OPTICAL CHARACTER RECOGNITION IN INFORMATION PROCESSING (RECONNAISSANCE OPTIQUE DES CARACTERES IMPRIMEES INFORMATIQUE RESUMES)

Ching Y. Suen
Sir George Williams University
Montreal 107, Quebec

ABSTRACT

The use of optical character recognition (OCR) speeds up the flow of processing printed information. This paper considers the application of OCR to the direct scanning of documents. Its characteristics and capabilities are examined. Factors affecting the recognition rate and methods of improving the performance are studied. The recognition of constrained handprints and the human factors involved are discussed. (L'utilisation de la méthode de reconnaissance optique des caractères(ROC) acceleré la procédure des données imprimées. Cette étude montre les applications de la méthode ROC dans le déchiffrage direct des documents. Les caracteristiques capacitées de cette méthode ainsi que les facteurs affectant la vitesse de reconnaissance et les méthodes d'amélioration de la performance y sont étudiés. Aussi la reconnaissance d'empreintes manuelles contraintes et les facteurs humain concernés y sont éxaminés également.)

INTRODUCTION

Optical character recognition (OCR) is the conversion of printed or written information into computer readable codes for subsequent processing. Its development stemmed mainly from the rapidly increasing volume of data to be processed by the computer.

The first OCR machine appeared in 1954. Despite their large number of potential applications, optical readers did not attract widespread interest in the nineteen sixties. High costs and limited capabilities were the two chief reasons. Most of these machines cost between $150,000 and $1,500,000 and could read only one or two specific character fonts. Acceptance of these earlier models of optical character readers was further hampered by the availability of keypunch machines and the ever-increasing number of keypunch operators.

Fortunately, technological advances in recognition techniques and digital electronic equipment in the past few years have cut down the cost of OCR components considerably and at the same time made possible an increase in OCR capacity. Most modern OCR machines can read at a high speed several common typewritten character fonts as well as some handprinted

symbols and numerals. As a result, the population of OCR machines has expanded rapidly. It has become one of the most promising techniques for overcoming the bottleneck of data input to the computer. Its applications include data capture of cash register tapes, sales slips, credit-card sales tickets, utility stubs and meter readings, bills and invoices, medicare claims, mail and a great variety of data records and commercial forms.

OPERATION OF THE OCR SYSTEM

A block diagram of the OCR system is shown in Fig. 1. Operation of the entire system is controlled by a console which consists of a teletype or a video display unit. Documents containing the required information are first fed into the document transport unit which in turn conveys each of them through the scanner. Flying spots and photocell arrays are the two most common types of optical scanners. The function of the recognition unit, that is, the digitization and recognition of characters, is based on the scanning of the contrasts in patterns of the printed characters against the paper background. The recognition process consists of the extraction of distinctive features of the character and matching them against tables containing the characteristics of a trained set of characters. Processed characters are classified into recognizable and unrecognizable groups; documents containing the latter will be directed into a reject stacker for subsequent amendment. In addition, numerous OCR systems are now equipped with editing facilities. Documents containing unrecognizable characters can be retrieved immediately and the operator can enter from a keyboard the correct data into the recognition unit. In some cases, a video display is used to show the processed characters on a screen. Unrecognized characters can thus be corrected on-line. Once a document has been read, the information it contained will be transferred into a storage device for further processing. Magnetic tapes, discs and paper printouts are the three most commonly used storage media.

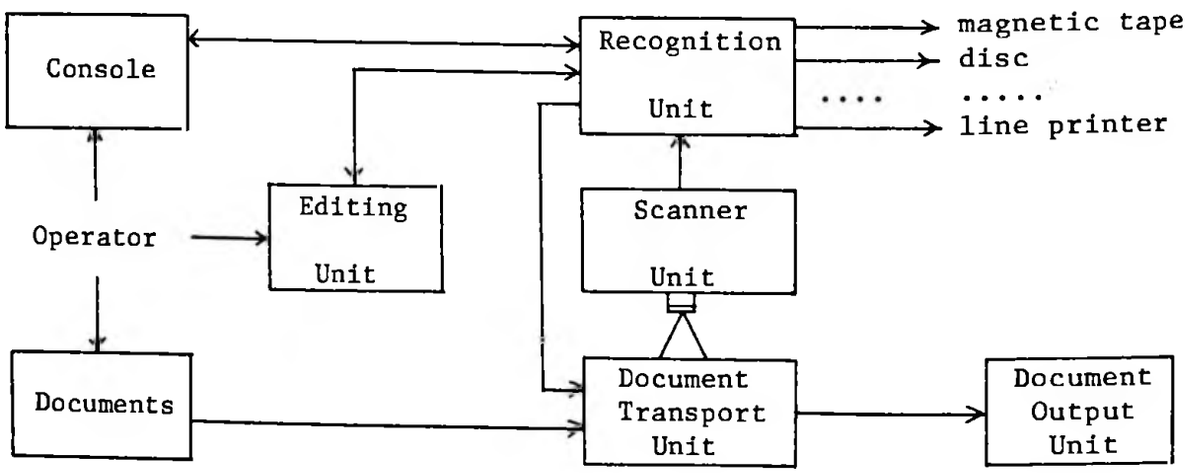

Fig. 1 Block diagram of the OCR system.

OPTICAL CHARACTER RECOGNITION

OCR CAPABILITIES AND CONSTRAINTS

Incorporating the recent advances in electronic technology and pattern recognition techniques, modern OCR machines can read a number of common typewritten fonts. To aid the manufacturers and the users, standdard OCR fonts have been developed; among these are the OCR-A font developed by the American National Standards Institute (1966) and the OCR-B font designed by the European Computer Manufacturers Association (1971). These standard fonts which have gained widespread acceptance, aim at achieving optimum machine readability while at the same time taking conventional character appearance into consideration. Style A emphasizes machine reading performance while style B emphasizes conventional appearance. Both styles include upper and lower case letters of the alphabet, numerals and special symbols. In addition to these standards, a number of fonts have been developed by individual manufacturers. Some machines in the market can read the OCR-A font only; others can read not only a variety of common machine printed fonts, but also handprinted numerals and a few special symbols. OCR machines which have only one reading head can process only one line of information at a time; those equipped with more than one reading head can process a corresponding number of lines at the same time. Error rates as low as 0.001 - 0.003% have been quoted by manufacturers. The average size of documents handled varies from 3.0" X 3.5" to 4" X 8"; some machines can even accommodate full size documents of 8.5" X 11.0". The rate at which documents can be processed depends on the amount of information carried, the design of the documents, the number of lines of characters and the spacing between them. The processing speed can vary from 20 up to 3,600 characters/sec. (Andersson, 1971).

Apart from its capacity of processing data, the optical reader can also be applied to read books and journals, digest papers, extract index terms and abstract articles. Here one faces more constraints than reading typewritten materials. First, the optical reader must be equipped with a dictionary having a large vocabulary of words. Second, it must have a multipurpose transport unit to handle the source materials properly. Third, it should be capable of reading virtually an unlimited number of fonts and characters of many different sizes. Microfilming is one way of alleviating some of the above problems and machines which can read microfilm materials are now available.

In order to facilitate recognition, documents must be prepared very carefully, and a number of constraints must be observed. The most important factor affecting the recognition rate is the formation of characters. To enhance the recognition rate, characters must be clean and sharp, have a uniform density and be free from voids and black-in spots. They must also have a low reflective index with sufficient contrast against the paper background. Similarly, the paper itself must be clean, smooth, opague and highly reflective. Also, paper stiffness and thickness

must be designed to maintain durability and to facilitate its conveyance by the transport unit. High-quality carbon ribbons are generally recommended. Typewriter keys must be clean. Apart from the above, the stroke width of characters, character separation and rotation, line and field separation, must all conform to specifications. Some manufacturers publish instruction booklets to help users in the design of documents to optimize OCR efficiency.

Type font variations do not generally present any difficulty to human character recognition. This is because understanding is facilitated by reading the characters in their context. In the case of OCR, machine recognition of characters depends solely on the analysis of their distinctive features. Although some machines have a learning capability to adapt to reading new fonts, the difficulty of recognizing the more than a hundred existing type font styles is self-evident. As a result, many machines are programmed to read only one or two type fonts. This places severe restrictions on many users since documents generally come from a variety of sources. A relatively high rejection rate can also be expected from documents which have undergone a large amount of handling due to the accumulation of dirt smudges and creases.

ECONOMIC CONSIDERATIONS

In determining the economic feasibility of employing OCR as a means of processing data, several factors must be taken into consideration. These are the volume of data, throughput speed, character fonts, document size, form design; and above all, its cost as compared with other methods of inputing the data.

Keypunching is probably the greatest rival of OCR in information processing and several reports have been published describing their relative merits, see for example, Andersson (1971), Tierney (1972) and Sheinberg (1968). These analyses indicate that OCR is economically feasible when the amount of data to be processed exceeds half a million cards. The cost for processing the above amount of data by the keypunch method is equivalent to the employment cost of about 9 to 12 full-time keypunch operators plus the rent for keypunch machines and their acessories. With increasing labour costs, rising volume of data and decreasing costs of digital equipment, the break-even point will be considerably lowered in the future. Since OCR minimizes manual intervention of the recorded data, it is therefore less susceptible to transcription errors.

The high efficiency of OCR has the advantages of quick throughput, short billing cycle and quick information and sale transfer. As a result, OCR has attracted a large number of users such as utility companies, banks, post offices, government agencies and other large enterprises. These organizations which handle large volumes of data have reported immense savings through the use of OCR.

OPTICAL CHARACTER RECOGNITION

RECOGNITION OF HANDPRINTED CHARACTERS

The ability of reading handprints is a desirable feature in OCR since it can eliminate the expensive and error-prone process of data transcription. This capacity is especially useful in the many circumstances where typed documentation is either impossible or impracticable, e.g. outdoor data collecting and meter reading. It is particularly useful in processing handwritten programs.

Although many OCR machines can read multifont and intermixed font characters, they can read only a few handprinted characters. Since each writer has his own style of writing, handprints are less tractable than multifont characters and recognition of the former must be based upon general character configurations rather than simple matching processes. Current recognition techniques have been described by Harmon (1972).

One way of facilitating handprint recognition is to develop a standard such that the person entering the data can write in a style as close as possible to the given standard. One such standard is now in preparation at the American National Standards Institute (1972). Some manufacturers have also produced their own set of stylized characters. This includes all the numerals and special characters C, S, T, X, Z, +, - and /.

To facilitate machine recognition and to aid the writer in printing characters suitable in shape and size, pre-printed boxes are used as guidelines. The user is instructed to write the characters neatly within these boxes with a sharp black-lead pencil (e.g. HB pencils). Pre-printed boxes have a pitch of 4 to 5 characters per inch, their sizes vary from 0.16" X 0.20" to 0.24" X 0.32". There is also a limit on the stroke width, normally between 0.01" and 0.04". The slopes of both vertical and horizontal lines should also conform to specifications. The handprints should resemble closely the pre-printed samples, otherwise a high rejection rate can be expected.

Because of the rigidity of these rules and the difficulty of changing one's established style of writing, the writer usually has to undergo some kind of training process before he can print characters acceptable to the machine.

Through motivation and patience in training, successful recognition of appropriate printing styles of numeric handprints can be achieved. This has been reported by a number of organizations in diverse applications such as payroll, accounting, driver's license registration and meter reading.

OPTICAL CHARACTER RECOGNITION

HUMAN FACTORS INVOLVED

OCR machines are very sensitive to print degradation and document mutilation. In order to cut down the error and rejection rates and to improve the overall performance, co-operation of the users is necessary. This is particularly so in handprint recognition.

The results of a handprint experiment carried out at Cognitronics Corporation (1972) revealed that character deformity was the chief reason for rejects. Other factors include interference between over-sized characters, improper erasures, stray marks, etc. It was also shown that a large proportion of rejects was generated from the handprints of a small number of tested writers. In an experiment recently conducted by Suen (1973), approximately 2,000 handprinted characters were collected and analyzed by an OCR reader. The results indicated that characters written by subjects with a light stroke yielded a recognition rate 4.7% lower than those written with a heavy stroke. Erased characters yielded a recognition rate about 10% lower than those without erasure. The above results suggest that more research should be done to investigate human behaviour in handwriting. In order to avoid ambiguity in man-to-machine communication, harmony between machine specifications and human handwriting practices should be reached.

OCR PROSPECTS

Many OCR machines have demonstrated capabilities markedly superior to earlier models. Improvement in recognition techniques and logic designs, accompanied by the reduction in cost of components, have stimulated the growth in the use of OCR readers. Although numerous human factors remain unsolved in OCR recognition of handprints, satisfactory training programs have been developed by both manufacturers and users, attesting to the progress in handprint recognition.

With the continuous effort aimed at character and document standardization, both nationally and internationally, it is expected that a powerful impetus will be given to the introduction of more low-cost, special — purpose OCR machines. There is little doubt that OCR will become one of the most prominent technologies in information processing in the next twenty years.

REFERENCES

American National Standards Institute, Inc., "USA Standard Character Set for Optical Character Recognition," Standard X3.17, (1966).

_____, "Proposed American National Standard Character Set for Handprinting," Document no. X3A1/72-60, (1972).

OPTICAL CHARACTER RECOGNITION

Andersson, P. L., 1971, "OCR Enters the Practical Stage," Datamation, v. 17, no. 23, p. 22-27.

Cognitronics Corporation, "Handprinted Digit Reading under Conditions of Limited Control," Nov. (1972).

European Computer Manufacturers Association, "Standard ECMA-11 for the Alphanumeric Character Set OCR-B for Optical Recognition," Oct. (1971).

Harmon, L. D., 1972, "Automatic Recognition of Print and Script," Proceedings of the IEEE, v. 60, no. 10, p. 1165-1176.

Sheinberg, I., 1968, "Optical Character Recognition for Information Management," in: Kanal, L. N., ed., Pattern Recognition: Washington, D. C., p. 31-40.

Suen, C. Y., 1973, "Factors Affecting the Recognition of Handprinted Characters," in preparation.

Tierney, D. F., 1972, "Traditional OCR Problems Passé in Today's Market," Computerworld,(Nov.), p. 8.