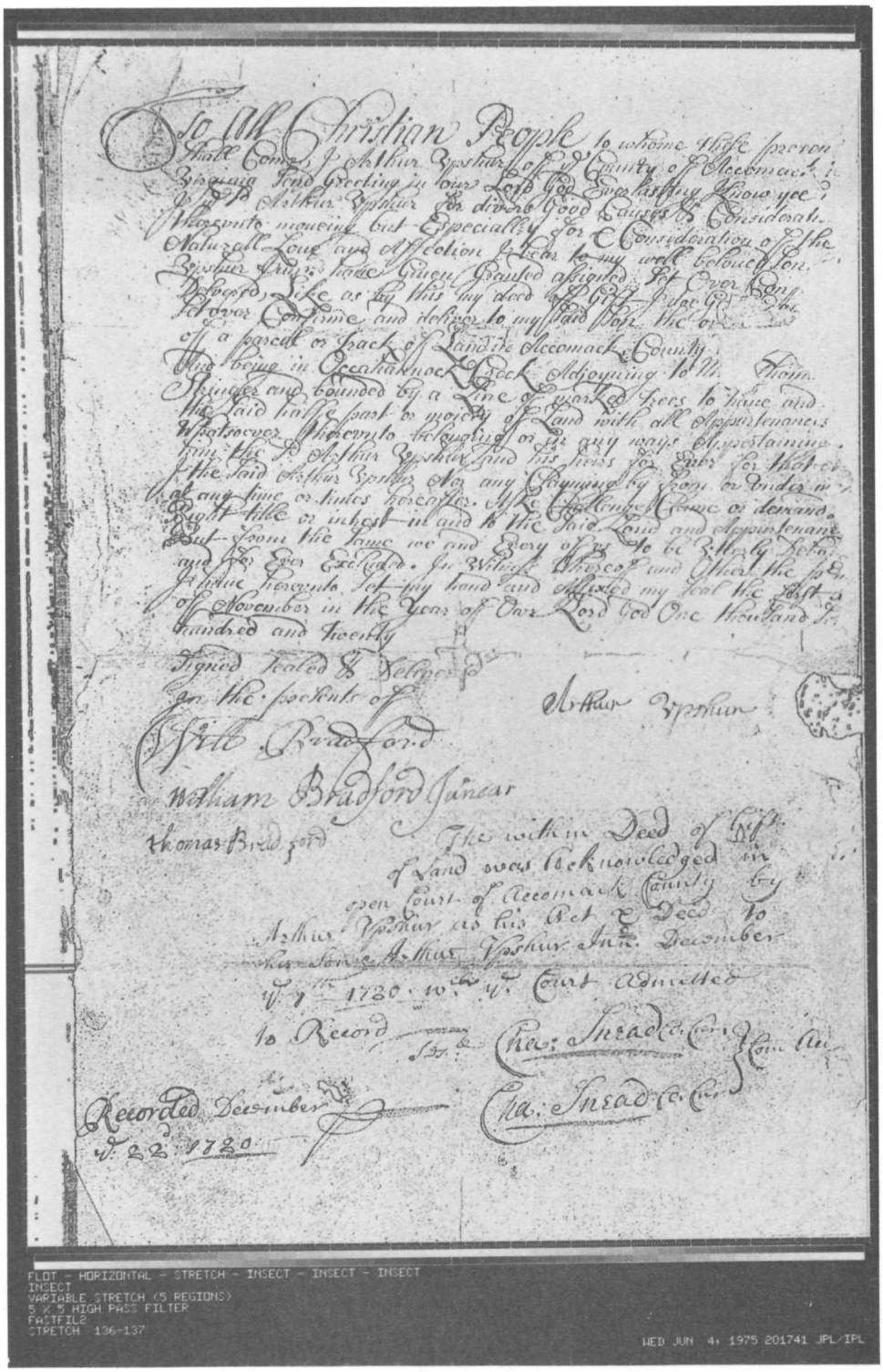
9. Mrs. Brally's document after digitization and contrast adjustment.



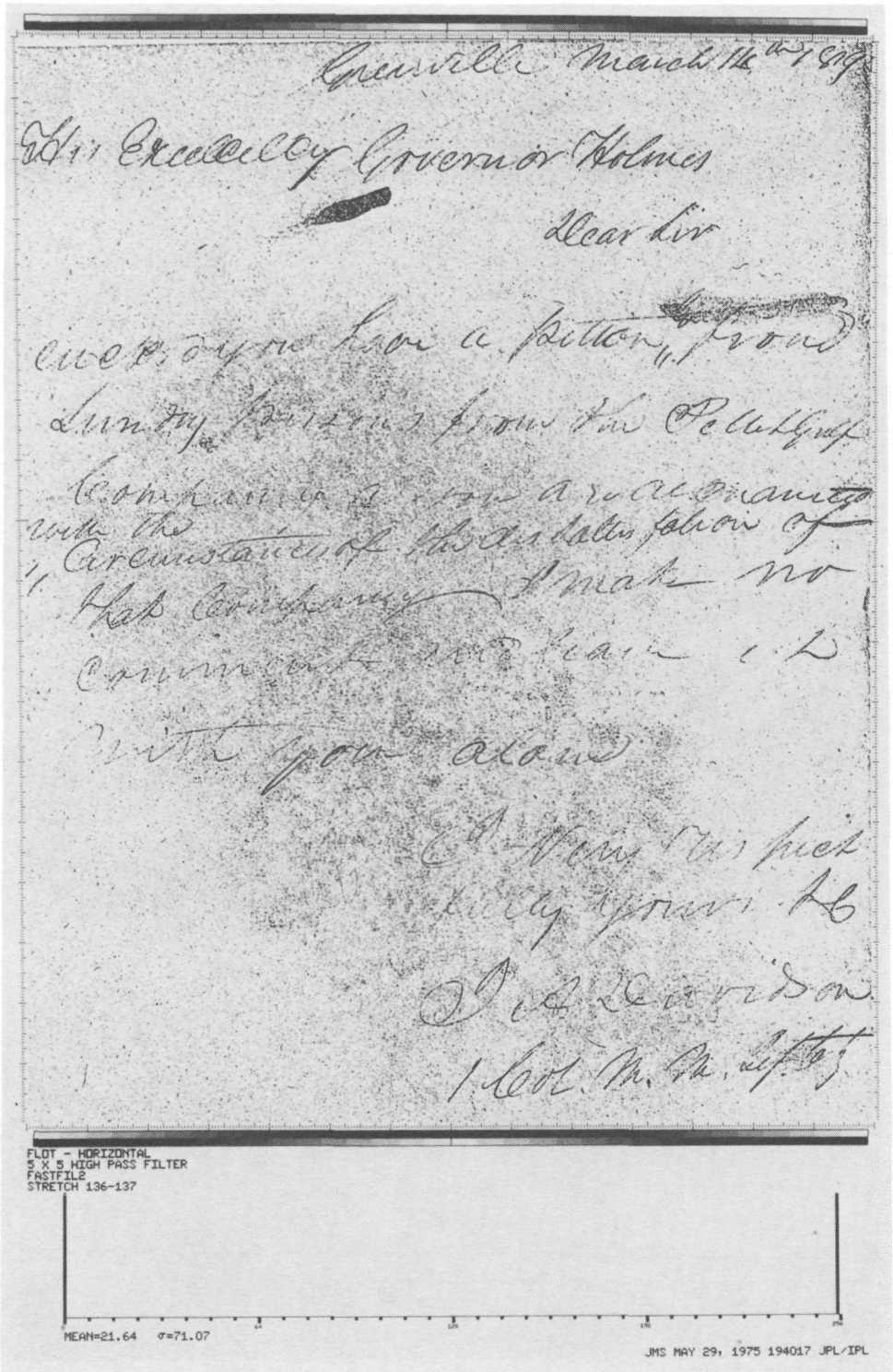
10. Mississippi document after digitization and contrast adjustment.



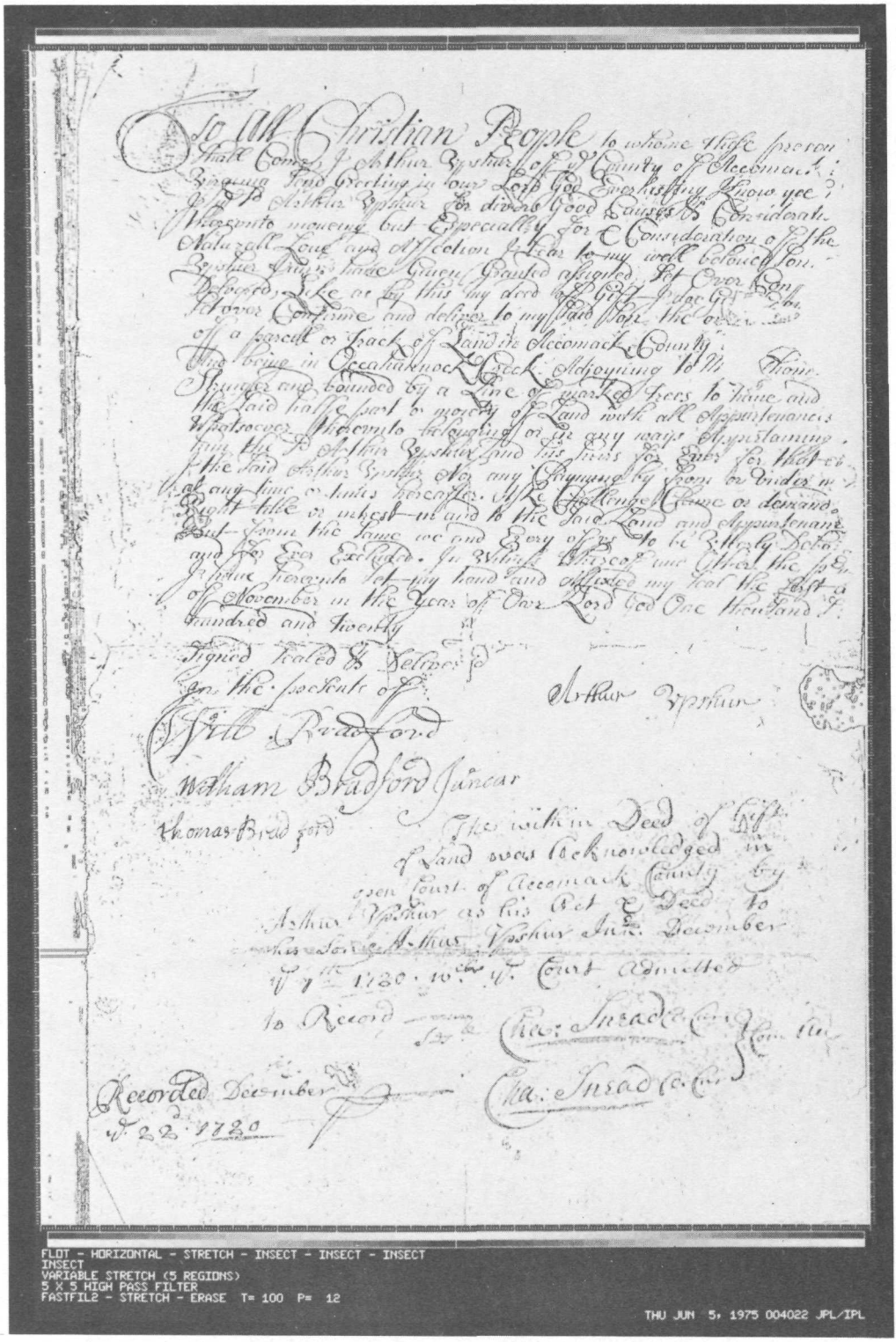
11. Mrs. Brally's document after performing independent contrast adjustment on five separate segments.



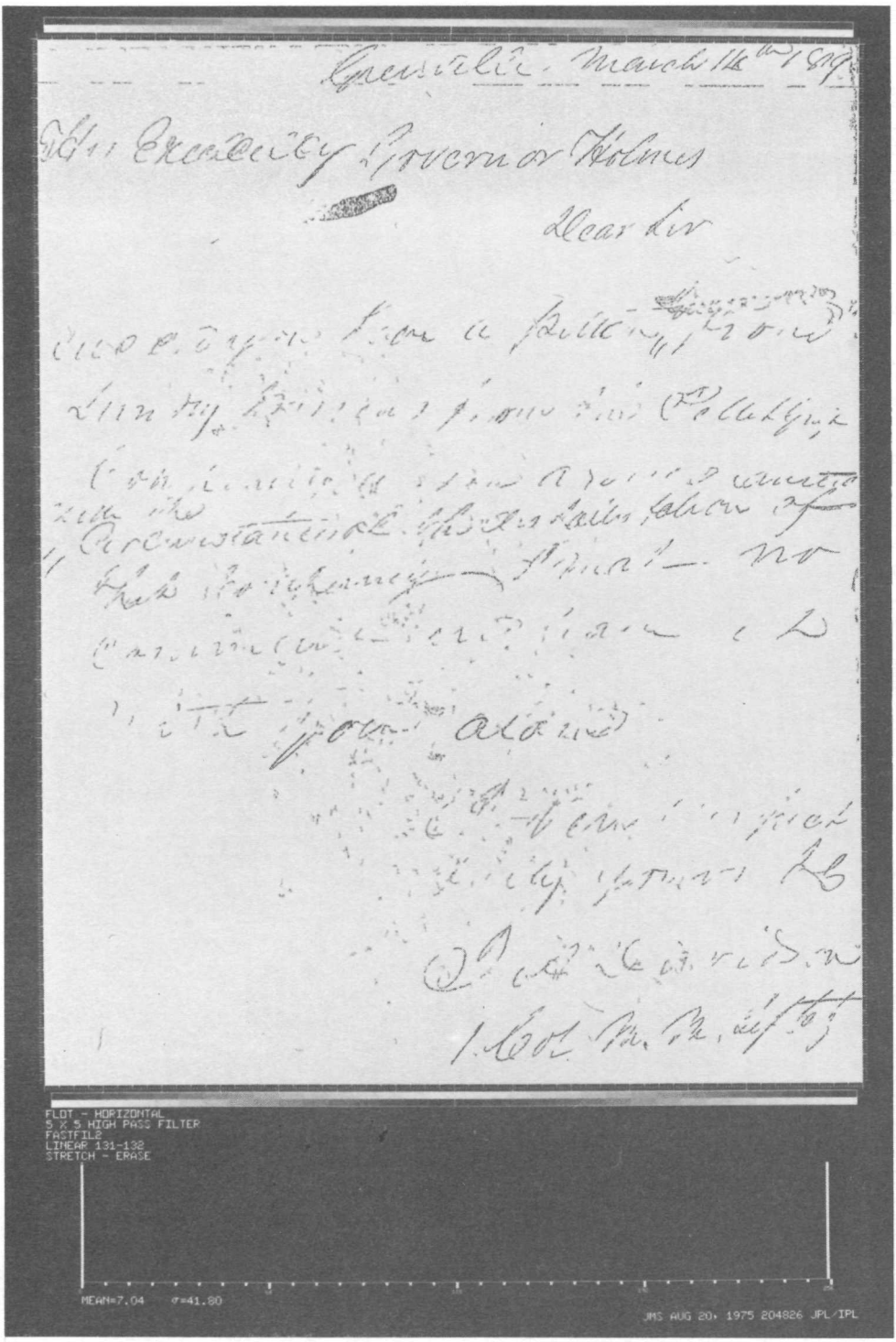
12. Edge sharpening applied to Mrs. Brally's document. Writing is essentially uniformly discernible at this stage, although noise remains.



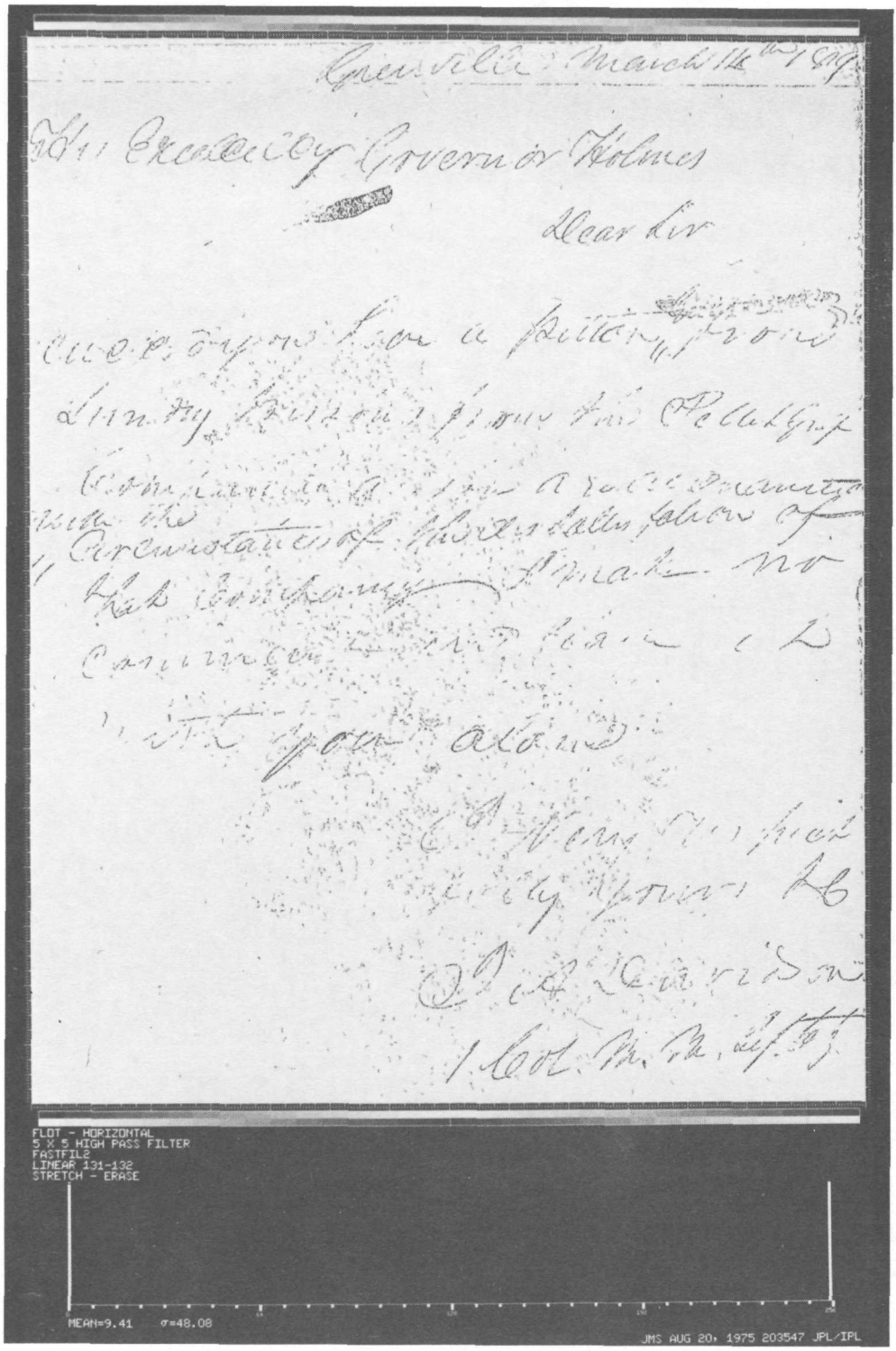
13. Edge sharpening applied to the Mississippi document. Here the writing remains difficult to separate from the noise in some areas.



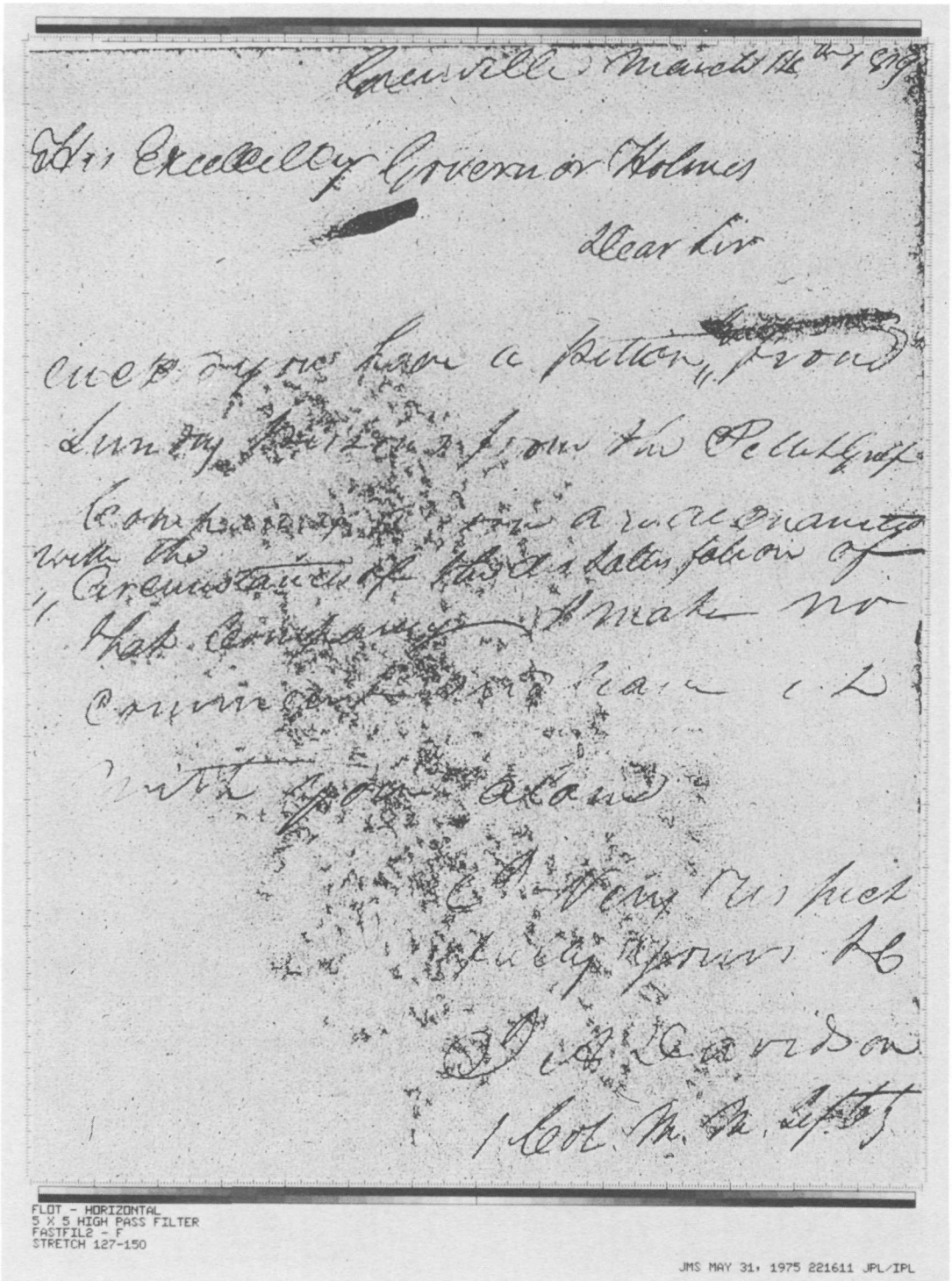
14. Mrs. Brally's document after noise cleaning procedure.



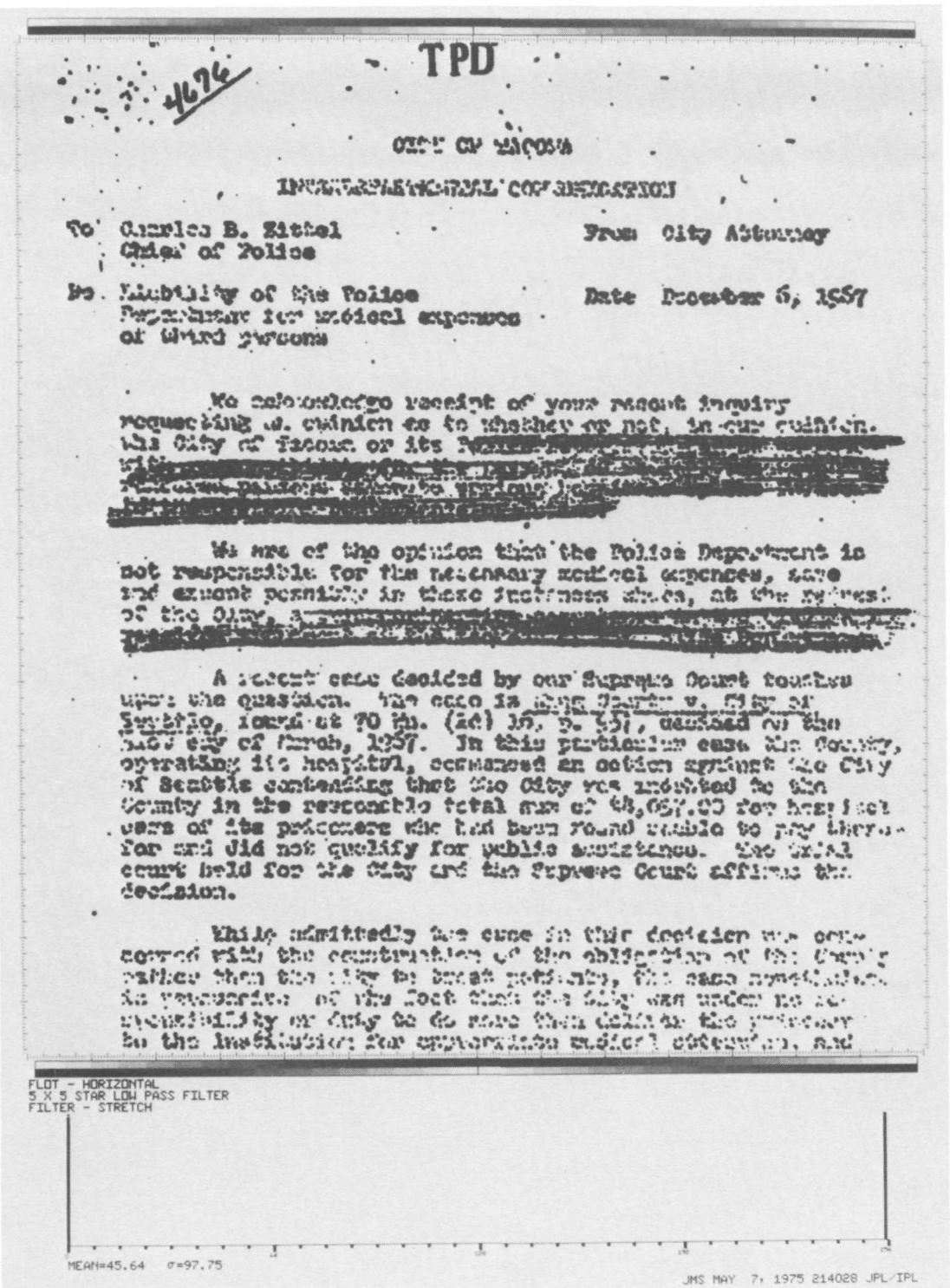
15. Mississippi document. Writing and noise are not as easily distinguished. Here a short perimeter value was used and, while almost all noise has been removed, so has much of the inscription.



16. A smaller perimeter value results in greater retention of inscription data, but more noise remains also.



17. Mississippi document after filtering to increase connectivity, followed by noise cleaning. A fraction of the edge-sharpened product of Figure 13 was superimposed on this final version.



18. Low-pass (smoothing) filter applied to the Tacoma Police Department document in an attempt to increase connectivity.