Scanning Recognition: A Practical Overview

A white paper by John Scaletta, Director of Business Development for Scan-Optics, Inc.

Introduction

This paper is intended to furnish the reader with the pros and cons of in-line and off-line recognition and how one can complement the other. It will also define the methods, terminology, and practical applications for OCR/ICR/OMR and will describe other recognition technologies, such as Barcodes and 2D Barcodes.

Why OCR?

Optical Character Recognition (OCR) was developed as a cost saving process which automated the manual data entry of information. OCR was the automated conversion of hand-written and/or pre-printed data into machine-readable code for computer processing. In the early days of computer processing, punch cards were used to enter data in computers. This was a very manual and time-consuming process. Intelligent Machines Research Corp developed the first OCR machine in 1952, but it wasn't until the early 1960's when companies like IBM, Addressograph-Multigraph and Recognition Equipment Incorporated introduced scanners that automated the OCR process. The early problems with OCR had to deal with both the variations in OCR printed typefaces and the forms that they were printed on.

In 1966, the American National Standards Institute created a standard typeface for the United States. This typeface was called OCR-A. It provided both the manufacturers and the end-users of OCR equipment with a standard typeface that didn't have conflicting character types. The letter "O" and the number "0" had distinct, non-conflicting typefaces. Other characters like the letter "Z" and the number "2" were also unique. In this regard, the reason that the OCR-A font was development was to improve the accuracy of reading data from forms. The final benefit from OCR was the speed at which character recognition took place. OCR characters could be read at 10,000 characters per second, where as a person manually entering data could only enter 2 to 3 characters per second.

So WHY OCR? It provides the end-user with improvements in data accuracy and speed, which results in overall data processing cost savings.

Note: Appendix A of this document describes the terminology used for OCR/ICR processing of documents. This paper will highlight some of this terminology in respect to what is needed to successfully employ OCR/ICR recognition.

Method of Input

Probably the most important part of OCR/ICR recognition process is the paper media and the ink used to print onto the paper media. The paper (document) should have the following characteristics:

- Zero rag content, no fluorescent additives or watermarks.
- Reflectance of light from the surface of paper at 80% or more.
- Opacity light reflectance of paper backed with black divided by paper backed with white at 80% or more.
- Nominal paper weight typically refers to the thickness of paper at 20 pd to 24 pd.
- Standard paper sizes:
    - American standard paper sheet sizes are 11 x 17 (two 8-1/2 x 11) or 22 x 17 (four 8-1/2 x 11).
    - European standard paper sheet sizes are A4 (11.69 x 8.43) and A3 (11.69 x 16.8).
    - No dirt, impurities, specks, wood pulp, etc.
    - Paper smoothness when measured with Sheffield instrument of an air-flow between 100 and 200.
    - The porosity Gurley reading for air or ink passing through paper should be between 10 and 95.

- Paper gloss or coating should be low gloss.
- Paper grain should be in the direction of travel of the paper through a system.

Ink applied to the paper (document) should have the following characteristics:

- Read inks must not reflect more than 50% as a measure of the paper background.
- Reflective inks must reflect more than 85% of the paper background, i.e., dropout ink, blind ink or non-readable ink - visible to the human eye but invisible to the image camera.

There are basically six types of OCR/ICR READ functions that a recognition scanner performs: machine print, handprint, mark sense, character destruct, reference marks, and barcode/patch code.

1. Machine Print

For machine print, OCR-A is the American ANSI standard recommended font for printed data and will read with an accuracy in excess of 99% when printed at the "Range x" print quality, as defined by the American National Standards Institute (ANSI) X3.99.

2. Handprint

Handprint is a free format style of entering data on a form. This information can be alphanumeric plus a few special characters. Handprint characters will read with accuracy in excess of 90% when printed in the character shapes and sizes as defined by the American National Standards Institute (ANSI) X3.45.

To assist in placing handprint information on a document, handprint guide boxes were created. Handprint boxes are printed in a color that is chosen from NON-READ INKS that match the scanner light source or filters. The reflectance of the handprint box must be 85% or less of the paper on which it is printed. The three most commonly used sizes of handprint box sizes are 4, 4-1/2, and 5 boxes per inches horizontally.

The measurements below refer to the “center to center” distance between boxes. The height for each can vary depending on the aesthetic appearance. Normally, vertically spacing is three boxes per inch.

Box Size		Width x Height
Minimum	- - - - -	0.200" x 0.230"
Nominal	- - - - -	0.225" x 0.275"
Large	- - - - -	0.250" x 0.300"

Box Border (Line Weight): For aesthetics, it recommended but not critical that a size of 0.100" be used. The minimum and maximum box border depends on the individual designing the form. But as a rule, the person designing the form should take into account that the larger the border the more likely the person filling out the form is to stay within the boxes. This is very important because characters printed outside of the handprint boxes increase the likelihood of misreads and rejects during scanning.

The Laser Handprint Guide Boxes allow you to print your own application forms on a laser printer rather than using a professional printing service. This feature also allows you to print the guide boxes for handprint, without using dropout inks.

Applications that run with dropout ink boxes will run with black boxes without degradation of recognition performance.

Xmax = Max. Dot Size = 0 .008"

Xmin = Min. Dot Size = 0 .0025"

Ymax = Max.Dot Spacing = 0.052"

Ymin = Min. Dot Spacing = 0.012"

Y/2 = Spacing Between rows of dots.

Generally the size of the Dot must not be larger than 10% of the handprint character width.

3. Mark Sense (OMR)

Just as Handprint boxes, all mark sense targets should be printed in a color that is chosen from NON-READ INKS that match the scanner light source or filters. Although most scanner recognition units (hardware or software) are not restrictive about the size and shape of the mark target, the person designing the form should be aware of the minimum dimensions. Always try to maximize the distance between marks because the recognition units need adequate room to distinguish one mark from the next.

All targets on a form should be of one uniform size and shape. Targets should be made large enough so they cannot be mistaken for extraneous dirt. The recommended minimum amount of targets that can be packed into one inch is 6. This means that the space between targets should be no less than 0.167". Target spacing determines what the size of the targets will be. Targets should never overlap.

4. Character Destruct

The character destruct recognition is used in "mark sense" type applications where the user is required to fill in one or more target boxes on a document and the outlines of the target box are not printed in dropout ink. The determination of whether a target is considered marked or not, is based upon the destruction of the character. Typically, the shape of the character destruct is the character "zero", oval or square.

5. Reference Marks

Printed reference marks or anchor points are recommended when processing handprint and OMR forms. Reference marks allow the system's recognition engine to adjust for variation in placement of targets from one printing to another. Variation in data placement with respect to the edge of a form is often caused by inaccuracy in the printing process. Because the reference mark or anchor point is printed with respect to the targets, variations in form cutting or printing will not affect reading performance.

<u>All Reference Marks:</u>

Should have a fixed width and height.

Should be unique in size, and larger than the average text character.

Must be printed in black ink.

It is highly recommended that all reference marks have at least a 0.250" clear area on all sides whenever possible. This makes for greater convenience when trying to find the center of the mark. The minimum stroke width should be no smaller than 0.014", this allows for capturing a solid video picture of the reference mark.

6. Barcode/Patch Code

Typically, barcodes are horizontal bars that are black and white with different widths representing alphanumeric data. Typical barcodes are 3 of 9, 2 of 5, Codabar, etc. Essentially, these are called one-dimensional barcodes, which tend to occupy a great deal of real estate for a limited number of characters. To represent 10 characters in a one-dimensional barcode may take up a rectangle 1/2" by 4". The same rectangle in a 2D barcode may represent 100 to 400 characters. As 2D barcodes become more acceptable there will be a trend toward barcode recognition in the future.

There are a few characteristics about 2D barcode that are important to remember when it comes to recognition. 2D barcode is also known as PDF 417. The basic assumption is that document images of 2D barcode must be scanned at 3 times the printed resolution to properly scan a barcode. The following table provides the scanning resolution requirement for the 2D X-Dimension in mils:

Barcode X-Dimension (mils)	Barcode X-Dim. (DPI	Required Scanning DPI
5	200	600
10	100	300
15	66.7	200
20	33.3	100

The Patch Code is a very simple barcode consisting of two types of long horizontal bars with a minimum length of 2.0 inches, and a maximum length equal to the document width. The characters of the Patch Code consist of 4 bars, vertically spaced by 0.08 inches. Patch codes are used to separate and/or index documents or transactions.

Form Identification (Form ID)

To successfully utilize OCR/ICR, it is essential to identify each form to be processed. When processing a single form, then the ID can be a default definition since it is only one form. Form identification only becomes important when processing multiple forms within a single job (intermixed forms processing). There are a number of techniques used to do this.

These types of forms are usually called structured forms:

- OCR numeric or alphanumeric characters printed on each form to provide a unique identification
- Unique pre-printed areas, logos, or blank (empty) form templates

Once the structured form has been identified, matching of each field is very fast since each field is always located in the same x,y location.

The most difficult types of form identification are on semi-structured and unstructured forms.

For both semi- and unstructured forms, there is no traditional Form ID for fast matching. The system must be able to learn from form to form, recognize "floating" field names and locate the data related to those field names. This can be a slow process and may result in false positive field results (substitutions).

Unstructured forms are the most difficult documents to apply recognition techniques against. The recognition techniques used to read unstructured forms are applied to the entire form. All information on the form is used to match against Forms Identification, patterns, tables, and the actual data to be recognized. This can also be a very slow process, but there are off-line recognition engines attempting to solve this most difficult process.

OCR/ICR Accuracy

After a form has been successfully identified and the fields found that need to be read, then the next important attribute of OCR/ICR is accuracy.

Machine printed characters (OCRA font) will read with accuracy in excess of 99% when printed at the "Range x" print quality, as defined by the American National Standards Institute (ANSI) X3.99. The substitution rate of this data is one in a hundred thousand or .001%.

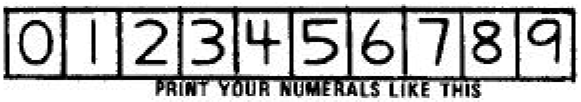
During the early days of printing machine-printed characters, printing was done from a ribbon-chain printer, where the alphanumeric and special characters are placed on a mechanical chain and printed through a rotating ribbon. As both the ribbon and the mechanical characters on the chain would wear, the printing quality would degrade thus producing poorly formed characters. This was typical of the print quality in the 60's, 70's, and 80's. Typewriters were also used quite a bit during this time period to place data on documents to be read using OCR/ICR. These typewriters exhibited similar problems to the chain printers. Another example of printing problems came with the advent of dot matrix printers. These printers were not very uniform in character style, and had problems with character sizing. With the use of in-line recognition engines, reading this degraded data could be done using multiple levels of gray scale, and techniques could be applied to take the best read result at these various levels of gray. This could not be done with off-line recognition engines since they could only handle black and white bi-tonal images.

Then came the advent of laser printers. The laser printer provided print quality that was very uniform and not missing any data. With the improvement of ink jet and laser printer versus the older style of dot matrix and chain printers, the off-line engine's black and white bi-tonal images could now be used to read machine-printed OCR data since the printing of the images improved drastically.

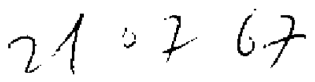
Today, the recognition accuracy rates on machine printed data has risen to the high levels of 99%, which has allowed off-line recognition systems to read as well as in-line recognition systems using just black and white bi-tonal images.

Pitch (spacing between characters) and Point size (character sizing) have always been critical to character recognition. The recommended character pitch for machine-printed data is 10 characters per inch and the point size recommendation is 12pts. The recommended the machine print font is OCR-A.

Improvements to reading handprint information have not progressed much over the past thirty years. It can be suggested how to print characters but it is still dependent upon the individual filling out the form. The following is an example of how handprint numbers should be printed:



The following example is an international date dd/mm/yy, which has been printed very light and in a European style where they flag their "1's" and cross their "7's":



Handprint characters will read with an accuracy rate in excess of 90% when printed in the character shapes and sizes as defined by the American National Standards Institute (ANSI) X3.45. The substitution rate for this type of data is less than one in a hundred or 1%. Uncontrolled handprint has, typically, a substitution rate of less than five in a hundred, or 5%. Again, handprint recognition rates are dependent upon the person filling out the forms.

In-Line Recognition versus Off-Line Recognition

There has always been a debate between vendors on the benefits of in-line recognition versus off-line recognition.

The conventional meaning of in-line recognition dealt with Optical Character Recognition (OCR) scanning equipment. This equipment consisted of electro-mechanical devices that processed documents/forms in real time. Documents/forms were picked by a feeder, transport to a read station (image dissector tube or diode array) which recognized the typewritten, hand-written, or mark targets, and then sorted the documents/forms into multiple pockets. These electro-mechanical scanners had the recognition hardware built right into the system (in-line) and did the recognition as the documents/forms were transported through the system. These systems were mainly developed during the 60's, 70's, 80's by Control Data, IBM, Recognition Equipment, ScanData, and Scan-Optics to name a few.

During the early 1990's, imaging systems started to move more toward in-line recognition capture of documents/forms. To process these images many off-line recognition processes were developed to read typewritten, hand-written or mark targets.

To this end, there are a number of pros and cons between in-line vs. off-line recognition and ways in which one can complement the other.

In-Line (Real-Time) Recognition has been an integral part of OCR/ICR recognition since the early 1960's. The following are a number of features that were standard in processing OCR/ICR data in-line:

- Image Enhancements
- Image Deskew
- Cropping
- De-speckle
- Grayscale
- Recognition Speeds
- OCR @ 10,000 char/sec
- ICR @ 7,500 char/sec
- Mark Sense @ 10,000 char/sec
- System Functions
- Batch and Transaction integrity handled at time of capture
- Document identification at the time of capture
- Intelligent Ink Jet Serializing for Audit Trail
- Outsorting of Documents based upon data recognition
- Process: Single Scanning, Image Enhancement, Form ID, Recognition, Output, and Sorting Process

The Off-Line Recognition consists of those features that required additional considerations:

- Image Enhancement
- Separate image enhancement process before recognition can be applied
- Recognition
- Typically recognition is done from black and white image (bi-tonal)
- Recognition speeds in the hundreds of char/sec
- System Functions
- Can't handle Batch or Transaction integrity
- Non-intelligent Ink Jet Serialization

- Can't sort documents based upon recognized data
- Additional document prep may be required to separate transactions, i.e., separator sheets (Patch Codes)
- Multiple processes required for Scanning, Image Enhancement, Form ID, Recognition, and Output

Combined Features

It is now possible to combine the features of both the In-Line Recognition and Off-Line Recognition. End users would find improvements in read rates and recognition accuracy for both OCR and ICR data. This would be accomplished in the correction of non-read characters (rejects); in the use of Voting Recognition Engine techniques; and in the use of applying Context Editing Rules.

Correction of Non-read Characters

The combination of features from In-Line recognition and Off-Line recognition can be used to correct non-read characters using a voting algorithm. The In-Line recognition can provide the non-read ASCII character value and that can be compared to the Off-Line recognition ASCII character value. If they match, then it can be corrected without operator intervention or entry. Additional voting methods are mentioned below.

Voting Engines

Voting Engines have been developed over the past few years to provide for improved recognition accuracy by using different recognition techniques (matrix matching, contour tracing, neural networks), and then voting the results of each technique. The following are some of the voting techniques and an example how each can be utilized.

- Safe: The result of the vote must be unanimous.
- Normal: The voting is unanimous if there is no conflict. Only engines that return a result are counted.
- Majority: This is a simple majority vote. If there is no "winner", the result s rejected.
- Order: The first engine (according to the order of engines) that is above the confidence threshold determines the result.
- Equalizer: The result is normalized by an algorithm, which uses the value of the engine results according to confidence levels.

The following is an example of each voting technique:

Engine	Result
1	25***8
2	2*5378
3	253478
4	2*34*8

Voting Method	Results
Safe	2****8
Normal	25**78
Majority	253478
Order	255378
Equalizer	253478

Context Editing

Another method used to provide verification, business rules, and unattended correction of data is called Context Editing. Context Editing can provide the following automatic (unattended correction):

- Names
- First Names
- Last Names
- First & Last Names
- Addresses
- Street Addresses
- City, State, and Zip Codes

Each Context Edit field(s) contains a dictionary Table of all possible ASCII values, and confidence (recognition) values, plus field processing and edits rules. Software logic uses the 1st, 2nd, 3rd recognition choices for each character and compares all combination to the dictionary Tables. First and/or Last Name (as single field, or two separate fields) use these dictionary Tables to automatically non-read characters or substitution characters. The United States City, State, and Zip (3 fields are combined) to do the similar unattended correction and then the United States Streets Addresses uses the results form the City, State, and Zip correction.

For the Context Edit process, Scan-Optics has produced a First Name dictionary containing 14,000 first names covering 97% of the U.S. population. The Last Name dictionary contains 133,000 last names covering 94% of the U.S. population.

The Context Edit for processing Street Address contains 13 million entries from the U.S. Postal Service Database. Each street is analyzed according to the U. S. Postal Service standards. The street number and street name are compared to entries in the directory, and are automatically matched to the city, state, & zip combination to correct OCR/ICR rejects and substitutions in the street address. Again, this is an unattended process.

Again using the U. S. Postal Database, which contains 76,000 City, State, and Zip entries, the Context Edit process can correct OCR/ICR rejects and substitutions without operator intervention.

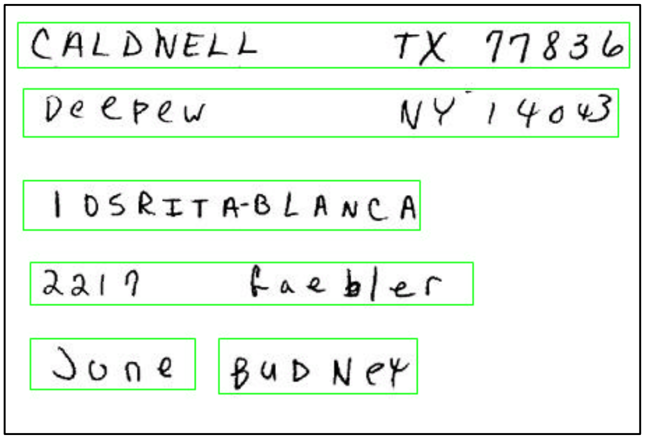
This greatly improves both the machine-print and handprint field recognition results.

Other Tables are:

- Canadian Names
- Canadian Postal Codes
- Ship Names
- Hotel Names
- Airline Names
- Combined US+Canadian Tables

Any "sparse" table (where the set of all possible combinations includes more illegal then legal values) can successfully be used to utilize the Context Edit processing.

Sample of Context Edit



The first example fixes the letter "W" in CALDWELL.

CALDWELL TX ***not CALDNELL***

The second example removes the extra letter "E" in DEPEW.

DEPEW NY 14043 ***not DEEPEW***

The third example separates the street number from the rest of the address name.

105 RITA-BLANCA ***not 10 SRITA-BLANCA***

The fourth example corrects the street name.

2217 GAEBLER ***not 2217 Faebler***

The final example fixes a person's last name.

BUDNEY ***not BUDNET***

Practical Applications

There are many applications that can utilize OCR technology for the reading of handwritten and machine printed information, plus optical mark targets. The following is just a sample list of applications:

- Accounts Receivable
- Airline Tickets
- Cash Resister Tapes

- Check Reconciliation
- Claims Handling
- Credit Card Sales Slips
- Customer Billing
- Driver's License Renewals and Registrations
- Insurance Premium Payments
- Inventory
- Gas, Electrical, Water Meter Reading
- Mortgage Payments
- Order Entry
- Payroll Processing
- Proxy Processing
- Questionnaires, Surveys, and Score Testing
- Retail Sales Slips
- Route Delivers (Proof of Delivery)
- Stock Transfer
- Subscription Renewals
- Tax Forms
- Time Cards
- Waybills

Appendix A

Terminology

ADF - Automatic Document Feed

Auto-Cropping - After deskewing a document, hardware or software will automatically crop the scanned image to the actual size of the document

Barcode - Typically, horizontal bars that are black and white with different widths representing alphanumeric data. Typical barcodes are 3 of 9, 2 of 5, Codabar, 2D, etc.

Bi-tonal Image - An image that contains black and white pixels

CCD - An acronym for Charge Coupled Device

Compression - The use of mathematical expression to reduce the size of a captured image. There are a number of compression algorithms, i.e., CCITT G3/4 for Bi-tonal images, JBIG for grayscale images, and JPEG for color images

Cropping - Creating an image portion from actual size of the document

Deskew - Either through hardware or software, the ability to correct (no skew) an image that a scanner had been skewed during feeding and image capture

Dots per Inch - The capture camera resolution of a scanner, or fax device, and output capability of a printing device

Drop out boxes - Non-read boxes used for hand printed characters.

Duplex Scanning - Scanning the front-side and back-side of a document

Filters - A special colored glass (blue, red, and Infra-red) that covers the camera lens, or filters can be done electronically if images are capture in color.

Font - Style or shape of printed information

Grayscale - Shade of black using in capturing an image. Grayscale can be expressed in 3-bit, 4-bit, 8-bit or 16-bit.

Handprint - Hand written characters ranging from 0-9 and A-Z, including the + and – characters scan-optics, inc.

Scanning Recognition: A Practical Overview

ICR - An acronym for Intelligent Character Recognition

IICR - An acronym for Imaging and Intelligent Character Recognition

Image - Electronic digital reproduction of a document

Inks - Visual Aid for printed media, instructions, and guides

IPM - Images per minute

ISIS - Industry Standard Interface Specification used as a scanner software interface

JBIG - Joint Bi-level Image Experts Group grayscale image compression

JPEG - Joint Photographic Expert Group color image compression

Landscape mode - Scanner document orientation where the document width is larger than the document length

Machine print - Characters generated by a printer and/or typewriter.

Mark sense targets - An enclosed shape such as a circle, oval or square that is used to define the presence or absence of a mark

Noise - Small particle or dot that appear on a document that will be imaged by a scanner

Paper - Media for inputting or capturing information.

Patch Code - The Patch Code is a very simple barcode consisting of two types of long horizontal bars with a minimum length of 2.0 inches, and a maximum length equal to the document width. The characters of the Patch Code consist of 4 bars, vertically spaced by 0.08 inches. Patch codes are used to separate and/or index documents or transactions.

PPM - Pages per minute

Pixel - The small component of a scanner camera for imaging a document. This is usually defined as one pixel equals one dot.

OCR - An acronym for Optical Character Read (originally, Optical Mark Reading)

Portrait mode - Scanner document orientation where the document length is larger than the document width

Reference Mark - A dark mark printed on the document to provide accurate frame of reference for locating fields on the document

Resolution - This is camera imaging capability of a scanner usually defined in pixels per inch or dots per inch

Scanner - The electro-mechanical device that transports, images, and stacks documents

SCSI - Small Computer System Interface used as a scanner hardware interface

Simplex Scanning - Scanning the front-side of a document

Skew - A scanner sometimes creates skew during feeding and image capture. Skew is related to the distortion of an image

Threshold - Scanners that capture in grayscale provide black and white output. The threshold is the grayscale point where a pixel or dot is considered either black or white

TWAIN - A scanner interface used by a scanner for document imaging

Scan-Optics, Inc. is a leader in applying technology to high-speed imaging, recognition, data capture, and archive and retrieval solutions. The growth of the Company's product line and the diversification of its services since its incorporation in 1968 reflects Scan-Optics' ability to respond with innovation and technical expertise to the rapidly changing business requirements of our customers. Scan-Optics' ability to offer customized and integrated system solutions has helped companies all over the world meet their productivity and profitability objectives. Offices and service representatives to support our products are located throughout the United States, and supplemented worldwide by select distributors in over forty countries.

Corporate Headquarters
169 Progress Drive • Manchester, CT 06040
Tel: 800-745-6001 • Fax: 860-645-7995

European Office
Unit 5 • Brookside • Colne Way, Watford
Hertfordshire WD247QJ • UK • Tel: 44-1923-819581
Fax: 44-1923-212633

www.scanoptics.com