6. Optical Character Recognition (OCR) Technology

6.1 Overview

What is OCR

OCR (Optical Character Recognition) also called Optical Character Reader is a system that provides a full alphanumeric recognition of printed or handwritten characters at electronic speed by simply scanning the form. More recently, the term Intelligent Character Recognition (ICR) has been used to describe the process of interpreting image data, in particular alphanumeric text.

Function of OCR

Forms containing characters images can be scanned through scanner and then recognition engine of the OCR system interpret the images and turn images of handwritten or printed characters into ASCII data (machine-readable characters).

Therefore, OCR allows users to quickly automate data capture from forms, eliminate keystrokes to reduce data entry costs and still maintain the high level of accuracy required in forms processing applications.

Features of OCR

The technology provides a complete form processing and +documents capture solution. Usually, OCR uses a modular architecture that is open, scaleable and workflow controlled. It includes forms definition, scanning, image pre-processing, and recognition capabilities.

What is ICR

Intelligent Character Recognition (ICR) is the module of OCR that has the ability to turn images of hand written or printed characters into ASCII data. Sometimes OCR is known as ICR.

The Different of ICR, OCR and OMR

ICR and OCR are recognition engines used with imaging; while OMR is a data collection technology that does not require a recognition engine. Therefore, basically OMR can not recognize hand-printed or machine-printed characters. However, in the OCR technology, answer for question in “tick” or “mark” is also known as OCR.

Comparisons Between OMR and OCR/ ICR

1. Optical Mark Reader (OMR)

a. Forms

An OMR works with specialized document and contains timing tracks along one edge of the form to indicate scanner where to read for marks. It also contains form ID

marks, which look like black boxes on the top or bottom of a form. The cut of the form is very precise and the bubbles on a form must be located in the same location on every form.

b. Storage

With OMR, the image of a document is not scanned and stored.

c. Accuracy

Considering that OMR is simpler than OCR, and if the forms and the system is designed properly, then OMR has accuracy more than of OCR.

2. OCR/ ICR

a. Forms

OCR/ ICR is more flexible since no timing tracks or block like form IDs required. In addition, the image can float on a page. The ICR/ OCR technology uses registration mark on the four-corners of a document, in the recognition of an image. Respondents place one character per box on this form.

The color is very important because the use of drop color reduces the size of the scanner's output and enhances the accuracy.

b. Storage/ retrieval

If document needs to be electronically stored and maintained, then OCR/ ICR is needed. Because with OCR/ ICR technologies, images can be scanned, indexed, and written to optical media.

Comparison Table

Items	OCR/ICR	OMR
Handprint recognition	Y	N
Machine print recognition	Y	N
Recognition of checks and "X"s	Y	Y
Requires timing tracks/ form IDs	N	Y
Requires registration marks	Y	N
Electronic image storage and retrieval	Y	N

6.2 System Requirement

Hardware Requirement

PC Computers with minimum capacity:
Processor: Pentium 200 MHz
RAM: 32 MB
Disk: 4 GB
Form modules are designed to operate in a batch processing, run under LAN and PC based platforms and take full advantage of the graphical user interface and 32 bit processing power available with Windows 95.

Scanner Requirement

OCR scanners with minimum capacity:
Duplex scanning
Speed: 60 sheets/ min
Automatic Document Feeder (ADF): Scanning can take a significant amount, and the system lets user scan up without doing the OCR.

Software Requirement

Software:
OCR with ICR capability software
Questionnaire Design Software

6.3 Advantages and Disadvantages

Advantage Using Images Rather Than Paper Questionnaires

- Theoretically imaging deals with no paper works that leads NSO for not carrying of questionnaires to and from the workstation; clear desks; quicker processing; and no storage of questionnaires near operators.
- There were significant savings in costs and efficiencies by not having the paper questionnaires
- The scanning and recognition of questionnaires allowed us to efficiently manage and plan the rest of the processing workload. Once the questionnaire are recognized we knew how much work (repair work, edit failures) we have to do.
- Reduced long term storage requirements because questionnaires could be destroyed after the initial scanning, recognition and repair.

When performing editing, which required some documents to be reprocessed, the electronic questionnaires could be found very quickly and sent back for reprocessing. This would be very time consuming if the questionnaires had to be located from a physical storage area and they are randomly scattered throughout. The more often the paper questionnaires are moved from the storage area the more

likely they will not be put back in the correct place

Disadvantages

- **Accuracy**

  While OCR technology can be effective in converting handwritten or typed characters, it does not give as high accuracy as of OMR for reading data, where users are actually marking forms.

- **Additional workload to data collectors**

  OCR has severe limitations when it comes to human handwriting. Characters must be hand-printed with separate characters in boxes.

6.4 Operation and Management

OCR Process Stages

a. Document Scanning process

Scanning speed will be determined by the quality of the scanner machines, the size of non-drop out color. Paper quality, cleanness, weights, Right setting of the OCR system.

b. Recognizing process

The recognizing process is to interpret images. The right memory (dictionary) and the configuration threshold will determine the accuracy of interpretation of the ICR.

c. Verifying Process

To compare the value of the interpreted image with the real image of the form.

Processing Order

Processing can be in geographic order or in random order. That is, we could scan questionnaires as soon as they arrive and then hold them to be processed regionally or allow them to be processed in some other order

Image Manipulation

- Electronic questionnaires can be sent to specialist operators then back to the original operator if necessary
- With electronic questionnaires the same questionnaire can be worked on simultaneously by two or more persons
- Electronic questionnaires are readily available for post census analysis (easier access to questionnaires)
- Parts of various questionnaires on screen at once for inter record editing
- Can view the relevant field book entry on screen in conjunction with questionnaires. Helpful for coding

	and editing
Coding Assistance	• The problem are needing operator attention can be highlighted to make it easier for the operator to identify • Only the questions relevant to the coding or editing problem were shown on screen although all other questions and questionnaires for that dwelling are available top the operator. This is particularly useful when editing between questionnaires. • Can use images of questions that will not be captured (scanned but not recognized) to help the coding process. For example, light pencil. In addition the operator can magnify images so that characters not discernible to the naked eye can be read. This lead to better data quality.
	The appropriate software ensures that the data is validated as the forms are read. Each form is checked to ensure that it has been fed in the right way round and not skewed. There is also a check that ensures that all the selections on a form that should have been filled in have been and that two marks have not been entered where only one mark should be. It is also possible to distinguish between intended marks and marks that have been erased.
OMR Scanner Speed	There are many factors that determine the throughput of an application such as number of characters per page, number of different document types, and legibility of handwriting that will affect the throughput
Skew	Each document is moved from an automatic feeder into a scanner and angel of skew is sometimes introduced.
De-skew	Analyze the image bit-map, calculates and returns the angle of skew up to +/-25. Example. De-skew often refer to %, which is the pixel shift. 10% is a 20-pixel shift in a line of 200 pixels or one tenth of an inch in an inch long line.
Landscape Detection and Auto Rotation	When batches of paper are scanned using an automatic feeder all the pages are fed in the same direction, landscape detection will automatically detect and rotate appropriate images 90 degrees.
White Page Detection	Normally, a double-sided scanner creates two images per scanners page. However, if the back or front page is blank, there is no need to store this image. White page detection allows the user to avoid storing blank page.

Automatic Image Registration

The image registration module calculates registration coordinates for a zone within an image. These calculated values could then be compared to the expected registration point values and image position correction phase initiated.

De-Speckle and Shade Removal

Sometimes scanners pick up colors and shades creating unwanted black dots in the image. The image enhancement module removes unwanted dots from the image reducing storage requirements up to 50% and substantially improving OCR.

Character Enhancer

This module enhances or decreases the darkness of printed characters, because the accuracy of OCR is less if characters are too fat. The character enhancer enhances the characters by filling-in missing dots, or removing extraneous dot. For example, dotted character formed by matrix printer can be corrected.

Cost Savings

- Estimated that imaging saved up to 2% of the total cost of the census compared to using paper document
- Staffing levels were reduced. In 1991 there were 70 data entry operators employed to do capture and 50 coding operators working one shift per day for 5 days per week.

Automatic processes to improve recognition rates

Recognition engines provide raw probability tables of the confidence with which a character has been recognized. Additional automatic processes can be used to improve the recognition rates, for example:

- Contextual editing and correction (e.g. state, dates where there is knowledge of the expected date period.)
- Dictionary look-ups
- Adaptive matching - comparing results for one field against other fields
- Trigram analysis
- Language techniques (e.g. for most English language questions, where the string "QV" is recognized, it would assume that it should be “QU”.
- Multiple recognition engines and Voting

Voting techniques

There are different methods of using Voting techniques to improve recognition rates. These include the use of:

- different engines with different capabilities
- the same engine with different settings
- the same engine with different character sets

Multiple engines

Multiple (two or more) recognition engines can be used to improve the recognition rate, reduce the repair load and reduce substitution errors. The benefit of using multiple engines depends on the quality of recognition achieved with the existing engine. If the recognition engine used already provides a high recognition level with a low substitution error rate, there might not be much benefit from using multiple engines and the cost of the additional engine(s) would need to be taken into account.

Using "voting" among two (or more) engines in parallel to determine the most likely character can be useful for numerics where a high degree of confidence is required. Alternate recognition may be useful for alphas where a second recognition engine might identify those characters the first engine could not identify with the required level of confidence. Some recognition engines are better at recognizing numerics, others better at alphas, and different engines may be better at recognizing specific characters.

Learning

Some ICR software may provide the facility to enable "learning", that is, the provision of on-line reports of recognition accuracy to provide immediate evaluation of voting results to tune and improve the recognition results. Neural networks may also be used in the learning process

A thorough testing program is necessary to determine whether multiple recognition engines and voting will improve recognition results and add to the efficiency of the ICR operation.

6.5 Questionnaire Design and Preparation

Drop Out Color

- Drop out color, usually red, is the color facility in OCR system that allows the system to pick up only the meaningful information from an OCR form.
- The system doesn't need to know the values including tick boxes written in the drop out color.
- The OCR system only needs to see the black parts, and compares them to specifications to see parts that are filled or written.

Characters or Marks?

- Considering the speed of the data capture process and to reduce rates, it is advisable to use marks or “ticks” as much as possible

How to Obtain Good Results of Scanning

- Select adequate paper quality
- A reliable printing press
- Appropriate ink, considering drop out color, for the questionnaires
- Use paper heavier than 80 grams per square meter can help avoid paper crashes or over read the other side of a single page

Form Design Advises

- Number items to be included in a form. The more items included in a form, the more information can one get, however, it could lead to ineffectiveness
- Design size of boxes for each character answer so it will not too small or to big
- Define drop out color properly
- Do not forget to use of registration marks
- Pre-print the codes near the place where the box for ticks are located
- Maintain consistent pattern in which the information to be collected will be located
- Do not disturb the visibility of the ticks and marks with titles, labels or instructions. This technique will also help the staff when manual editing done
- Avoid putting "answers" of one field to another page of the questions. Because if the enumerator change the position of the page many times, then the probability of paper to be folded will be high
- Avoid using open ended questions

6.6 OCR Field Operation

Training for Collection and Processing Staff

The intensive training covers subjects such as:

- Basic software, scanner operations, including installation and troubleshooting.
- Applications with emphasis on the development of custom applications including: configuring non-standard forms
- Pre-marking of forms, use of overprinting customize forms
- Processing of surveys
- Crating custom outputs file formats

Field Operation Supplies	Handling of the OCR documents should be highly careful. Therefore, survey enumerators should be accompanied with needed supplies such as a documents-bag, several sharp HB or BB pencils, good erasers, etc.
Reasons of Error-Reading of OCR	• Bad condition of the form because of dirt, folded, crumple, etc • Forms fed into OCR scanner are not straight (at an angle) • Forms are incompletely filled
Reduce Error-Reading of OCR	• Checking the questionnaires for completeness and consistencies • Preparation of own memory (dictionary) • Defining permissible margins of OCR reading errors
Particular Care in Writing Numbers or Alphabetic	• One box contains only one character • Characters should not extend outside designated boxes • Unnecessary lines of characters such as points, decorative strokes, hooks, etc. are prohibited • Strokes should not be ended with flourishes or extensions • All lines should be connected without breaks • All lines or dots should be pressed with the same pressure
Value Checking Steps	Verify that the information captured by OMR is the same with the questionnaire
Control for Blank	If the information is blank, what type of control must be taken. Type of control steps should be taken if the information image is partial or no information to assure the quality of generated files
Missing Questionnaire	Make sure that the entire questionnaires are scanned completely, no missing and no duplication as well. Therefore control procedures including to produce control tables to compare with manual work.

Case Study 10
OCR in the ABS

Optical Character Recognition (OCR) is currently used by ABS for a number of small data collections such as labor survey and social survey. Overall, the results have been encouraging, but in qualified and marginal way. The setup overheads of small survey have been significant, and social survey (disability) had a number of alphabetic fields that required coding and repair overheads were high than expected.

At current ABS cost-recovery rates the net cost of OCR compared with more traditional capture options varies from small (unquantified) gains to small (though significant) losses. Part of the problem has been the evolving nature of the OCR service, part has been the usual learning curves, and part has been the difficulty interfacing the OCR subsystems with the rest of our statistical processing infrastructure. All these factors can be expected to improve over time, and there is a number of interesting development in commercial OCR systems.

The Labor Statistics Center has used existing OCR facilities for two surveys (Major Labor Costs and Employment Earnings and Hours), and is considering using it in a number of other surveys. A recent report on their experience suggests that OCR is suitable for surveys with a large number of data items and form definitions as these produce saving during operation. It also suggests that forms should be designed (differently) for OCR. The movement to image based rather than paper based form storage was appreciated, and staff adapted to the new way of working well. However, cost saving are only expected once thing have settled down.

In the 2001 Population Census, ABS has decided to replace OMR-based system to the OCR-based system. They are currently negotiating with suppliers for OCR equipment. They also expect to use a combination of OCR and Automatic coding to reduce both publication time frames and processing costs, and expect to replace paper distribution with image distribution thereby reducing paper handling overheads.

For more information called: Dr. Rob Edmondson, Director, Technology Services Division, Client Relation Manager for Population Statistics Group and Methodology Division, Australian Bureau of Statistics, email: rob.edmondson@abs.gov.au

Case Study 11
OCR in BPS-Statistics Indonesia

BPS-Statistics Indonesia (BPS) has started utilizing the imaging technology since 1971 for processing the 1971 Population Census and the 1973 Agriculture Census. The imaging technology, OMR technology at that time, however, was abandoned. There are several reasons why BPS had to abandon the technology, among others, the high quality of paper requirement, the high printing capacity, and also because the innovation of the data entry machines using personal computers have made the OMR technology is not worth to perform.

Population census was conducted by BPS in every 10 years. Even though population census has been conducted four times (in 1961, 1971, 1980, and 1990), however BPS-Statistics Indonesia was unable published the detail information up to small statistics area (village level). The detail information could only publish up to province and district levels.

In the next Population Census year 2000 (new millennium) BPS-Statistics Indonesia plan collect detail information up to small statistics area, 15 information items in the questionnaire should be filled out by the individual of population in Indonesia.
The decision to make small statistics area is realized, then data capture technology should be chosen that could support the system. For that reason BPS has determined to utilize OCR technology, based on the consideration that all the constraints that faced by the OMR technology could be overcome. Therefore, an intensive study has been carried to see how far the system can be applied.

In this study, an OCR system has been procured under the assistance of JICA, Japan. In this system, OCR software has been applied, i.e, NCS Nestor Reader and OCR. Scanner Fujitsu was used for scanning the questionnaire. Several pilots have been carried out in Bogor District, Tangerang District, Bekasi District, Cianjur District and East Java Province. From this study found that the acceptance rate of mark was very high, numeric was around 90 per cent and alphabetic was very low. This is due to high variation of character type written by the interviewers. Therefore, this study concluded that in the 2000 Population Census, only mark and numeric character would be applied. And after the intensive study we come up with a decision that the 2000 Population Census will be taken cared using OCR Technology.

For carrying out the 2000 Population Census, about 80 OCR systems are planned to be installed throughout the country. The system consists of Duplex Scanner, Computer with CD writer and software. Fortunately, an overseas donor supports the idea and wishes to give grant for provisioning the systems.

For more information called: Dr. Sihar Lumbantobing, Director, Statistics and Computer Training Center, BPS-Statistics Indonesia, email: sihar@mailhost.bps.go.id, http://www.bps.go.id