Journal of the Arkansas Academy of Science

Volume 44 Article 32

1990

Image Enhancement of Faded Historic Documents

Robert Crisp Jr.
University of Arkansas, Fayetteville

Barry C. Wells
University of Arkansas, Fayetteville

GENERAL NOTES

IMAGE ENHANCEMENT OF FADED HISTORIC DOCUMENTS

The need to preserve historic documents is generally recognized and few would deny that these documents become more valuable with the passage of time. Unfortunately, as time passes, the document becomes faded and/or tattered making it less and less legible. Methods are available to increase the life of the original, but these often inhibit access. Facsimiles are often used to circumvent this problem. But many facsimiles procedures have limitations; for example, a first generation copy is better than a second generation copy and successive copies lose information. Digital (or computerized) copies overcome many of the obstacles encountered with the present methods of preservation.

"Digital computer image enhancement has proven highly useful in numerous fields. Remote sensing, astronomy, and medical diagnostics in particular have experienced quantum leaps in performance through the introduction of this technology". (Asmus, John F., "Digital Image Processing in Art Conversion", *BYTE*, March 1987, p. 151). With a digital copy, the document can be viewed as having an extremely fine grid of points. These points or dots (often called pixels) can be turned on or off in a simple "line art" copy, or they may have different intensities, or gray scales, to provide a copy of black and white photographs. Finally, they may have both intensity and color to provide a full color copy. These computer generated copies have several advantages: 1) they do not exhibit degradation over time other than the risk (which can be minimized) of losing the copy or corrupting the data; 2) a tenth generation copy is as good as a first generation copy; 3) they can be easily exchanged by using electronic media or electronic transmission; and 4) most importantly, they lend themselves to computer indexing and retrieval so they may be retrieved in milliseconds and viewed on video screens.

Traditional methods of image enhancement involve "gray scale reversal, stretching of gray levels, modification of the histogram to the desired function, and so on". (Chellappa, Rama and Alexander A. Sawchuck, *Digital Image Processing and Analysis*, Vol. 1, 1985, p. 426). Several areas of enhancement have emerged. These consist of optical character recognition, satellite imaging and medical photography, to name a few. As most of these areas are unrelated, what is available today is a collection of algorithms and functions which may or may not apply to a given image.

A primary obstacle of digital copies has been the very large amount of data (and associated cost) required for a copy. Several developments have occurred which lessen this problem: 1) computers are much faster and the speed is doubling about every 2 years; 2) the cost of storing digital data is dropping by half every 2 years. The development of data compression algorithms also significantly reduce the data storage requirements. Taken together, digital archives are a viable alternative today, and a near certainty for the future.

This paper explores the unique problems associated with digitizing old documents that contain handwritten script. Additionally, a method is suggested for accurately capturing the image data.

This project consists of an interdisciplinary approach to providing an alternate method of document preservation. Several branches of the humanities, coordinated with computer science and engineering at the University of Arkansas in Fayetteville will be involved.

The proposed project task force is centered around a newly formed History Technology Laboratory at the University of Arkansas. This group consists of a board which contains representation from engineering and the humanities. The humanities reprsentatives make up a majority of the group. History, library science and literature represent some disciplines within the humanities in this group.

Many books and documents address the topic of image processing. Most of the existing image processing systems are expensive and difficult to access and use; however, many of the techniques and algorithms used in existing systems and described in current books can be utilized in this project.

The National Archives and the Library of Congress are developing a similar type of imaging system. Recent contacts with personnel at the Library of Congress indicate that there are some complexities in their system. These larger systems typically are few in number and complicated to use. One of the goals of this project will be to produce a prototype which will be cost effective and easy to use. The prototype should have the capability of recovering difficult to read text and producing a facsimile which would be representative of the original document.

This work was performed on a 386 based machine with a high resoluton (1,024 X 768) monitor. The output is from a 300 DPI HP Laserjet. Higher output densities are available but are not yet implemented. Alternate storage devices (tape, optical disk, etc.) are being evaluated and efforts are underway to find or create an image database front end.

A typical old document contains nothing other than freehand script. Typically, the quality of the paper and the legibility of the text vary throughout the entire length and width of the document. Thus, some parts of the page will be clearly legible while other parts may be totally illegible. Refer to Fig. 1. "A basic tool utilized in performing subjective enhancement and image analysis is the image histogram. The histogram reveals the distribution of digitized intensity within an image....". (Green, William B., *Digital Image Processing, A Systems Approach*, 2nd ed., p. 64). Since each document has such variations as described above, difficulty often exists when applying a single algorithm to the entire histogram. Any enhancement scheme which applies to the entire document must take this into account.

Existing systems are rarely applicable in this situation. Optical character recognition, satellite and medical imaging contribute concepts and general guides to script enhancement but do not provide a total solution. Algorithms which affect the entire page can be utilized to improve the quality of a document to some degree. The algorithms utilized on the letter in Fig. 1 were (in order): 1) adjusted brightness, 2) adjusted contrast, 3) performed linear equalization and 4) sharpened the image. The algorithms produced Fig. 2. Notice that there are 2 horizontal areas splitting the document into thirds. These are folds in the paper. To make matters worse, at some time in the past scotch tape was applied to approximately two-thirds of these areas starting at the left edge. Figure 3 displays a blow up of part of the lower band. (The vertical and horizontal measurements were altered automatically by the software in an attempt to fit the picture on a single page.) Figures 4 and 5 represents a best attempt to clarify the words underneath the scotch tape. The letters barely become readable in Fig. 6.

The intent is to create a proximity algorithm to locate and enhance line segments. Figure 7 shows the desired end result. Lines will typically have characteristics which distinguish them from other marks. Lines will normally be of a consistent gray level in a continuous fashion and further will not be "very wide". These features will allow the algorithm to discern the difference between the background and the line.

The first 75 (¼ inches at 300 DPI) rows of pixels are read into the computer memory from the image file. The entire document cannot be read at once due to hardware and software requirements. The algorithm must then perform 2 functions. The first executes the necessary steps required to locate a line segment. The second "traces" the line segment and records the locations of those pixels which are recognized as part of a line into a table. The remaining pixels are then turned to white. The information is written back out to disk after performing these 2 steps. The script may then be enhanced through the use of readily available software by adjusting brightness and contrast levels.

Script detection on older manuscripts has been discussed with relationship to existing enhancement systems. In older manuscripts, existing algorithms proved inadequate. The need for new algorithms led to the formation of an History/Technology Laboratory. The purpose of the laboratory in regards to this project was to establish algorithms which are specific to script detection. Investigation of present technology led to the development of a proximity algorithm which locates and "traces" a line segment. This paper also included an example of an enhanced document.

ROBERT M. CRISP, JR., Computer Science Engineering Department, Engineering 313, University of Arkansas, Fayetteville, AR 72701 and BARRY C. WELLS, Acxiom, Inc., GENESIS Program, University of Arkansas, ERC 214, Fayetteville, AR 72701.

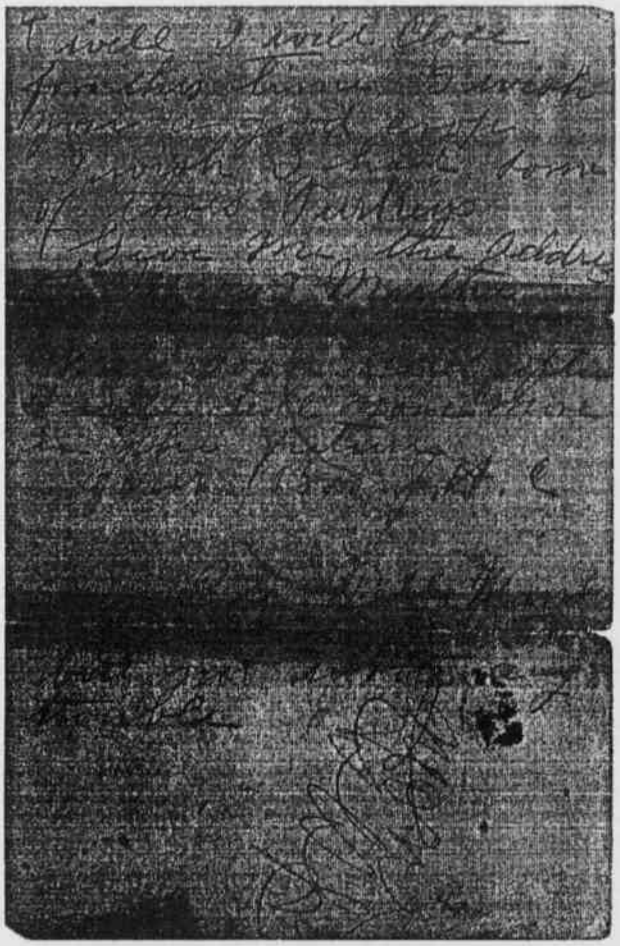


Figure 1. Original document produced in 1900

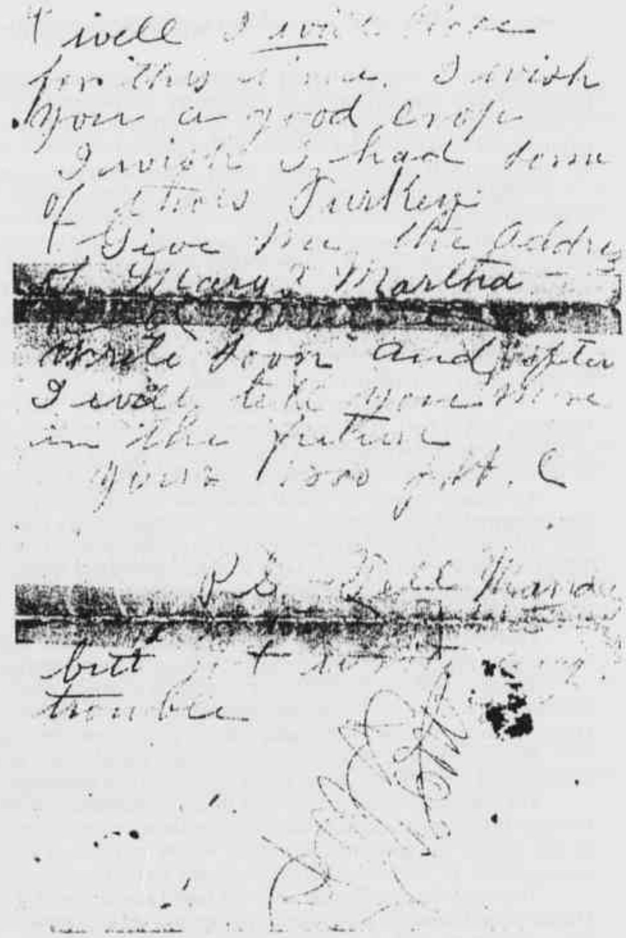


Figure 2. Original document after adjusting brightness and contrast, applying linear equalization and sharpening the image.

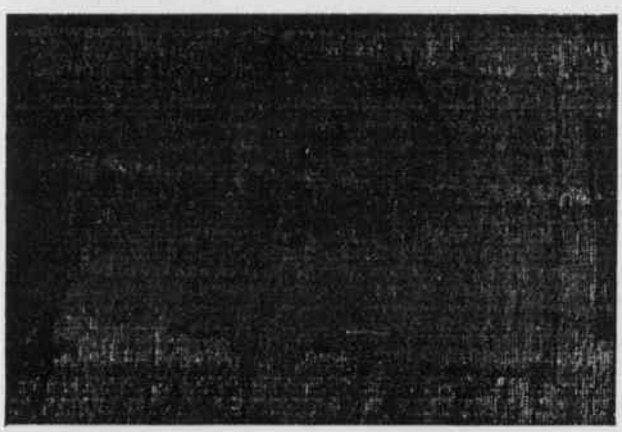


Figure 3. Small section of original document

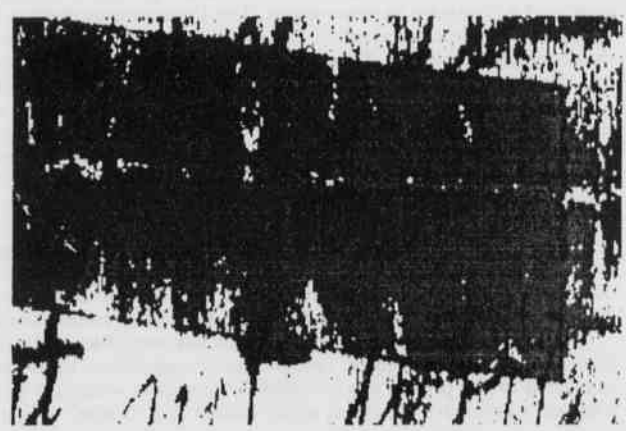


Figure 4. Figure 3 after applying threshold filter

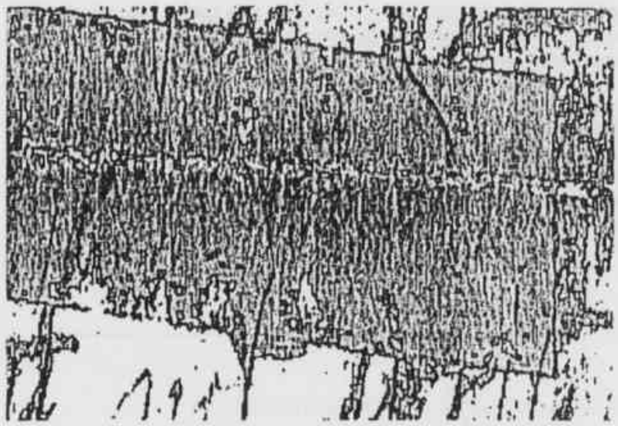
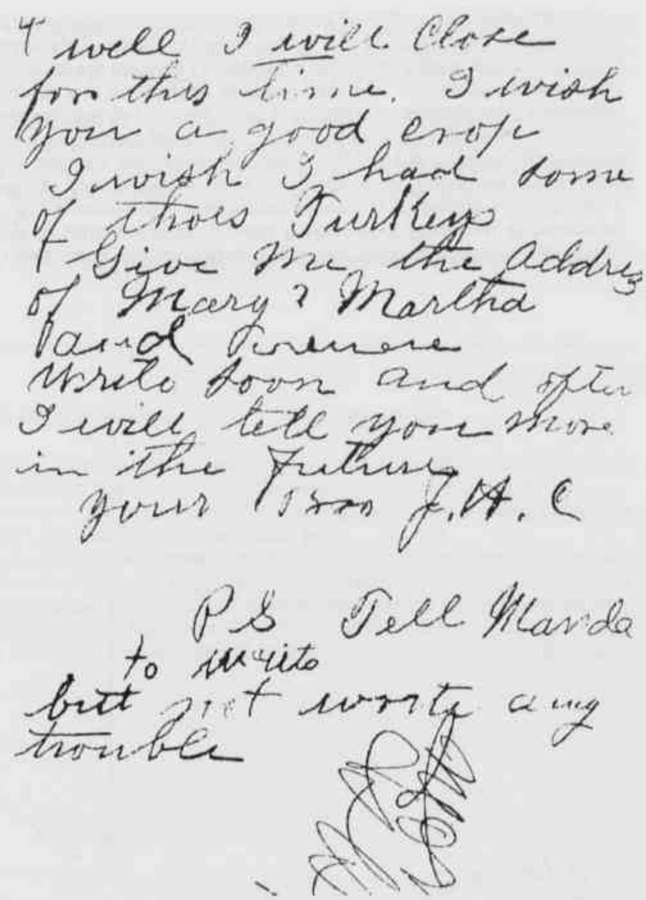


Figure 5. Figure 4 after applying equalization and sharpness

Figure 6. Figure 5 after applying contrast and brightness



+ well I will Close
for this time. I wish
you a good crop
I wish I had some
of thoes Turkeys
+ Give me the addres
of Mary & Martha
and Lorenore
write soon and often
I will tell you more
in the future
your Bro J.A.C

P.S. Tell Marda
to write
but not worry any
trouble

Figure 7. Desired end result after applying proximity algorithm

THE IMPACT OF MICROPROCESSOR PROTECTED MODE PROGRAMMING ON UNDERGRADUATE EDUCATION IN ENGINEERING TECHNOLOGY

The purpose of this paper is to examine some of the rapid changes that have recently occurred (and are continuing to occur) in microcomputers and the impact of these changes on both faculty and students in engineering, technology, and related fields — such as physics, biology, mathematics, and chemistry.

A microcomputer is any computer that uses a microprocessor for its central-processing-unit (CPU). In fact, a "microprocessor" can be described or defined as a CPU on one integrated circuit or "chip." Because of their small size and low cost, microprocessors, which were first developed in 1971, have revolutionized the computer industry. Before the advent of microprocessors, all computers were generally classified as main-frame or mini computers. Since main-frame and mini computers are usually very large and expensive, their resources are almost always shared by several users. By using microprocessors for the central processing unit, manufacturers were able to develop much smaller and less expensive computers. By 1981, improvements in capabilities of microprocessors led to the development and introduction of the now famous IBM Personal Computer (PC). The PC was different from existing mini computers and main-frame computers because it was intended for use by one person (single-user), and the software operating systems developed for the PC's (PC-DOS and MS-DOS) further limited the PC's to one user application program at a time (single-tasking). Performance improvements in existing characteristics such as operating frequency, address and data bus size, and chip integration are considered *evolutionary* changes in microcomputers. For example, the operating frequency of microcomputers has increased from one megahertz (MHz) to 33 MHz in the last 10 years. While this is a significant increase in operating frequency, 33 MHz is not even close to the current state-of-the-art in supercomputers such as the Cray computers which operate in the 300 MHz range. Of course, the Cray does not use