Evaluation of Page Quality Using Simple Features

by

Luis R. Blando

A thesis submitted in partial fulfillment
of the requirements for the degree of

Master of Science

in

Computer Science

Department of Computer Science
University of Nevada, Las Vegas
November 1994

The thesis of **Luis Ricardo Blando** for the degree of **Master of Science** in **Computer Science** is approved.

Chairperson, Junichi Kanai, PhD.

Examining Committee Member, Thomas A. Nartker, PhD.

Examining Committee Member, Kia Makki, PhD.

Graduate Faculty Representative, Gennady Bachman, PhD.

Dean of the Graduate College, Ronald W. Smith, PhD.

University of Nevada, Las Vegas
November 1994

ABSTRACT

A classifier to determine page quality from an Optical Character Recognition (OCR) perspective is developed. It classifies a given page image as either "good" (i.e., high OCR accuracy is expected) or "bad" (i.e., low OCR accuracy expected). The classifier is based upon measuring the amount of white speckle, the amount of broken pieces, and the overall size information in the page. Two different sets of test data were used to evaluate the classifier: the Sample 2 dataset containing 439 pages and the Magazines dataset containing 200 pages. The classifier recognized 85% of the pages in the Sample 2 correctly. However, approximately 40% of the low quality pages were misclassified as "good." To solve this problem, the classifier was modified to reject pages containing tables or less than 200 connected components. The modified classifier rejected 41% of the pages, correctly recognized 86% of the remaining pages, and did not misclassify any low quality page as "good". Similarly, it recognized 86.5% of the pages in the Magazine dataset correctly and did not misclassify any low quality page as "good" without any rejections.

Contents

List of Tables

List of Figures

Acknowledgements

This research was supported in part by a grant from the United States Department of Energy.

I would like to thank Dr. George Nagy (Rennselaer Polytechnic Institute) and Dr. Horst Bunke (Universtat Bern) for multiple suggestions and ideas to improve this work.

I am also very grateful to Dr. Thomas Nartker and Steve Rice of ISRI for their encouragement and support throughout the whole project. Many members of UNLV's CS faculty were very important in my two years here. I am particularly thankful to Dr. Kia Makki and Dr. Angelo Yfantis for their support. I also want to thank Dr. Gennady Bachman for agreeing to review and evaluate my work.

I would also like to thank all other ISRI members who helped with the project. I am specially grateful to Juan Gonzalez who did some of the programming and to Angela Molnar who proofread and corrected earlier versions of this work.

This work would have not been possible without the help of Dr. Junichi Kanai. He proved to be not only an outstanding researcher but an incredible human being. I consider myself very lucky to have been his disciple and I extend my gratitude to him for the countless hours he devoted to this project in particular and to my professional formation in general.

Last, but certainly not least, I would like to thank my wife, Laura, for her unconditional love and support.

Chapter 1

Introduction

Optical Character Recognition

Digital information storage has become commonplace mainly because of the growth of computer technology. There is a growing trend among publishers to offer digital versions of their products. Nevertheless, as stated in [7], printed versions of documents will always be needed and, more importantly, there will always be a need to convert these printed documents into their digital counterparts.

> A *document image* is a visual representation of a printed page, such as a journal article page, a magazine cover, a newspaper page, etc. Typically, a page consists of blocks of text, i.e. letters, words, and sentences that are interspersed with half tone pictures, line drawings, and symbolic icons. A digital document image is a two-dimensional numerical array representation of a document image obtained by optically scanning and raster digitizing a hard copy document. It may also be an electronic version that was created in that form, say, for a bit-mapped screen or a laser printer [15].

The process of transforming a printed document image into a digital document consists in the spatial sampling and simultaneous conversion of light photons to electric signals. This process is carried out by a "scanner", which in essence divides the

printed page into small pixels[1] and samples a light value for each of these pixels on the page. This value is then thresholded against a pre-set value to determine whether or not that particular pixel will be considered "filled"[2].

> Scan resolution is very important. The width of a typical character stroke is about 0.2mm (0.008 inch), with some of the widest strokes up to about 1mm. A 10-point character measures about 0.5mm (0.014 inch) between ascender and descender lines. A sampling rate of 240 ppi[3] corresponds to about 0.1 mm/pixel, which guarantees that at least one pixel will fall totally within the stroke [15][4].

Two of the main advantanges of having textual information rather than page images stored are the possibility of searching large amounts of information and the ease in retrieving only what is relevant to a query. There are several ways of querying a body of information; a discipline that studies these related aspects is called *Information Retrieval*. In order to generate the information to accomodate these queries, the image file is not enough. The image file only contains a digital representation of the "look" of a printed page but lacks understanding of any of its contents. Since information retrieval requires the contents in order to perform the retrieval, a way to extract these contents of the digital image is necessary.

Desktop Publishing (DTP) applications comprise another important reason to have information stored digitally. To accomodate the task of edition and modification a DTP system must handle the textual representation of the information in order to edit and format it. Having only the information as a picture prevents the DTP

[1] Acronym for PICture ELement.

[2] Global and adaptive thresholding mechanisms are possible. However, a discussion of these is beyondthe scope of this work.

[3] Pixels per inch.

[4] In this thesis, images were scanned at 300 ppi.

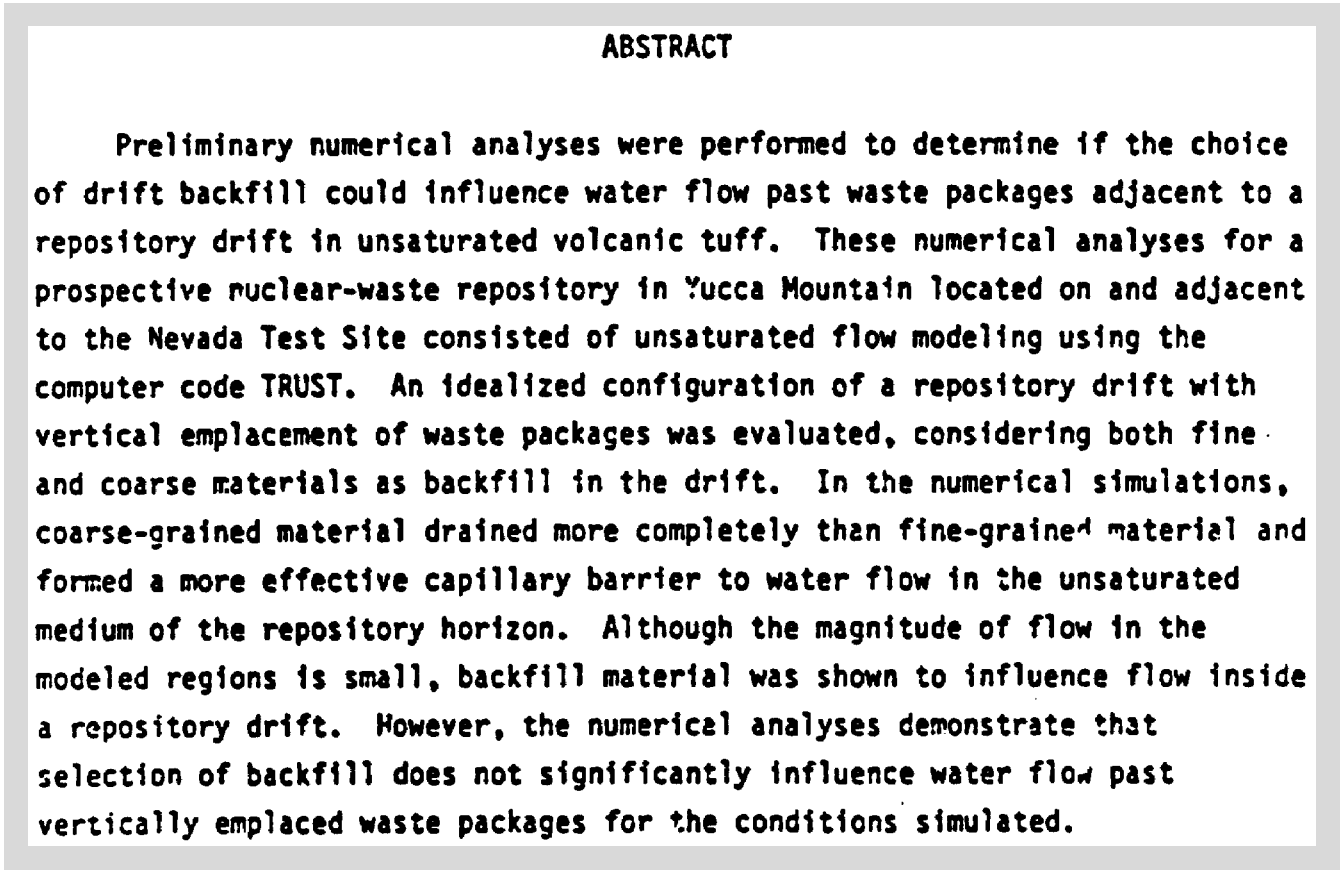
application from doing any editing or layout formatting because the image format is not suited for these operations.

There are other reasons motivating the extraction of contents from an image page besides the possibility of retrieval, cataloguing, and DTP. One very important aspect is that, in general, the digital version of the *contents* of a page usually occupy less space than the image file. Furthermore, if the content is textual information and the representation selected is a text file (as is usually the case), the content can be electronically mailed and distributed, not to mention modified, whereas any of these tasks would be difficult at best if working with the image file only.

Optical Character Recognition (OCR) is the process by which a page image is transformed into a text file. The purpose of the whole OCR process is to recognize the letters, words, and symbols printed on a page. Presently, there are many commercial OCR systems in use.

OCR systems usually first receive a page image as input, then they segment out characters, and finally they recognize these characters. Additionally, OCR systems may use spell checkers or other lexical analyzers that make use of context information to correct recognition errors and resolve ambiguities in the generated text. The output of the OCR process is a text file, corresponding to the printed text in the image file. Figures 1.1 and 1.2 show an example of a small image file and its corresponding OCR output, respectively.

OCR devices are usually very good at recognizing text from clean images; however, they make errors. A closer look at Figure 1.2 will reveal many recognition errors. The *performance* of an OCR device is measured in terms of its *character accuracy*. To define character accuracy, the number of insertions (i), substitutions (s), and deletions (d) required to correct the OCR output to agree with the "correct" text are measured. The accuracy is computed (where c is the number of characters in the correct text



ABSTRACT

Preliminary numerical analyses were performed to determine if the choice of drift backfill could influence water flow past waste packages adjacent to a repository drift in unsaturated volcanic tuff. These numerical analyses for a prospective nuclear-waste repository in Yucca Mountain located on and adjacent to the Nevada Test Site consisted of unsaturated flow modeling using the computer code TRUST. An idealized configuration of a repository drift with vertical emplacement of waste packages was evaluated, considering both fine and coarse materials as backfill in the drift. In the numerical simulations, coarse-grained material drained more completely than fine-grained material and formed a more effective capillary barrier to water flow in the unsaturated medium of the repository horizon. Although the magnitude of flow in the modeled regions is small, backfill material was shown to influence flow inside a repository drift. However, the numerical analyses demonstrate that selection of backfill does not significantly influence water flow past vertically emplaced waste packages for the conditions simulated.

Figure 1.1: Sample Image File

```
ABSTRACT

Preliminary numerical analyses were performed to determine if the choice
of drift backfill could influence water flow past waste packages adjacent to a
repository drift in unsaturated volcanic tuff. These numerical analyses for a
prospective nuclear-waste repository In "ucca Mountain located on and adjacent
to the Nevada Test Site consisted of unsaturated flow modeling using the
computer code TRUST. An idealized configuration of a repository drift with
vertical emplacement of waste packages was evaluated, considering both fine.
and coarse materials as backfill in the drift. In the numerical simulations,
coarse.grained material drained more completely than fine.grainefl wiaterial arid
fon::ed a more effective capillary barrier to water flow in the unsaturated
medium of the repository horizon. Although the magnitude of flow in the
modeled regions is small, backfill material was shown to influence flow inside
a repository drift. However, the numerical analyses deii'onstrate that
selection of backfill does not significantly influence water floe past
vertically emplaced waste packages for the conditions' simulated.
```

Figure 1.2: OCR Output for Sample Image File

UNLV-ISRI OCR Accuracy Report Version 4.0		
1129	Characters	
19	Errors	
98.32%	Accuracy	
0	Marked Errors	
1137	Generated Characters	
0	Marks	
0	False Marks	
Errors	Marked	Correct-Generated
3	0	{m}-{ii'}
3	0	{rm}-{n::}
2	0	{-}-{.}
2	0	{d}-{fl}
2	0	{m}-{wi}
2	0	{n}-{ri}
1	0	{Y}-{"}
1	0	{i}-{I}
1	0	{w}-{e}
1	0	{}-{'}
1	0	{}-{.}

Table 1.1: Sample Image OCR Accuracy Report

file) [12]:

$$Character\ Accuracy = \frac{c - (i + s + d)}{c}$$

The Information Science Research Institute (ISRI) has developed a set of tools to automate the measurement of character recognition accuracy from the OCR generated output [14]. Table 1.1 shows the number of OCR-generated errors from Figure 1.2 and the character accuracy[5].

Measuring OCR accuracy has become the universally accepted way of rating OCR devices' performance [11, 12, 13]. It is a good measure because, among other things,

[5]Generated using ISRI's experimental environment [14]

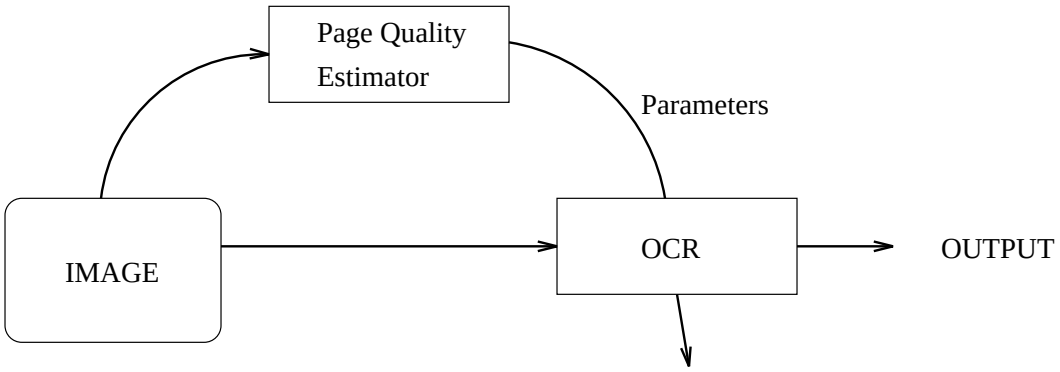


Figure 1.3: Adaptive OCR Algorithms Architecture

it correlates nicely with the end-user's perspective. Higher accuracy means better recognition and less work (cost) to correct OCR-generated errors.

Needs for Estimating Page Quality

Estimating page quality for any given image would be beneficial for several applications:

- **Controlling adaptive image processing for OCR**. The existence of a way to automatically evaluate the quality of any given image would be essential for an adaptive image-enhancement algorithm. The algorithm would iteratively produce an image to be graded by the page quality estimator, which in turn would feedback the noise type or the degree of noise present in the image to the adaptive algorithm to generate the next (better) iteration of the image.

- **Adaptive OCR algorithms.** An image quality estimator would be essential to the operation of adaptive OCR algorithms since it could set the parameters for the OCR engine according to the quality of the page that it is about to process (Figure 1.3).

- **Reducing rekeying costs.** As will be shown in this work, page/image quality is a direct cause of OCR errors. Therefore, estimating page quality can also

provide an estimation of OCR accuracy. The minimum acceptable OCR accuracy for large-scale OCR operations is in the range of 95%-98% [4]. Correcting the errors on a page with less than 95% accuracy is more costly than retyping the page from scratch. A hypothetical "OCR-accuracy estimator" would act as a filter, classifying pages and filtering out those that would be better off rekeyed manually. In large-scale OCR environments, such a filter would represent substantial cost savings, since often the whole process is automated and the cost of manually rekeying a page after it has been processed implies disrupting the normal flow of the entire system.

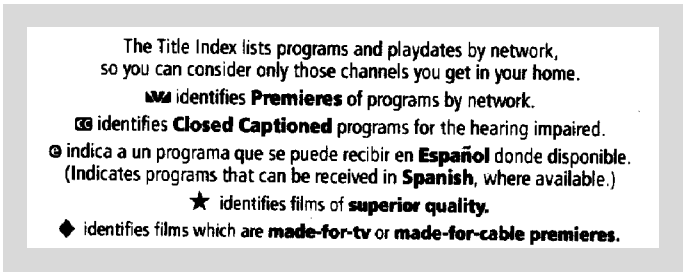
Difficulty in Predicting OCR Accuracy

The task of predicting OCR accuracy is very complex. Furthermore, to the best of the author's knowledge, no previous work has been done in this area; therefore, no reference titles can be given.

OCR algorithms seem to be affected by a myriad of different problems. However, the following three general "problem groups" can be identified:

- **Typographical problems.** Unusual fonts or complex typesetting often cause problems for OCR algorithms. Figures 1.4 and 1.5 shows some examples of complex images and their resulting OCR output.

- **Linguistic problems.** As stated in [5], linguistic content can also be problematic for OCR algorithms. OCR systems use lexicons to solve recognition ambiguity. This greatly improves OCR accuracy, but can also be a drawback when proper names, acronyms, or other words not likely to be in the lexicon are part of the text to be recognized. In these cases, the number of lexicon-dependent corrections are reduced to a minimum. Furthermore, an exaggerated sensitivity to the lexicon could be the cause for the introduction of new errors

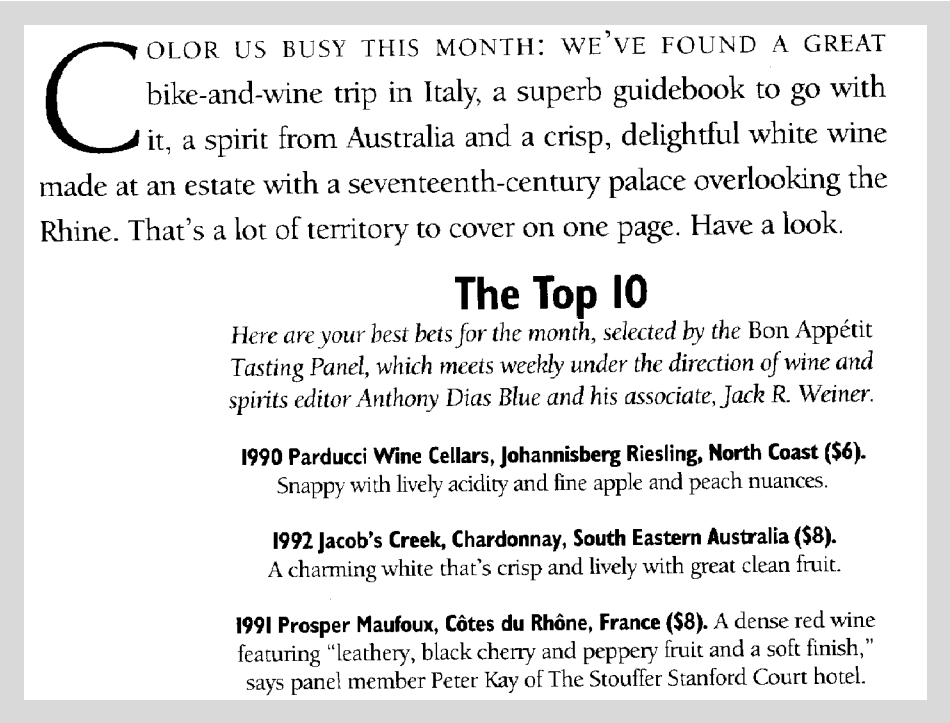



```
The Title Index lists programs and playdates by network,
so you can consider only those channels you get in your home.
tarn identifies Pimmiems of programs by network.

U identifies closed Capuoned programs for the hearing impaired.
* indica a un programa que se puede recibir en Espahol donde disponible.
(Indicates programs that can be received in Spanish, where available.)
* identifies films of Superior qudity

* identifies films which are  adfrfo tv or made-forccable prewieres.
```

Figure 1.4: Typographical Problems for OCR Algorithms (1/2)



COLOR US BUSY THIS MONTH: WE'VE FOUND A GREAT bike-and-wine trip in Italy, a superb guidebook to go with it, a spirit from Australia and a crisp, delightful white wine made at an estate with a seventeenth-century palace overlooking the Rhine. That's a lot of territory to cover on one page. Have a look.

The Top 10

Here are your best bets for the month, selected by the Bon Appétit *Tasting Panel, which meets weekly under the direction of wine and spirits editor Anthony Dias Blue and his associate, Jack R. Weiner.*

1990 Parducci Wine Cellars, Johannisberg Riesling, North Coast ($6). Snappy with lively acidity and fine apple and peach nuances.

1992 Jacob's Creek, Chardonnay, South Eastern Australia ($8). A charming white that's crisp and lively with great clean fruit.

1991 Prosper Maufoux, Côtes du Rhône, France ($8). A dense red wine featuring "leathery, black cherry and peppery fruit and a soft finish," says panel member Peter Kay of The Stouffer Stanford Court hotel.

```
OLOR US BUSY THIS MONTH: Nk7E'VE FOUND A GREAT

bike-and-wine trip in Italy a superb gaidebook to go with
it, a spirit from Australia and a crisp, delightful white wine
made at an estate with a seventeenth-century pAace overlooking the
Rhine. That's a lot of territory to cover on one page. Have a look.

TheTop10
Mete air your best betsfot. die uionth, selected fr die Ban Apperit
Tasting Panel, wilicli aleets weekI, under die dijeetion of wine and
spitits edit()r Antliony Dicis Blite and his associate. Jack R. 'Veiner.

1990 Parducci Wane Cellars, Johannisberg Riesling. North Coast ($6).
Snappy 'vith lively acidit,, and fine apple and peach nuances.

1992 Jacob's Creek, Chardonnay, South Eastern Australia ($8).
A charming white that's crisp and fively "4th great clean fruit.

1991 Prosper Maufoux, C6tes du Rh&ne, France ($8). A dense red  irie
featuring "leathery., black cherry and pepprry fruit and a soft finish,"
says panel member Peter Kay of The Stouffer Stanford Court hotel.
```

Figure 1.5: Typographical Problems for OCR Algorithms (2/2)

(i.e., an OCR device with an exaggerated sensitivity may modify an otherwise correctly recognized word to force it to match the system's lexicon.)

- **Image problems.** In the OCR community, it is well known that image defects directly affect the performance of OCR algorithms. Furthermore, image problems account for the majority of OCR errors [10].

Image defects constitute the bulk of the problems associated with OCR algorithms (see Chapter 2 of this thesis). Therefore, the focus of this work is on the detection of image problems. By better understanding image defects and subsequently implementing OCR algorithms that are sensitive to these type of problems, it could be possible to achieve acceptable accuracy ranges (95%-98%, [4]) for most printed pages. To achieve near perfect (99.5%-100%) recognition, however, typographical as well as linguistic problems would have to be addressed.

Problem Description

The objective of this research is to develop a classifier for predicting OCR accuracy by measuring image defects. In other words, this classifier measures image *quality* from an OCR perspective.

An ideal output of such a quality metric would be the actual accuracy any given OCR device would attain on the page. This is a very complex problem because of the rapid progress of OCR technology as well as the complexity and number of features needed for such a task. Therefore, this work will concentrate on the design and development of a binary classifier. The output of the system should be a label of "Good" or "Bad", depending on the accuracy that page would attain if processed through an OCR device. "Good" means the page image is clean and has an expected OCR accuracy of at least 90%, whereas "Bad" pages may have different degrees of noise in their images and the expected OCR accuracy for them would be below 90%.

Assumptions

In order to limit the scope of this project, the research has been limited to solving the problem for a subset of the documents that are normally processed by OCR devices. However, the subset selected is a major portion of the usual OCR pages and the results of this work can be extended to handle a much more varied set of pages.

The type of pages the classifier will be designed for have the following characteristics:

- White background and black letters (no color)
- Previously segmented pages. The pages have been manually segmented into "text", "table", "caption", "header/footer", and other types of zones depending on the contents. The classifier presented in this thesis extracts its features from "text" zones *only*.
- No artistic fonts

This work will consider a page to be "Good" if its median OCR-accuracy (calculated from a set of accuracies from different OCR devices) is equal to or higher than 90%. Conversely, a page will be labeled "Bad" if its accuracy falls below this 90% threshold.

Description of the Work

The remainder of this thesis consists of five chapters. In the next chapter, related works in this area and other approaches to the evaluation of page quality are briefly presented. The image features used to determine page quality and the design of the classifier are described in Chapter 3. Chapter 4 introduces the test dataset and describes the results obtained along with an analysis. Ideas for future work and conclusions are presented in Chapter 5 and 6, respectively.

Chapter 2

Related Work

OCR Difficulty Evaluation

After conducting an extensive literature survey and consulting with various researchers in the OCR field, no previous work similar to the one presented here could be found. Therefore, alternative approaches to OCR difficulty evaluation were investigated. These works, even though not aimed at finding page quality metrics, are closely related to this project's scope.

Mindy Bokser presents a complete view of the problems associated with trying to recognize letters from a page image [3]. According to her work, touching and broken (split) characters seem to be the most important source of OCR problems. Regarding OCR technology, she acknowledges that

> The best products do a good job on clean documents, but they all degrade in performance –some more gracefully than others– as document quality (or scanner quality) degrades [3].

Similarly, Nartker et al [8] identified broken and touching characters as the leading cause of OCR errors. Table 2.1[1] summarizes estimated OCR problems obtained from a 240-page test. *Page quality errors* account for 83.9% of the total number of errors in the set, whereas errors caused by other factors account for the rest.

[1] Reproduced from [8] with permission

Problem Category	**Number of Errors**	**Percent of Total Errors**
Broken Characters	1872	52.1
Touching Characters	734	20.4
Noise / Speckle	122	3.4
Skew (or curved baseline)	49	1.4
Broken & Touching	186	5.2
Broken & Noise	9	0.3
Broken & Skew	33	0.9
Touching & Noise	2	0.1
Touching & Skew	10	0.3
Similar Symbols (1,l O,0)	207	5.8
Wrong Case	12	0.3
Stylized Characters	46	1.3
Introduced Spaces	79	2.2
Dropped Spaces	39	1.1
Unknown Cause	196	5.5
Total	**3596**	**100.0**

Table 2.1: Distribution of Estimated OCR Problems

Jenkins and Kanai [5] studied the influence of lexical factors on OCR performance. They controlled image quality and typographical features by creating synthetic images and using them as input to OCR devices. Based on their results, they suggested that linguistic factors, apart from image-related factors, also affect OCR performance, since most current OCR products incorporate a system lexicon to resolve character recognition ambiguity. Along this idea, the number of *stopwords*[2] was identified as a factor in OCR accuracy, since they are more likely to be included in the system's lexicon than *non-stopwords*.

An important point to be made is that a high percent of the errors are due to relatively few causes, which ultimately correspond to image quality, whereas a large number of other factors represent a relatively low number of errors. Therefore,

[2]Common words not normally used in text retrieval searches, such as "a", "of", "the".

methods and solutions for image quality problems can and will have a direct impact on OCR performance from an end-user's point of view, since by fixing page-quality related errors the accuracy rate can increase considerably.

Other Approaches

Other approaches to the measurement of print quality for OCR systems include:

- **Physical approaches.** Throssell and Fryer [16] and Bohner et al [2] proposed mechanical systems to measure print quality as defined by ISO Recommendation 1831 (1968). These two works date back to the mid 1970s, when OCR systems were not popular except in commercial/financial institutions. As a result, both papers concentrated on ways to define print quality for OCR-A and OCR-B character sets. Their approach is to construct a high resolution scanning device to calculate Print Contrast Signal values, which are then used to rate each individual character according to the ISO recommendation. These approaches are not practical for current OCR needs, not only because of the cost associated in building these special scanning devices, but also because current OCR environments are omnifont; both of these approaches are very limited in selection of font-type and the fonts used must be known beforehand.

- **Using OCR output.** A popular way to estimate page difficulty for OCR output is simply to process the image first and then use the reject and/or suspect markers in the OCR output to estimate page quality. The drawback of this approach is that it is completely dependent on the OCR device being used. Furthermore, this approach is dependent on the capabilities of that particular OCR device to produce reject/suspect markers. If the OCR device does not produce reject/suspect markers, or if it does so very poorly, this method is useless.

- **Using spell checkers.** Another approach to estimate page quality would be to examine the OCR output using a spell checker to see how many words are not found in the dictionary. The problem with this approach lies in that, for many types of data, no words will be found in the lexicon. Proper names, acronyms, and numerical data are all examples of types of data that can not be corrected by simply using lexicon lookup. As a result, a metric that measures, for instance, the number of non-found words would underestimate the accuracy when presented with this "non-standard" type of data.

Chapter 3

Classifier Design

The design and development of a classifier based solely on simple image features will be presented.

Justification for Using Simple Features

In this project, only simple image features are used to design the classifier. The reasons behind this constraint are as follows:

- **Cost**. The classifier will act as a filter for pre-processing pages in a large-scale OCR production environment. Therefore, the filter must be fast and not become the bottleneck of the system. By restricting the features to only simple measurements, the resulting speed will be adequate.

- **Independence from OCR Technology**. The focus of this research is to be able to determine "image defects" instead of "character recognition" defects. Ideally, the set of features used by the page quality clasifier would be orthogonal to those used by OCR algorithms. Using only simple metrics as features guarantees that the classifier will not be mimicking an OCR device, since much more complex features are required for this later purpose.

- **No Previous Work**. Since there has been no previous work in this area and no previous approaches to this problem, simple features are chosen to understand

the classifier's behavior as much as possible, and thus lay the grounds for future research in the area.

Feature Selection Process

Features used by the classifier are identified in three steps. First, types of image features associated with image defects are studied using a small set of concept exploration data. Secondly, preliminary measurements for the features are constructed along with the initial version of the classifier's logic. Finally, a training dataset is formed and from there, the final form of the features' metrics as well as the classifier rules are determined.

Concept Exploration Dataset

ISRI's "Sample 1" Database[1] consists of 240 pages selected at random. Only the text portions of each page were zoned and then each image was processed by six OCR devices [11].

For the *concept exploration dataset* 10 pages were selected from the Sample 1 database. Sample 1 was divided into 3 quality groups [11] and 5 pages were selected from groups 1 and 3, respectively. Table 3.1 lists the Concept Exploration Dataset pages with their assigned "Good" or "Bad" labels.

Concept Exploration Observations

After visually examining this small dataset very closely, the following observations were made:

Observation 1. Pages with characters whose strokes are thick tend to have many of their characters touching. This touching causes OCR errors [3, 10]. Another

[1]See [11] for more information about ISRI and its databases

Page-ID	# of Chars	Label
2002-011	2447	Good
5207-005	2484	Good
5319-008	2996	Good
5329-013	2413	Good
5657-079	2728	Good
1970-002	2440	Bad
5752-003	3136	Bad
5768-025	2427	Bad
5770-009	1229	Bad
5777-015	2292	Bad

Table 3.1: Concept Exploration Dataset Image List

Figure 3.1: White-Speckle in Fat Characters

by-product of "fat" characters is that, often, the holes ("lakes") in letters like "a", "e", etc, get filled up completely or present only a minimal white portion in the center. This last fact is also a known cause of OCR errors since, for instance, many times letters like "e" are classified as "c" because of filled lakes. A metric that could capture the existence of these "minimally open" holes would be a good way to measure the image quality of fat characters. Figure 3.1 shows an example of these type of characters.

Observation 2. Pages with light characters or low contrast usually have their characters broken in pieces [3]. These pieces tend to be small and could have almost any shape. A metric that could weight the existence of these "broken pieces" would be a good estimator of image quality for broken-character pages. Figure 3.2 shows a portion of a broken-characters image.

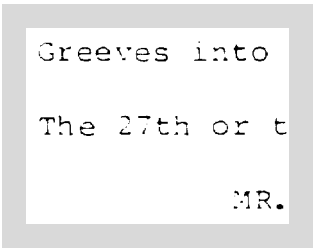


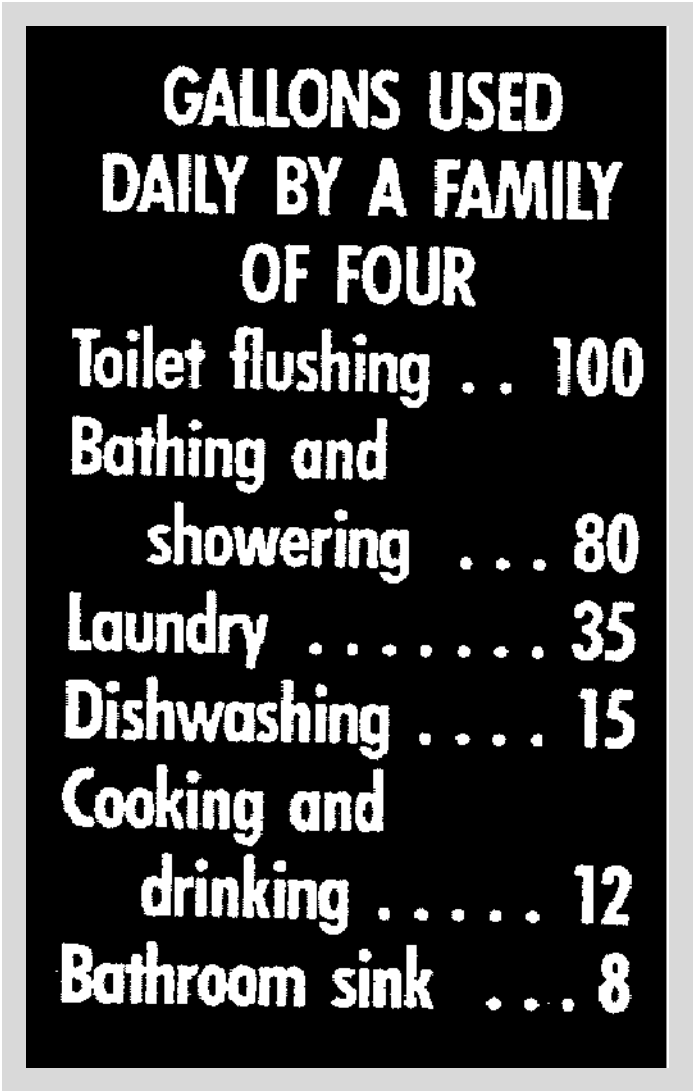
Figure 3.2: Broken Characters

Observation 3. Pages with "inverse video" (white letters on black background, see Figure 3.3) or with unusual typesetting (Figure 3.4) tend to produce more OCR errors. Some type of threshold on the font size information would be a good way of predicting the quality of the page from an OCR point of view.

Observation 4. Pages with characters that have gaps in their stroke are usually problematic for OCR algorithms. These gaps are usually very small in comparison to the stroke-width. Figure 3.5 shows the image of a real word with many broken characters with arrows pointing to these "micro-gaps."

Observation 5. Pages with characters that are not touching each other but occupy the same horizontal space or pages with fragmented/broken characters tend to produce more OCR errors. These type of characters produce Connected Component boxes (see the *Connected Components* section below) that overlap each other (See Figure 3.6). This type of characteristic is commonplace in pages with *italic* or *slanted* typefaces and in pages with seriously fragmented characters.

Observation 6. The degree of skew of a page is also a good predictor of OCR performance. As shown in [12], more than one degree of skew can cause problems for OCR algorithms.



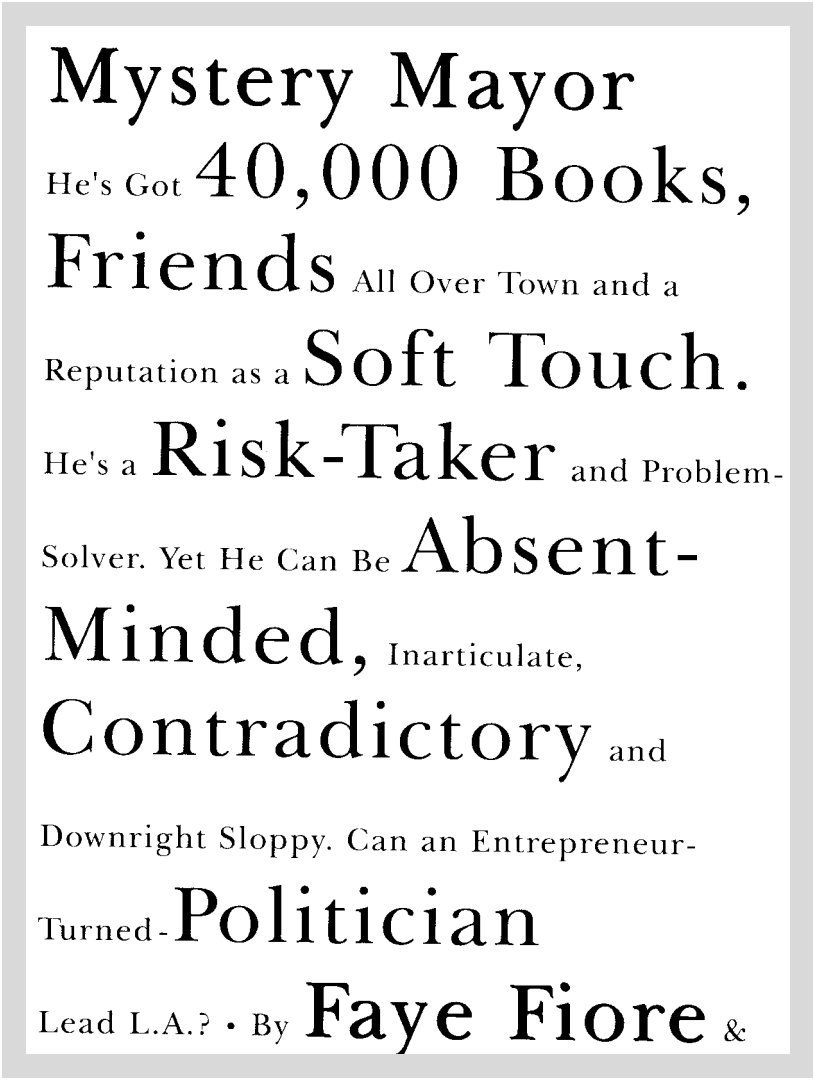

```
g - g jj


* *

I *

$ $ f
```

Figure 3.3: Inverse-Video Image and Corresponding OCR output




```
 Mystery Mayor
He’s Got 40,000 Books,
F ’ d
 rl e n s All Over Town and a
.epu ion Soft Touch.
 ’ k-Tak
He’s aR1 se r and Problem-
Solve . Yet He C n Be Ab S e n t-
M’ d d
 in e , Inarticulate,
Gontradictory and
Downright Sloppy. Can an Entrepreneur-

 Pol’ ’ ’
Turned-ltlclan
Lead L.A.? . By Faye FI ore
```

Figure 3.4: Unusually Typesetted Image and Corresponding OCR Output

Figure 3.5: Micro-Gaps in Broken Characters

Figure 3.6: Overlapping CC Boxes in Slanted and Broken Chars

In this research, the first three observations were selected for further study and ultimately used in the classifier. The rest of the observations can probably lead to very good page quality features but were eliminated from consideration because of the complexity involved in measuring them.

Connected Components

Since the characteristics selected as important are geometric in nature, the Connected Components Data (CC) [1] of the image will be used as the basic data element instead of the image pixels. The rationale is that the same degree of information can be obtained from the CC information as from the image data itself, but with a great gain in simplicity and performance.

The basic concepts behind the construction of the CC data are the following:

- An **8-connected component** is a set of neighboring pixels of the same color such that any pixel in the set can be reached from any other pixel in the set by only passing through pixels contained in the set. The eight possible directions (N, S, W, E, NE, SE, SW, NW) can be used to travel within the set. Figure 3.7 shows the difference between 8-connected and 4-connected components.

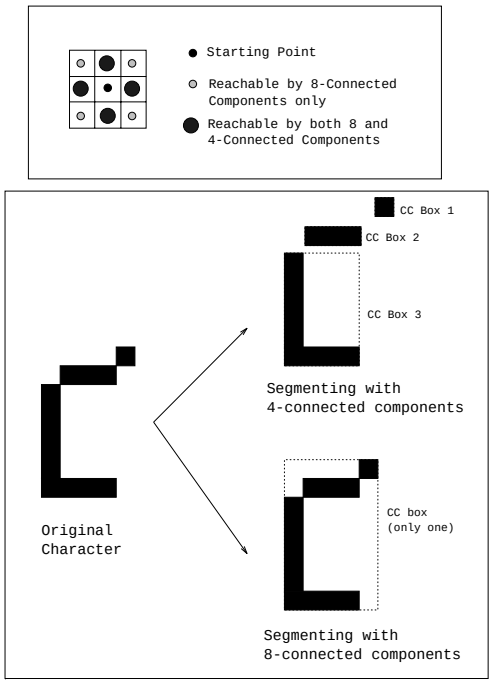


Figure 3.7: Connected Components

- A **connected component box**, also known as the *minimum bounding rectangle*, is the smallest rectangle that completely encloses an 8-connected component. Figure 3.7 exemplifies this concept.

- A **connected components file** is the collection of the starting position (x,y coordinates) and the size (width and height) information for each and all of the connected components in an image or part of an image.

- A **Width-Height Map (WH-Map)** is a 3D frequency distribution of the contents of a connected components file with the CC box width and height as the axis. Figure 3.8 shows the 3D histogram and Figure 3.9 shows a closed contour representation of the WH-Map for a typical image.

Feature Metrics Design

The design of the metrics based on the CC data and used to measure the characteristics metioned before is presented. The *white speckle*, *broken chars zone*, and *size*

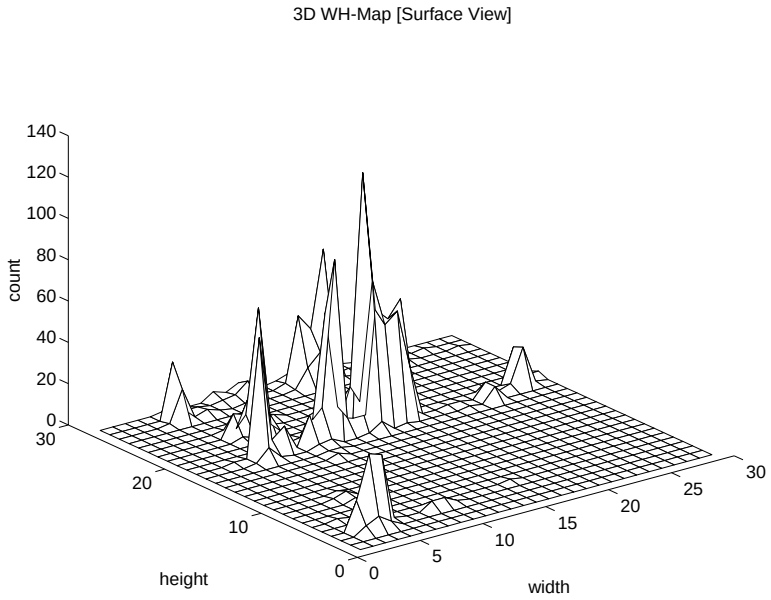


Figure 3.8: Width-Height Map (Surface View)

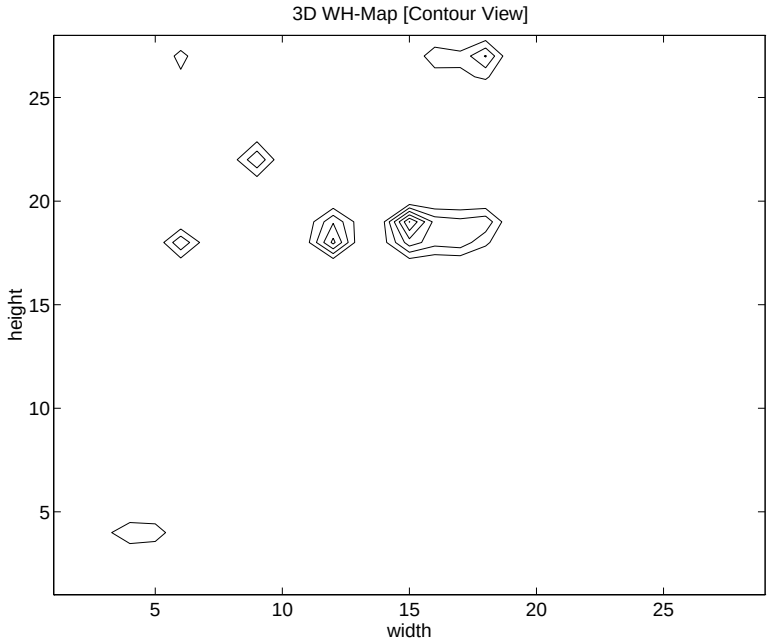


Figure 3.9: Width-Height Map (Closed Contours View)

metrics are introduced.

White Speckle

To detect minimally open holes (Observation 1), the White Speckle metric was designed. White speckle is defined as any 8-connected white region whose size is less than or equal to 3 pixel high and wide.

The *White Speckle Factor* is defined as:

$$White\ Speckle\ Factor = \frac{Number\ of\ White\ CCs \leq 3 \times 3}{Total\ number\ of\ White\ CCs}$$

This metric weights the amount of white-speckle present. We expect the image quality to go down as this ratio goes up. Likewise, a page with a low white speckle factor would probably have its lakes wide open and that, provided there are no other problems with the image, would translate to high OCR accuracy.

It is important to point out that this metric is not appropiate for small typesizes. For small sizes, this metric would incorrectly consider normal "lakes" in letters to be white speckle, since their size is below 3x3 pixels. Furthermore, our training data does not include small fonts. As a result, the effect of this feature on pages containing small fonts must be investigated.

Broken-Chars Zone

The other important problem for OCR algorithms seems to be broken characters. The Broken Character Factor is designed to measure the amount of broken characters in a given image (Observation 2). In general, the sizes and shapes of character fragments vary widely. Thus, their CC boxes will have many different widths and heights. In the WH-Map of a page with broken characters, these "broken" CC boxes will appear near the (width=0, height=0) vertex of the graph. Furthermore, taking

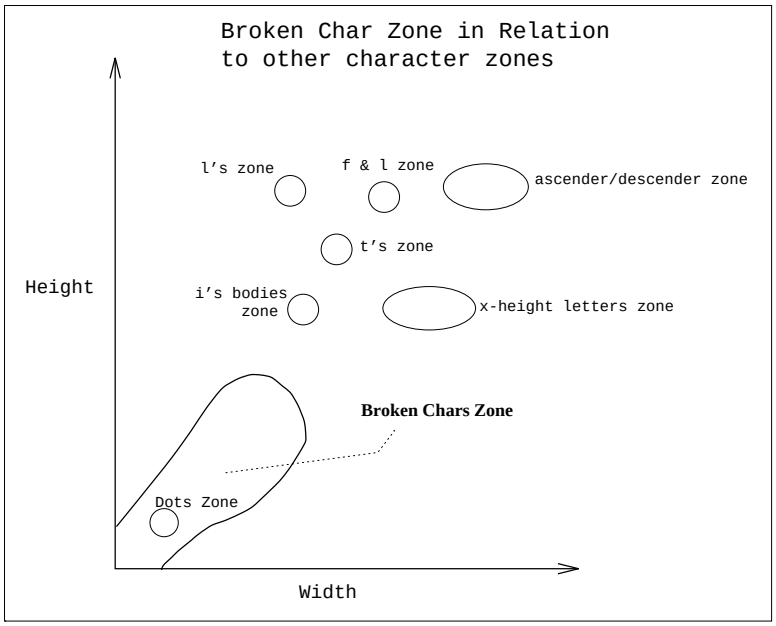


Figure 3.10: Broken Char Zone and Other Char Zones

into account the variations of their shapes, both "wide" and "tall" boxes are expected. Therefore, broken character pages present a "broken char zone" in the WH-Map as shown in Figure 3.10.

It is important to note that the broken char zone is designed to collect all small connected components. These small connected components are mostly the product of broken characters but can also be dots and, in small typesizes, other small legal characters. From a second look at Figure 3.10, it is observed that the location of the "Dots zone" is completely inside the broken char area. This means that all the dots in the page, such as a period and the dot of 'i', will generate connected components that will fall inside the broken characters zone.

A density measurement is sensitive to the distribution of characters in the page. Therefore, it is not a realiable estimator of the number of broken characters' pieces present in the image, because a page containing a large number of dots or other small legal characters would have a high density in the Broken Characters Zone. Therefore, the *coverage* of the broken char zone is of interest instead of its density.

To measure the degree of covering of the Broken Char Zone the following method is used: the zone is divided into square cells, at a rate of one per square pixel; then, the CC boxes are allocated to these cells according to their width and height. After all the CC boxes are allocated, the *Broken Char Factor* is computed as:

$$Broken\ Char\ Factor = \frac{Number\ of\ Cells\ Occupied}{Number\ of\ Cells}$$

A measure such as this one effectively removes the error of considering a zone with a large amount of dots or other small characters as a broken-char page suspect.

The broken char zone must be defined independent of any font-specific characteristics so that it can be reliable when used in pages with different fonts and typesizes. To define the broken char zone, a way of normalizing its dimensions and registering it inside the WH-Map must be determined. A standard way of registering planar information is to determine a single point in the plane from the available data and then define all subsequent plane mappings with regard to this "anchor point". Two approaches for determining the anchor point were examined:

- The most frequent width/height values in the connected components data for a page is selected as the reference point.
- The average width/height values are used as the reference point.

Results are described in the section "Determining Threshold Values".

After defining the reference point, the shape and boundaries of the broken chars zone must be defined. As suggested by experimental observations, the general shape of the broken chars zone should be a rectangle aligned with the width-height diagonal and thick enough to allow for "wide" and "tall" broken pieces.

Size Information

In addition to the two previous measures, the classifier incorporates two more preventive measures based on the connected components' size.

The rationale behind these heuristics is that pages which contain too many big connected components (black or white) are more OCR error prone than those that do not.

Large black connected components *throughout the whole page* can be the result of touching characters, a very large font, or complex vertical touching patterns. All of these characteristics pose difficulty to OCR algorithms.

Large white connected components, similarly, can be the product of large fonts, inverse video or complex touching patterns.

The classifier measures this information by taking the maximum of the average width and average height of the CCs on a page, for both black and white connected components.

Preliminary Set of Rules

Based on the concept exploration phase, the preliminary set of rules for the classifier were:

If $WhiteSpeckleFactor > SomeThreshold_A \rightarrow$"Page is Bad"

If $BrokenZoneFactor > SomeThreshold_B \rightarrow$"Page is Bad"

If $MaxAvgBlackCC > SomeThreshold_C \rightarrow$"Page is Bad"

If $MaxAvgWhiteCC > SomeThreshold_D \rightarrow$"Page is Bad"

Determining Threshold Values

In order to obtain values for the thresholds a complete test was conducted on the training dataset. This section describes the training dataset and the results obtained in relation with each of the metrics designed in the previous phase.

Training Dataset

Because of the geometric nature of the features, a more heterogeneous training dataset was needed. The concept exploration dataset lacked several font and pitch combinations; the features proposed can be affected by size variations. The *training dataset* was constructed with the following characteristics:

- Twenty four pages total, with 12 "good" and 12 "bad", where the meaning for "good" is an OCR median accuracy of at least 90%.
- Three pages were re-used from the concept exploration dataset and 21 new ones were selected from ISRI's "Sample 2"[2] database.
- The pages were selected based on their median OCR accuracy (see below), font type and pitch. All the combinations were constructed and 3 pages were selected for each combination (see Table 3.3). The median accuracy was computed from the output of eight OCR devices (see Table 4.1) except for pages *2002-011*, *5207-005* and *5319-008* that were processed by the devices listed in Table 3.2.
- Whenever possible, pages containing at least 500 characters were selected[3].
- Only text zones were considered. Tables and graphs were ignored.

The reason for using median accuracy instead of the mean accuracy is that the median measure is a more stable metric, since it is not affected by abrupt lows or highs in accuracy for any device. The mean value, on the other hand, would be affected by such a behaviour and thus would render an accuracy value that is not representative of the "general" accuracy OCR devices have on the page.

[2]Consult [12] for detailed information on the "Sample 2" database
[3]In some of the more unusual font/pitch combinations this was not possible

Vendor	Version Name	Version #
Caere Corp.	Caere OCR	132
Calera Recognition Systems, Inc.	Calera MM600	4
ExperVision, Inc.	ExperVision RTK	3.0
OCRON, Inc.	OCRON Recore	3.0
Recognita Corp. of America	Recognita Plus DTK	2.00.D12
Xerox Imaging Systems, Inc.	XIS ScanWorX API	10

Table 3.2: OCR Devices Processing pages 2002-011, 5207-005 and 5319-008

Conclusions from Training Test

After evaluating the training dataset closely, the needed thresholds and other information were determined.

White Speckle

Based on observations of the training set, the following conlusions are drawn:

- The 3x3 pixels limit on the size of the white speckle connected components was reasonable and constant throughout the whole test set.
- More than 10% of white speckle would generally translate into a page being difficult because of fat and/or touching characters.

Reference Point

The average values (as opposed to the most frequent value, as done in [6]) were chosen as the reference ("anchor") point because of their stability.

Tests performed using the most frequent value would fail in the presence of a page with many small or large connected components (as in an index page for example, where many dots can be present in relation to the number of letters). Since a stable reference to some point in the WH-Map is needed, the average values should be.

Page-ID	#Chars	Label	Pitch	Font Type	Accuracy
0151-105	1269	Good	Fixed	Sans Serif	99.932
1367-152	2233	Good	Fixed	Sans Serif	99.886
5804-060	1770	Good	Fixed	Sans Serif	99.976
2306-043	2340	Good	Fixed	Serif	99.929
5945-102	1973	Good	Fixed	Serif	100.00
6582-095	3023	Good	Fixed	Serif	100.00
0648-013	2119	Good	Proportional	Sans Serif	99.527
6293-017	286	Good	Proportional	Sans Serif	98.521
6654-023	565	Good	Proportional	Sans Serif	98.933
2002-011	2447	Good	Proportional	Serif	99.725
5207-005	2484	Good	Proportional	Serif	99.785
5319-008	2996	Good	Proportional	Serif	99.775
5020-009	874	Bad	Fixed	Sans Serif	71.826
5034-039	2181	Bad	Fixed	Sans Serif	81.709
6684-009	2239	Bad	Fixed	Sans Serif	80.267
5375-004	1183	Bad	Fixed	Serif	82.585
5993-006	1810	Bad	Fixed	Serif	77.422
6272-086	929	Bad	Fixed	Serif	78.911
5623-019	1602	Bad	Proportional	Sans Serif	67.619
1662-034	312	Bad	Proportional	Sans Serif	88.759
6831-001	250	Bad	Proportional	Sans Serif	85.811
5258-113	1106	Bad	Proportional	Serif	58.283
5258-170	3832	Bad	Proportional	Serif	71.771
5649-063	3691	Bad	Proportional	Serif	75.615

Table 3.3: Training Dataset Image List

Therefore, the reference point is calculated:

$$Ref_X = \overline{width} = \frac{\sum_{i=1}^{\#CCs} Width_i}{\#CCs}$$

$$Ref_Y = \overline{height} = \frac{\sum_{i=1}^{\#CCs} Height_i}{\#CCs}$$

Broken Chars Zone

The boundaries of the broken chars zone were defined as shown in Figure 3.11, where the percentage values are taken over the value of the reference point on that

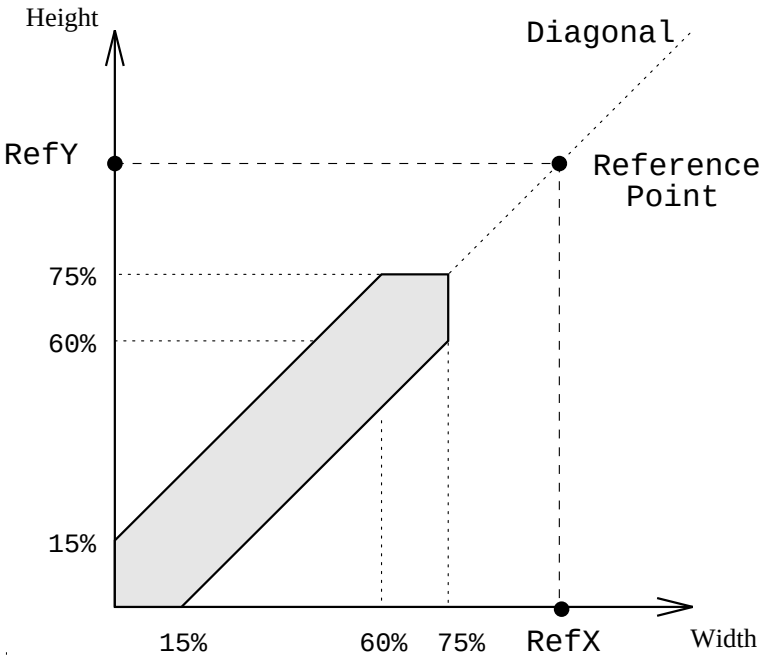


Figure 3.11: Broken Chars Zone Coordinates Definition

axis.

This zone is subdivided into cells at a rate of one cell per pixel in each direction and connected components are allocated to the cells according to their width and height.

From the observations on the training data, a broken chars zone 70% or more filled is a very strong indicator of the prescence of too many broken characters in the page, and thus poor OCR accuracy.

Size Information

The thresholds for the size cutoff were determined to be:

- 40 pixels for black connected components
- 30 pixels for white connected components

An additional rule was included in the white connected components case in order to rule out inverse video. The $\frac{Black\ CC}{White\ CC}$ ratio is tested and, if the number of black

connected components is less than 50% more than the number of white connected components, then the page could be inverse video and is therefore labeled as "Bad" provided it also complies with the 30 pixels threshold.

Final Set of Rules

The final set of rules for the classifier is therefore:

Features Measured

- $WhiteSpeckleFactor = \frac{White\ CCs<3x3}{Total\ of\ White\ CCs}$
- $BrokenZoneFactor = \frac{Num.\ of\ BZ\ cells\ filled}{Num.\ of\ BZ\ cells}$
- $MaxAvgWhiteCC = Max(\overline{Width_{white}}, \overline{Height_{white}})$
- $MaxAvgBlackCC = Max(\overline{Width_{black}}, \overline{Height_{black}})$
- $BWRatio = \frac{Num.\ of\ Black\ CCs}{Num.\ of\ White\ CCs}$

Rules

1. If $WhiteSpeckleFactor \geq 10\% \longrightarrow$ "Bad"
2. If $BrokenZoneFactor \geq 70\% \longrightarrow$ "Bad"
3. If $MaxAvgBlackCC \geq 40\ pixels \longrightarrow$ "Bad"
4. If $MaxAvgWhiteCC \geq 30\ pixels\ AND\ BWRatio < 1.5 \longrightarrow$ "Bad"

Summary

The design and development of the classifier have been presented. The concept exploration dataset was used to identify potentially good indicators of image defects from an OCR point of view. A subset of these indicators were selected to be included

in the classifier because of their simplicity. Tentative metrics were proposed to measure these indicators and the training dataset was used to determine the actual form of these metrics. Finally, from the metrics developed, a set of heuristic rules was put together to implement the classifier's logic.

Chapter 4

Results and Analysis

This chapter presents the classifier architecture and experimental results and analysis.

Classifier Testing Architecture

This section presents the classifier basic processing model, the testing environment model, and the report methodology used.

Basic Processing Model

Figure 4.1 shows modules in the classifier. The `ccomp` program generates the black and white connected components from a TIFF image file. The two CC files are then read by the `clas` program which calculates the features, applies the classification rules, and generates the results file, from where the reports are then extracted.

The accuracy value from the OCR processing of the image is used only for generating the output tables and is not used by the classifier's logic in any other way.

To automate the testing of a large number of images, the following steps are followed:

- Create a list of all the image-names that the test dataset will contain.
- Iterate over this list generating the connected components data files (two per image, –black and white–).

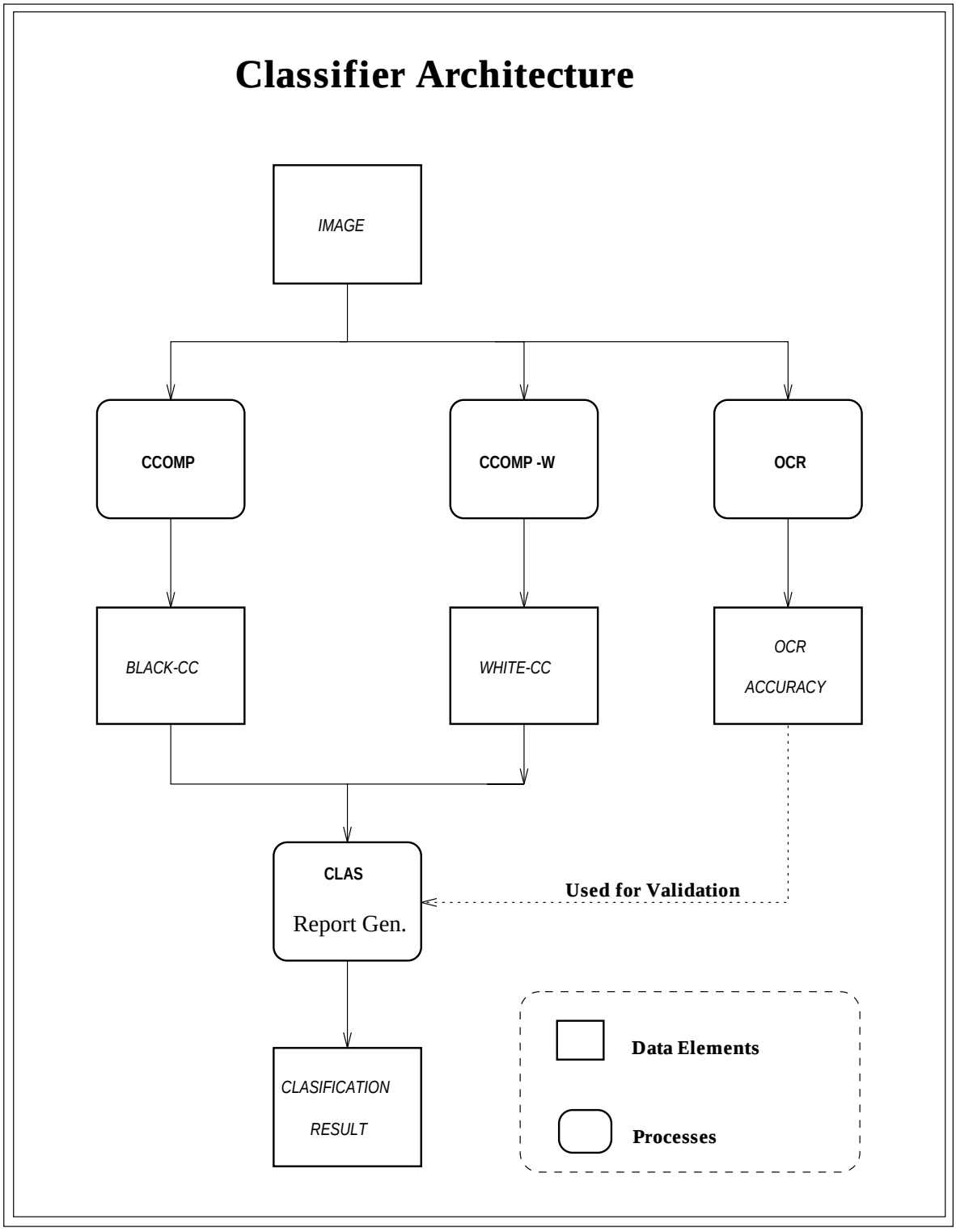


Figure 4.1: Classifier Logic Architecture

- Iterate over this list classifying each page and outputting the results to one file.
- Derive tables from the output file and the (independently tested) OCR accuracy values.

The reports and confusion matrices are generated automatically from the results file. The scripts to perform these tasks are written in the PERL programming language [17]. The connected components finder is written in C, as is the feature extractor from the CC data. The whole process is driven by a PERL script which produces the result file.

Test Data Set

The test data set consists in 439 pages that were taken from ISRI's Sample 2 Document Database. The set of pages have the following characteristics:

- Mostly black-on-white pages, although there are some pages containing white-on-black text zones.
- Some of the pages contain tables

The pages were processed by eight OCR devices (see Table 4.1)[1]. The median OCR accuracy was computed for each page from the results of these eight devices and that was the accuracy value used to label the pages as "Good" or "Bad".

Unfiltered Results

The test dataset was processed by the classifier. In this experiment, a "good" page is defined as a page where the median OCR accuracy is equal to or higher than 90%. Appendix A presents the complete results of this test. Table 4.2 shows the

Vendor	Version Name	Version #
Caere Corp.	Caere OCR	109
Calera Recognition Systems, Inc.	Calera MM600	mm24su
Cognitive Technology Corp.	Cognitive Cuneiform	0.8
CTA, Inc.	CTA TextPert DTK	1.2.9
ExperVision, Inc.	ExperVision RTK	2.0
OCRON, Inc.	OCRON Recore	2.0.5
Recognita Corp. of America	Recognita Plus DTK	2.0β.BC3
Xerox Imaging Systems, Inc.	XIS ScanWorX API	2.0β3

Table 4.1: OCR Devices Processing the Test Data

True ID	Recognized	
	Good	Bad
Good	349	53
Bad	**15**	22

Table 4.2: Confusion Matrix for all 439 pages (good = 90%)

confusion matrix for this test. The classifier worked as expected in most cases, even though it made 15 misclassifications of "Bad" pages as "Good".

The Error Rate is calculated:

$$\begin{aligned} Error\ Rate &= \frac{B \rightarrow G + G \rightarrow B}{Total\ Number\ of\ Classified\ Pages} \\ &= \frac{15 + 53}{439} \\ &= 0.15 \end{aligned}$$

Error Analysis

The 15 "B $\rightarrow$ G" misclassifications (Table 4.2) were carefully examined. The page images did not present substantial degradation, confirming the results of the classifier. Figure 4.2 shows excerpts from some of these images where the quality can be appreciated.

[1]Reproduced from [12] with permission.

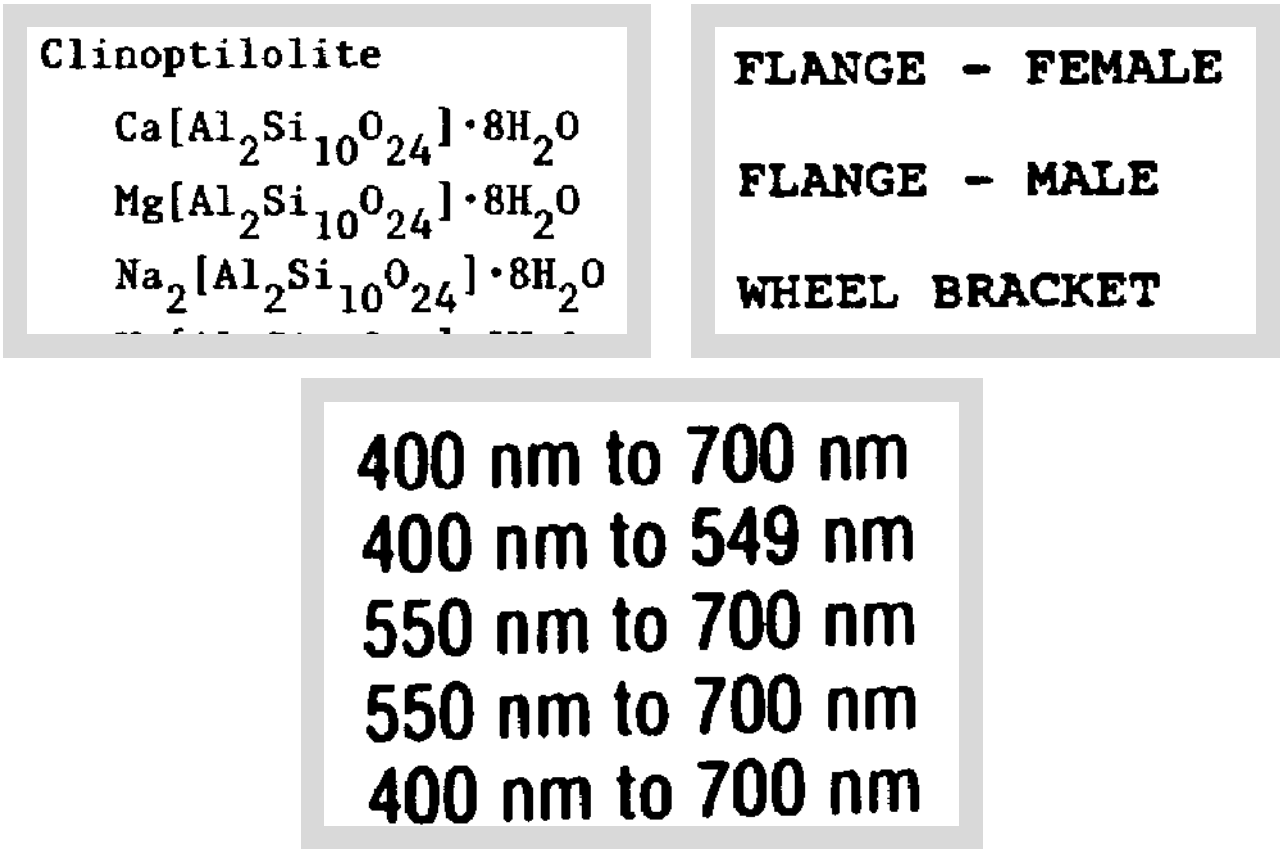


Figure 4.2: Excerpts from B→G misclassified images (magnified)

IS-AS	[0-80]	(80-90]	(90-95]	(95-98]	(98-99]	(99-100]
B - B	8	14	0	0	0	0
B - G	**5**	**10**	**0**	**0**	**0**	**0**
G - B	0	0	16	21	10	6
G - G	0	0	28	79	68	174

Table 4.3: Results by Accuracy - All 439 pages (good = 90%)

All but 4 of these 15 "problematic" pages contained tables in them. The OCR output for many of these tables were illegible and generally useless. Figures 4.3 and 4.4 show a clean table image and its associated OCR output. Tables pose special problems to OCR devices.

The classifier labeled all these 11 pages containing tables "Good" because, from an image defects point of view, it could not find enough evidence to give the pages a "Bad" label. The contents of these pages, however, suggest that table-generated OCR errors are special and, therefore, are not related to image-generated OCR errors. After evaluating all 11 "Bad $\rightarrow$ Good" misclassified "table pages", the following observations are in order:

Table Observation 1. There is a marked difference in OCR performance among different OCR devices when handling pages with tables. Table 4.4 lists the 11 B$\rightarrow$G misclassified pages. It can be seen that, in general, devices in the left part of the table tend to do much better than those on the right part of the table. Because of this variability, we can no longer assume that a page with median accuracy $< 90\%$ is a bad page, since the OCR results can be radically different depending upon the OCR device used.

Table Observation 2. Character recognition measures are not enough for Table-OCR evaluation. When tables are present, an OCR user is not only interested in the contents (i.e., characters) on the table but also in the table's overall

TABLE LXVIII

THERMODYNAMIC DATA FOR ZEOLITES

Mineral	ΔG°_f (kcal/mole) TG[a]	Modified TG[a]	Nriagu[b]	Chen[c]	Experimental ΔG°_f (kcal/mole) SUPCRT[d]	Robie and Waldbaum[e]
Analcime						
$Na[AlSi_2O_6]\cdot H_2O$	-747.3	-741.0	-732.5	-734.9	-738.1	-734.3
Wairakite						
$Ca[Al_2Si_4O_{12}]\cdot 2H_2O$	-1502.0	-1479.0	-1474.0	-1477.8	-1477.8	--
Laumonite						
$Ca[Al_2Si_4O_{12}]\cdot 4H_2O$	-1620.4	-1597.4	-1586.6	-1591.1	-1597.0	--
Clinoptilolite						
$Ca[Al_2Si_{10}O_{24}]\cdot 8H_2O$	-3084.8	-3061.8	-3047.8	-3045.8	--	--
$Mg[Al_2Si_{10}O_{24}]\cdot 8H_2O$	-3061.5	-3051.2	-3035.7	-3024.5	--	--
$Na_2[Al_2Si_{10}O_{24}]\cdot 8H_2O$	-3077.4	-3064.8	-3038.8	-3040.2	--	--
$K_2[Al_2Si_{10}O_{24}]\cdot 8H_2O$	-3090.0	-3090.0	-3062.8	-3065.8	--	--
Heulandite						
$Ca[Al_2Si_7O_{18}]\cdot 6H_2O$	-2352.6	-2329.6	-2317.2	-2318.4	--	--
Mordenite						
$Na[AlSi_5O_{12}]\cdot 3H_2O$	-1479.5	-1473.2	-1463.1	-1463.2	--	--
$K[AlSi_5O_{12}]\cdot 3H_2O$	-1485.8	-1485.8	-1475.1	-1476.2	--	--

Figure 4.3: Clean Table Image

```
TABLE LXVIII
THERMODYNAMIC DATA FOR ZEOLITES
AG Experimental AG,
<Kcal/mole> <Kcal/mole>
MAneral TG MoAifieA TG' NrAagu cKen' Rotxe anA
SCPCRT w-I* -
 -I-*--
N-  AIS*,o, .H,o -T T.   -   .c -   .s -   .s -  s.   -   .
w- --Kit.
c-  AI,si,o,,l.2H,o - soz. o - * a.c -    .n -  *  . s - 4TT. s --
Laumonite
Ca (AI,Si.O,,) * 4H,O - 1620.4 -1597.4 -1586.6 -1591. I -1597.0 --
Cliaoptilolite
" ' '2':10 241 '8H,o -so84.8 - 061. s - o47.8 -so45.8 -- --
"" '2':l0 241 '8H,0 - 061.5 - 051.2 - 0 5.7 - 024.5 -- --
"*,IAI,sil00241 *8H,O - 077.4 - 064.8 -aoaa.8 - 040.2 -- --
'2' '2':l0 241 "",o - 090.0 - 090.0 -s062.8 - 065.8 -- --
Heulan*i .
'* IAI,':7 181 '6H,o -2 52.6 -2 29.6 -2 17.2 -2 18.4 -- --
MorAen te
N-  Alsi,0,,1 *SH,o -1479.5 -147 .2 -146 . I -146 .2 -- --
" ":5 121 ' H, -1485.8 -1485.s -1475. I -1 76.2 -- --
```

Figure 4.4: OCR Output for Clean Table Image

layout. This information is critical for tables with empty cells, where if the OCR algorithm does not generate the proper layout, neighboring cell values can be wrongly assigned to these empty cells. A character recognition measure such as the one used at ISRI, is not sophisticated enough to check table formatting and thus not well suited for the kind of accuracy evaluation in consideration. The development of new ways to measure Table-OCR accuracy is beyond the scope of this work, but should be addressed if table-generated OCR output is to be evaluated.

Table Observation 3. Numeric tables seem to present more difficulty to OCR algorithms than textual tables and normal text. The majority of the errors found in the OCR output for the evaluated tables stemmed from numerical data. Substitutions of "0" by "O", "1" by "l" were among the most common. Some OCR devices are strongly biased towards textual information and, therefore, make

	OCR Accuracy							
Page-ID	**Dev.A**	**Dev.B**	**Dev.C**	**Dev.D**	**Dev.E**	**Dev.F**	**Dev.G**	**Dev.H**
6347-112	91.070	98.100	92.970	27.830	-10.450	68.280	22.410	33.900
1124-024	92.190	91.380	95.240	68.670	62.300	65.260	39.140	78.990
2136-083	87.620	71.150	75.150	85.960	82.750	77.390	69.590	69.400
5347-145	87.670	94.990	83.830	94.740	80.230	57.180	48.810	76.590
1674-088	94.730	89.720	83.820	79.230	85.390	83.350	84.660	66.280
1796-095	74.200	91.440	95.050	78.480	86.360	85.160	82.350	89.300
1674-138	93.600	94.830	89.130	88.590	78.420	79.660	85.520	79.980
1060-223	94.070	86.480	90.630	88.020	75.210	91.220	86.480	62.510
2306-093	84.690	85.150	91.180	87.240	88.630	82.600	89.560	89.330
6546-011	94.210	97.520	92.980	91.120	83.680	85.330	82.230	84.090
5830-240	95.910	85.340	86.640	95.610	90.230	92.920	68.890	86.240

Table 4.4: B→G Misclassified Tables, listed by device

these kind of errors in purely numeric data. Furthermore, OCR devices cannot use lexicon-based correction for numerical data.

It is important to make clear that none of these three observations are dependent on page quality. Furthermore, after visually inspecting the images, tables that presented poor image quality were in general correctly flagged as "Bad" by the classifier and their OCR output was subject to the normal image quality-related errors (in addition to the special problems posed by tables and mentioned above). Similarly, pages labeled as "Good" by the classifier were visually inspected and found to indeed have high image quality. Some of these pages were considered "Bad" in the experiment because their low OCR accuracy stem from the special characteristics of tables mentioned above and not from image defects. Furthermore, many of the "B→G" pages would not be misclassified if the devices used in the experiment had been a selected subset of the ones used. This variability in OCR output is not usual in "normal" textual pages.

True ID	Recognized		Rejects
	Good	Bad	Table
Good	257	42	103
Bad	**4**	18	15

Table 4.5: Confusion Matrix - Tables Filtered - (good = 90%)

Therefore, a *reject region* was established to filter-out all the pages containing tables. These pages would need to be processed by another type of classifier since the difficulty they present to OCR algorithms does not result from image quality, but from the complexity of their contents and layout. Table 4.5 shows the confusion matrix with the addition of the *Table Reject* column. All the pages with tables in them were not processed by the classifier and were assigned to the Table Reject column.

The Error and Reject Rates are now:

$$Error\ Rate = \frac{4 + 42}{321} = 0.14$$

$$Reject\ Rate = \frac{103 + 15}{439} = 0.27$$

After examining the results analyzed by the number of connected components on the page (see Table 4.6), it can clearly be seen that a reject zone for any page with 200 connected components or less has to be implemented.

The rationale behind this decision takes root in that this classifier is based in measured ratios. Pages with a low number of connected components are not "stable" enough to present credible ratios, since a little variation can result in a very high (or low) ratio. Therefore, having a cutoff number is a requirement to make the classifier robust.

The results obtained after applying this new filter are shown in Table 4.7, where no "B→G" misclassifications exist.

IS-AS	[0-100]	(100-200]	(200-300]	(300-400]	(400-500]	(500++]
B - B	6	1	4	0	0	7
B - G	**2**	**2**	**0**	**0**	**0**	**0**
G - B	6	1	2	2	1	30
G - G	29	17	9	5	4	193

Table 4.6: Results by #CCs - Tables Filtered - (good = 90%)

True ID	Recognized		Rejects	
	Good	Bad	Table	#CCs
Good	211	35	103	53
Bad	**0**	11	15	11

Table 4.7: Confusion Matrix - Filtered on Tables and #CCs (good = 90%)

These last results produce the following Error / Reject Rates:

$$Error\ Rate = \frac{0 + 35}{257} = 0.136$$

$$Reject\ Rate = \frac{103 + 15 + 53 + 11}{439} = 0.41$$

In this case, however, there are no "B $\rightarrow$ G" misclassifications, which was a desired goal with regards to quality control.

There are a considerable number of "G$\rightarrow$B" misclassifications. After evaluating the misclassified pages as well as the triggered rules that produced the misclassifications, the following considerations are in order:

- Rule #1, the White Speckle Factor, seems to be oversensitive and is producing the bulk of the errors (see Table 4.8). On the other hand, the same rule is correctly classifying Bad pages (see Table 4.9). Thus, a quick solution for this rule is not obvious. A complimentary rule or, better yet, an improved way to detect touching characters may be needed. It is to be noted that this rule was

G→B	
Rule	**Count**
1	27
2	3
3	5

Table 4.8: Rules Triggered for G→B Classification

not designed to handle small fonts and several of the misclassified pages have fonts with size < 10pt in them. This rule could be regarded as a necessary but not sufficient condition for the existence of fat and/or touching characters.

- Rule #3, the Black Size rule, is the second largest cause of errors in the G→B misclassifications. Furthermore, it does not uniquely flag a Bad page as such (Table 4.9). Therefore, based on the test dataset results, this rule could be eliminated altogether and the number of misclassifications (of any kind) would be lowered by 5.

- The classifier is definitely biased towards filtering out all bad pages. This would need to be revised in order to provide a finer degree of control. This behavior is, however, appropiate in a large scale OCR environment, where it is critical not to let any bad page slip by the filter in order not to incur in error-correction costs.

Higher Good Thresholds

Until now, a page has been considered "Good" for OCR purposes if it produces an OCR output with at least 90% accuracy. In [4] it is suggested that a page should be considered "good" only if it is in the 95%-98% accuracy range, depending on its textual contents' difficulty. The classifier was therefore run twice, assuming a "good threshold" of both 95% and 98%, respectively. The results follow.

B→B	
Rule	**Count**
1	3
2	4
3	0
4	1
1 and 2	2
1 and 3	1

Table 4.9: Rules Triggered for B→B Classification

True ID	**Recognized**	
	Good	**Bad**
Good	202	28
Bad	**9**	18

Table 4.10: Confusion Matrix - Filtered by Tables and #CCs (good = 95%)

Good Threshold = 95% Results

Table 4.10 shows the confusion matrix for a threshold of 95% after having filtered out all pages containing tables and/or less than 200 connected components.

The error rate is:

$$Error\ Rate = \frac{9 + 28}{257} = 0.144$$

It is to be noticed that this error rate is not too much higher than the one at 90% threshold, but now there are 9 "B → G" misclassifications.

Tables 4.11 and 4.12 display these results analyzed by accuracy and number of connected components.

Good Threshold = 98% Results

The confusion matrix for a good = 98% threshold is shown in Table 4.13, and the accuracy and number of CCs view are given in Tables 4.14 and 4.15 respectively.

IS-AS	**[0-80]**	**(80-90]**	**(90-95]**	**(95-98]**	**(98-99]**	**(99-100]**
B - B	4	7	7	0	0	0
B - G	**0**	**0**	**9**	**0**	**0**	**0**
G - B	0	0	0	16	8	4
G - G	0	0	0	33	42	127

Table 4.11: Results by Accuracy - Filtered by Tables and #CCs (good = 95%)

IS-AS	**(200-300]**	**(300-400]**	**(400-500]**	**(500++]**
B - B	4	0	0	14
B - G	**2**	**2**	**0**	**5**
G - B	2	2	1	23
G - G	7	3	4	188

Table 4.12: Results by #CCs - Filtered by Tables and #CCs (good = 95%)

The error rate is:

$$Error\ Rate = \frac{42 + 12}{257} = 0.21$$

True ID	**Recognized**	
	Good	**Bad**
Good	169	12
Bad	**42**	34

Table 4.13: Confusion Matrix - Filtered by Tables and #CCs (good = 98%)

IS-AS	[0-80]	(80-90]	(90-95]	(95-98]	(98-99]	(99-100]
B - B	4	7	7	16	0	0
B - G	**0**	**0**	**9**	**33**	**0**	**0**
G - B	0	0	0	0	8	4
G - G	0	0	0	0	42	127

Table 4.14: Results by Accuracy - Filtered by Tables and #CCs (good = 98%)

IS-AS	(200-300]	(300-400]	(400-500]	(500++]
B - B	6	2	0	26
B - G	**2**	**2**	**2**	**36**
G - B	0	0	1	11
G - G	7	3	2	157

Table 4.15: Results by #CCs - Filtered by Tables and #CCs (good = 98%)

Magazine Data

Two hundred magazine pages were also run through the classifier to test its performance. Magazine pages are very different to standard "document-type" pages because they often contain artistic fonts, graphs, color, etc. The difference between these pages and the ones used to create the classifier make the magazine dataset a perfect choice for testing the classifier's performance in a different environment.

The magazine dataset consists of 200 pages taken from the top 100 magazines in the US, according to their circulation. Two pages were randomly selected from each magazine and each page was clipped out, scanned in, and the *truth* text file was generated. All the pages were manually zoned. All parts of the page zoned except for commercial advertisments and pictures [13]. Table 4.16 lists the types of zones used in the preparation of this data[2]. Each page was processed by 6 OCR devices and, as with the test dataset, the median accuracy was computed. Table 4.17 lists the devices used.

[2]Reproduced from [13] with permission

Zone Type	# of Zones	# of Chars.
"Main body" Text	1072	630441
Table	8	5462
Caption	153	25403
Footnote	2	655
Header/Footer	179	4173
Total	1414	666134

Table 4.16: OCR Devices Processing the Magazine Data

Vendor	Version Name	Version #
Caere Corp.	Caere OCR	132
Calera Recognition Systems, Inc.	Calera WordScan	4
Electronic Document Technology	EDT ImageReader	2.0
ExperVision, Inc.	ExperVision RTK	3.0
Recognita Corp. of America	Recognita Plus DTK	2.00.D12
Xerox Imaging Systems, Inc.	XIS OCR Engine	10

Table 4.17: OCR Devices Processing the Magazine Data

All 200 pages were processed by the classifier without pre-filtering. The confusion matrices for "good thresholds" of 90%, 95% and 98% are shown in Tables 4.18, 4.19, 4.20, respectively. Appendix B contains the complete classification results for all 200 pages.

The error rates for each good threshold are:

$$Error\ Rate\ (90\%) = \frac{0 + 27}{200} = 0.135$$

$$Error\ Rate\ (95\%) = \frac{13 + 17}{200} = 0.15$$

$$Error\ Rate\ (98\%) = \frac{45 + 4}{200} = 0.245$$

True ID	Recognized	
	Good	Bad
Good	159	27
Bad	**0**	14

Table 4.18: Confusion Matrix for 200 Magazine Pages (good = 90%)

True ID	Recognized	
	Good	**Bad**
Good	146	17
Bad	**13**	24

Table 4.19: Confusion Matrix for 200 Magazine Pages (good = 95%)

True ID	Recognized	
	Good	**Bad**
Good	114	4
Bad	**45**	37

Table 4.20: Confusion Matrix for 200 Magazine Pages (good = 98%)

The classifier did very well on the 90% threshold because it did not incur in any B→G misclassifications. As expected, its performance degraded at the higher thresholds.

The classifier was then modified as suggested by the results obtained with the test dataset. Consequently, the Black CC Size rule (Rule #3) was disabled and the magazine dataset was again run through the classifier. The confusion matrix is presented in Table 4.21 and, contrary to what was expected, the error rate went up and the number of misclassifications remained the same (compare Tables 4.18 and 4.21). Based on the results, Rule #3 works well for the magazine dataset but poorly for the test dataset. This preliminary evidence suggests that generalizations cannot be made about the behavior of the classifier in a different environment.

True ID	Recognized	
	Good	**Bad**
Good	159	27
Bad	**1**	13

Table 4.21: Conf. Matrix for Magazine Dataset and Modified Classifier

Summary

After the implementation of a reject region to filter out tables and small zones, the classifier was able to correctly detect all pages with OCR accuracy of less than 90% in the Test Dataset. Some misclassifications of Good pages as Bad were incurred, but the overall error rate was consistently below 15%.

Reject regions had to be implemented because the current version of the classifier is not able to detect defects on images containing a very low number of connected components (characters) and because table-generated OCR errors do not appear to be directly related (dependent) to image quality.

The system degraded gracefully as the cutoff threshold for "good" and "bad" labels was moved up. This is to be expected, mainly because this classifier uses only simple features. A more complex approach is needed to differentiate accuracies in the 95% and above region.

The classifier also processed a completely different dataset, the Magazine dataset. It performed flawlessly in filtering out bad pages at the 90% threshold.

The simple features selected have proven to be useful in detecting image quality to a certain level of detail. The results indicate that the classifier logic would be applicable not only to pages conforming to the type it was created for, but also to other types of pages and possibly to all pages. Further testing is required to validate this last hypothesis since improved features would be required to increase the level of detail the classifier must be able to detect.

Chapter 5

Future Work

The classifier presented in this work is the first attempt in the page-quality metrics arena and, therefore, very simple. This chapter presents new ideas to extend and enhance this research.

Statistical Pattern Recognition

Instead of a heuristic approach, a more traditional statistical approach could be used. Issues to explore in such a case would be:

- **More training data.** In this research, only 24 pages were used as the training dataset. In order to implement a statistical classifier, more data is needed for the training and designing phases.

- **Risk concept.** The introduction of the risk concept can be included and the thresholds updated according to the related specifications. This approach would be the correct way to handle the two different misclassifications (B $\rightarrow$ G and G $\rightarrow$ B) with two different weights. In this thesis, the B$\rightarrow$G misclassification has been determined to be of higher risk than the other type of classification; this determination has driven the design of the classifier. However, in order to cope with higher classifier accuracy, a better risk model is required.

- **Confidence.** Instead of producing just a single binary output, it would be interesting to provide degrees of confidence to the classifier responses. A response of "Good" or "Bad" would be followed by a degree of "certainty".

- **Quantizied output.** Instead of reporting "good" or "bad", the degree of image defects for each would be more convenient. In such a model, a rate of 0 would mean, for instance, "no defect" while a rate of 1 would be "severe defect". Rates for touchiness and brokeness should be reported separately as they could both be present within some pages.

Features Observed but not Used

The use of new and more complex features would be a key component in a production-type classifier. Specifically if the user is interested in higher "good thresholds", more features will be needed for the filter to use.

While performing the research for this thesis, several features were discovered which could be useful in a full-blown system. These features along with a brief explanation of their significance follow:

- **Overlapping.** The amount of overlapping between two neighboring connected components' boxes could serve as an indicator of both the font complexity and of deformed and/or broken characters (see Observation 5 in Chapter 3).

- **Skew angle.** The degree of skew of an image could very well be a strong indicator of the quality of the image. If the skew degree is more than a certain threshold [13] then the page should be considered "Bad".

- **Width distribution.** In order to determine the amount of touchiness in a page, the width distribution would probably have to be calculated and information compiled from that calculation.

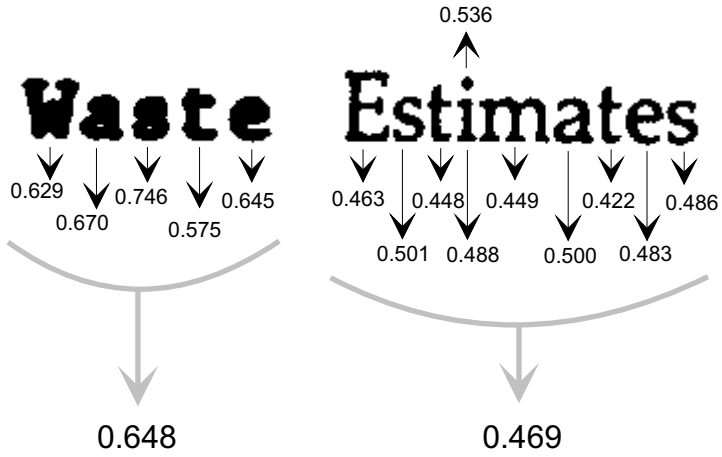


Figure 5.1: Black Density for Connected Components

- **Micro-Gaps.** In this work, very small white blobs were used to estimate the degree of thickness (and therefore touchiness) the characters have. In the same vein, a metric to detect micro-gaps[1] would account for a large amount the broken characters and also for the degree of character complexity (see Observation 4 in Chapter 3).

- **Filled CC boxes.** A way to detect completely filled lakes in letters like "e", "a", etc, would be to measure the black density inside a CC box of certain (minimum) dimensions. Figure 5.1 shows an example of this metric in action for two real-word string of characters. The black densities of each connected component, and for the collection of connected components, are shown for a "fat" and a "normal" character strings.

- **Deformed contours.** The degree of complexity of the contours of a character could also be a very good predictor of the font complexity and the paper/scanner quality. Figure 5.2 shows a "well-formed" and a "deformed" character.

[1] Very thin white-space separating a character's stroke

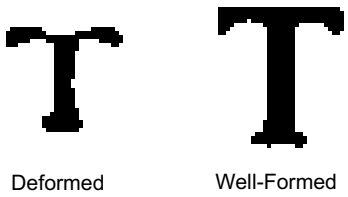


Figure 5.2: Deformed and Well-Formed Characters

New Datasets

In this work, only one type of page has been concentrated on. In order to produce a full-blown system, a more heterogeneous dataset must be devised. Among the kind of data that would certainly be needed are:

- **Faxed documents.** The poor resolution and the amount of line noise introduced in faxed documents make this type of data ideal for page quality classification. Reseach is underway at ISRI to address this issue.

- **Foreign Language Documents.** In any classifier based on "normal" vs. "abnormal" ratios, any change in the "correct" character set can be a major problem, since the ratios can and often do change. It is therefore of great importance to tune the classifier for the language of choice. Note that since lexicographical features were not used, nor was the OCR output relied upon, changes for alphabets with characters highly similar to English (like Spanish) should be minimal. For other languages (i.e., Japanese, Chinese, Arabic); however, different features will have to be studied since the characters in these languages are radically different to the ones used in English.

Chapter 6

Concluding Remarks

A first attempt to a page quality estimator has been presented. The classifier is based upon measuring simple features from the connected components' data of a page image. Only parts of the image that contain textual information were used to design and train the classifier.

The features used are the *white speckle factor*, the *broken zone factor* and *size information*. The white speckle factor measured the density of small white connected components, and was designed to capture minimally open lakes in pages with very bold/fat characters. The broken zone factor measures the coverage of an area in the width-height map of the connected components' data, presumably populated by the broken pieces of the characters in a broken-characters' page. The size information metrics measure the maximum average black and white connected components' size as well as the ratio of black to white connected components in order to rule out pages with unusually large fonts and/or inverse video.

After testing the classifier on a 439-page test dataset, it was observed that tables were not correctly processed. Further study on the table pages showed that tables present special difficulty for OCR algorithms. Specifically, a great difference in performance on tables among OCR systems was discovered. The need for a table-specific OCR accuracy evaluation model was acknowledged and the importance of numeric

data in table-related OCR errors was emphasized. Because of table problems, a reject region was established to filter out all pages with tables. After that, the classifier was able to correctly filter out all bad pages, with the exception of four. After evaluating these four pages, a new reject region based in the number of connected components was defined to filter out all pages with less than 200 connected components. The classifier is based in computing densities and ratios and, therefore, needs a minimum number of data to work reliably.

From the analysis of the results it was observed that the white speckle factor is not a very robust feature, since it breaks down in the prescence of small font and other special circumstances. The need for a better metric was therefore recognized. The broken character zone factor, on the other hand, proved to be very robust and worked fairly well in all the tests performed. After evaluating some of the size information rules, we observed that one was not being triggered to detect bad pages and was the cause of many misclassifications. The removal of this rule could enhance the classifier.

A new dataset was put together to further test the classifier. Two hundred magazine pages, with radically different characteristics from pages in the previous dataset, were assembled and processed by the classfier. The pages were processed with and without one of the size information rules. The results showed that, for this particular dataset, the rule works fine in classifying a bad page as such, and it did not produce any misclassifications. On the other hand, after removing the rule and re-testing, errors were introduced in the classification. This evidence suggests that the classifier, as designed, is data dependant and it must be tuned for the kind of data it will be processing.

Testing was performed assumming that a good page has a median OCR accuracy of 90% and above; the results hereby described are based upon this assumption. However, testing was also performed at the 95% and 98% levels. In these cases, the

performance of the classifier degraded gracefully. The conjecture is that more complex features are required to identify subtler image defects and, even then, there are errors that are not related to image quality that will not be captured by the classifier.

A number of features that were observed but not used are presented. Among them, the black density, micro-gap detection and skew are very promising page-quality related indicators. Similarly, testing on fax and foreign language documents is identified to be a requirement in producing a full-blown image quality classifier. For this purpose, a heuristic binary decision system such as the one presented is not sophisticated enough. A more standard statistical approach should be taken.

This work has presented an image-quality estimator based on very simple features. The classifier was able to attain a stable error rate of approximately 14% in two different datasets. The contribution of this research lies in being the first approach at such a classifier and in establishing a base for future research to build on.

Appendix A

Classifier Results for Test Dataset

Page Characteristics				Measured Features					Classifier Logic		
PageID	NCC	T	Acc.	White Speckle	Broken Zone	M B	M W	B / W Ratio	Rules	R	ClasAs
0101-003	872	Y	99.239	0.000000	0.125000	22	17	2.71651		Y	GOOD
0103-040	4303	N	99.851	0.002380	0.296875	22	12	2.55979		N	GOOD
0103-083	4289	N	99.829	0.000580	0.265625	22	12	2.48782		N	GOOD
0103-091	4285	N	99.794	0.001297	0.259259	21	12	2.77886		N	GOOD
0108-090	4314	N	99.828	0.002454	0.259259	21	12	2.64663		N	GOOD
0110-099	1621	Y	99.285	0.001656	0.175258	24	20	2.68377		Y	GOOD
0111-003	244	N	69.369	0.005650	0.187500	25	46	1.37853	4	N	BAD
0112-094	2652	N	99.967	0.000000	0.219178	21	12	2.86084		N	GOOD
0112-431	112	N	100.000	0.000000	0.152174	19	60	2.80000		Y	GOOD
0113-013	2690	Y	95.693	0.048356	0.428571	16	13	5.20309		Y	GOOD
0113-310	2278	N	99.745	0.002466	0.246575	21	13	2.80888		N	GOOD
0113-381	2360	N	99.758	0.002375	0.301370	21	13	2.80285		N	GOOD
0113-445	3028	N	99.721	0.002788	0.246575	21	12	2.81413		N	GOOD
0122-003	1606	N	99.091	0.002677	0.169355	27	11	2.14993		N	GOOD
0146-281	1591	Y	96.567	0.003802	0.283333	23	17	3.02471		Y	GOOD
0147-038	2804	N	99.885	0.000000	0.226415	22	13	2.81244		N	GOOD
0147-079	292	N	99.690	0.000000	0.112903	22	43	3.07368		N	GOOD
0147-400	720	N	97.527	0.000000	0.141026	22	16	2.80156		N	GOOD
0148-123	607	N	99.382	0.005000	0.180328	21	28	3.03500		N	GOOD
0148-271	2061	Y	99.643	0.000000	0.241935	22	14	2.88655		Y	GOOD
0148-337	2562	N	99.801	0.000000	0.188679	22	12	2.97907		N	GOOD
0151-127	68	N	100.000	0.000000	0.055556	21	80	3.23810		Y	GOOD
0151-163	2937	N	99.818	0.001019	0.283019	22	13	2.99388		N	GOOD
0158-010	1486	N	99.675	0.000000	0.315068	21	14	2.72161		N	GOOD
0161-030	2338	N	99.923	0.000000	0.338462	21	13	2.90074		N	GOOD
0161-056	3756	N	98.789	0.002419	0.441176	18	11	3.02903		N	GOOD
0166-009	3044	N	99.897	0.003581	0.230769	21	13	2.72516		N	GOOD

Page Characteristics				Measured Features					Classifier Logic		
PageID	NCC	T	Acc.	White Speckle	Broken Zone	M B	M W	B / W Ratio	Rules	R	ClasAs
0168-031	2719	N	99.967	0.001011	0.276923	21	13	2.74924		N	GOOD
0188-005	2762	N	99.868	0.001092	0.338462	21	13	3.01528		N	GOOD
0199-384	1421	Y	94.578	0.051395	0.322581	28	17	2.08664		Y	GOOD
0199-646	1674	N	98.912	0.003540	0.171875	24	14	2.96283		N	GOOD
0201-041	153	N	97.771	0.057692	0.078947	19	58	2.94231		Y	GOOD
0201-132	337	N	97.165	0.305556	0.134615	21	29	3.12037	1	N	BAD
0202-138	1687	N	97.973	0.013889	0.178571	26	12	2.34306		N	GOOD
0203-027	2672	N	99.847	0.004484	0.172840	23	11	2.39641		N	GOOD
0203-075	1888	Y	99.310	0.006527	0.134021	24	15	2.46475		Y	GOOD
0203-109	2995	N	99.687	0.005040	0.209877	23	11	2.15623		N	GOOD
0206-007	2393	Y	99.496	0.008026	0.319149	18	15	3.84109		Y	GOOD
0207-018	2495	N	99.691	0.016598	0.144444	23	11	2.07054		N	GOOD
0214-036	1448	N	99.712	0.025180	0.129630	28	13	2.60432		N	GOOD
0216-228	2800	N	98.479	0.005263	0.366667	23	15	2.94737		N	GOOD
0216-256	1084	N	97.762	0.005602	0.229167	25	24	3.03641		N	GOOD
0216-262	2322	N	97.814	0.018570	0.291667	25	12	2.15599		N	GOOD
0219-092	2094	N	99.208	0.018610	0.148148	22	11	2.59801		N	GOOD
0220-030	1035	Y	98.652	0.008264	0.062500	26	16	2.85124		Y	GOOD
0224-042	81	N	98.851	0.029412	0.061728	23	72	2.38235		Y	GOOD
0232-018	2542	N	99.817	0.023810	0.137931	24	9	1.89137		N	GOOD
0232-070	83	N	92.778	0.000000	0.041667	23	94	3.07407		Y	GOOD
0641-059	87	N	93.889	0.031250	0.148148	23	72	2.71875		Y	GOOD
0651-008	140	Y	95.000	0.026316	0.150000	23	56	3.68421		Y	GOOD
0651-013	687	Y	97.327	0.024390	0.122093	34	29	3.35122		Y	GOOD
0656-027	63	N	99.275	0.000000	0.027397	27	85	2.73913		Y	GOOD
0668-004	720	N	99.814	0.000000	0.090909	22	14	2.66667		N	GOOD
0668-062	39	N	93.023	0.000000	0.014815	32	90	2.29412		Y	GOOD
0672-233	2412	N	99.449	0.000000	0.377358	22	13	2.90953		N	GOOD
0683-004	622	N	98.684	0.004484	0.315789	19	39	2.78924		N	GOOD
0685-048	1716	Y	96.470	0.001916	0.534247	21	21	3.28736		Y	GOOD
0699-030	2987	N	99.747	0.002885	0.250000	22	13	2.87212		N	GOOD
0725-024	1116	N	99.671	0.035955	0.135802	22	13	2.50787		N	GOOD
0725-026	832	Y	99.148	0.036697	0.115385	22	20	2.54434		Y	GOOD

Page Characteristics				Measured Features					Classifier Logic		
PageID	NCC	T	Acc.	White Speckle	Broken Zone	M B	M W	B / W Ratio	Rules	R	ClasAs
0729-021	74	N	96.753	0.037037	0.041667	22	80	2.74074		Y	GOOD
0729-187	2058	N	99.161	0.006588	0.188889	23	11	2.71146		N	GOOD
0743-013	346	N	98.039	0.012821	0.222222	22	23	2.21795		N	GOOD
0765-005	3736	Y	99.691	0.000000	0.500000	12	14	7.19846		Y	GOOD
0765-018	1736	N	99.821	0.002886	0.138889	23	19	2.50505		N	GOOD
1024-011	153	N	100.000	0.000000	0.038961	26	57	2.68421		Y	GOOD
1040-032	764	Y	95.440	0.015686	0.094340	22	19	2.99608		Y	GOOD
1051-001	97	N	95.327	0.459016	0.114583	36	42	1.59016	1	Y	BAD
1060-103	174	N	100.000	0.000000	0.071429	19	74	3.00000		Y	GOOD
1060-146	1071	Y	98.357	0.000000	0.123457	25	25	3.10435		Y	GOOD
1060-223	776	Y	87.248	0.000000	0.288889	18	16	3.07937		Y	GOOD
1081-024	1393	N	97.608	0.000000	0.308824	21	18	2.99570		N	GOOD
1081-057	1162	N	97.387	0.000000	0.222222	21	23	2.80676		N	GOOD
1091-002	1197	N	98.941	0.031185	0.342857	17	13	2.48857		N	GOOD
1110-076	142	N	98.944	0.000000	0.060606	25	39	3.02128		Y	GOOD
1111-014	1821	N	97.556	0.047214	0.193548	28	10	1.40944		N	GOOD
1112-023	2306	Y	96.637	0.018950	0.500000	21	19	3.36152		Y	GOOD
1124-024	947	Y	73.833	0.066890	0.235294	20	14	3.16722		Y	GOOD
1132-108	472	Y	92.623	0.017143	0.151515	27	19	2.69714		Y	GOOD
1148-088	2090	N	99.654	0.001205	0.247423	24	12	2.51807		N	GOOD
1164-047	326	N	93.629	0.000000	0.138889	22	24	2.93694		N	GOOD
1210-026	2551	N	97.106	0.003384	0.581818	22	16	4.31641		N	GOOD
1210-323	2518	N	92.702	0.017921	0.695652	19	12	3.00836		N	GOOD
1227-006	445	N	98.346	0.099448	0.180328	20	35	2.45856		N	GOOD
1238-006	3558	Y	91.790	0.112880	0.596154	19	9	1.28726	1	Y	BAD
1241-063	2544	N	99.550	0.000000	0.225806	23	12	2.56452		N	GOOD
1241-101	2922	N	98.772	0.011905	0.365854	18	10	2.67582		N	GOOD
1249-069	4395	N	99.581	0.121649	0.344262	21	10	2.26546	1	N	BAD
1275-200	2133	N	97.439	0.001639	0.483871	23	16	3.49672		N	GOOD
1279-003	1865	N	98.157	0.042980	0.262295	21	18	2.67192		N	GOOD
1335-003	1465	N	99.503	0.024876	0.250000	20	12	2.42952		N	GOOD
1339-086	1514	Y	96.902	0.000000	0.205479	20	14	2.33642		Y	GOOD
1343-180	2339	N	99.715	0.019406	0.177419	22	10	2.67009		N	GOOD

Page Characteristics				Measured Features					Classifier Logic		
PageID	NCC	T	Acc.	White Speckle	Broken Zone	M B	M W	B / W Ratio	Rules	R	ClasAs
1353-032	1811	Y	97.648	0.001828	0.382353	21	17	3.31079		Y	GOOD
1354-083	414	N	97.742	0.000000	0.065934	28	51	2.81633		N	GOOD
1356-083	1558	Y	99.398	0.004357	0.222222	23	23	3.39434		Y	GOOD
1356-205	2288	N	99.884	0.000000	0.277778	22	12	2.44184		N	GOOD
1360-074	809	Y	94.835	0.003663	0.096296	28	30	2.96337		Y	GOOD
1367-022	2700	N	99.376	0.008717	0.266667	24	15	3.36239		N	GOOD
1367-074	2447	N	99.604	0.008178	0.169231	26	15	2.85864		N	GOOD
1367-239	2904	N	99.655	0.003933	0.184211	26	14	2.85546		N	GOOD
1368-074	566	Y	98.841	0.000000	0.104167	28	24	2.41880		Y	GOOD
1371-063	370	N	99.645	0.000000	0.111111	23	33	2.68116		N	GOOD
1383-058	1387	N	99.268	0.045510	0.144444	23	11	1.70603		N	GOOD
1391-111	77	N	97.802	0.000000	0.132353	21	71	3.08000		Y	GOOD
1398-004	377	N	99.765	0.055944	0.088889	29	22	2.63636		N	GOOD
1399-001	1361	N	92.513	0.132883	0.169643	31	21	3.06532	1	N	BAD
1414-215	2080	N	99.845	0.000000	0.273973	21	13	2.85714		N	GOOD
1484-028	280	Y	94.776	0.025641	0.025000	66	27	2.39316	3	Y	BAD
1485-078	74	N	100.000	0.000000	0.030769	21	61	2.38710		Y	GOOD
1486-060	114	N	99.206	0.000000	0.055556	22	57	2.85000		Y	GOOD
1489-019	2839	N	99.823	0.005325	0.153846	21	12	3.02343		N	GOOD
1494-070	1997	N	97.981	0.015038	0.500000	21	18	3.00301		N	GOOD
1516-050	1199	N	99.505	0.002604	0.159420	25	27	3.12240		N	GOOD
1522-008	1435	N	99.499	0.028623	0.144444	23	12	2.56708		N	GOOD
1522-101	1390	N	99.743	0.006263	0.123457	23	13	2.90188		N	GOOD
1522-116	62	N	100.000	0.000000	0.055556	22	81	2.58333		Y	GOOD
1522-157	87	N	98.404	0.000000	0.061538	21	88	3.48000		Y	GOOD
1539-012	994	Y	92.318	0.023346	0.272727	26	33	3.86770		Y	GOOD
1542-010	1106	N	96.764	0.007331	0.072993	34	9	1.62170		N	GOOD
1550-001	3308	N	97.745	0.218013	0.604167	20	14	2.78451	1	N	BAD
1551-015	1194	N	99.887	0.007126	0.131579	26	15	2.83610		N	GOOD
1552-032	1399	Y	99.706	0.000000	0.211111	23	16	2.14242		Y	GOOD
1553-021	1433	Y	99.610	0.018557	0.197531	22	17	2.95464		Y	GOOD
1553-053	3004	N	99.926	0.003182	0.264706	21	10	2.38982		N	GOOD
1558-064	69	N	63.793	0.333333	0.068627	24	38	1.76923	1	Y	BAD

Page Characteristics				Measured Features					Classifier Logic		
PageID	NCC	T	Acc.	White Speckle	Broken Zone	M B	M W	B / W Ratio	Rules	R	ClasAs
1561-002	2520	N	95.533	0.121602	0.526316	19	14	3.60515	1	N	BAD
1569-009	514	N	99.301	0.000000	0.109589	25	16	2.16878		N	GOOD
1570-042	124	N	97.744	0.000000	0.084746	25	45	2.81818		Y	GOOD
1580-295	759	Y	99.014	0.000000	0.118644	25	30	2.86415		Y	GOOD
1599-087	252	N	99.635	0.000000	0.015385	26	31	2.76923		N	GOOD
1601-003	1529	N	92.615	0.235556	0.246914	25	9	2.26519	1	N	BAD
1624-063	668	Y	95.131	0.010239	0.074534	30	30	2.27986		Y	GOOD
1632-003	1082	N	99.501	0.002632	0.214286	25	16	2.84737		N	GOOD
1634-164	1898	N	99.865	0.001224	0.177215	24	10	2.32313		N	GOOD
1649-020	110	N	95.868	0.000000	0.024096	27	76	2.68293		Y	GOOD
1662-085	2920	Y	93.069	0.296852	0.294872	22	10	2.18891	1	Y	BAD
1662-092	4497	N	97.047	0.238573	0.444444	21	10	2.50669	1	N	BAD
1665-028	28	N	97.561	0.055556	0.000000	28	49	1.55556		Y	GOOD
1665-083	99	N	96.903	0.000000	0.009174	30	134	3.41379		Y	GOOD
1674-088	1707	Y	84.238	0.069405	0.230769	21	14	2.41785		Y	GOOD
1674-138	1650	Y	87.056	0.097978	0.333333	20	14	2.56610		Y	GOOD
1675-199	802	Y	92.320	0.010563	0.338710	23	20	2.82394		Y	GOOD
1675-209	1443	Y	96.199	0.005415	0.416667	21	17	2.60469		Y	GOOD
1676-008	806	Y	97.385	0.093248	0.283951	25	18	2.59164		Y	GOOD
1676-120	1824	N	91.632	0.275912	0.187500	24	10	2.66277	1	N	BAD
1676-346	2287	N	97.556	0.130597	0.415094	22	10	2.84453	1	N	BAD
1693-032	1759	N	98.794	0.000000	0.250000	22	13	2.73988		N	GOOD
1696-037	4141	N	98.916	0.000767	0.558824	18	11	3.17805		N	GOOD
1696-084	4064	N	95.972	0.000000	0.655172	17	20	3.69119		N	GOOD
1707-005	1651	N	97.687	0.000000	0.117647	27	14	3.02381		N	GOOD
1711-029	2307	N	98.922	0.030405	0.343750	22	8	1.94848		N	GOOD
1711-077	59	N	86.538	0.117647	0.029703	25	55	1.73529	1	Y	BAD
1717-039	173	N	99.756	0.000000	0.068966	24	38	2.27632		Y	GOOD
1719-007	1518	Y	99.074	0.003868	0.227273	25	23	2.93617		Y	GOOD
1723-157	2338	N	99.233	0.001376	0.459016	20	12	3.21596		N	GOOD
1723-194	275	Y	95.122	0.011628	0.166667	20	31	3.19767		Y	GOOD
1732-001	682	N	95.913	0.158482	0.211111	26	17	1.52232	1	N	BAD
1742-157	406	Y	96.916	0.000000	0.068966	31	70	2.68874		Y	GOOD

Page Characteristics				Measured Features					Classifier Logic		
PageID	NCC	T	Acc.	White Speckle	Broken Zone	M B	M W	B / W Ratio	Rules	R	ClasAs
1752-007	54	N	100.000	0.000000	0.054545	22	58	2.16000		Y	GOOD
1786-032	574	Y	90.459	0.009709	0.061538	20	17	1.85761		Y	GOOD
1788-030	122	N	98.438	0.000000	0.041667	21	101	4.06667		Y	GOOD
1796-095	581	Y	85.762	0.013953	0.085202	35	25	2.70233		Y	GOOD
1829-015	2167	N	99.875	0.000000	0.230769	26	15	2.91263		N	GOOD
1830-102	979	N	99.527	0.005747	0.206186	24	18	2.81322		N	GOOD
1834-036	1304	N	99.414	0.000000	0.146667	24	26	2.82251		N	GOOD
1852-024	1208	Y	98.061	0.000000	0.250000	19	14	2.33205		Y	GOOD
1852-095	2770	Y	99.564	0.000000	0.560976	17	13	3.05402		Y	GOOD
1864-019	1809	N	99.246	0.012891	0.328571	23	16	3.33149		N	GOOD
1871-016	5243	N	97.656	0.004678	0.529412	18	11	3.06608		N	GOOD
1896-033	2941	N	99.984	0.002899	0.230769	21	11	2.84155		N	GOOD
1896-369	2534	Y	99.928	0.001065	0.291667	22	15	2.69862		Y	GOOD
1901-001	204	N	93.510	0.098039	0.134615	36	59	2.00000		N	GOOD
1940-007	1537	N	99.371	0.009464	0.160920	24	13	2.42429		N	GOOD
1993-274	3078	N	98.169	0.005376	0.687500	20	11	3.30968		N	GOOD
2007-044	3413	N	98.810	0.095541	0.452830	22	14	3.62314		N	GOOD
2010-055	250	N	99.457	0.000000	0.126437	24	33	3.04878		N	GOOD
2010-335	595	N	91.391	0.145740	0.079470	30	11	1.33408	1	N	BAD
2024-057	1368	N	99.588	0.000000	0.159420	25	14	2.28000		N	GOOD
2024-255	23	N	100.000	0.000000	0.018018	28	54	1.53333		Y	GOOD
2029-193	2323	N	92.332	0.000000	0.105263	24	15	3.93063		N	GOOD
2029-339	282	Y	97.507	0.000000	0.112676	24	31	2.82000		Y	GOOD
2029-438	1060	Y	93.979	0.000000	0.159722	28	18	3.09942		Y	GOOD
2029-489	938	Y	95.503	0.101533	0.093567	32	13	1.79693	1	Y	BAD
2032-035	59	N	68.519	0.250000	0.152174	19	55	2.10714	1	Y	BAD
2042-004	1274	Y	98.779	0.025362	0.139130	27	21	2.30797		Y	GOOD
2057-045	158	N	96.154	0.022222	0.161290	22	51	3.51111		Y	GOOD
2059-004	1757	N	99.179	0.023166	0.144231	29	10	2.26126		N	GOOD
2060-033	74	N	98.718	0.000000	0.075758	25	74	2.84615		Y	GOOD
2060-105	607	N	97.587	0.047619	0.267606	24	18	3.21164		N	GOOD
2070-034	604	Y	99.201	0.011905	0.155340	27	18	2.39683		Y	GOOD
2083-020	1901	Y	98.945	0.002911	0.323944	24	13	2.76710		Y	GOOD

Page Characteristics				Measured Features					Classifier Logic		
PageID	NCC	T	Acc.	White Speckle	Broken Zone	M B	M W	B / W Ratio	Rules	R	ClasAs
2092-058	7	N	100.000	0.000000	0.002674	56	93	1.75000	3	Y	BAD
2093-006	4094	Y	99.586	0.001372	0.500000	20	11	2.80796		Y	GOOD
2094-052	3297	N	99.233	0.000931	0.291667	21	11	3.06983		N	GOOD
2095-013	1801	N	99.477	0.010606	0.274510	22	9	2.72879		N	GOOD
2095-026	70	N	96.739	0.066667	0.022222	26	39	1.55556		Y	GOOD
2096-140	2729	N	99.399	0.017897	0.234375	24	13	3.05257		N	GOOD
2096-229	811	N	97.500	0.049853	0.133333	26	19	2.37830		N	GOOD
2096-349	1114	N	99.541	0.000000	0.172840	23	19	2.91623		N	GOOD
2104-010	2362	N	99.883	0.000000	0.144737	26	15	2.94514		N	GOOD
2117-040	2413	Y	99.070	0.007264	0.369231	21	14	2.92131		Y	GOOD
2117-178	2628	N	99.583	0.001004	0.222222	22	13	2.63855		N	GOOD
2117-276	911	N	99.552	0.012539	0.338235	21	22	2.85580		N	GOOD
2117-329	1720	Y	99.814	0.000000	0.215385	21	14	2.79221		Y	GOOD
2118-011	1569	N	99.887	0.000000	0.133333	23	12	2.78191		N	GOOD
2118-036	658	N	97.722	0.123684	0.235294	20	13	1.73158	1	N	BAD
2136-083	897	Y	76.267	0.007958	0.264368	24	18	2.37931		Y	GOOD
2137-003	1994	N	95.803	0.147609	0.442308	21	17	4.14553	1	N	BAD
2177-035	1630	N	99.382	0.013133	0.250000	22	13	3.05816		N	GOOD
2253-078	1769	N	99.525	0.000000	0.296296	23	13	2.92881		N	GOOD
2306-093	387	Y	87.935	0.022857	0.216667	29	57	2.21143		Y	GOOD
5008-016	1775	N	94.848	0.157837	0.270833	25	8	0.96942	1	N	BAD
5011-006	5145	N	99.645	0.000000	0.621622	16	10	2.95520		N	GOOD
5017-039	83	N	98.876	0.000000	0.016667	23	99	3.77273		Y	GOOD
5017-041	2022	Y	98.477	0.000000	0.186916	26	18	3.27184		Y	GOOD
5020-295	114	N	68.129	0.071429	0.031746	28	51	2.71429		Y	GOOD
5025-016	2632	N	99.844	0.001112	0.215385	26	15	2.92770		N	GOOD
5028-047	2685	N	99.431	0.002146	0.204819	27	14	2.88090		N	GOOD
5039-013	3730	N	98.753	0.049218	0.235294	24	11	1.71573		N	GOOD
5039-041	1528	N	98.919	0.006234	0.131313	23	14	1.90524		N	GOOD
5043-059	2769	N	99.457	0.000000	0.333333	21	14	2.72808		N	GOOD
5045-074	2312	Y	98.485	0.002561	0.285714	21	12	2.96031		Y	GOOD
5050-054	106	N	99.561	0.000000	0.028986	25	53	2.65000		Y	GOOD
5064-030	4138	N	96.975	0.011402	0.468085	18	10	2.77532		N	GOOD

Page Characteristics				Measured Features					Classifier Logic		
PageID	NCC	T	Acc.	White Speckle	Broken Zone	M B	M W	B / W Ratio	Rules	R	ClasAs
5065-041	1022	Y	97.762	0.002809	0.250000	21	18	2.87079		Y	GOOD
5067-007	1234	Y	99.515	0.000000	0.151786	26	14	2.03630		Y	GOOD
5069-022	1663	N	99.646	0.001942	0.308824	21	14	3.22913		N	GOOD
5074-052	77	N	97.753	0.000000	0.032967	28	76	3.34783		Y	GOOD
5081-037	2458	N	99.670	0.006543	0.273973	21	12	2.68048		N	GOOD
5086-072	1321	Y	95.996	0.007782	0.151515	19	13	2.57004		Y	GOOD
5092-057	1721	N	87.079	0.107527	0.810811	16	13	3.70108	12	N	BAD
5099-068	880	Y	96.928	0.055072	0.029777	65	22	2.55072	3	Y	BAD
5103-020	1392	Y	99.526	0.001880	0.230769	26	16	2.61654		Y	GOOD
5107-033	2	N	100.000	0.000000	0.000000	32	37	1.00000	4	Y	BAD
5109-001	1980	Y	97.643	0.005479	0.349398	27	21	2.71233		Y	GOOD
5112-097	68	N	53.526	0.513514	0.014493	25	24	0.91892	1	Y	BAD
5112-101	71	N	94.872	0.133333	0.052632	14	42	2.36667	1	Y	BAD
5126-002	5159	N	98.811	0.194579	0.392157	18	9	1.91571	1	N	BAD
5129-022	1957	N	97.922	0.005814	0.403846	20	18	2.84448		N	GOOD
5138-025	1642	N	99.437	0.011392	0.186667	24	13	2.07848		N	GOOD
5138-044	1595	N	99.700	0.011331	0.266667	24	16	2.25921		N	GOOD
5140-040	847	Y	98.574	0.000000	0.192308	24	23	3.05776		Y	GOOD
5146-014	67	N	89.726	0.026316	0.078125	22	62	1.76316		Y	GOOD
5171-018	4342	N	98.675	0.000641	0.758621	17	10	2.78155	2	N	BAD
5179-007	1994	N	96.475	0.054588	0.446154	21	12	2.31591		N	GOOD
5182-078	1374	Y	95.270	0.000000	0.215686	22	17	3.68365		Y	GOOD
5184-131	1368	N	98.990	0.014463	0.320755	22	15	2.82645		N	GOOD
5184-172	1249	N	98.864	0.022371	0.358491	22	13	2.79418		N	GOOD
5184-404	741	N	68.517	0.160517	0.231481	28	23	1.36716	1	N	BAD
5184-421	1356	N	98.456	0.011686	0.264368	24	16	2.26377		N	GOOD
5192-089	1544	Y	99.203	0.007018	0.166667	25	15	2.70877		Y	GOOD
5195-054	869	Y	98.815	0.028571	0.142857	25	18	2.25714		Y	GOOD
5195-077	541	Y	97.649	0.010526	0.113821	30	30	2.84737		Y	GOOD
5195-102	2150	N	99.896	0.001295	0.222222	22	13	2.78497		N	GOOD
5210-003	1178	N	99.401	0.012121	0.218182	22	15	2.37980		N	GOOD
5245-036	1662	N	99.868	0.000000	0.294118	21	12	2.67203		N	GOOD
5245-093	1819	N	99.580	0.000000	0.250000	21	12	2.51243		N	GOOD

Page Characteristics				Measured Features					Classifier Logic		
PageID	NCC	T	Acc.	White Speckle	Broken Zone	M B	M W	B / W Ratio	Rules	R	ClasAs
5252-009	1216	Y	93.960	0.004425	0.297297	26	18	2.69027		Y	GOOD
5253-042	22	N	96.154	0.000000	0.000000	58	220	2.75000	3	Y	BAD
5265-137	1649	Y	90.922	0.052301	0.163934	21	15	3.44979		Y	GOOD
5279-013	4	N	0.000	0.000000	0.000000	23	121	4.00000		Y	GOOD
5303-003	2387	Y	98.822	0.019231	0.419753	22	12	2.41599		Y	GOOD
5314-020	2537	N	98.845	0.010476	0.215385	20	11	2.41619		N	GOOD
5324-011	3559	N	99.656	0.000738	0.306667	24	13	2.62657		N	GOOD
5338-022	2670	N	99.359	0.005825	0.346154	19	10	2.59223		N	GOOD
5347-145	2293	Y	82.026	0.022863	0.094118	29	9	2.27932		Y	GOOD
5351-009	1981	N	99.637	0.012784	0.173333	24	12	2.81392		N	GOOD
5363-009	1243	N	97.761	0.002212	0.390625	22	11	2.75000		N	GOOD
5365-013	3185	N	99.252	0.002849	0.352941	21	12	3.02469		N	GOOD
5367-007	2875	Y	84.242	0.103896	0.833333	12	18	9.33442	12	Y	BAD
5378-013	1980	N	99.852	0.003386	0.137931	24	12	2.23476		N	GOOD
5378-099	77	N	98.980	0.031250	0.034783	27	74	2.40625		Y	GOOD
5380-020	1423	N	99.599	0.001908	0.296875	22	13	2.71565		N	GOOD
5380-053	1247	Y	98.786	0.000000	0.345455	22	21	3.41644		Y	GOOD
5384-005	212	N	98.246	0.013514	0.117647	24	33	2.86486		N	GOOD
5384-122	22	N	82.143	0.153846	0.003953	48	86	1.69231	1 3	Y	BAD
5385-003	844	Y	99.024	0.000000	0.090909	26	23	2.64577		Y	GOOD
5385-373	13	N	96.667	0.000000	0.015625	22	69	2.16667		Y	GOOD
5385-491	13	N	100.000	0.000000	0.014706	21	64	1.85714		Y	GOOD
5394-028	1564	N	98.919	0.080386	0.272727	27	12	2.51447		N	GOOD
5408-004	2556	Y	99.113	0.000000	0.115385	20	14	4.65574		Y	GOOD
5412-028	956	Y	97.212	0.002941	0.216495	24	23	2.81176		Y	GOOD
5415-006	1477	Y	97.823	0.000000	0.129630	21	26	3.92819		Y	GOOD
5417-003	272	N	98.476	0.000000	0.086420	25	61	3.16279		N	GOOD
5424-007	255	N	44.481	0.285714	0.923077	10	186	18.21429	12	N	BAD
5435-166	3166	Y	94.709	0.014587	0.804878	18	16	5.13128	2	Y	BAD
5449-105	744	Y	98.742	0.000000	0.065789	34	28	3.41284		Y	GOOD
5455-024	2261	N	98.265	0.001364	0.581395	19	12	3.08458		N	GOOD
5460-011	2950	N	96.142	0.002677	0.758621	17	11	3.94913	2	N	BAD
5471-022	2506	N	99.682	0.004082	0.229508	21	10	2.55714		N	GOOD

Page Characteristics				Measured Features					Classifier Logic		
PageID	NCC	T	Acc.	White Speckle	Broken Zone	M B	M W	B / W Ratio	Rules	R	ClasAs
5480-004	6540	N	98.450	0.120077	0.689655	17	6	2.09414	1	N	BAD
5482-030	2000	N	97.851	0.007576	0.541667	21	10	3.03030		N	GOOD
5482-038	1741	N	93.521	0.016736	0.604167	20	12	3.64226		N	GOOD
5487-044	3400	N	99.441	0.000000	0.441860	19	13	2.88625		N	GOOD
5487-095	2187	N	99.099	0.000000	0.382353	18	13	2.99589		N	GOOD
5487-219	2048	N	99.307	0.007742	0.312500	20	14	2.64258		N	GOOD
5487-514	2225	N	99.359	0.000000	0.325581	19	12	2.68072		N	GOOD
5487-641	2337	N	99.541	0.000000	0.302326	19	13	2.73333		N	GOOD
5489-006	1246	N	99.396	0.033395	0.326923	20	17	2.31169		N	GOOD
5499-017	495	N	95.988	0.000000	0.310345	17	16	2.96407		N	GOOD
5515-011	2656	Y	98.853	0.004301	0.171053	26	14	2.85591		Y	GOOD
5518-002	2289	N	98.324	0.003676	0.545455	22	14	2.80515		N	GOOD
5575-112	733	Y	97.528	0.000000	0.000000	28	24	2.13703		Y	GOOD
5578-013	105	N	98.305	0.000000	0.041667	21	74	3.18182		Y	GOOD
5588-036	205	N	92.469	0.014286	0.050000	24	52	2.92857		N	GOOD
5592-012	5356	Y	91.175	0.007036	0.794118	18	16	4.71064	2	Y	BAD
5593-060	1208	Y	97.348	0.011111	0.358491	18	12	2.68444		Y	GOOD
5611-007	3083	Y	92.889	0.001389	0.489362	18	14	4.28194		Y	GOOD
5611-010	2569	Y	97.866	0.009070	0.344262	21	15	2.91270		Y	GOOD
5611-013	3206	Y	97.866	0.012079	0.676471	21	14	2.76618		Y	GOOD
5624-002	188	N	95.602	0.111111	0.163934	20	46	2.32099	1	Y	BAD
5629-006	1757	N	88.354	0.021631	0.759259	21	15	2.92346	2	N	BAD
5648-007	1189	N	98.548	0.008602	0.323529	21	23	2.55699		N	GOOD
5650-002	1617	N	99.533	0.001387	0.281250	15	8	2.24272		N	GOOD
5650-044	2197	N	99.746	0.010959	0.208333	22	12	3.00959		N	GOOD
5655-026	162	N	98.619	0.000000	0.114754	20	42	3.17647		Y	GOOD
5657-011	22	N	92.000	0.000000	0.055556	20	83	3.14286		Y	GOOD
5665-008	2770	N	92.311	0.002660	0.736842	19	20	7.36702	2	N	BAD
5668-041	2200	N	99.698	0.000000	0.166667	22	12	2.73973		N	GOOD
5677-018	543	N	99.840	0.000000	0.148148	22	17	2.39207		N	GOOD
5677-020	1449	N	99.911	0.000000	0.185185	23	13	2.34846		N	GOOD
5678-084	754	N	98.705	0.004329	0.294118	18	20	3.26407		N	GOOD
5680-035	1575	Y	99.309	0.002789	0.163934	20	14	2.19665		Y	GOOD

Page Characteristics				Measured Features					Classifier Logic		
PageID	NCC	T	Acc.	White Speckle	Broken Zone	M B	M W	B / W Ratio	Rules	R	ClasAs
5680-493	2256	N	99.785	0.003555	0.327869	21	13	2.67299		N	GOOD
5712-042	2288	N	97.088	0.000000	0.564516	22	12	2.99476		N	GOOD
5713-137	1929	N	99.049	0.177198	0.391304	19	10	2.64973	1	N	BAD
5713-160	2670	N	98.977	0.042042	0.395349	19	10	2.67267		N	GOOD
5713-246	2639	N	98.988	0.048544	0.488372	19	10	2.84682		N	GOOD
5715-181	2018	N	94.395	0.080916	0.512195	18	12	3.08092		N	GOOD
5719-004	4431	N	98.937	0.155290	0.509091	22	10	2.52048	1	N	BAD
5721-021	2580	N	99.378	0.026733	0.250000	21	12	2.55446		N	GOOD
5727-096	472	Y	87.201	0.158940	0.212121	27	79	1.56291	1	Y	BAD
5727-105	359	Y	88.310	0.172691	0.145161	27	97	1.44177	1 4	Y	BAD
5727-109	312	Y	92.489	0.198238	0.126126	28	75	1.37445	1 4	Y	BAD
5730-024	2840	N	95.733	0.091463	0.604651	19	12	2.88618		N	GOOD
5738-002	4963	N	99.249	0.006701	0.442623	20	11	2.55825		N	GOOD
5752-021	2112	N	86.385	0.236364	0.689655	17	10	3.84000	1	N	BAD
5770-034	3	N	100.000	0.000000	0.000000	32	115	3.00000		Y	GOOD
5784-001	2978	N	97.968	0.127943	0.419355	23	14	3.04811	1	N	BAD
5784-006	2652	N	99.233	0.154746	0.370968	22	14	3.44863	1	N	BAD
5804-093	552	N	100.000	0.000000	0.105263	26	20	2.66667		N	GOOD
5805-003	30	N	94.118	0.000000	0.000000	63	232	3.33333	3	Y	BAD
5809-003	2898	Y	97.172	0.016423	0.157895	12	11	5.28832		Y	GOOD
5809-049	88	N	80.114	0.111111	0.156863	18	63	2.44444	1	Y	BAD
5816-029	1771	N	98.817	0.225122	0.403846	21	11	2.88907	1	N	BAD
5816-115	284	N	88.576	0.478723	0.241379	17	25	3.02128	1	N	BAD
5820-095	86	N	100.000	0.000000	0.007812	32	85	2.45714		Y	GOOD
5830-018	1440	N	98.273	0.008811	0.228916	27	14	3.17181		N	GOOD
5830-144	1325	Y	98.089	0.000000	0.160000	24	21	2.12340		Y	GOOD
5830-158	1222	Y	94.607	0.004608	0.222222	23	20	1.87711		Y	GOOD
5830-164	1250	Y	96.813	0.000000	0.166667	23	21	2.39464		Y	GOOD
5830-240	1052	Y	88.435	0.006928	0.295082	21	23	2.42956		Y	GOOD
5836-082	6966	N	96.952	0.045694	0.675676	15	7	1.92911		N	GOOD
5836-175	6982	N	97.644	0.080982	0.648649	15	6	1.82393		N	GOOD
5837-041	2389	N	99.440	0.004535	0.115942	25	14	2.70862		N	GOOD
5842-060	2199	N	92.740	0.258446	0.416667	20	13	3.71453	1	N	BAD

Page Characteristics				Measured Features					Classifier Logic		
PageID	NCC	T	Acc.	White Speckle	Broken Zone	M B	M W	B / W Ratio	Rules	R	ClasAs
5856-026	1388	Y	43.860	0.240741	0.888889	12	124	25.70370	12	Y	BAD
5860-014	1974	N	86.114	0.015306	0.862745	18	17	5.03571	2	N	BAD
5871-003	155	N	96.629	0.040541	0.045226	39	50	2.09459		Y	GOOD
5878-001	5421	N	82.525	0.003774	0.888889	15	30	20.45660	2	N	BAD
5881-274	1722	N	97.890	0.001887	0.576923	20	26	3.24906		N	GOOD
5916-021	67	N	95.833	0.000000	0.098765	22	72	3.04545		Y	GOOD
5916-266	2277	N	98.456	0.000000	0.172840	25	13	2.76000		N	GOOD
5916-318	2381	N	98.980	0.001193	0.367816	24	13	2.84129		N	GOOD
5920-046	84	N	98.095	0.000000	0.018519	28	42	1.75000		Y	GOOD
5921-071	1361	Y	93.912	0.010889	0.296296	21	15	2.47005		Y	GOOD
5922-086	1613	N	99.657	0.001802	0.236111	22	13	2.90631		N	GOOD
5925-025	2854	Y	93.193	0.073120	0.472727	22	15	2.93924		Y	GOOD
5926-001	183	N	80.819	0.066667	0.184713	33	38	2.03333		Y	GOOD
5935-149	851	Y	95.829	0.005970	0.151515	26	27	2.54030		Y	GOOD
5945-013	2163	N	99.789	0.007926	0.279412	21	16	2.85733		N	GOOD
5945-058	2741	N	99.835	0.000000	0.279412	21	14	2.77992		N	GOOD
5946-031	1150	N	97.589	0.088578	0.297872	26	13	2.68065		N	GOOD
5946-205	599	N	95.308	0.078603	0.340426	26	21	2.61572		N	GOOD
5951-009	6518	N	95.724	0.218832	0.630435	19	8	2.13845	1	N	BAD
5951-019	6585	N	98.001	0.125082	0.615385	20	8	2.15056	1	N	BAD
5952-007	2269	N	99.524	0.013187	0.156863	29	11	2.49341		N	GOOD
5954-061	652	N	98.430	0.023569	0.147727	25	20	2.19529		N	GOOD
5955-103	2094	Y	98.645	0.016371	0.265625	24	16	2.85675		Y	GOOD
5959-094	277	N	98.406	0.011429	0.071429	25	19	1.58286		N	GOOD
5964-054	1437	N	98.564	0.051075	0.424658	21	10	1.93145		N	GOOD
5964-211	2271	N	99.740	0.001957	0.307692	21	10	2.22211		N	GOOD
5965-092	1801	N	97.259	0.005263	0.430556	22	14	2.36974		N	GOOD
5965-133	1697	N	98.651	0.008677	0.218750	22	12	1.84056		N	GOOD
5967-093	2161	N	99.851	0.009816	0.250000	22	13	2.65153		N	GOOD
5975-102	1248	N	98.925	0.024609	0.364583	25	18	2.79195		N	GOOD
5992-099	2274	N	99.575	0.011455	0.328767	21	15	2.60481		N	GOOD
6008-006	348	N	96.649	0.000000	0.054983	45	50	2.67692	3	N	BAD
6008-014	262	Y	91.129	0.040000	0.027512	70	49	2.62000	3	Y	BAD

Page Characteristics				Measured Features					Classifier Logic		
PageID	NCC	T	Acc.	White Speckle	Broken Zone	M B	M W	B / W Ratio	Rules	R	ClasAs
6008-038	68	Y	98.684	0.000000	0.003704	68	68	2.61538	3	Y	BAD
6090-018	411	N	98.404	0.014925	0.050000	57	43	3.06716	3	N	BAD
6168-002	161	N	97.632	0.000000	0.022222	28	38	3.57778		Y	GOOD
6177-054	1116	N	99.303	0.002841	0.262626	27	28	3.17045		N	GOOD
6272-068	1410	N	83.617	0.000000	0.804348	19	24	5.57312	2	N	BAD
6286-013	1133	Y	99.088	0.005305	0.175824	30	22	3.00531		Y	GOOD
6294-061	61	N	97.143	0.000000	0.030303	25	86	2.65217		Y	GOOD
6294-118	1669	Y	99.654	0.017572	0.233333	26	17	2.66613		Y	GOOD
6294-145	1280	N	99.675	0.007246	0.136986	25	18	3.09179		N	GOOD
6310-007	202	N	96.330	0.017241	0.038339	49	61	3.48276	3	N	BAD
6346-056	157	N	83.784	0.135135	0.156250	22	33	2.12162	1	Y	BAD
6347-112	1732	Y	51.092	0.036585	0.000000	16	49	10.56098		Y	GOOD
6477-004	183	N	100.000	0.000000	0.013158	26	44	3.00000		Y	GOOD
6478-017	256	N	99.656	0.000000	0.057971	25	48	2.90909		N	GOOD
6478-021	2105	N	99.373	0.256270	0.130435	25	12	2.29553	1	N	BAD
6504-018	279	Y	98.485	0.028037	0.044068	47	37	2.60748	3	Y	BAD
6504-025	622	N	98.217	0.019802	0.054795	41	25	3.07921	3	N	BAD
6544-008	800	N	95.770	0.085470	0.127389	33	30	3.41880		N	GOOD
6546-011	429	Y	88.223	0.000000	0.333333	23	51	4.37755		Y	GOOD
6546-016	238	Y	90.669	0.000000	0.172414	31	65	3.40000		Y	GOOD
6560-002	809	N	98.160	0.000000	0.103448	24	14	1.86836		N	GOOD
6575-007	2204	Y	99.399	0.011516	0.305085	19	14	4.23033		Y	GOOD
6577-020	1283	Y	99.439	0.005479	0.168224	26	17	3.51507		Y	GOOD
6577-040	2950	N	99.679	0.003976	0.347222	22	12	2.93241		N	GOOD
6580-011	54	N	100.000	0.000000	0.011494	24	65	2.00000		Y	GOOD
6580-018	1731	Y	99.751	0.004076	0.134615	24	15	2.35190		Y	GOOD
6580-040	2551	N	99.760	0.003899	0.244444	23	12	2.48635		N	GOOD
6580-091	1129	Y	99.135	0.008772	0.176471	24	20	2.47588		Y	GOOD
6583-055	676	Y	97.172	0.029126	0.045752	38	24	2.18770		Y	GOOD
6585-025	2067	N	99.477	0.002165	0.211111	23	10	2.23701		N	GOOD
6585-232	2225	N	99.809	0.002220	0.222222	23	11	2.46948		N	GOOD
6585-388	2604	N	99.790	0.000906	0.177778	23	11	2.35870		N	GOOD
6585-593	1435	Y	97.126	0.004587	0.314286	17	14	3.29128		Y	GOOD

Page Characteristics				Measured Features					Classifier Logic		
PageID	NCC	T	Acc.	White Speckle	Broken Zone	M B	M W	B / W Ratio	Rules	R	ClasAs
6654-049	270	N	97.735	0.034091	0.016588	55	50	3.06818	3	N	BAD
6716-005	3007	N	98.372	0.000981	0.500000	23	17	2.95093		N	GOOD
6720-001	273	N	72.408	0.216495	0.077626	40	92	2.81443	1 3	N	BAD
6723-086	2393	N	96.781	0.004274	0.471264	24	18	2.55662		N	GOOD
6790-011	1674	Y	97.013	0.066451	0.285714	15	11	2.71313		Y	GOOD
6809-135	1552	N	99.590	0.001931	0.142857	25	12	1.49807		N	GOOD
6809-167	1233	N	99.302	0.003645	0.104167	25	12	1.49818		N	GOOD
6809-236	456	N	98.762	0.000000	0.052083	25	20	1.42056		N	GOOD
6848-001	392	N	90.590	0.048276	0.262774	34	44	2.70345		N	GOOD
6904-008	2085	N	99.662	0.006457	0.130952	25	11	1.68281		N	GOOD
6905-002	2251	N	98.410	0.072664	0.254237	25	13	2.59631		N	GOOD
6907-009	1554	N	97.123	0.117750	0.265625	24	17	2.73111	1	N	BAD
6908-007	625	N	96.936	0.121951	0.130435	25	19	3.04878	1	N	BAD
6917-003	1703	N	98.019	0.008475	0.218750	24	17	2.88644		N	GOOD
6928-005	549	N	99.511	0.000000	0.092593	21	26	2.55349		N	GOOD
6955-005	717	N	92.661	0.012397	0.440678	19	18	2.96281		N	GOOD

Observations:

PageID. Identification code for the page (internal to ISRI).

NCC. Number of connected components.

T. Page contains tables (Y/N).

Acc. Median OCR accuracy as used for the experiment.

White Speckle. Value of the white speckle factor for the page.

Broken Zone. Value of the broken zone factor for the page.

MB. Maximum average size (width or height) for black connected components.

MW. Maximum average size (width or height) for white connected components.

B/W Ratio. Number of black CC to white CC ratio.

Rules. Rules triggered in case of a "BAD" classification.

R. Page considered a "reject" (Y/N).

ClasAs. Page classified as (GOOD/BAD).

Appendix B

Classifier Results for Magazine Dataset

Page Characteristics				Measured Features					Classifier Logic		
PageID	NCC	T	Acc.	White Speckle	Broken Zone	M B	M W	B / W Ratio	Rules	R	ClasAs
8000-012	3319	N	93.000	0.424149	0.800000	15	21	5.13777	12	N	BAD
8000-027	5667	N	90.030	0.075000	0.714286	12	103	70.83750	2	N	BAD
8001-044	4931	N	99.680	0.001292	0.368421	19	14	3.18540		N	GOOD
8001-055	5603	N	99.420	0.000000	0.229167	24	15	2.80852		N	GOOD
8002-037	9416	N	90.700	0.089992	0.857143	7	16	7.30489	2	N	BAD
8002-060	3730	N	99.830	0.000000	0.346154	21	14	2.61571		N	GOOD
8003-033	2197	N	96.700	0.406179	0.368421	19	10	2.51373	1	N	BAD
8003-075	6498	N	78.660	0.000000	0.833333	15	65	38.22353	2	N	BAD
8004-029	10122	N	97.150	0.019211	0.769231	9	13	5.11729	2	N	BAD
8004-035	1728	N	97.050	0.123518	0.359375	22	12	1.70751	1	N	BAD
8005-032	1015	N	96.620	0.005540	0.465116	16	25	2.81163		N	GOOD
8005-125	1975	N	92.950	0.012987	0.323529	21	15	1.97303		N	GOOD
8006-030	3568	N	98.320	0.006150	0.348837	19	17	4.38868		N	GOOD
8006-078	1939	N	99.200	0.181435	0.255814	19	20	2.72714	1	N	BAD
8007-026	1291	N	98.460	0.000000	0.278689	21	19	3.32732		N	GOOD
8007-047	2518	N	99.000	0.001217	0.250000	23	17	3.06326		N	GOOD
8008-024	3241	N	96.280	0.019571	0.365385	21	18	3.02050		N	GOOD
8008-052	4628	N	99.270	0.007260	0.278689	20	14	2.79976		N	GOOD
8009-018	6965	N	96.390	0.208479	0.441176	18	11	2.70906	1	N	BAD
8009-032	3393	N	77.410	0.404924	0.714286	12	9	2.88031	12	N	BAD
8010-050	1982	N	99.140	0.000000	0.379310	17	24	5.44505		N	GOOD
8010-097	2984	N	97.290	0.000000	0.793103	17	46	11.93600	2	N	BAD
8011-004	3148	N	98.070	0.061705	0.319444	22	13	2.00254		N	GOOD
8011-012	3371	N	96.360	0.021308	0.355556	23	10	0.98395		N	GOOD
v8012-112	4450	N	99.770	0.001276	0.302326	19	12	2.83982		N	GOOD
8012-113	2178	N	99.740	0.000000	0.232558	19	13	3.04190		N	GOOD
8013-494	4390	N	91.660	0.098233	0.400000	20	11	2.28170		N	GOOD

Page Characteristics				Measured Features					Classifier Logic		
PageID	NCC	T	Acc.	White Speckle	Broken Zone	M B	M W	B / W Ratio	Rules	R	ClasAs
8013-497	4346	N	91.770	0.009841	0.519231	21	16	2.03655		N	GOOD
8014-054	4637	N	98.540	0.023118	0.261905	21	14	2.74867		N	GOOD
8014-070	4333	N	99.770	0.000634	0.270833	20	14	2.74762		N	GOOD
8015-056	1689	N	97.360	0.047002	0.370968	22	24	2.73744		N	GOOD
8015-128	2424	N	96.310	0.013126	0.383333	23	21	2.89260		N	GOOD
8016-092	6035	N	99.440	0.000487	0.461538	20	14	2.93817		N	GOOD
8016-228	7918	N	99.780	0.001790	0.395349	19	10	2.83494		N	GOOD
8017-007	606	N	100.000	0.000000	0.078947	19	12	3.07614		N	GOOD
8017-022	1489	N	99.380	0.002179	0.263158	19	23	3.24401		N	GOOD
8018-038	3493	N	99.170	0.006831	0.462963	20	11	2.38593		N	GOOD
8018-089	2403	N	98.770	0.000000	0.269231	20	13	2.63198		N	GOOD
8019-061	3029	N	94.180	0.050445	0.367647	20	12	1.79763		N	GOOD
8019-097	1285	N	98.420	0.001479	0.150685	21	13	1.90089		N	GOOD
8020-022	3724	N	98.730	0.018057	0.567568	16	17	3.20206		N	GOOD
8020-134	51326	N	87.580	0.011215	1.000000	3	13	23.98411	2	N	BAD
8021-028	3641	N	99.530	0.000000	0.250000	21	17	3.50096		N	GOOD
8021-056	2200	N	98.920	0.007680	0.327869	21	19	3.37942		N	GOOD
8022-028	5207	N	99.360	0.000983	0.213115	21	14	2.55998		N	GOOD
8022-074	2160	N	98.110	0.000000	0.217391	19	14	3.23353		N	GOOD
8023-017	1641	N	99.070	0.009245	0.375000	22	16	2.52851		N	GOOD
8023-084	294	N	97.380	0.003610	0.063063	28	16	1.06137		N	GOOD
8024-015	4222	N	89.490	0.138596	0.800000	6	14	7.40702	12	N	BAD
8024-029	2443	N	98.350	0.005330	0.254545	22	14	2.60448		N	GOOD
8025-062	2987	N	96.680	0.102204	0.457143	17	12	2.99299	1	N	BAD
8025-067	5172	N	83.380	0.104297	0.678571	14	14	2.31513	1	N	BAD
8026-014	698	N	98.820	0.015326	0.162791	19	13	2.67433		N	GOOD
8026-018	3003	N	98.590	0.013672	0.659574	18	21	5.86523		N	GOOD
8027-088	5022	N	99.310	0.002330	0.346154	21	13	2.92487		N	GOOD
8027-147	2448	N	98.410	0.016474	0.470588	18	14	4.03295		N	GOOD
8028-052	2381	N	98.940	0.008457	0.296296	20	14	2.51691		N	GOOD
8028-053	828	N	98.720	0.009346	0.260870	25	21	2.57944		N	GOOD
8029-052	2315	N	99.670	0.001105	0.269231	20	11	2.55801		N	GOOD
8029-076	2783	N	99.460	0.000000	0.131148	20	11	2.42211		N	GOOD

Page Characteristics				Measured Features					Classifier Logic		
PageID	NCC	T	Acc.	White Speckle	Broken Zone	M B	M W	B / W Ratio	Rules	R	ClasAs
8030-040	3795	N	99.120	0.002753	0.271186	19	13	2.61184		N	GOOD
8030-070	629	N	89.970	0.282857	0.327103	26	23	1.79714	1	N	BAD
8031-162	1658	N	92.640	0.002976	0.300000	23	21	2.46726		N	GOOD
8031-232	3253	N	93.850	0.000000	0.543478	19	19	3.00369		N	GOOD
8032-017	2078	N	98.830	0.022049	0.265625	22	19	2.69520		N	GOOD
8032-035	325	N	98.890	0.007752	0.078125	22	21	2.51938		N	GOOD
8033-046	2741	N	98.980	0.001134	0.250000	20	15	3.10771		N	GOOD
8033-106	1630	N	98.940	0.000000	0.109375	22	14	2.57911		N	GOOD
8034-023	2107	N	98.390	0.001978	0.046358	30	18	2.08408		N	GOOD
8034-098	1659	N	99.270	0.021407	0.229508	20	15	2.53670		N	GOOD
8035-180	2314	N	98.240	0.010601	0.387097	23	18	2.72556		N	GOOD
8035-208	7402	N	93.390	0.008372	0.882353	18	18	6.88558	2	N	BAD
8036-076	2797	N	99.510	0.009285	0.192308	21	13	2.59703		N	GOOD
8036-089	5918	N	99.800	0.002381	0.326923	20	13	2.81810		N	GOOD
8037-021	5238	N	95.600	0.029427	0.781250	15	13	2.65753	2	N	BAD
8037-032	4210	N	98.920	0.009180	0.480769	20	13	2.76066		N	GOOD
8038-014	4966	N	99.380	0.013393	0.647059	18	14	3.16709		N	GOOD
8038-077	5407	N	99.550	0.005011	0.395349	19	13	3.01058		N	GOOD
8039-011	2095	N	99.720	0.014684	0.169355	28	21	3.07636		N	GOOD
8039-030	1766	N	99.540	0.004808	0.287671	21	25	2.83013		N	GOOD
8040-124	2607	N	99.490	0.004197	0.187500	20	12	2.73557		N	GOOD
8040-128	5914	N	97.050	0.015760	0.352941	18	11	3.00661		N	GOOD
8041-044	2944	N	97.970	0.003272	0.360656	20	17	3.21047		N	GOOD
8041-087	3481	N	92.520	0.011938	0.491803	21	19	3.19651		N	GOOD
8042-067	5268	N	99.180	0.001536	0.327869	20	12	2.69739		N	GOOD
8042-104	1355	N	99.040	0.000000	0.092593	20	11	2.46364		N	GOOD
8043-020	1629	N	99.200	0.000000	0.109091	22	16	2.73782		N	GOOD
8043-022	1966	N	98.960	0.005312	0.187500	22	16	2.61089		N	GOOD
8044-060	4807	N	98.270	0.000000	0.414634	18	19	7.28333		N	GOOD
8044-080	2565	N	96.480	0.006250	0.517241	17	27	16.03125		N	GOOD
8045-043	2098	N	96.870	0.521368	0.453125	24	15	1.49430	1	N	BAD
8045-096	3415	N	99.560	0.000000	0.461538	15	12	3.53886		N	GOOD
8046-038	1175	N	96.350	0.000000	0.197452	33	29	2.51068		N	GOOD

Page Characteristics				Measured Features					Classifier Logic		
PageID	NCC	T	Acc.	White Speckle	Broken Zone	M B	M W	B / W Ratio	Rules	R	ClasAs
8046-078	415	N	97.090	0.023810	0.185484	28	37	2.47024		N	GOOD
8047-018	256	N	75.630	0.030303	0.031841	80	79	2.58586	3	N	BAD
8047-027	2475	N	98.150	0.000000	0.229508	21	16	2.49748		N	GOOD
8048-136	4901	N	99.020	0.008495	0.309524	20	13	2.97391		N	GOOD
8048-171	756	N	99.700	0.000000	0.190476	21	13	2.68085		N	GOOD
8049-020	794	N	77.300	0.437158	0.204545	25	11	1.08470	1	N	BAD
8049-135	440	N	93.590	0.173228	0.121212	27	12	1.73228	1	N	BAD
8050-048	3772	N	99.180	0.003865	0.259259	25	15	1.82222		N	GOOD
8050-078	3813	N	98.910	0.001162	0.219178	25	19	2.21557		N	GOOD
8051-026	1657	N	97.710	0.003442	0.581395	19	15	2.85198		N	GOOD
8051-156	1780	N	97.830	0.005772	0.180328	20	15	2.56854		N	GOOD
8052-019	2800	N	99.780	0.000000	0.115385	20	10	3.02376		N	GOOD
8052-109	3259	N	99.240	0.007722	0.200000	20	12	2.51660		N	GOOD
8053-034	3736	N	98.820	0.003342	0.436364	22	14	2.49733		N	GOOD
8053-070	5746	N	99.270	0.738872	0.360656	20	4	0.65751	1	N	BAD
8054-035	2547	N	93.590	0.038916	0.500000	19	13	1.76998		N	GOOD
8054-045	2132	N	96.830	0.082090	0.271605	22	15	1.76783		N	GOOD
8055-029	1779	N	99.600	0.000000	0.153846	20	14	2.76242		N	GOOD
8055-066	1524	N	98.900	0.000000	0.163934	21	15	2.76087		N	GOOD
8056-024	1931	N	99.300	0.004219	0.294118	18	16	4.07384		N	GOOD
8056-030	551	N	96.650	0.025157	0.140625	24	59	3.46541		N	GOOD
8057-038	1826	N	99.150	0.027417	0.222222	23	16	2.63492		N	GOOD
8057-106	2171	N	98.580	0.001221	0.135593	19	16	2.65079		N	GOOD
8058-045	523	N	97.110	0.283951	0.142857	16	16	3.22840	1	N	BAD
8058-056	4529	N	98.990	0.098971	0.397059	20	11	2.21901		N	GOOD
8059-043	4134	N	99.290	0.004397	0.307692	19	12	2.59673		N	GOOD
8059-056	2914	N	98.760	0.005396	0.500000	19	14	2.62050		N	GOOD
8060-086	4426	N	99.290	0.000000	0.213115	20	15	2.76798		N	GOOD
8060-088	4028	N	99.410	0.001418	0.326923	20	16	2.85674		N	GOOD
8061-278	4028	N	98.240	0.030683	0.488372	19	14	2.80893		N	GOOD
8061-404	4919	N	98.270	0.039431	0.517241	17	13	3.17970		N	GOOD
8062-011	1869	N	97.970	0.019830	0.351852	21	13	2.64731		N	GOOD
8062-020	3032	N	97.260	0.016935	0.484375	22	14	2.56732		N	GOOD

Page Characteristics				Measured Features					Classifier Logic		
PageID	NCC	T	Acc.	White Speckle	Broken Zone	M B	M W	B / W Ratio	Rules	R	ClasAs
8063-092	2366	N	97.860	0.001033	0.315068	21	17	2.44421		N	GOOD
8063-147	2599	N	96.460	0.002584	0.442623	21	19	3.35788		N	GOOD
8064-125	2780	N	99.240	0.012685	0.403846	20	17	2.93869		N	GOOD
8064-177	2223	N	99.520	0.000000	0.187500	21	15	2.86100		N	GOOD
8065-007	438	N	91.500	0.000000	0.586207	17	26	7.55172		N	GOOD
8065-095	1616	N	95.530	0.335714	0.442308	21	34	5.77143	1	N	BAD
8066-061	6763	N	99.670	0.000396	0.304348	19	12	2.68054		N	GOOD
8066-140	5123	N	99.820	0.000000	0.239130	19	12	2.70343		N	GOOD
8067-036	43198	N	81.710	0.037415	1.000000	3	12	18.36650	2	N	BAD
8067-039	1191	N	98.690	0.004762	0.194444	22	20	2.83571		N	GOOD
8068-062	22457	N	80.470	0.170124	1.000000	3	10	10.35362	12	N	BAD
8068-089	2308	N	98.340	0.068337	0.450980	18	12	2.62870		N	GOOD
8069-050	2349	N	99.320	0.000000	0.365385	21	14	2.73458		N	GOOD
8069-110	2425	N	99.320	0.001072	0.196721	21	13	2.59914		N	GOOD
8070-043	838	N	99.860	0.008357	0.088235	29	19	2.33426		N	GOOD
8070-047	426	N	97.610	0.016393	0.123288	25	17	2.32787		N	GOOD
8071-093	3352	N	99.410	0.068297	0.532258	22	12	2.82631		N	GOOD
8071-109	2990	N	97.180	0.054152	0.574074	21	13	2.69856		N	GOOD
8072-082	4837	N	98.780	0.002162	0.370370	20	12	2.61459		N	GOOD
8072-166	2882	N	98.060	0.006321	0.529412	18	14	3.64349		N	GOOD
8073-010	782	N	99.550	0.000000	0.117647	18	12	3.17886		N	GOOD
8073-034	2248	N	98.400	0.003490	0.279070	19	16	3.92321		N	GOOD
8074-020	2544	N	97.380	0.346700	0.353846	21	10	1.56940	1	N	BAD
8074-063	1073	N	96.160	0.009036	0.288462	19	18	3.23193		N	GOOD
8075-060	7084	N	91.860	0.473896	0.833333	12	8	2.18845	12	N	BAD
8075-166	3563	N	94.930	0.412612	0.620690	17	13	1.76912	1	N	BAD
8076-013	26	N	100.000	0.444444	0.026667	24	74	2.88889	1	Y	BAD
8076-018	211	N	94.130	0.023529	0.047619	34	46	2.48235		N	GOOD
8077-112	30799	N	91.210	0.066341	1.000000	3	18	21.50768	2	N	BAD
8077-151	2769	N	98.760	0.031111	0.323529	18	18	3.07667		N	GOOD
8078-014	4816	N	96.400	0.000000	0.689655	17	21	11.25234		N	GOOD
8078-048	3193	N	94.990	0.000000	0.696970	16	37	28.25664		N	GOOD
8079-018	2734	N	99.370	0.005274	0.245283	22	13	2.88397		N	GOOD

Page Characteristics				Measured Features					Classifier Logic		
PageID	NCC	T	Acc.	White Speckle	Broken Zone	M B	M W	B / W Ratio	Rules	R	ClasAs
8079-099	926	N	94.180	0.083516	0.155963	30	24	2.03516		N	GOOD
8080-032	3533	N	99.610	0.011445	0.161290	22	15	2.52718		N	GOOD
8080-068	1448	N	99.670	0.003578	0.240741	21	17	2.59034		N	GOOD
8081-045	3601	N	99.780	0.008053	0.395349	19	13	2.63616		N	GOOD
8081-060	3566	N	99.350	0.800268	0.302326	19	3	0.47770	1	N	BAD
8082-024	2071	N	84.580	0.163454	0.322581	23	13	1.59676	1	N	BAD
8082-039	737	N	94.410	0.334081	0.457143	15	11	1.65247	1	N	BAD
8083-056	256	N	95.820	0.000000	0.173913	19	70	5.33333		N	GOOD
8083-096	2803	N	98.850	0.000000	0.368421	19	14	3.12486		N	GOOD
8084-015	3869	N	97.370	0.011484	0.370370	22	11	1.70892		N	GOOD
8084-182	8101	N	97.510	0.002784	0.384615	15	11	3.22235		N	GOOD
8085-120	1571	N	93.380	0.355049	0.414286	23	17	2.55863	1	N	BAD
8085-128	4192	N	76.980	0.557967	0.888889	12	11	2.59888	12	N	BAD
8086-033	3446	N	99.480	0.034822	0.250000	20	13	2.60863		N	GOOD
8086-040	2792	N	99.070	0.004888	0.278689	21	15	2.72923		N	GOOD
8087-054	3667	N	99.250	0.003082	0.254237	19	14	2.82512		N	GOOD
8087-136	1803	N	99.450	0.001610	0.239130	19	13	2.90338		N	GOOD
8088-052	5944	N	97.530	0.005510	0.827586	17	32	16.37466	2	N	BAD
8088-061	3191	N	99.150	0.000000	0.278689	20	14	2.68829		N	GOOD
8089-006	2110	N	95.260	0.075962	0.361111	22	13	2.02885		N	GOOD
8089-018	1564	N	95.290	0.066766	0.400000	21	21	2.32047		N	GOOD
8090-043	1239	N	97.510	0.000000	0.368421	19	14	3.57061		N	GOOD
8090-047	4894	N	96.660	0.014257	0.730769	15	17	4.98371	2	N	BAD
8091-044	2895	N	99.260	0.000000	0.274194	22	14	2.63661		N	GOOD
8091-160	6587	N	99.740	0.000000	0.279070	19	13	2.72754		N	GOOD
8092-065	2604	N	99.220	0.002022	0.145455	25	16	2.63296		N	GOOD
8092-081	1121	N	96.510	0.009615	0.226415	22	19	3.59295		N	GOOD
8093-154	4900	N	99.520	0.004079	0.232558	19	12	2.85548		N	GOOD
8093-311	7517	N	99.580	0.000708	0.292683	18	10	2.66277		N	GOOD
8094-033	1250	N	68.770	0.534694	0.653846	16	13	5.10204	1	N	BAD
8094-052	1673	N	93.480	0.038817	0.541667	23	24	3.09242		N	GOOD
8095-046	4278	N	98.800	0.058960	0.250000	23	15	2.47283		N	GOOD
8095-083	2943	N	99.740	0.000000	0.196078	22	13	2.45046		N	GOOD

Page Characteristics				Measured Features					Classifier Logic		
PageID	NCC	T	Acc.	White Speckle	Broken Zone	M B	M W	B / W Ratio	Rules	R	ClasAs
8096-003	3106	N	96.770	0.007508	0.323077	21	16	2.33183		N	GOOD
8096-021	3589	N	98.850	0.002219	0.444444	20	16	2.65459		N	GOOD
8097-027	673	N	98.630	0.000000	0.122222	23	15	2.25084		N	GOOD
8097-064	1003	N	99.020	0.002227	0.177778	23	17	2.23385		N	GOOD
8098-011	1312	N	99.060	0.000000	0.282609	19	12	2.69959		N	GOOD
8098-073	46529	N	80.350	0.050940	1.000000	3	11	18.23237	2	N	BAD
8099-045	2642	N	96.310	0.093842	0.441860	19	15	2.58260		N	GOOD
8099-052	4498	N	99.330	0.009202	0.365385	21	16	2.75951		N	GOOD

Observations:

PageID. Identification code for the page (internal to ISRI).

NCC. Number of connected components.

T. Page contains tables (Y/N).

Acc. Median OCR accuracy as used for the experiment.

White Speckle. Value of the white speckle factor for the page.

Broken Zone. Value of the broken zone factor for the page.

MB. Maximum average size (width or height) for black connected components.

MW. Maximum average size (width or height) for white connected components.

B/W Ratio. Number of black CC to white CC ratio.

Rules. Rules triggered in case of a "BAD" classification.

R. Page considered a "reject" (Y/N).

ClasAs. Page classified as (GOOD/BAD).

Bibliography

[1] Ballard, Dana and Brown, Christopher. Computer Vision. *Prentice Hall Publishers.*

[2] Bohner, M. et al. "An Automatic Measurement Device for the Evaluation of the Print Quality of Printed Characters." *Pattern Recognition, vol. 9, pp. 11-19.*

[3] Bokser, Mindy. "Omnidocument Technologies." *Proceedings of the IEEE, Vol. 80, No. 7, July 1992.*

[4] Dickey, Lois A. "Operational Factors in the Creation of Large Full-Text Databases." *DOE Infotech Conference, Oak Ridge, TN, May 1991.*

[5] Jenkins, Frank and Kanai, Junichi. "The Use of Synthesized Images to Evaluate the Performance of Optical Character Recognition Devices and Algorithms." *SPIE, Vol. 2181 Document Recognition (1994) (pp 194-203)*

[6] Masayuki Okamoto and Akira Myazawa. "An Experimental Implementation of a Document Recognition System for Papers Containing Mathematical Expressions", in *Structured Document Image Analysis by H.S. Baird, H. Bunke and K. Yamamoto (Eds). Springer Verlag, 1992.*

[7] Nagy, G. et al. "A Prototype Document Image Analysis System for Technical Journals." *IEEE Computer Magazine, July 1992, pp. 10-22.*

[8] Nartker, T.A. et. al. "A Preliminary Report on UNLV/GT1: A Database for Ground-Truth Testing in Document Analysis in Document Analysis and Character Recognition." *Proceedings of Symposium on Document Analysis and Information Retrieval, ISRI, 1992. (pp 300-315).*

[9] Nartker, T.A. "On the Need for Information Metrics." *Keynote presentation, 1994 Symposium on Electronic Imaging Science & Technology, San Jose, California, February 1994.*

[10] Nartker, T.A. et al. "A Preliminary Report on OCR Problems in LSS Document Conversion." *Nuclear Waste Management Conference, Las Vegas, NV, April 1992.*

[11] Rice, S., Kanai, J. and Nartker T. "A Report on the Accuracy of OCR Devices." *Technical Report ISRI TR-92-02, University of Nevada, Las Vegas, March 1992.*

[12] Rice, S., Kanai J. and Nartker T. "An Evaluation of OCR Accuracy." *Technical Report ISRI TR-93-01, University of Nevada, Las Vegas, April 1993.*

[13] Rice, S., Kanai J. and Nartker T. "The Third Annual Test of OCR Accuracy." *Information Science Research Institute 1994 Annual Report. pp. 11-38.*

[14] Rice, S. "The OCR Experimental Environment, Version 3." *Information Science Research Institute 1993 Annual Report, pp. 83-86.*

[15] Srihari, Sargur N. "Document Image Understanding." *Proceedings of ACM-IEEE Computer Society, 1986. Fall Joint Computer Conference, Dallas, Texas, November 2-6, 1986.*

[16] Throssell, W.R. and Fyrer, P.R. "The Measurement of Print Quality for Optical Character Recognition Systems." *Pattern Recognition, vol. 6, pp. 141-147.*

[17] Wall, Larry and Schwartz, Randal. "Programming PERL." *O'Reilly and Associates, Inc.*