AUTOMATED PROOFING OF LARGE PHARMACEUTICAL DOCUMENTS FOR MINUTE FLAWS[1]

Michael Negin, Ph.D.
MNEMONICS, INC.
Mt. Laurel, New Jersey

ABSTRACT

Proofing of large pharmaceutical documents is a formidable challenge requiring hours of human proofreading time. This presentation will include a systems approach to the solution to this problem through the use of computer image processing. These large documents require the processing of images up to 600 Megabytes and to be able to find all flaws within a few minutes. The presentation will review the requirements for such a system and will present a practical solution along with some performance results.

MNEMONICS' approach to the automation of proofreading is to provide a computer vision replacement for the human visual system so that minute flaws are efficiently found with high detectability performance, i.e., high specificity and high sensitivity. Once this is achieved, a human inspector can then quickly review the resulting list of flaws.

This paper describes a system for automated proofreading that can find the minute flaws in a 24 by 24 inch document (scanned at 600 dpi) in less than one minute. This performance is about 95% faster than human proofreaders. The system also compares favorably to other computer automated proofing systems by factors on the order of five to ten times speed improvement with improved detectability.

For reference in this paper, the system that has been designed and implemented by MNEMONICS is ***AVIA*** – Automated Visual Information Analysis.

[1] Presented at the Automated Imaging Association's Machine Vision for Pharmaceutical & Medical Application Workshop, October 14, 2004, Princeton, NJ.

OVERVIEW

INSPECTING LARGE IMAGES FOR SMALL DEFECTS

In order to understand the difficulty of the inspection problem, it is helpful to understand the magnitude of the problem faced by human proofreaders and the requirements for a machine vision system.

Human Proofreading

- **Document and image sizes** can range up to 36 by 48 inches sampled at 600 dpi. This corresponds to images that are about 600 Mbytes.
- For pharmaceutical product information sheets, there can be upwards of 200,000 characters on a 36 by 48 inch sheet.
- A human being is typically asked to find a partially or completely missing decimal point. This corresponds to finding a defect of about 12 pixels in size or about 1 part in 50 million (1:50,000,000)
- This process takes a human inspector about 4 hours or more to complete. The task is daunting in that the human must remain alert, vigilant, and must pay attention to the most minute details. Defects on the order of ½ of a decimal point are very easy to miss, and require very highly skilled human proofreader.

Machine Vision Proofreading System

- The inspection must be automated in terms of finding differences, i.e., the human must be out of the loop during the flaw finding process. The machine vision system must replace the human visual system during this part of the process.
- The inspection process must present suspicious areas to the human in an orderly and logical fashion so that final decisions as to defect severity are easily performed and logged for report generation. Visualization tools to assist the inspector must be intuitive and easy to use.
- The system must provide an extremely high detectability performance, i.e., a very high true positive rate and a very low false positive rate. Without a high detectability index, the human could be subjected to as much work as manual proofreading if many false positives are called out for review.
- Images are large – up to 600 Mbytes. This corresponds to 2,000 times the size of a standard machine vision camera image (~300,000 pixels)
- Flaws are as small as 12 pixels are approximately ½ the size of a decimal point (at 600 dpi) and must be reliably detected.

- The source material typically is obtained from a press sheet that is printed on very thin paper stock. When the paper is taken off the press, the sheets can relax and shrink non-uniformly. This causes numerous local distortions whose position and extent cannot be predicted in advance. This means that standard linear image processing methods cannot be successfully applied.
- Ink bleeding through from the reverse side of the document is a frequent occurrence. This causes numerous artifacts that can easily be misinterpreted as flaws.
- Multiple media types. The sheets to be inspected have a variety of potential sources against which they have to be compared. The source documents can be from electronic files, printed proofs, printing plates, blue line documents, and other printed sheets. An automated system must be able to handle any of these reference sources.

SYSTEM CONSIDERATIONS

The following three figures show the ***SYSTEM ORGANIZATION***, a ***SAMPLE OUTPUT***, and ***DOCUMENT PRODUCTION FLOW***.

The ***SYSTEM ORGANIZATION*** shows each of the major components. Conceptually, the system is straightforward in its configuration. The complications arise in the details of how each part of the process is implemented and then how the system functions as an entity.

The essence of the system is to have a representative master or standard document for comparison. The **standard document** can be an electronic file such as might be generated by a word processing package or graphic design package that can produce a PDF, PS, or EPS file. In this case, the electronic file must be converted and altered in appearance to include the variations that would be expected if the document were printed and then scanned. Without these conversions, the electronic files are of little use as standards. Issues such as character appearance or character placement vagaries must also be considered, as these are frequent occurrences in electronic documents (i.e. what you see on the computer screen is not always what you get on the printed page). Physical media can also be used as **standard documents**. Various physical media can include printed proofs, printing plates, positive and negative transparencies, blue lines (blue prints), and press sheets. These **standard documents** are then used to compare against the **sample documents**. Other complicating factors are that the sample documents may have multiple standards (i.e., multiple source documents) that were used to generate the printed document, and there may be multiple copies of the source or standards on the printed page.

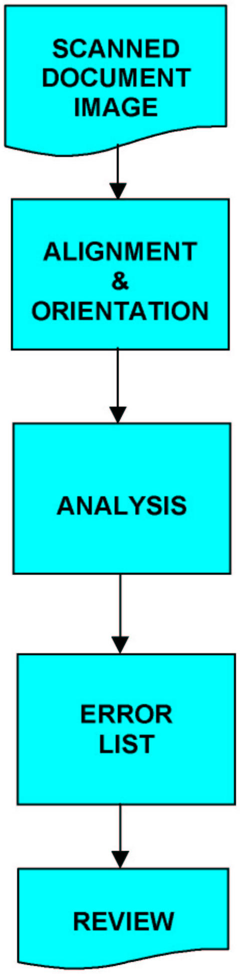
The system must automatically correct for brightness variations, internal document distortion, ink bleed through, rotation, translation, and stretch of the documents.

The system must also have an efficient means for producing reports for further review and archiving. ***AVIA*** provides reports that are compatible with Internet browsers that can easily be shared on the Internet and intranets. Reports can also be generated as PDF documents.

The ***DOCUMENT PRODUCTION FLOW*** is a critical aspect of the successful operation of the entire system. A critical step in performing successful proofreading is to be able to qualify a **standard document**. This is shown in the flow chart.

SYSTEM ORGANIZATION

Sample image to be proofread
AVIA Document Scanner
AVIA
PC & Software



A large format scanner is used to read in the product information document.

AVIA's specialized proprietary software aligns and orients the document and the reference document. The software also corrects for paper stretch and bleed through.

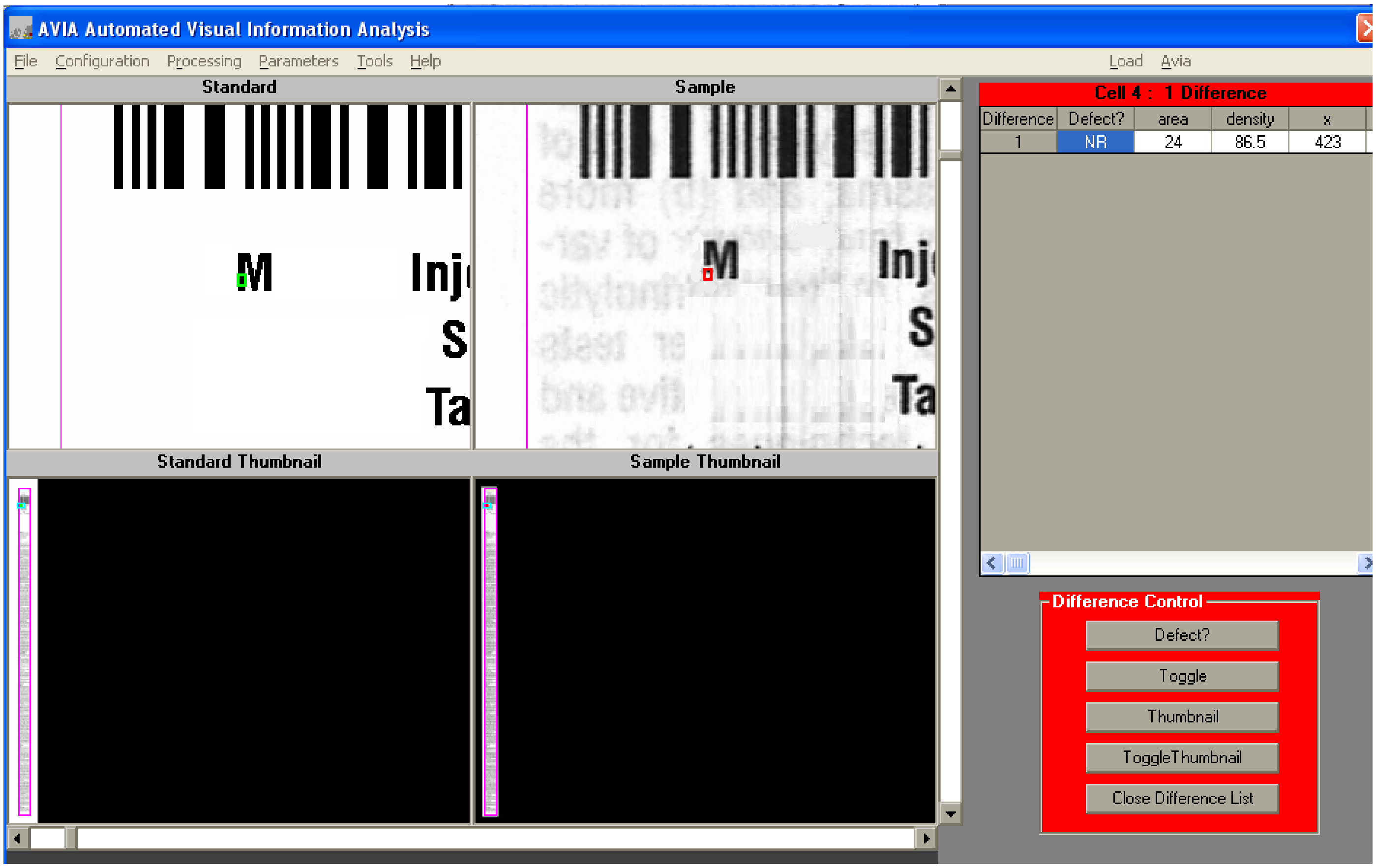
The system inspects the characters in the document as well as graphics, chemical formulae, charts, and pictures. All disparities greater than a user-specified size and intensity are located and marked as flaws.

AVIA ranks and tabulates flaws. An Internet compatible (HTML) document is produced for human review and subsequent document management.

A human operator reviews the generated error report, annotates the report, and initiates corrective action as needed. A final report can be generated for record keeping.

Sample Output

The panel below illustrates a sample output screen shot of a small flaw that ***AVIA*** has detected. In this case, the sample was a multiple set of copies of the standard document. This panel shows Cell 4 of the twelve cells that were inspected. Note the small region of the “M” that is missing the sample. In this case the Standard image was derived from a PDF file and the Sample image was obtained from a scanned press sheets.

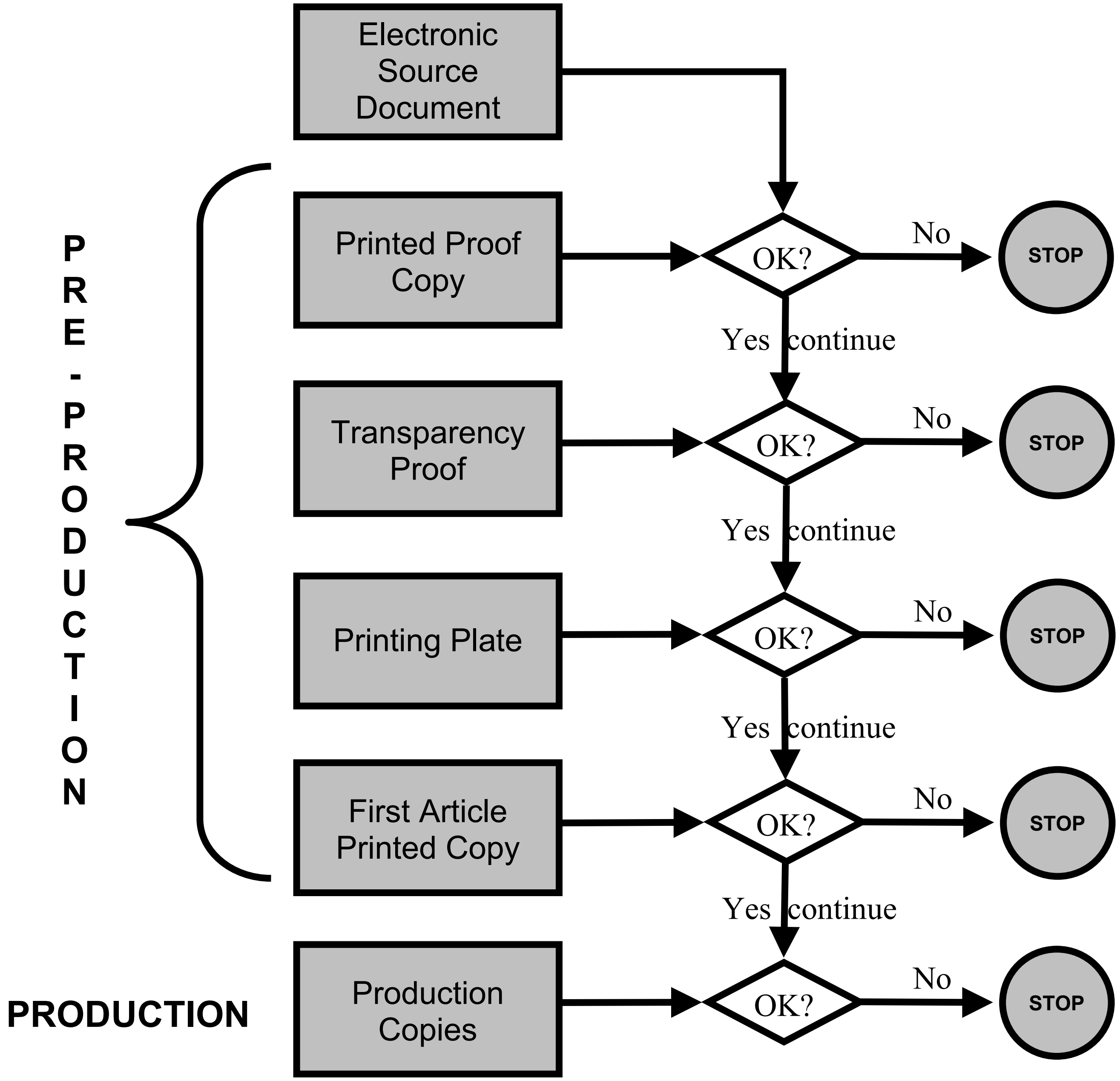


Cell 4 : 1 Difference

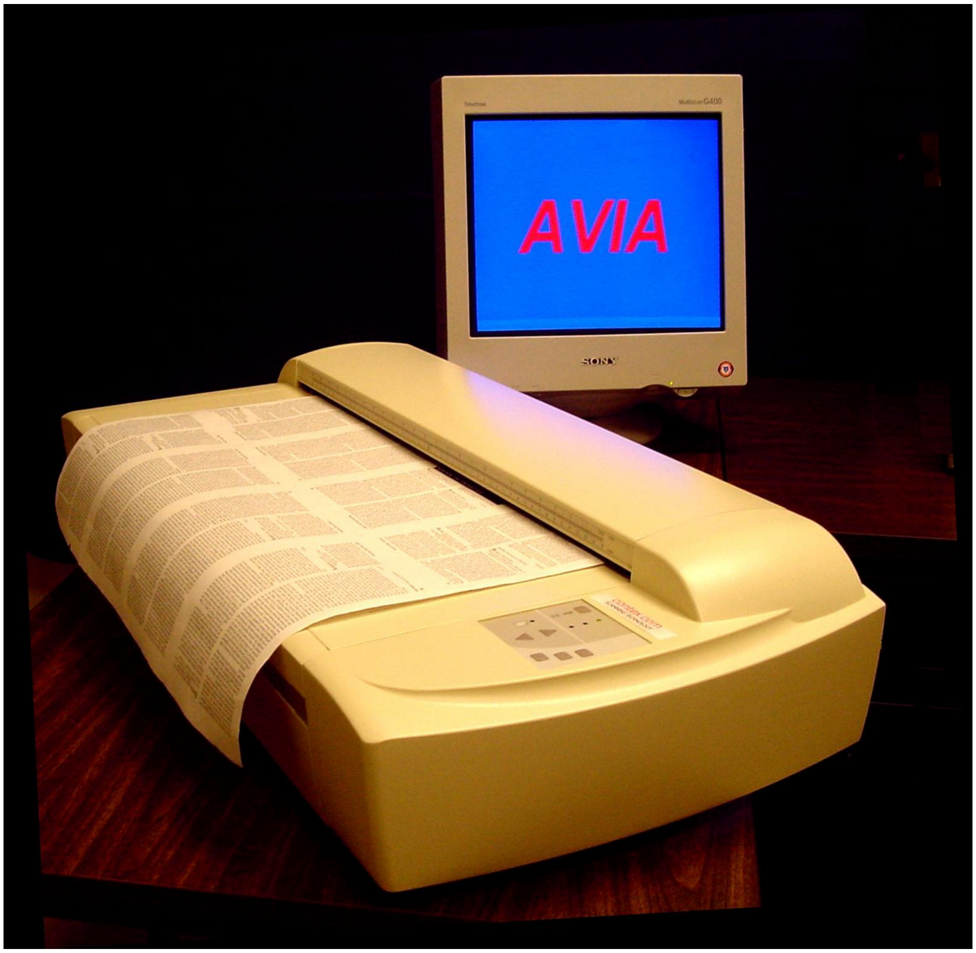
Difference	Defect?	area	density	x
1	NR	24	86.5	423

Document Production Flow

AVIA can be used in all phases of production document generation, from pre-press electronic document comparison and preparation to print quality inspection during production runs.



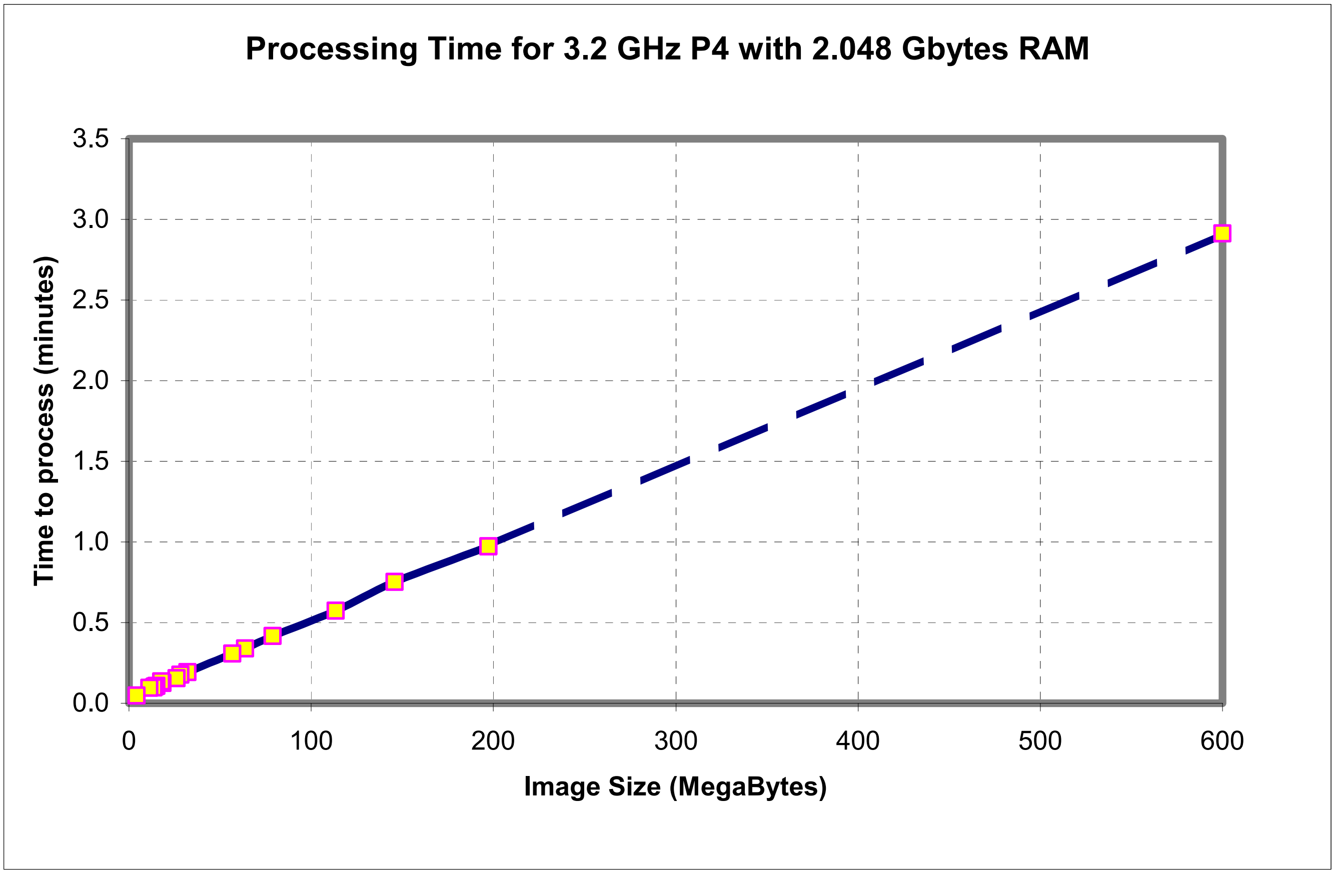
The picture below shows an ***AVIA*** system with a 25-inch wide scanner.



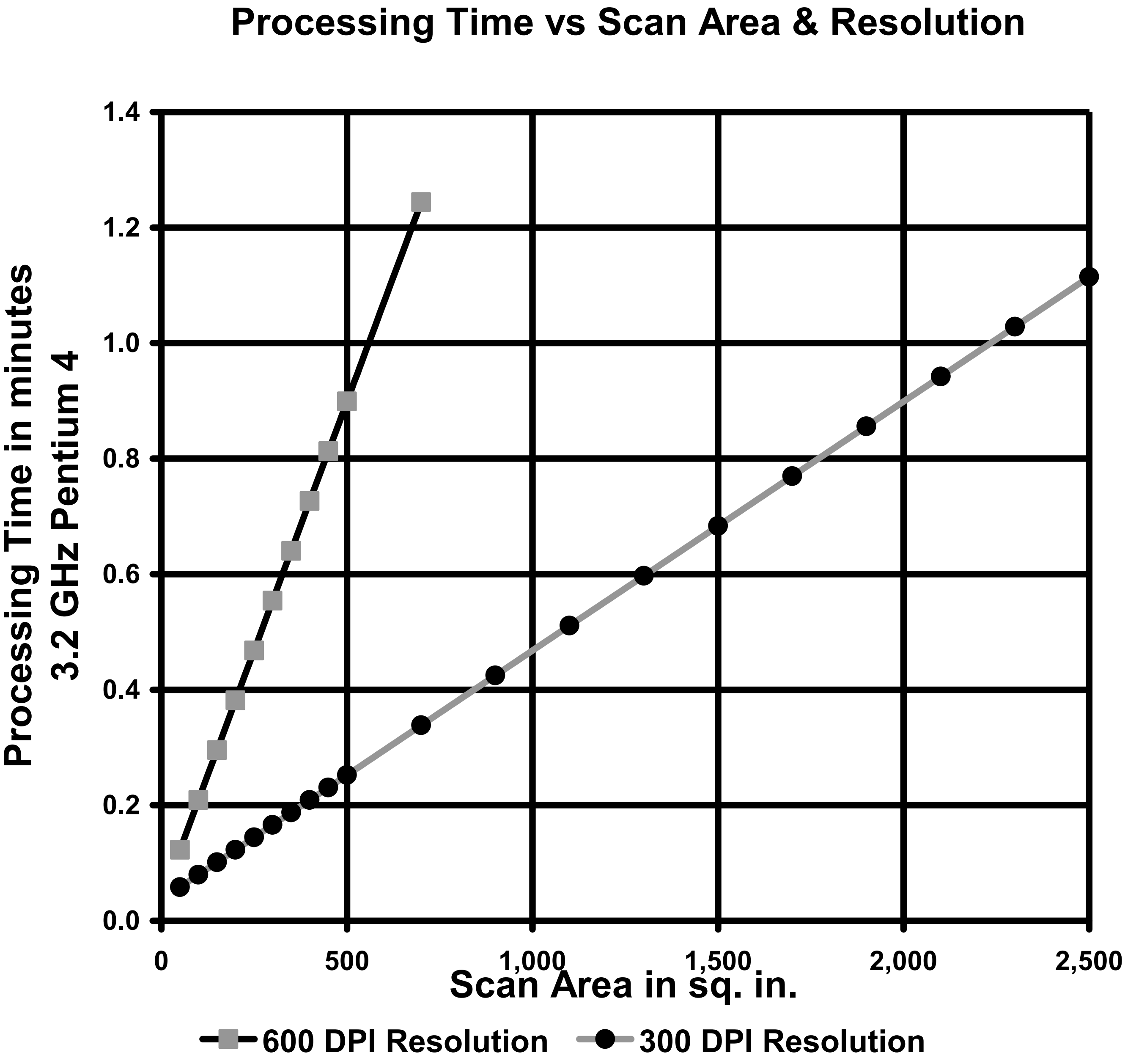
SYSTEM PERFORMANCE

The chart below shows system performance as measured for images up to 200 Mbyte and is extrapolated to 600 Mbyte images. The system has been benchmarked on a variety of WINTEL processors and is essentially linear with processor speed. As processors increase in speed, the process time will be proportionally decreased.

The software is designed to be compatible for use with parallel processors. This means that proportional speed increases are also available by using more processors. This is probably not required for off-line applications, but it is expected that on-line proofreading will eventually be in demand.



The chart below shows document size in area units vs. Processing time. From this chart, one can estimate the performance of the computerized proof reading process used by ***AVIA***.



CONCLUSION

From a proofreading efficiency and throughput perspective, this paper discussed the rationale and justification for the development of an automated proofreading system capable of detecting minute flaws for large pharmaceutical documents. This was based primarily on the difficulty of the proofreading task for human proofreaders.

In addition to the human efficiency considerations, there are other reasons for the introduction of automation into proofreading, such as:

- Improved consumer information and consumer safety resulting from a better quality product. Improved quality of pharmaceutical documents.
- Provide a common supply chain tool for use throughout the production cycle. Since ***AVIA*** is a software comparison engine, it can be used in document development, initial proofing, verifying that printing plates are correctly produced, checking transparencies and first article production, checking press sheets, and inspecting product inserts and outserts.
- Lower cost of document production. With the introduction of automated proofreading of printed production documents, a sample inspection from the printing line can be practically realized. Proofreading samples of production documents every 10 to 15 minutes can trap printing errors and reduce waste as well as prevent defective product from being shipped and then become subject to recall.
- Shorter development times for product information documents. By having electronic automation in the proofing process, many of the current tasks that require transfer of paper documents can be eliminated. By exchanging information electronically along with automated proofreading, the document development process can be reduced on the order of 80% simply by eliminating physical document transfer.
- Increased potential for on-line print quality inspection of large documents.
- Reduction of potential product liability by preventing defective product information sheets from being distributed.
- Ordinary human proofreaders can be transformed into excellent proofreader by virtually eliminating the human vision discrimination task.