Using Optical Character Recognition to Identify Legibility of Non-western Languages

By: Jennifer Owen

Graphic Communication Department

College of Liberal Arts

California Polytechnic State University

2011

Using Optical Character Recognition to Identify Legibility of Non-western Languages

Jennifer Owen

Graphic Communication Department, March 2011

Advisor: Kevin Cooper

Abstract

The purpose of this study is to see if it is viable to use Optical Character Recognition (OCR) to determine legibility of non-western languages. The results will show that OCR can be used to inspect defects in order to waste fewer products. A large amount of waste is an issue in pharmaceutical printing due to the need for products to be one hundred percent accurate. Waste increases with non-western languages when printers do not know that language and cannot determine legibility themselves.

The research method used was the scientific method. A paragraph of text taken from pharmaceutical packaging was printed in both English and Kanji. Kanji is a form of Japanese and was used as the test language because of its complex characters. The text was printed using flexography at four different impression levels. Each set of samples was then run through OCR.

Results show that the percent wasted was too great using OCR alone. More advanced OCR is recommended and further testing is needed to determine the best combination of OCR and on-press imaging.

Table of Contents

List of Tables

Chapter I – Introduction

Many printers in the United States are printing products that are used overseas. When printing a language that is not understood by anyone at a facility, it becomes difficult to determine the legibility of those characters. A representative from one such company that prints packaging for hospital sterilizers in 11 different languages mentioned that they have the most difficulty printing Kanji on plastic substrates using flexography. This study will show that it is best to use optical character recognition software to identify print defects in non-western languages.

Kanji is a form of Japanese writing that originates from Chinese characters and typically contains pictograms. There is currently no specification regarding tolerances of readability for printing Kanji, which leaves the printers guessing whether or not the characters are readable. The problem with this system is that a print defect has the potential of changing one Kanji character into another. Many pharmaceutical companies that run into this problem use on-press cameras to make comparisons between the running product and previously accepted product. This process is known as halftone comparison or dot-to-dot comparison and results in a large amount of waste because the product it is rejected without question if it is not exactly the same.

Flexography is a relief printing process, meaning the image is raised on the printing plate. To transfer the image from the plate cylinder to the substrate, pressure is put against the plate by running the substrate between the plate cylinder and impression cylinder. This requires a precise amount of pressure to transfer the image correctly. This amount of pressure is referred to as impression. Optical character recognition (OCR) software reads scanned images and turns them

into editable documents. OCR can be used to read printed Kanji samples and find acceptable variations in print defective products.

The purpose of this study is to see if it is viable to use OCR to determine legibility of non-western languages. Using Kanji as the basis for research, the results will show that OCR can be used to inspect defects in order to waste fewer products.

Chapter II – Literature Review

Written Japanese incorporates three different alphabets: Hiragana, Katana, and Kanji. Hiragana and Katana are often combined and referred to as Kana. Kanji originates from Chinese characters and typically contains pictograms that are more complicated than Kana. The characters represent ideas or words, instead of syllables or letters, and have different meanings when combined with other Kanji characters. Many Kanji characters can also be read and pronounced differently based on context (The Kanji Site). Over 50,000 Kanji characters exist; however, in 1981 the Japanese government introduced a list "which includes 1,945 regular characters, plus 166 special characters used only for people's names. Government documents, newspapers, textbooks and other publications for non-specialists use only these Kanji" characters (Ager).

When printing Kanji in the United States, it can be difficult to determine how print defects compromise legibility. One stray dot or smear has the potential to change the Kanji character. Typical flexography defects result from the pressure between the plate cylinder and the impression cylinder being either too low or too high. When the pressure is too low, the entire image is not transferred; and when the pressure is too high, there can be dot gain and rings of ink around the image called halos.

These defects are especially an issue in the pharmaceutical industry where many products could mean life or death. Although the human mind is brilliant and capable of deciphering text that is distorted to a certain degree, pharmaceutical companies cannot take the risk associated with someone not being able to read instructions or dosages. To avoid risk and ensure quality, many

pharmaceutical companies use cameras on the production line to make comparisons with previously accepted and rejected products. “Systems for high-speed production line inspection enable insights into processes that are running too fast for the human eye to follow” (Vaczek). However, “the advantages that automated inspection offers in higher accuracy and labor cost savings must be balanced with the potential for lost productivity deriving from high false reject rates. ‘An enemy of acceptance of machine vision technology is the impact on productivity, from rejected product that should have passed as good product,’ says Michael Soborski, director of inspection solutions, central engineering group, Systech International” (Vaczek). In other words, too much product is labeled as unacceptable because it is not a dot-for-dot match to previously accepted product, even though it is still adequate. Variations are acceptable “from a human-readable standpoint, but if you have a font recognition engine that is not tolerant of those variations, you will get false rejects” (Vaczek).

To avoid wasting product, some pharmaceutical companies have added advanced self-monitoring to their quality checks. Cameras are still used to pull out rejected product, but an operator then checks those products for legibility. For example, Symetix, a capsule and soft gel integrity company, uses a self-monitoring system for inspection of soft gel capsules. Although this process is done for the capsule itself and not the printing, the imaging concept is the same. The system looks for variations in size, shape, color, etc. “When a rogue capsule or defective product is identified, the system automatically removes the problem from the product stream. Pictures of all rouge-classified product are presented to the operator and stored in a batch file” (Vaczek). The operator then checks the product for human-readability and accepts or rejects the product. An additional feature of separate checking is the operator’s ability “to save an image of

a failed product, and flag and post images of subsequent similar failures, for continuous identifying of what has failed and why" (Vaczek).

Although it is possible to hire a Kanji reader, that method is subjective. Companies must have a quantifiable system to identify defects. To avoid hiring a linguistics expert, companies can add optical character recognition (OCR) software to quality checks. For example, Systech International "has released next-generation software for optical character recognition that the company says is more responsive to normal acceptable manufacturing variances" (Vaczek). To accommodate for normal variations and defects, "the company streamlined the font-training process, which develops character libraries used to score inspected images" (Vaczek). The system takes common print variations and "predicatively puts those variations into the font library, which makes the font training faster and more user-friendly" and helps to avoid false rejects (Vaczek).

OCR is a "method for the machine-reading of typeset, typed, and, in some cases, hand-printed letters, numbers, and symbols using optical sensing and a computer" ("Optical Character Recognition"). When a document is scanned, the light reflected by the text is "recorded as patterns of light and dark areas…" ("Optical Character Recognition"). "The OCR software then processes these scans to differentiate between images and text and determine what letters are represented in the light and dark areas" (Lals). The term 'optical' is actually "a bit misleading, as modern OCR software does not use optical character recognition, but actually uses digital character recognition" (McGuigan). The reason for the confusion is that as technology advanced,

the two fields merged, adopting the more commonly known name of optical character recognition.

When OCR software first emerged, it "required training the program on a specific font before it could be accurately input" (McGuigan). Each font had to be entered in and stored for the software to recognize it. Scanning, a document that used a font that was not stored would result in a high level of errors, as the software would use the closest match that it could find. "Early OCR software was used in a wide range of applications, with major corporations using it to read credit card imprints in the 1950's, and the United States Postal Service using it to sort mail since the mid-1960's" (McGuigan).

Newer OCR software adds "multiple algorithms of neural network technology to analyze the stroke edge, the line of discontinuity between the text characters, and the background" (Lals). Taking into account irregularities of ink and paper, "each algorithm averages the light and dark along the side of a stroke, matches it to known characters and makes a best guess as to which character it is. The OCR software then averages or polls the results from all the algorithms to obtain a single reading" (Lals). These "more intelligent systems are now the norm. The methods used are now relatively static, with only a little bit of research going into developing entirely new methods, and most research going into refining existing procedures to make them ever more accurate" (McGuigan).

Chapter III – Research Methods and Procedures

The purpose of this study was to establish if it is viable to use optical character recognition (OCR) software to determine legibility of non-western languages. More specifically, OCR was tested for recognition of pharmaceutical printing. The current process of determining legibility with image comparison results in a large amount of waste. To show that OCR is a better tool for testing legibility, the scientific method was used.

As defined in Dr. Harvey Levenson's book titled *Some Ideas about Doing Research in Graphic Communication*, the scientific method involves five steps. These steps are: identify and define the problem; formulate a hypothesis; collect, organize and analyze data; formulate conclusion; and repeat, verify, and modify the research. The first step, identify and define the problem, has already been completed. The problem is the amount of waste that is created using the current method of determining legibility. For step two, formulate a hypothesis, my hypothesis is that OCR is more accurate and will reduce waste.

For step three, organize and analyze data, samples were printed, scanned, and run through an OCR software program. The original artwork was from a company that prints pharmaceutical packaging using Flexography. The artwork was edited down to two paragraphs of text containing a warning about product use and storage. One paragraph was in English and the other in Kanji. The samples were printed on a Mark Andy 2200. Four different types of samples were printed. The first had perfect impression. These samples would have passed a dot-for-dot image comparison. The second had too low impression. The impression during the press run was turned

down until the image became lighter and, in some cases, disappeared completely. The third had too high impression. The impression on the press was turned up until halos began to appear. The fourth had the impression turned up as high as possible. One hundred samples of perfect impression, high impression, and maximum impression were collected, while ninety-six samples of low impression were collected. The difference in sample size occurred because less low impression samples being printed. After printing, the samples were separated by language. To ensure accuracy, both the English and Kanji samples were run through the same scanner and OCR software. The scanner used was a Fujitsu Scan Snap S510M and the OCR used was Adobe Acrobat Professional. To easily track samples, each one was split up by its sample group and language, and then labeled with its own sequence number, one to one hundred.

For step four, formulate conclusion, the OCR results were checked to confirm accurate character recognition. To repeat and verify, step five of the scientific method, individuals read the samples and confirmed legibility. Conclusions were made based on three different result scenarios:

1) Samples were legible based on image comparison and confirmed humanly readable, but not recognized by OCR.
2) Samples were not legible based on image comparison, but recognized by OCR and confirmed humanly readable.
3) Samples were considered not legible by all three checks.

Chapter IV – Results

The purpose of this study is to see if it is viable to use OCR software to determine legibility of non-western languages. Kanji was used as the basis for research due to its unique and complicated characters. Both English and Kanji samples were printed on a Mark Andy 2200 press, scanned on a Fujitsu Scan Snap S510M, and run through Adobe Acrobat Professional. Four different groups of samples were collected: perfect impression, decreased impression, increased impression, and maximum impression. These four settings were used to represent common press behavior. To verify the findings, individuals unaware of the OCR results then looked at the samples and determined legibility. The results of each test were then compared and a conclusion was made based on what percentage of waste could have been saved by OCR.

Perfect impression was considered to be the optimum printing level. All text was visible and contained no halos or smearing. These samples would have passed a halftone dot comparison with complete accuracy. The results of the perfect impression samples are shown in the following table:

Perfect Impression English						Perfect Impression Kanji					
Sample	Reader	Adobe	Sample	Reader	Adobe	Sample	Reader	Adobe	Sample	Reader	Adobe
1	x		51	x	x	1	x		51	x	x
2	x		52	x	x	2	x		52	x	
3	x	x	53	x		3	x		53	x	
4	x		54	x	x	4	x		54	x	
5	x		55	x	x	5	x		55	x	
6	x	x	56	x	x	6	x		56	x	
7	x	x	57	x	x	7	x		57	x	
8	x		58	x	x	8	x		58	x	
9	x	x	59	x	x	9	x		59	x	
10	x	x	60	x	x	10	x		60	x	
11	x	x	61	x	x	11	x		61	x	
12	x	x	62	x	x	12	x		62	x	
13	x	x	63	x		13	x		63	x	
14	x	x	64	x	x	14	x		64	x	
15	x	x	65	x	x	15	x		65	x	
16	x	x	66	x	x	16	x		66	x	
17	x	x	67	x	x	17	x		67	x	
18	x	x	68	x	x	18	x		68	x	
19	x	x	69	x	x	19	x		69	x	
20	x	x	70	x	x	20	x		70	x	
21	x	x	71	x		21	x		71	x	
22	x	x	72	x	x	22	x		72	x	
23	x		73	x	x	23	x		73	x	
24	x		74	x	x	24	x	x	74	x	
25	x	x	75	x	x	25	x	x	75	x	
26	x		76	x	x	26	x		76	x	
27	x		77	x	x	27	x		77	x	x
28	x	x	78	x	x	28	x		78	x	
29	x	x	79	x		29	x		79	x	x
30	x	x	80	x	x	30	x		80	x	x
31	x	x	81	x		31	x		81	x	
32	x	x	82	x	x	32	x	x	82	x	
33	x	x	83	x	x	33	x		83	x	
34	x	x	84	x	x	34	x		84	x	
35	x		85	x	x	35	x		85	x	
36	x	x	86	x	x	36	x		86	x	
37	x		87	x	x	37	x		87	x	
38	x	x	88	x	x	38	x		88	x	
39	x		89	x	x	39	x		89	x	
40	x	x	90	x	x	40	x		90	x	
41	x	x	91	x	x	41	x		91	x	x
42	x	x	92	x	x	42	x		92	x	
43	x	x	93	x	x	43	x		93	x	
44	x	x	94	x	x	44	x	x	94	x	x
45	x	x	95	x		45	x		95	x	
46	x	x	96	x	x	46	x		96	x	
47	x	x	97	x	x	47	x		97	x	
48	x	x	98	x	x	48	x		98	x	
49	x	x	99	x	x	49	x	x	99	x	
50	x	x	100	x	x	50	x		100	x	

On each table, an “x” indicates a sample that was determined legible. This table shows that out of one hundred English samples, a reader determined that one hundred were legible, and OCR determined that eighty-two were legible. Out of one hundred Kanji samples, a reader determined that one hundred were legible, and OCR determined that eleven were legible.

For samples with decreased impression, the impression was reduced until the text became light. On some samples, parts of the text disappeared completely. These samples would not have passed a halftone comparison test. The results of the decreased impression samples are shown in the following table:

Low Impression English						Low Impression Kanji					
Sample	Reader	Adobe	Sample	Reader	Adobe	Sample	Reader	Adobe	Sample	Reader	Adobe
1	x		51	x		1			51		
2	x		52	x	x	2			52	x	
3	x	x	53	x		3			53		
4	x	x	54	x	x	4			54		
5	x	x	55	x		5			55	x	x
6	x		56	x		6			56	x	
7			57	x		7			57		
8	x		58			8			58		
9	x		59			9			59	x	
10	x		60	x		10			60	x	
11	x		61	x	x	11			61		
12	x		62			12			62		
13			63	x	x	13	x		63	x	
14	x		64	x	x	14	x		64	x	
15	x		65	x		15			65	x	
16	x		66	x	x	16			66		
17	x		67	x	x	17	x		67		
18	x		68	x	x	18	x		68		
19	x	x	69	x	x	19			69	x	
20			70	x		20			70		
21	x	x	71	x		21			71		
22	x		72	x	x	22			72	x	
23			73	x	x	23			73		
24			74	x		24	x		74		
25	x	x	75			25			75	x	
26	x	x	76			26			76	x	
27			77	x	x	27	x		77		
28			78	x		28	x		78		
29	x	x	79			29	x		79	x	
30	x	x	80	x	x	30	x		80	x	x
31	x		81			31			81	x	
32	x		82			32			82	x	
33	x		83	x	x	33	x		83		
34	x	x	84	x		34	x		84		
35	x		85			35			85	x	
36	x	x	86			36			86	x	
37	x		87	x		37			87	x	
38			88	x	x	38	x		88	x	
39	x		89	x		39	x		89		
40	x	x	90	x		40			90		
41			91	x		41			91	x	x
42			92	x	x	42	x		92	x	
43	x		93	x	x	43			93		
44	x		94	x	x	44			94		
45	x		95	x	x	45	x		95	x	
46	x		96	x		46	x		96		
47	x					47					
48	x					48					
49	x					49					
50	x	x				50					

This table shows that out of ninety-six English samples, a reader determined that seventy-six were legible, and OCR determined that that thirty-two were legible. Out of ninety-six Kanji samples, a reader determined that thirty-nine were legible, and OCR determined that three were legible.

The third test group was samples that had the impression increased until halos began to appear. Had these samples been run through a halftone comparison, they would not have been exact matches and would have been wasted. The results of the increased impression samples are shown in the following table:

High Impression English						High Impression Kanji					
Sample	Reader	Adobe	Sample	Reader	Adobe	Sample	Reader	Adobe	Sample	Reader	Adobe
1	x	x	51	x	x	1	x		51	x	
2	x	x	52	x	x	2	x		52	x	
3	x	x	53	x	x	3	x		53	x	x
4	x	x	54	x	x	4	x		54	x	
5	x	x	55	x		5	x	x	55	x	
6	x	x	56	x	x	6	x		56	x	
7	x		57	x	x	7	x	x	57	x	
8	x	x	58	x	x	8	x	x	58	x	
9	x	x	59	x	x	9	x		59	x	
10	x	x	60	x	x	10	x		60	x	x
11	x	x	61	x	x	11	x		61	x	
12	x	x	62	x	x	12	x		62	x	
13	x	x	63	x	x	13	x		63	x	
14	x	x	64	x	x	14	x		64	x	
15	x	x	65	x	x	15	x		65	x	
16	x	x	66	x	x	16	x		66	x	
17	x	x	67	x	x	17	x	x	67	x	
18	x	x	68	x	x	18	x		68	x	
19	x	x	69	x		19	x		69	x	
20	x	x	70	x	x	20	x		70	x	
21	x	x	71	x	x	21	x		71	x	
22	x	x	72	x	x	22	x		72	x	
23	x	x	73	x		23	x	x	73	x	
24	x	x	74	x	x	24	x	x	74	x	
25	x	x	75	x		25	x		75	x	x
26	x	x	76	x		26	x	x	76	x	
27	x		77	x	x	27	x		77	x	
28	x	x	78	x	x	28	x		78	x	
29	x		79	x	x	29	x		79	x	
30	x	x	80	x	x	30	x		80	x	
31	x	x	81	x		31	x	x	81	x	
32	x	x	82	x	x	32	x		82	x	
33	x	x	83	x		33	x		83	x	
34	x	x	84	x	x	34	x		84	x	
35	x	x	85	x	x	35	x		85	x	
36	x	x	86	x	x	36	x		86	x	
37	x		87	x		37	x	x	87	x	
38	x	x	88	x	x	38	x		88	x	x
39	x	x	89	x	x	39	x		89	x	
40	x	x	90	x	x	40	x		90	x	
41	x	x	91	x		41	x		91	x	x
42	x	x	92	x	x	42	x		92	x	
43	x	x	93	x		43	x		93	x	
44	x	x	94	x	x	44	x		94	x	
45	x	x	95	x		45	x		95	x	
46	x	x	96	x	x	46	x		96	x	
47	x	x	97	x	x	47	x		97	x	
48	x	x	98	x	x	48	x		98	x	
49	x	x	99	x	x	49	x		99	x	
50	x	x	100	x		50	x		100	x	

This table shows that out of one hundred English samples, a reader determined that one hundred were legible, and OCR determined that eighty-four were legible. Out of one hundred Kanji samples, a reader determined that one hundred were legible, and OCR determined that fourteen were legible.

The final group of samples had the impression set as high as possible. These samples would not have passed a halftone comparison and would have been wasted. The results of the maximum impression samples are shown in the following table:

Max Impression English						Max Impression Kanji					
Sample	Reader	Adobe	Sample	Reader	Adobe	Sample	Reader	Adobe	Sample	Reader	Adobe
1	x		51	x		1			51		
2	x		52	x		2			52		
3	x		53	x		3	x		53	x	
4	x		54		x	4			54		
5	x		55	x		5			55		
6	x		56			6			56		
7	x		57	x		7			57		
8			58	x		8			58	x	
9	x		59	x		9			59		
10			60	x		10			60		
11	x		61	x		11			61		
12			62	x		12			62		
13	x		63	x		13			63		
14	x		64	x	x	14			64		
15	x		65	x		15			65		
16	x		66	x		16			66	x	
17	x		67	x		17			67	x	
18	x		68	x		18			68		
19	x		69	x		19	x		69		
20			70			20			70		
21	x		71	x	x	21			71		
22			72		x	22	x		72		
23	x		73	x		23	x		73		
24			74			24	x		74	x	
25	x		75	x		25			75	x	
26		x	76			26			76		
27	x	x	77	x		27			77		
28	x	x	78			28			78		
29	x		79	x		29			79	x	
30			80	x		30	x		80		
31	x		81	x		31			81		
32	x		82	x		32			82		
33	x		83	x		33			83	x	
34	x		84	x		34			84		
35	x		85	x		35			85		
36	x		86	x		36	x		86		
37	x	x	87	x		37			87	x	
38			88			38			88	x	
39	x		89	x		39			89		
40			90			40	x		90		
41	x		91	x		41	x		91		
42			92		x	42			92		
43	x		93	x		43			93	x	
44	x		94	x		44			94		
45	x		95	x		45			95		
46	x		96	x		46			96		
47	x		97	x		47	x		97		
48	x		98	x		48			98		
49	x		99	x		49			99		
50	x		100			50			100		

This table shows that out of one hundred English samples, a reader determined that seventy-eight were legible, and OCR determined that nine were legible. Out of one hundred Kanji samples, a reader determined that twenty-one were legible and OCR determined that zero were legible.

The results of each test were then examined and a percentage of waste saved was established based on the number of samples determined legible. The results of the Kanji samples and their representative percentages are shown in the following table:

Kanji Samples

	Number Printed	Reader Legibility	% Saved by Reader	OCR legibility	% Saved by OCR
Perfect Impression	**100**	100	100%	11	(89%)
Low Impression	**96**	39	40.63%	3	3.13%
High Impression	**100**	100	100%	14	14%
Maximum Impression	**100**	21	21%	0	0%

Although the perfect impression samples would pass a halftone comparison and not need to be run through OCR, the chart shows that with OCR alone, 89% percent of the samples would have been wasted. Meanwhile, 3.13% of the low impression samples would have been saved and 14% of the high impression samples would have been saved; but none of the maximum impression samples would have been saved.

The results of the English samples and their representative percentages are shown in the following table:

English Samples

	Number Printed	Reader Legibility	% Saved by Reader	OCR legibility	% Saved by OCR
Perfect Impression	**100**	100	100%	82	(18%)
Low Impression	**96**	76	79.17%	32	33.33%
High Impression	**100**	100	100%	84	84%
Maximum Impression	**100**	78	78%	9	9%

Were OCR to be used alone, the chart shows that 18% of perfect impression samples would have been lost, 33.33% of the low impression samples would have been saved, 84% of the high impression samples would have been saved, and 9% of the maximum impression samples would have been saved.

Chapter V – Conclusion

The result of the tests done on both language samples show that the amount of waste created was greater than the amount of waste saved. Therefore, OCR by itself is not a viable tool for checking legibility of non-western languages. Using OCR alone would only increase the amount of waste the Pharmaceutical printing companies have.

The OCR software used in this test was a simple version with few settings and basic functions. For practical use, more advanced OCR software is needed. Since there is "no such thing as perfect OCR… choosing a program to buy comes down to extra features: multi-lingual support, one-touch scan and conversion integration, automatic PDF conversion, and whole-word recognition across specialized disciplines like legal and medical fields" (McGuigan). In this instance, having a software package with broader language availability or specialized recognition should increase results.

One example of broader language availability comes from IRIS, a company that creates solutions for document and information management. IRIS has its own OCR program called Readiris, which has an Asian language version. This package contains additional forms of Japanese, Chinese, Korean, and even Hebrew that are not available in most programs (I.R.I.S.). Another example comes from Nuance, a company that provides speech, document, and imaging solutions. Nuance's product, Omnipage, claims an accuracy rate 50% greater than most programs. Omnipage also includes "recognition dictionaries for financial, legal, and medical specialties [to] ensure the most accurate conversion of important industry-specific terms" (Nuance).

Another issue when choosing OCR is price. The cost of OCR has a broad range and generally increases with features and accuracy. Readiris Pro 12 Asian costs $249.00 and Readiris Corporate 12 Asian costs $589.00 (I.R.I.S.). Omnipage 17 for at home use costs $149.99. Omnipage Professional 17 costs $499.99 (Nuance).

Although OCR alone was not successful in reading non-western languages, OCR in combination with other on-press imaging systems may save a more significant amount of waste. More advanced OCR is recommended and further testing is needed to confirm the best combination of on-press imaging and OCR.

Bibliography

Ager, Simon. “Japanese Kanji.” *Omniglot*. 2010. Web. 21 October 2010. http://www.omniglot.com/writing/japanese_kanji.htm

I.R.I.S. 2009. Web. 1 March 2011. http://www.iriscorporate.com/c2-1-17/I-R-I-S----OCR-Software--Document-Management-Solutions-and-Complex-IT-Infrastructure.aspx

Lals, Sami. “QuickStudy: Optical Character Recognition.” *Computerworld*. 29 July 2002. Web. 21 October 2010. http://www.computerworld.com/s/article/73023/Optical_Character_Recognition

Levenson, Harvey. *Some Ideas about Doing Research in Graphic Communication.* 2001. p19-21

McGuigan, Brendan. “How do I Choose the Best OCR Software?” *WiseGeek*. 08 September 2010. Web. 21 October 2010. http://www.wisegeek.com/how-do-i-choose-the-best-ocr-software.htm

Nuance. 2011. Web. 2 March 2011. http://www.nuance.com/

“Optical Character Recognition.” *Columbia Electronic Encyclopedia, 6th Edition* (2009). Academic Search Elite. EBSCO. 21 October 2010. http://web.ebscohost.com/ehost/detail?vid=1&hid=14&sid=9984f025-9da9-4847-8996-89c108e9dece%40sessionmgr4&bdata=JnNpdGU9ZWhvc3QtbGl2ZQ%3d%3d#db=afh&AN=39025742

The Kanji Site. 2001. 20 October 2010. Web. http://www.kanjisite.com/index.htm

Vaczek, David. “INSPECTION SYSTEMS: Looking Closer at Inspection.” *Pharmaceutical & Medical Packaging News*. 29 May 2008. Web. http://www.pmpnews.com/article/inspection-systems-looking-closer-inspection