White Paper

Automatic Classification Improves Processes and Yields Major Reductions in Scanning Costs

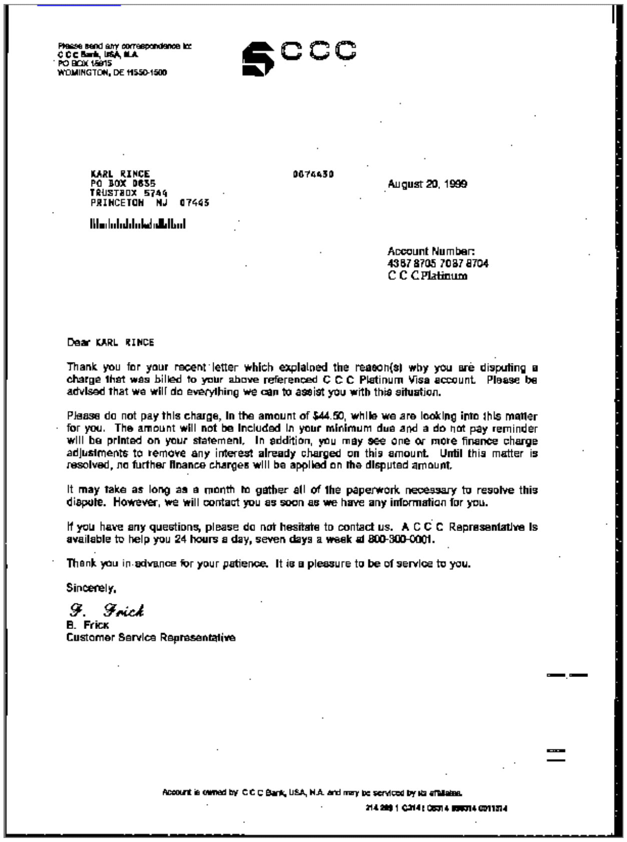
By Harvey Spencer

**Presented by
Océ Document Technologies
4330 East West Highway
Suite 304
Bethesda, MD
Tel: 301-652-9732
www.odt-oce.com**

Capturing Documents is Expensive

When document imaging and management systems were first introduced back in the 1980's, it was conventional to distribute the scanners into departments. As with other office equipment a seated operator opened a folder, removed the contents, and scanned each page, viewing it on a high resolution screen to locate and key the index data that was needed to retrieve it.

Consider the following piece of correspondence (figure 1) which was sent to an individual employee regarding a charge on his corporate credit card. This may have been in a folder or placed in a stack of documents to be scanned by the employee's company into a document management system. Some of the documents in the stack are single pages, some are multiple pages. There may be a mix of correspondence, notes, forms and other documents in the stack.



Please send any correspondence to:
C C C Bank, USA, N.A.
PO BOX 15615
WOMINGTON, DE 11550-1500

CCC

KARL RINCE
PO BOX 0635
TRUSTBOX 5744
PRINCETON NJ 07445

0074430

August 20, 1999

Account Number:
4387 8705 7087 8704
C C C Platinum

Dear KARL RINCE

Thank you for your recent letter which explained the reason(s) why you are disputing a charge that was billed to your above referenced C C C Platinum Visa account. Please be advised that we will do everything we can to assist you with this situation.

Please do not pay this charge, in the amount of $44.50, while we are looking into this matter for you. The amount will not be included in your minimum due and a do not pay reminder will be printed on your statement. In addition, you may see one or more finance charge adjustments to remove any interest already charged on this amount. Until this matter is resolved, no further finance charges will be applied on the disputed amount.

It may take as long as a month to gather all of the paperwork necessary to resolve this dispute. However, we will contact you as soon as we have any information for you.

If you have any questions, please do not hesitate to contact us. A C C C Representative is available to help you 24 hours a day, seven days a week at 800-300-0001.

Thank you in advance for your patience. It is a pleasure to be of service to you.

Sincerely,

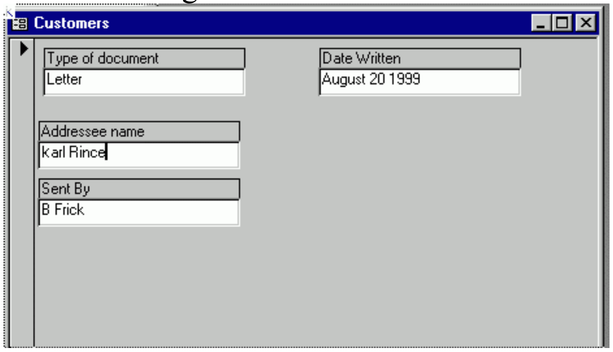
F. Frick
B. Frick
Customer Service Representative

Account is owned by C C C Bank, USA, N.A. and may be serviced by its affiliates.
214 289 1 C314 [illegible]

Figure 1: Example scanned letter

In this case this letter is from the Credit Card Issuer in relationship to a dispute on a charge. Let's assume that the document is indexed for retrieval or processing by type of document (i.e. letter), by date, by who it is from and who it was addressed to. Other fields or classifications may also be needed. So for example depending on the document's usage, some systems may also require the capture of the subject matter i.e. dispute over Visa charge or subject of the complaint.

The traditional way to handle this is to scan it with the operator viewing the image to see its topic, locating the relevant fields and entering them on a screen in a separate window as shown in figure 2.



Type of document: Letter

Date Written: August 20 1999

Addressee name: karl Rince

Sent By: B Frick

Figure 2: Index data entry screen

Having entered the relevant information, the operator then moves to the next scanned page, and so it continues.

This is not the most effective way to work. However this method has persisted in low volume departmental applications and is probably still the predominant way of working. With the move to low speed scanners or digital copiers for departmental scanning of small volumes or single documents, this method is set to continue for a while.

The problem is that it is a labor intensive model that slows the scanning down to the speed of the operator -- to prepare the documents (as he takes them out of the folder), place them one at a time in the feeder or on the flatbed, press scan, view the image(s) and key enter the retrieval information. Using this method a 50 page/minute scanner could easily be running at an effective rate of 4 or 5 pages/minute.

The Solution is to Insert Indexing Control Documents

Service Bureaus and others who need to scan high volumes of paper in a cost effective manner, needed to find a more effective batch process with automated indexing. One improvement was to sort the documents by type, and use templates to identify the index fields to maximize an operator's performance. But the operator still has to view each image and key the index data. An average trained key entry operator costs around $18/hr. If an average set of retrieval indexes consists of 20 characters in a single page document, the cost of indexing say 1,000 documents manually is $66. When dealing with 20,000 or more documents a day, this becomes substantial.

One solution had already been pioneered by Kodak through the use of Patch Cards (with their high speed microfilmers). These are specially coded sheets containing bold black stripes -- which designated an index break. Patch cards could be inserted at three different levels, roughly approximating a document, a folder, and a file.
Using the letter example above in an H-R application these might be "a *letter*" as the document, "*expenses*" as the folder and the recipient "*Karl Rince*" as the file. Although

somewhat limited, these Patch Cards allowed the high volume service bureau to streamline the capture process into a more efficient production line process --

1. extract the documents from their folders and prepare them for moving at high speed through a transport -- this included unfolding, removing staples and repairing rips
2. feed large batches of mixed paper documents as fast as possible through the capture device
3. remove the patch cards and restore the papers into their original state in the folders assuming this was what the customer wanted

When high speed paper scanners first appeared, these Service Bureaus wanted to run the same sorts of process efficiencies. The software industry responded by adding Patch Card reading as above and went further. They allowed for printing the inserted control cards with barcodes and / or OCR'able fields. By presorting documents into batches by type of document, the service bureaus could optimize the settings on the scanners and use predefined templates to locate the index fields.

It substantially improved performance of the scanners, allowing them to run at close to rated speed -- a factor of 10 or more. With these efficiencies the costs of preparation easily get absorbed as a data preparation person will prepare around 1,000 pages/hour at an average of $15.73 loaded labor cost. But users have ended up spending 50% of their time sorting documents by type and inserting separator pages as shown below.

Activity	% of time spent
Sort by Type	34%
Inserting Separator Pages	16%
Photocopying	9%
Sorting by Darkness	7%
Repair / Taping Documents	6%
Other	10%
Staple Removal	18%

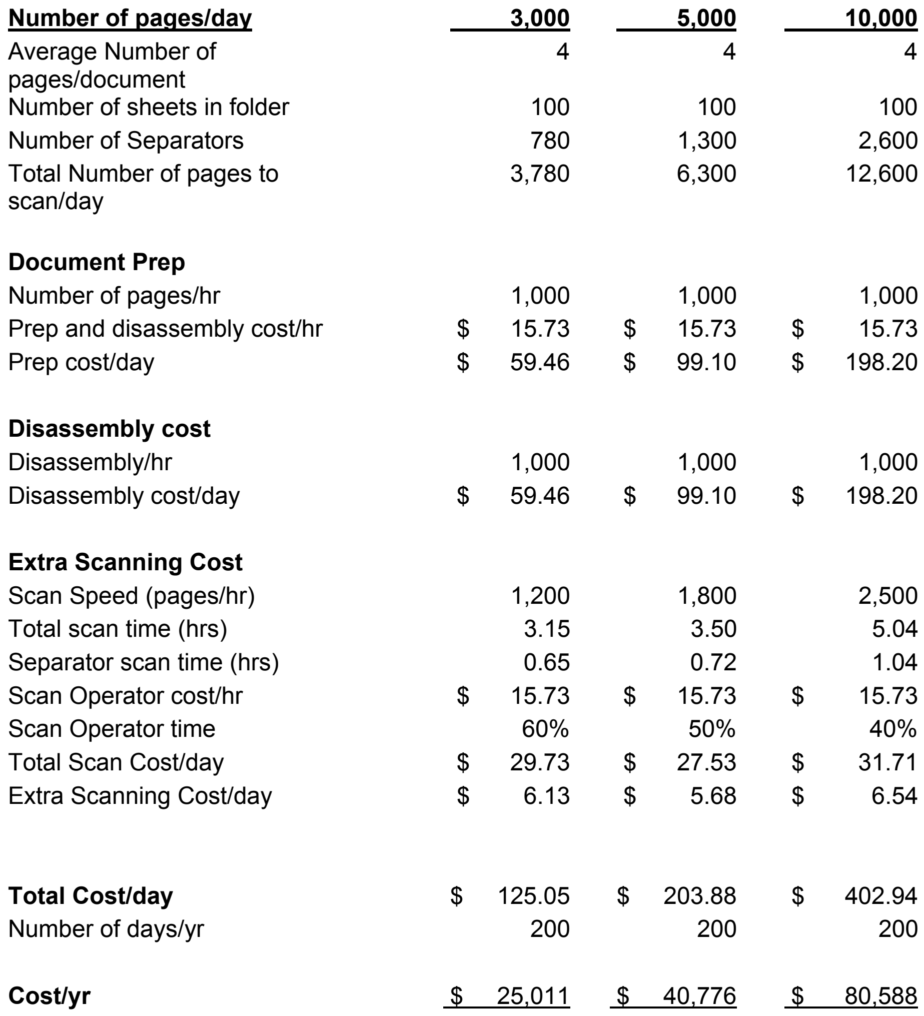
Figure 3: Source: TAWPI Forms Processing and Image Capture Study -- 2000

Hardware Improvements Reduce Preparation

A number of scanner vendors have taken the next step to reduce the need for preparation. By improving their autofeeders and transports to accept wider varieties and qualities of documents and including devices to ensure that double feeds could not go through undetected. Color scanning and auto sharpening thresholding products such as Kofax' VRS have reduced the need to sort by darkness.

Automatic Classification Eliminates the Remaining Preparation Costs.

Still high volume users are spending thousands of dollars a year on preparation and there is an additional load placed on the scanners because of the increased volume of separator sheets. Figure 3 analyses these costs under 3 different scenarios based on 3,000, 5,000 and 10,000 sheet of paper to be scanned per day. I have assumed in each case that there is a set of folders each containing approximately 100 sheets and that the average document contains four pages to allow for multi-page correspondence, reports etc. Of course some will have one page, some will have two and others may have five or more. Each document is indexed by two levels -- a document type, and a folder.



Number of pages/day	3,000	5,000	10,000
Average Number of pages/document	4	4	4
Number of sheets in folder	100	100	100
Number of Separators	780	1,300	2,600
Total Number of pages to scan/day	3,780	6,300	12,600
Document Prep			
Number of pages/hr	1,000	1,000	1,000
Prep and disassembly cost/hr	$ 15.73	$ 15.73	$ 15.73
Prep cost/day	$ 59.46	$ 99.10	$ 198.20
Disassembly cost			
Disassembly/hr	1,000	1,000	1,000
Disassembly cost/day	$ 59.46	$ 99.10	$ 198.20
Extra Scanning Cost			
Scan Speed (pages/hr)	1,200	1,800	2,500
Total scan time (hrs)	3.15	3.50	5.04
Separator scan time (hrs)	0.65	0.72	1.04
Scan Operator cost/hr	$ 15.73	$ 15.73	$ 15.73
Scan Operator time	60%	50%	40%
Total Scan Cost/day	$ 29.73	$ 27.53	$ 31.71
Extra Scanning Cost/day	$ 6.13	$ 5.68	$ 6.54
Total Cost/day	$ 125.05	$ 203.88	$ 402.94
Number of days/yr	200	200	200
Cost/yr	$ 25,011	$ 40,776	$ 80,588

Figure 4: Analysis of costs of batch scanning with separators

In a mixed batch of documents, a separator page has to be inserted to tell the batch scanning software where each document starts and where a new folder starts. The preparation clerk has to insert and remove hundreds of extra sheets each day. In figure 4, it can be seen that with an average volume of 3,000 pages/day, this costs $25,000 a year -- with 10,000 pages/day it costs over $80,000.

The Initial Solutions Introduced Extra Costs

Software vendors initially came up with ideas to reduce patch card insertion such as putting barcode stickers on the documents or using color codes. But both methods kept the need for the preparers to look at each document and had other disadvantages which has limited their usefulness:-

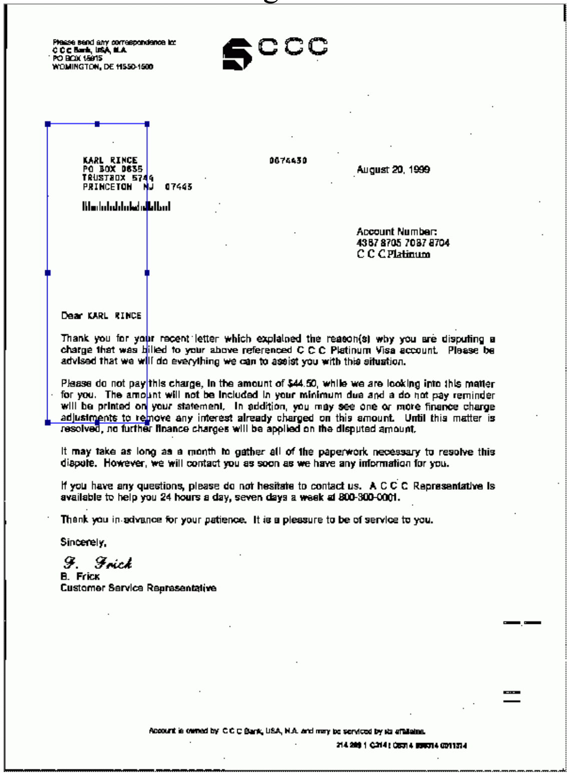
- Barcodes cover a small area on the document which could be important, like initials on a legal document and they can be difficult to remove

- Color is interpreted differently by each scanner and background colors can change the highlighting and produce a completely different and unintelligible color.

Software Auto Classification Is the Answer
Auto classification eliminates the time taken to sort by type and insert and remove separator pages. Automatic indexing can then use the classification which also finds and extracts the index fields. It starts with pattern recognition combined with the latest voting OCR, auto classification to mimic the manual operator's decision making -- the need to look at the image, decide what it is, where the fields to enter are located and then it can use its OCR software to capture the data. But, being an automated process , it can do it much quicker.

Using just the four fields shown in Figure 2 as an example, Oce's DOKuStar and other products like it, can automatically classify the document type and extract and convert the fields, placing them into the index areas without human interaction.

The way it works it is to allow the system to search for key phrases, patterns or rules. It's very similar to the way the human operator decides on how to index the documents. So for example a letter can be classified by the fact that it contains both or either the salutation word "Dear" and the word(s) that precede the signature -- "Sincerely", "Yours", "Best", "Regards" etc. A company name can be found on the basis of a logo match. These 'search criteria' can be set up to look in sub areas of the document so as to maximize performance. The word 'Dear' for example in a letter occurs towards the left of the document underneath the senders return address. So the search is set up to look at the left side of each image as shown within the rectangle banded area in figure 5:-



Please send any correspondence to:
C C C Bank, USA, N.A.
PO BOX 15915
WILMINGTON, DE 11550-1500

CCC

KARL RINCE
PO BOX 0635
TRUSTBOX 5749
PRINCETON NJ 07443

0074430

August 20, 1999

Account Number:
4387 8705 7087 8704
C C C Platinum

Dear KARL RINCE

Thank you for your recent letter which explained the reason(s) why you are disputing a charge that was billed to your above referenced C C C Platinum Visa account. Please be advised that we will do everything we can to assist you with this situation.

Please do not pay this charge, in the amount of $44.50, while we are looking into this matter for you. The amount will not be included in your minimum due and a do not pay reminder will be printed on your statement. In addition, you may see one or more finance charge adjustments to remove any interest already charged on this amount. Until this matter is resolved, no further finance charges will be applied on the disputed amount.

It may take as long as a month to gather all of the paperwork necessary to resolve this dispute. However, we will contact you as soon as we have any information for you.

If you have any questions, please do not hesitate to contact us. A C C C Representative is available to help you 24 hours a day, seven days a week at 800-300-0001.

Thank you in advance for your patience. It is a pleasure to be of service to you.

Sincerely,

F. Frick
B. Frick
Customer Service Representative

Account is owned by C C C Bank, USA, N.A. and may be serviced by its affiliates.
214 289 1 C3141 C6314 [illegible] 0011374

Figure 5: Example zone search

In figure 5, the line boundary shows the approximate area in which to search the images. A phrase list and rules associated with this provide the search criteria and action to perform. In this case, once the word "Dear' is found, the system can be pretty certain that it is dealing with a letter and place that in the 'type of document' field in figure 2. It can then extract name of the addressee -- in this case Karl Rince, which is next to it, and place that into the 'Addressee name' field.

In a similar way the other classification and indexing retrieval fields can easily be set up. By enabling exclusion lists, pattern matching, lists of specifically allowed and disallowed characters, edit patterns, logical 'ands' and 'ors' etc. and even in some cases specific code associated with a phrase, the user's system can be set up to analyze each of the scanned documents, classify them by type(s) and extract the indexes.

The Bottom Line is Process Improvement and Labor Savings

Auto Classification and Indexing, known as Intelligent Document Recognition (IDR) works like a human to search the images in order to find known data elements that classify documents and can automatically extract the indexes. It eliminates the need to insert, scan and remove batch control and indexing sheets. In high volume environments it can save the user thousands of dollars a year in document preparation and scanning. But it can also replace the basic scanning, view and index method in use in lower volume applications. Using automatic classification methods, the operator can place a document or groups of documents into a scanner's or copier's autofeeder and press ***start***, secure that the system will analyze the content of each page, classify each document according to rules set up by the integrator and index it so that it can easily be found.. Automatic classification is set to become the standard for all scanning jobs in the future.