OCR, MAGICBOOK AND MAGICMATH: A DESCRIPTION OF THE TECHNOLOGIES AND AN EXPLANATION OF THEIR INDUSTRIAL APPLICATION

William J. Ray, Ph.D.*

KEYWORDS: OCR, Typesetting, Neural, Network, Prepress

ABSTRACT

The conversion of printed material to digital data has been heavily studied from, particularly, the perspective of pattern recognition and image signal processing. It is the purpose of this research to briefly examine two new approaches to optical character recognition (OCR) production systems: 1) a post scanning, post OCR application that provides a perfect or near perfect conversion of printer or offset printed derived originals to page described, text editable digital data and 2) a neural network based system that provides highly accurate conversion of geometrically complex documents..

* Group InfoTech, Inc., East Lansing, Michigan. wjr@gat.phm.msu.edu

INTRODUCTION

This document examines the use of Optical Character Recognition (OCR) systems in the work flow for commercial reproduction of books and documents. We examine, specifically, techniques to enhance OCR for reliable, inexpensive use in the reproduction of back list publishing.

THE RATIONAL OF CONVERSION

This is an age of conversion. Society is converting from analog based data storage technology (e.g., paper and film) to digital storage technology. How data are *delivered* to the consumer is less impacted by the storage technology (and, of less interest to us here) than the manufacturing process of getting data to a *deliverable*.

There is good reason for this conversion. Digital origin of delivered data provides several benefits to the data provider, among these are:

1) Uniformity of data storage -- Essentially all newly created text based and image based data now arise as digital originals. This presents two possible logical storage fates for such data -- one can either store all data (or the significant subset of data to be retained) as digital files or all data as physical files (paper or film). We argue later that the most cost efficient storage form is digital. However, while it is fairly obvious how one converts a file into paper, it is not so obvious how one reliably converts exclusively analog data into digital files.

2) Accessibility and retrievability of data -- Paper or film storage and retrieval are well understood, however, such storage is expensive as both physical environmental and space cost requirements are significant. Further, the granularity, cost and timeliness of retrieval is nontrivial. For those areas where re-use of data is contemplated, exclusive use of digital data -- specifically editable text data as opposed to page image data -- is more cost effective than either mixed data types or exclusively analog storage of data.

3) Data reuse and re purposing -- Clearly, one of the driving factors in the digitizing movement is data reuse and data re purposing. In the publishing industry alone the ability to reuse data means that the constraints of the mechanical printing process are significantly lessened. Choice of press -- or even the use of a conventional press -- is generally a run length cost versus cost per folio analysis. Starting from an essentially zero prep cost (from the second printing on) allows the publisher to contemplate profitable, very short run printings.

DIGITAL IMAGE VERSUS DIGITAL GEOMETRY

It is common currency in document processing technology to accept page images as opposed to text geometry pages (e.g. page described, text editable pages) as the end product of a conversion process. From the printers perspective we might consider this process analogous to a digital "opticopy". This document is still, fundamentally, an "analog" document as it cannot be altered other than by the use of image manipulation tools. Further, its' value in re purposing is limited by being an image.

This technique is cumbersome from the perspective of the printer as whatever errors exist in the image cannot, practically, be fixed. Further, the documents type quality is subject to the quality of the scanning process and, by extension, subject to the quality of the original from which the scan was made. Generally, these constraints do not produce type quality levels that are acceptable to even the poorest of modern mass printing techniques.

Finally, work flow considerations mitigate against an image approach. In the PostScript process, assembling a two hundred page book from image data, while somewhat simplified by being only images, is time consuming and awkward.

TEXT CONVERSION TECHNIQUES

There are only two real methods to digitally capture existing textual data -- rekeying and Optical Character Recognition (OCR):

Rekeying is the obvious first pass solution as it requires no particular extension to standard word processing and typesetting technology. Expense, however, is an issue. The actual rekeying cost can be significantly decreased by the use of off shore labor and by employing a dual entry work flow. Dual entry allows an efficient first pass error check by comparison of the resultant files for difference. The underlying assumption, here, is that the probability of overlapping error is low.

Under all circumstances, however, re-entry techniques require re-proofing and re-typesetting of the document. This generally means that whatever we save by off shore re-entry work is lost (and then some) by the manually intense process of validation. Typical final document error rates are similar to those found in original typesetting -- e.g., about 1 error per 100 pages (or .0005% on a character basis assuming an average of 2000 characters per page).

OCR techniques offer an alternative to the re-entry work flow by supplanting the manual re-entry process with character recognition software. At first blush this would seem to be an advantage as labor costs are reduced and timeliness enhanced by not needing off shore work. However, the very high intrinsic error rate associated with standard OCR techniques introduces an editing nightmare that results in the simple OCR process being much more expensive than keyboarding.

THE PROBLEM OF OCR

The problem of high OCR error rates is illustrated by a simple calculation. If we assume a *correct* conversion rate of 99% on high quality originals then, on an average page of 2000 characters, about 20 character errors will occur *per page*. This presents a formidable problem in the editing process and sets up the proofreading step as an even more critical and expensive stage.

Rice et al (1) in the 1995 ISRI accuracy report of OCR engines notes that, on the 817,946 character Department of Energy (DOE) test sample, the best conversation rate was only 96.62% correct conversion. The DOE sample represents a random cross section

of documents classified into quality types ranging from 1 (best) to 5 (worst). Typically, a quality of 1 represents a first generation document on bright white paper with even letter spacing (e.g. no kerning) while quality type 5 represents an "n"th generation document with faded or broken type. Typically the largest number of character errors (up to 70%) occurring in a DOE sample conversion are found in pages of quality type 5.

Typeset documents of the type generally seen in backlist publication generally range from type 3 to type 5 in quality . Conservatively, this represents an error range of about 40 to 180 character errors per page.

Error analysis by character error, however, leaves out a significant element of the conversion of formatted documents -- e.g. the page geometry data contained in the document. Errors here are at least as significant as those of spelling and punctuation. Manually recreating the format and type specifications of a document essentially places the publisher in the position of saving only a fraction of pre-printing production costs as opposed to publishing a new document.

The purpose of converting the back list via OCR is to leverage the value of as much prior prepress work as possible while adding as little increase in overhead as possible when compared to the traditional analog (i.e. page opticopy) methods. Thus a document can print at each press run as if it were a newly typeset document which no generational loss in type quality.

Given these data, we approached the problem from two perspectives; 1) a near term technique leveraging existing recognition technology with a post processing step (MagicBook) and 2) a longer term technique utilizing an entirely new method of recognition (MagicMath)

THE MAGICBOOK TECHNIQUE

Figure 1 is schematic of the MagicBook Technique. The SCANNING step includes a modest amount of signal processing (e.g. de spoting and de skewing), however, it was found that much effort beyond this was not profitable. The OCR process

itself is a dual stream approach. Two commercial products are used to convert a single image file into two recognized rich text format (RTF) files. At the same time, a third pass is made through the image to capture certain geometric elements of the page structure.

Much as in the case of rekeying, it has been found that different recognition processes make different errors. Thus the post conversion MagicBook process compares the two resultant RTF files (the MAGIC COMPARE step of Figure 1) for spelling and unknown character symbols (commercial OCR packages either recognize a character or assign a little used symbol to unrecognizable characters). The difficulty here, of course, is synchronizing the text streams and parsing "in common" word tokens from that stream.

Once this is accomplished the texts are compared based upon a weighting table ("weight" in the figure) derived from characterizing the error types associated with each recognition process.

Two error types can result from the conversion process: type 1 or "knowable" errors -- e.g. errors that can be detected by simple spelling or positional rules, and type 2 or "unknowable" errors -- words that are spelled correctly but are not correct when compared to the original text and that cannot easily be detected from simple positional rules. One of the advantages of the weighted dual stream approach is that many type 2 errors can be mapped out due to know error proclivities of the particular conversion process. Thus, if one system is known to, say, have a higher probability of converting "cl" to "d" and that system converts a word as "down" when the other converts the same token as "clown" then we can reasonably assign a greater probability of correct conversion to "clown".

Weighting probabilities are not enough, however, to get a complete handle on the conversion process. Each word token needs to be analyzed within the context that it occurs -- much as if a human were reading it. We have applied a simple (about 65,000 rule) context tool using a method derived from language independent grammar (the MAGIC LIG step) to the text. This step recognizes that certain elements occur based upon

unambiguous rules. Thus, we can, as an example, mark italic and quoted segments based upon recognizable delimiters. We know that certain things must happen if something else has occurred and, while MagicBook cannot always correct recognized errors, we can mark known ambiguities.

The final step in MagicBook is the MAGIC EDIT process. Figures 2a and 2b are screen captures of this process. This step is user interactive and presents the operator with a formatted document showing various errors, probable errors or corrected errors in color code. This process was designed to be, as much as possible, a *permissive check* system i.e. we want the operator to confirm the correctness of changes that MagicBook has made by tabbing through the color element (thus changing the color to black type) rather than having to actively alter text or format. As the operator tabs through the page from color to color and the window displays the image of the original text.

The final product is an RTF file containing the converted text in the geometry and type face of the original document. Our objective, here, was a pragmatic engineering solution to text that was geometrically relatively simple and that, *in production*, required no proof reading.

Very high correct conversion rates have been demonstrated with this process -- ranging from 99.995% correct to 99.99975% (or 1 error in 100 to 200 pages) as measured by external three proof reader tests. Figures 3 and 4 illustrate a before and after for this conversion technique. Note that this page, used by permission of Academic Press, has a slightly different header and footer format due to client specification.

MagicBook is now a production application and has been used to digitize several multi book CD-ROMS, hundreds of backlist books and numerous historical documents for various libraries.

THE MAGICMATH TECHNIQUE

MagicMath is a completely new approach to OCR conversion. The original design specification deliberately attacked conversion of documents that were not usually considered good

candidates for OCR due to geometric complexity, non standard font usage (such as higher mathematics) and document systems using mixed language elements. At inception we considered this project to be "blue sky" research.

It appears that two general classes of recognition systems are commercially employed; feature detection type systems (Herz et al, 1994)) and neural network type systems. While several different neural network types have been used, most appear to be of the multi layer back propagation type (Romero et al, 1995).

On analysis we found that most OCR systems appear not to extract as much data from the scanned document page as might be the case. Both recognition techniques seem to have an optimal "notch" of scanned image resolution data of about 300 dpi. When presented with data of greater resolution these systems actually appear to diverge from a recognition solution rather than converge. Thus, our initial hypothesis was that by building a more complex neural network we might be able to characterize a large number of very high resolution scanned character token patterns and, thus, more accurately recognize similar occurrences.

As illustrated in Figure 5, scanning is done at either 600 or 1000 dpi and an explicit signal processing step is performed.

We employed a new type of self organizing network (NETWORK K in the illustration) to separate glyph elements. "N" dynamically assigned back propagation type networks bin similar glyph types.

The element ZADAH MACHINE is a fuzzy logical glyph to token assembly process that feeds the GLUE BOX process for word and geometry assembly. Resultant output (OUTPUT DRIVER) is currently in TeX, a mathematical typesetting language.

Figure 6 illustrates an image from a Mathematical Reviews abstract (used by permission). Figure 7 shows the same abstract having been recognized and converted to TeX. Figure 8 shows the abstract rendered by a TeX display tool.

Significant to this conversion technique is the fact the system regularly deals with type ranging from 6 points to 24 points on the same image. Further, the reader will note that there is no manual intervention in the process after scanning. Processing time is, however, significant -- at this time. An initial training period for the networks is required for type faces not previously encountered and not similar to those already trained.

As this is very new as of this writing we do not have enough data to explicitly claim an error rate. However, ignoring mathematical symbols that the system has not encountered before, once trained on a font class its' error rate appears to be *very* much less than that of MagicBook for type 1 documents.

CONCLUSION

An available, low error rate OCR production system has been described for geometrically straightforward documents. This system performs as an extension to off-the-shelf OCR recognition systems.

A new neural network based OCR engine has also been described. While not yet in production, this system offers the potential for conversion of not only simple but complex documents with very high accuracy and at extremely low costs.

The author would like to thank the engineering group at Group InfoTech, Inc. for an outstanding effort and our friends and colleagues at Mathematical Reviews for their interesting ideas and problems -- not to mention the research money for MagicMath. Thanks also go to David Strong and Ben Wong at RR Donnelley & Company for ideas and the patient analysis of the MagicBook process.

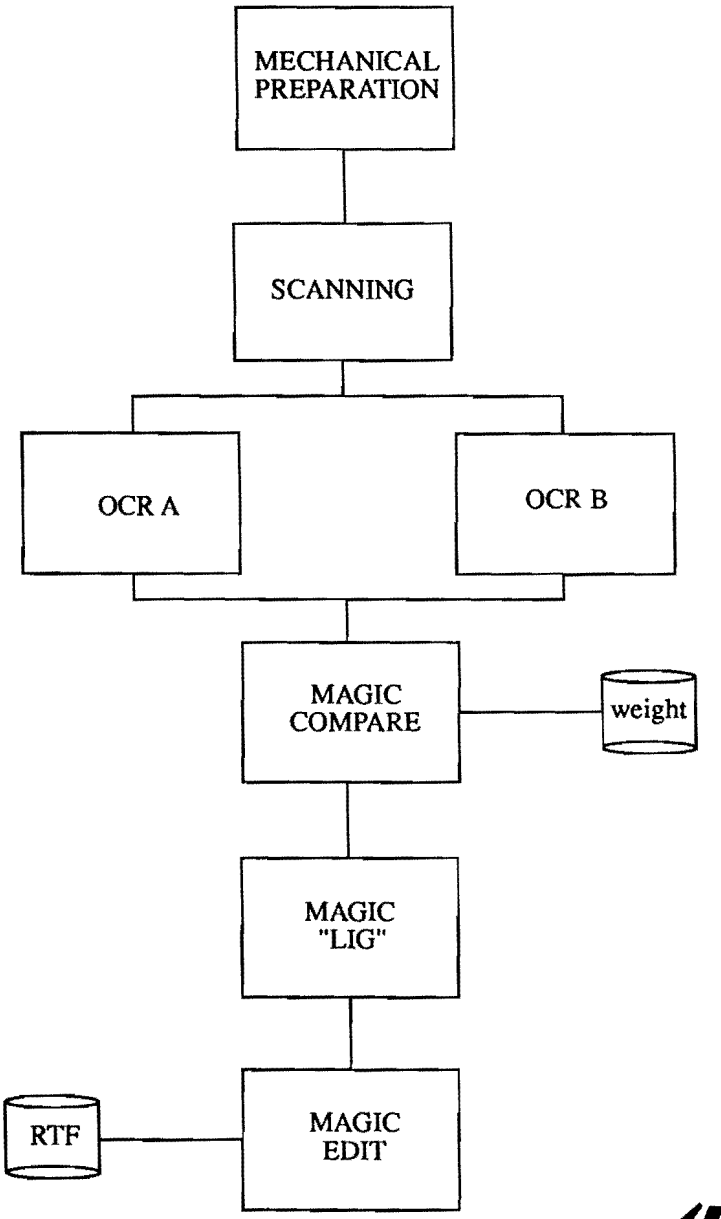
REFERENCES

1) Stephen V. Rice, Frank R. Jenkins and Thomas A. Nartker. The Fourth Annual Test of OCR Accuracy. UNLV Information Science Research Institute 1995 Annual Report, April, 1995.

2) Jacky Herz and Roger D. Hersch. Towards a universal auto-hinting system for typographic shapes. Electronic Publishing, 7(4), 251-260, 1994.

3) Richard Romero, Robert Berger, Robert Thibadeau and David Touretzky. Neural Network Classifiers for Optical Chinese Charactor Recognition. Fourth Annual Symposium on Document Analysis and Information Retrieval, Las Vegas, April, 1995.

Figure 1: The MagicBook Process



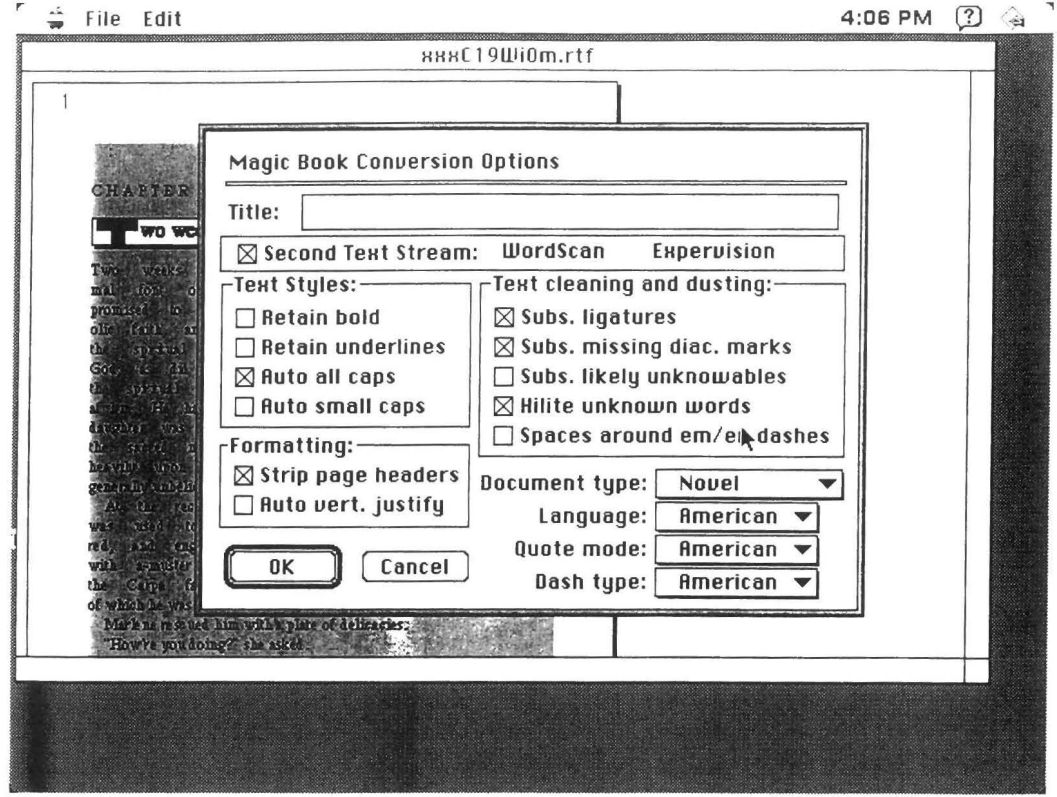


Figure 2a:
Screen Capture of Magic Edit
Showing Conversion Options

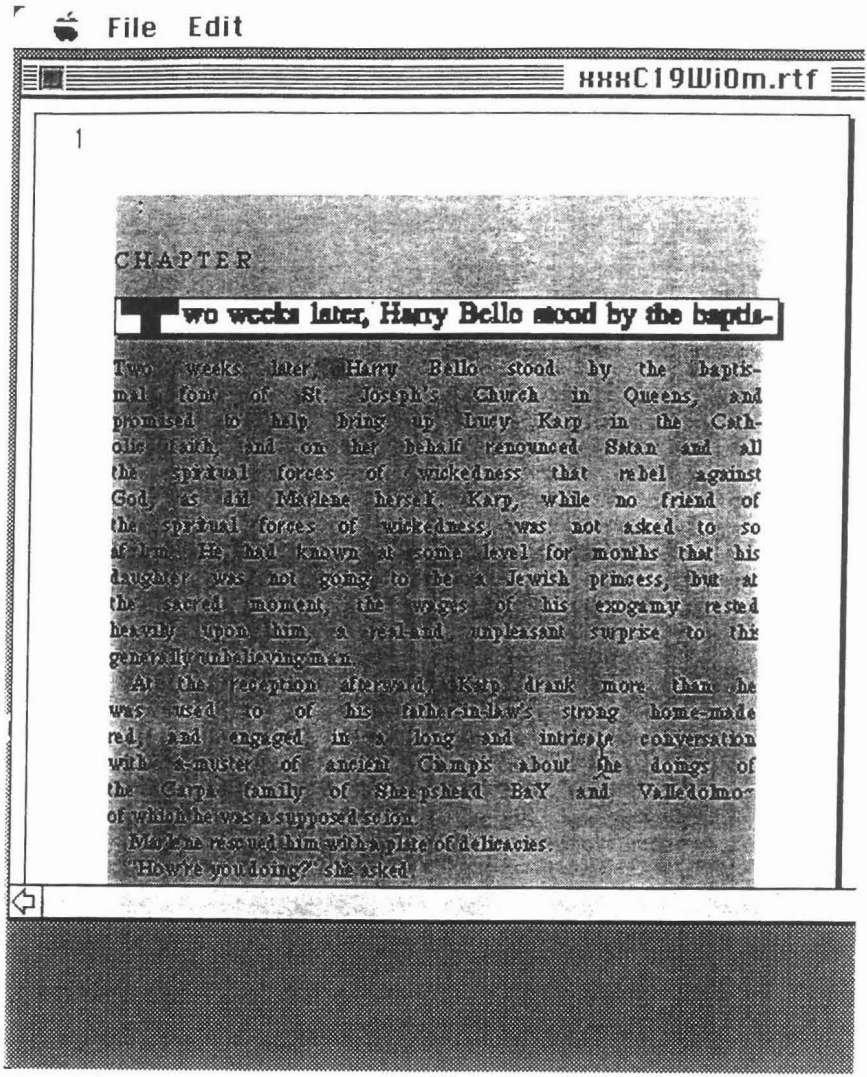


Figure 2b:
Screen Capture of Magic Edit
Showing Editing Process

distributions are, unfortunately, difficult to specify and do not have a standard form.

Expressing the goniometric distribution in terms of power per steradian, the goniometric diagram can be reformulated as a *maximum* power per steradian I_{max} scaled by a polar function ranging from 0 to 1. The polar scaling function $S(\vec{\omega})$ is defined to takes a direction, $\vec{\omega}$, away from the source and returns a value between 0 and 1.

Interpolation is required to obtain a value $S(\vec{\omega})$ from the goniometric diagram for a direction that does not lie on either of the two perpendicular planes defining the distribution. Languénou and Tellier [144] suggest the following method of interpolating smoothly between the given goniometric slices:

1. Project the direction $\vec{\omega}$ onto the two planes. For example, if the main axis is in the $+Z$ direction and the diagram depicts the XZ and YZ slices, then the projection of an arbitrary vector $\vec{\omega} = (x, y, z)$ yields the new vectors, $(x, 0, z)$ and $(0, y, z)$, with angles $\phi_x = \text{atan2}(x, z)$ and $\phi_z = \text{atan2}(y, z)$ off the Z axis.

2. Perform elliptic interpolation:

$$S(\vec{\omega}) = \sqrt{S_x(\phi_x)\cos^2\phi_x + S_y(\phi_y)\cos^2\phi_y} \tag{10.7}$$

3. Finally, divide the result by the maximum, I_{max}.

The form factor from a point light i to an element j can now be derived. Again, the form factor is proportional to the solid angle subtended by j from the point of view of the light and is scaled at each dA_j by $S(\vec{\omega})$:

$$F_{ij} = \int_{A_j} S(\vec{\omega})\frac{\cos\theta_j}{r^2} V(\mathbf{x}_i, \mathbf{x}_j)\, dA_j \tag{10.8}$$

where ω is a vector from the light sources to dA_j.

For general area lights, the goniometric diagram must be converted to luminance by dividing by the projected area of the source. For example if the light intensity, I, is given in terms of *candelas* (cd), then the luminance (cd/m^2) is given by

$$L_i(\theta) = \frac{1}{\cos\theta}\frac{I}{A_i} \tag{10.9}$$

In this case, the form factor must be integrated over the area A_i of the light and normalized by dividing by A_i,

$$F_{ij} = \frac{1}{A_i}\int_{A_i}\int_{A_j} S(\vec{\omega})\frac{\cos\theta_j}{r^2} V(\mathbf{x}_i, \mathbf{x}_j)\, dA_j \tag{10.10}$$

Figure 3:
Copy of Typeset Original

distribution are, unfortunately, difficult to specify and do not have a standard form.

Expressing the geniometric distribution in terms of power per steradian, the goniometric diagram can be reformulated as a *maximum* power per steradian I_{max} scaled by a polar function ranging from 0 to 1. The polar scaling function $S(\vec{\omega})$ is defined to takes a direction, $\vec{\omega}$, away from the source and returns a value between 0 and 1.

Interpolation is, required to obtain a value $S(\vec{\omega})$ from the goniometric diagram for a direction that does not lie on either of the two perpendicular planes defining the distribution. Languénou and Tellier [144] suggest the following method of interpolating smoothly between the given goniometric slices:

1. Project the direction $\vec{\omega}$ onto the two planes. For examples if the main axis is in the $+Z$ direction and the diagram depicts the XZ and YZ slices, then the projection of an arbitrary vector $\vec{\omega} = (x, y, z)$ yields the new vectors, $(x, 0, z)$ and $(0, y, z)$, with angles $\phi_x = \text{atan2}(x, z)$ and $\phi_z = \text{atan2}(y, z)$ off the Z axis.

2. Perform elliptic interpolation:

$$S(\vec{\omega}) = \sqrt{S_x(\phi_x)\cos^2\phi_x + S_y(\phi_y)\cos^2\phi_y} \tag{10.7}$$

3. Finally, divide the result by the maximum, I_{max}.

The form factor from a point light i to an element j can now be derived. Again, the form factor is proportional to the solid angle subtended by j from the point of view of the light and is scaled at each dA_j by $S(\vec{\omega})$:

$$F_{ij} = \int_{A_j} S(\vec{\omega})\frac{\cos\theta_j}{r^2} V(\mathbf{x}_i, \mathbf{x}_j)dA_j \tag{10.8}$$

where $\vec{\omega}$ is a vector from the light sources to dA_j.

For general area lights, the goniometric diagram must be converted to luminance by dividing by the projected area of the source. For example if the light intensity, I, is, given in terms of *candelas* (cd), then the luminance (cd/m^2) is given by

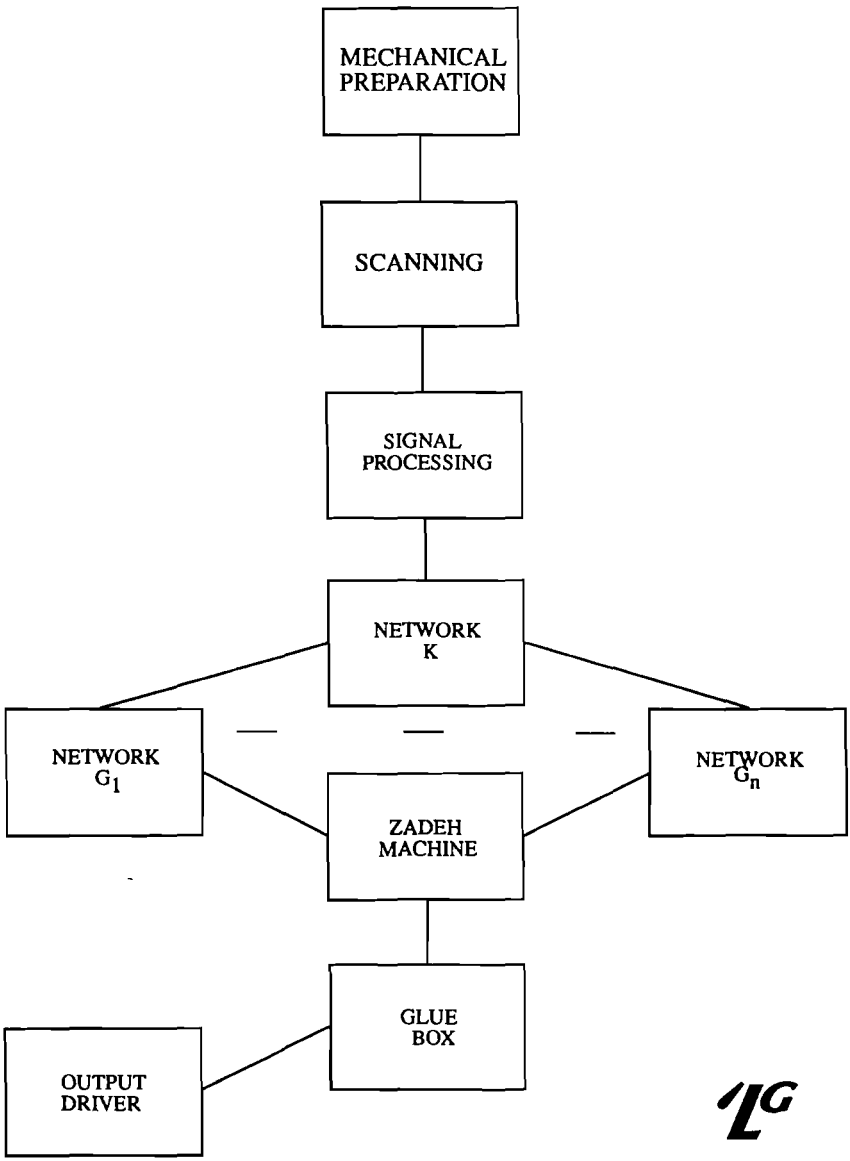
$$L_e(\theta) = \frac{1}{\cos\theta}\frac{I}{A_i} \tag{10.9}$$

In this case, the form factor must be integrated over the area A_i of the light and normalized by dividing by A_i,

$$F_{ij} = \frac{I}{A_i}\int_{A_i}\int_{A_j} S(\vec{\omega})\frac{\cos\theta_j}{r^2} V(\mathbf{x}_i, \mathbf{x}_j)dA_j \tag{10.10}$$

Figure 4:
Output of Magic Book

Figure 5: The MagicMath Process



Patry, Jean. Le théorème de Fuchs et les équations linéaires à coefficients périodiques. C. R. Séances Soc. Phys. Hist. Nat. Genève 59, 118–122 (1942). [MF 14203]

This note is concerned with differential equations of the form

$$(*) \qquad \sum_{m=0}^{n} (e_m + f_m e^{-ix} + g_m e^{ix}) d^m u/dx^m = 0.$$

If $f_n \neq 0$, or if $f_k = 0$ and $e_n \neq 0$, the transformation $z = e^{ix}$ reduces the equation to the form

$$z^n u^{(n)} + z^{n-1} P_{n-1}(z) u^{(n-1)} + \cdots + P_0(z) u = 0,$$

where the P's are functions which are holomorphic in the neighborhood of $z = 0$. The classical Fuchs theory is then applied to obtain n solutions of the form $u = \sum_{k=0}^{\infty} a_k z^{\mu+k} = \sum_{k=0}^{\infty} a_k e^{i(\mu+k)x}$. If $g_n \neq 0$, or if $g_k = 0$ and $e_n \neq 0$, the transformation $z = e^{-ix}$ can be used in a similar manner to obtain n solutions of the form $u = \sum_{k=0}^{\infty} a_k e^{i(\mu-k)x}$. It is shown that, in order that this procedure shall give series which converge for real values of x, it is necessary that the absolute values of the roots of the equation $g_n z^2 + e_n z + f_n = 0$ are both less than unity, or both greater than unity. *L. A. MacColl.*

Figure 6:
Abstract Image Printout

Patry, Jean. Le théorème de Fuchs et les équations linéaires à coefficients périodiques. C. R. Séances Soc. Phys. Hist. Nat. Genève **59**, 118-122 (1942). [MF 14203]

This note is concerned with differential equations of the form

$$(*) \qquad \sum_{m=0}^{n}(e_m + f_m e^{-ix} + g_m e^{ix})d^m u/dx^m = 0.$$

If $f_n \neq 0$, or if $f_k = 0$ and $c_n \neq 0$, the transformation $z = e^{ix}$ reduces the equation to the form

$$z^n u^{(n)} + z^{n-1}P_{n-1}(z)u^{(n-1)} + \cdots + P_0(z)u = 0,$$

where the P's are functions which are holomorphic in the neighborhood of $z = 0$. The classical Fuchs theory is then applied to obtain n solutions of the form $u = \sum_{k=0}^{\infty} a_k z^{\mu+k} = \sum_{k=0}^{\infty} a_k e^{i(\mu+k)x}$. If $g_n \neq 0$, or if $g_k = 0$ and $c_n \neq 0$, the transformation $z = e^{-ix}$ can be used in a similar manner to obtain n solutions of the form $u = \sum_{k=0}^{\infty} a_k e^{i(\mu-k)x}$. It is shown that, in order that this procedure shall give series which converge for real values of x, it is necessary that the absolute values of the roots of the equation $g_n z^2 + e_n z + f_n = 0$ are both less than unity, or both greater than unity. *L. A. MacColl.*

Figure 7:
Output from Magic Math