ECRM

JANUARY 1974

DOCUMENT M-101

5000 SERIES
AUTOREADER

INTRODUCTION TO OCR AND THE ECRM AUTOREADER

ECRM, INC. 205 BURLINGTON ROAD, BEDFORD, MASSACHUSETTS 01730

Telephone: (617) 275-1760 • Telegrams: AUTOREADER • Telex: 92-3349

NOTICE

INTRODUCTION TO OCR AND THE ECRM AUTOREADER

1.0 GENERAL

OCR, or Optical Character Recognition is an automatic means of entering human-readable information directly into a computer. It eliminates the need for keyboards and keyboard operators to translate typewritten information into computer code.

OCR technology is a natural complement to the computer industry's progress. The newer, larger computers of the 1960's had tremendous capacities for handling data. The cost of entering that data was staggering. Industry spent more than $3 billion on keypunch operator salaries in 1968 alone. OCR provided an accurate and cost effective alternative to both small and large companies. Data entry costs have been reduced as much as 90 percent with OCR and companies with as few as three keypunch operators have economically justified conversion to OCR techniques.

While OCR was applied successfully and profitably in many industries, its benefits could not be applied to the type composition field. Existing OCR devices could read only "clean" copy with no edit marks.

2.0 BACKGROUND

ECRM developed the world's first practical OCR system for the graphic arts industries. Designed specifically for type composition applications, it overcame the clean copy limitation. Writers and editors could continue to alter copy much as they have always done. They could add or delete characters, words or whole sentences. They could correct misspelled words. They could add typesetting commands. They could cut and paste copy into pages up to four feet long. And they could do all of this directly on the copy that was to be scanned.

The first ECRM installation, in 1970, revolutionized the performance standard of the type composition industry. The ECRM Autoreader equaled the keyboarding speed of 20 keyboard operators at a fraction of their error rate. Most significantly, the installation investment was justified by effectively doing the work of 20 operators at the cost of three operators.

3.0 OPERATION

News copy, classified ads, manuscripts and other copy for typesetting are routinely typed on an electric typewriter with a carbon ribbon. No special typing skills or paper are needed. The only requirement is the use of a type font such as Courier 12 which is recognizable to the OCR system. Copy is double or triple spaced to allow room for normal editing on the copy itself. Editing marks can be made with a pen using colored ink that is invisible to the OCR scanning device. Deletions are made with a black pen so the changes can be read by the scanner. Handwritten corrections are then typed in between the lines on the edited copy.

Formats are also typed on the copy. For example, "SH" indicates a sub-head and will cause the object machine to set the line boldface centered. The indent defines a paragraph and causes the proper codes (quad prior line and indent) to be inserted for all subsequent paragraphs on the page. '

The copy is then simply placed in the input hopper of the Autoreader. The system reads the text. Quadding, upper rail, lower rail, and other function codes are automatically inserted in proper sequence. The Autoreader's output is then sent to the hyphenation and justification (H&J) processor (Object Machine). This Object Machine may be a photocomposition machine which includes H&J capability. A paper tape or an optional wired interface from the Autoreader provides the input to a computer for the hyphenation and justification process. The justified tape is then ready for a linecaster or photocomposition machine.

4.0 TECHNICAL DESCRIPTION

4.1 *Scanning Technique.* **Optical scanning is accomplished through the ability of the Autoreader to sense varying degrees of reflected light. When there is sufficient contrast between the amount of light reflected from a typed character and the amount of light reflected from the surrounding background (blank paper), the Autoreader can detect and distinguish the shape of a character. ECRM's 5000 Series Autoreaders combine a laser beam optical system with the latest digital computer technology. These systems have few limitations which allows various typefaces, including foreign languages, to be recognized. The principle behind ECRM's recognition technique is a tiny beam of light from a laser. Light reflected from the paper surface strikes a photodetector which generates a corresponding electrical signal. This process produces a modulated electrical signal corresponding to the black and white areas on the typewritten page.**

4.2 Video Buffer. **This signal is sampled to break-up the black and white areas into a series of "black and white dots". The dots represent digitized video or picture elements (PELS) with white corresponding to a binary 0 and black to a binary 1. The digitized video is stored in a magnetic core memory (Video Buffer). Figure 1 shows a representative sample of a lower case letter "a" stored in the Video Buffer. The Video Buffer "height" is larger than the height of the largest possible character and its length is equal to the width of the longest typed line.**

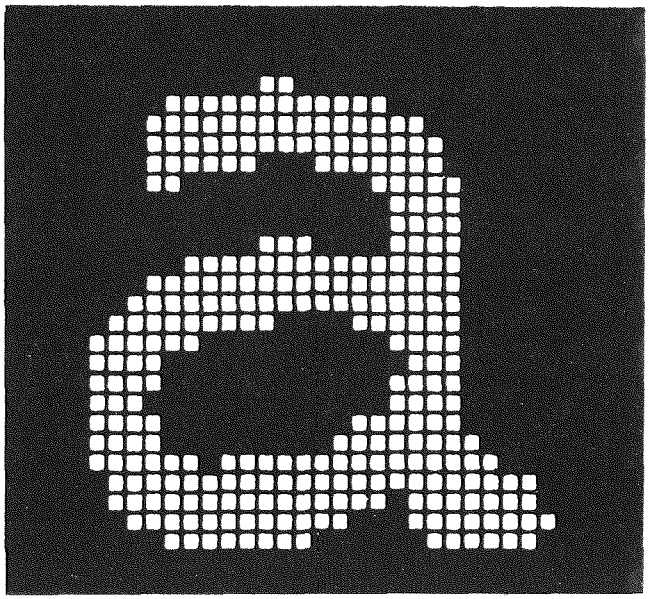

FIGURE 1. SAMPLE OF CHARACTER IN VIDEO BUFFER

4.3 Character Recognition. **Figure 2 shows the basic data flow from the Scanner to the output device. Video scan and storage are hardware operations using the DMA channel of a PDP-8 Computer. Scanning is accomplished by the laser beam which sweeps from left to right. At the end of each sweep, the stepping motor in the paper handling mechanism advances one step and the laser beam sweeps again. As the digitized video is stored in the Video Buffer, one scan line at a time, an Acquisition Scan software routine is initiated. When a character is detected, Acquisition Scan is terminated and a Boxing Routine is initiated. This establishes a rectangular box with North, South, East and West co-ordinates. Assuming sufficient lines have been scanned to detect the entire character, the box will circumscribe the character to be identified.**

Several measurements are made on the stored character image after it has been boxed. For example, two of these measurements are the height and width of the character. After the measurements are made, a table of allowed values is stored to determine the character which has the best fit. One way to visualize this process is to think in terms of a key and a set of locks. Each measurement corresponds to a tumbler on the lock. Imagine you are at the post office and that each mail box corresponds to one of the possible characters - an A or B or b etc. For simplicity, assume that the locks on each box are of a type used for bicycles where you set the combination on numbered wheels as indicated in Figure 3. The measurements made on the

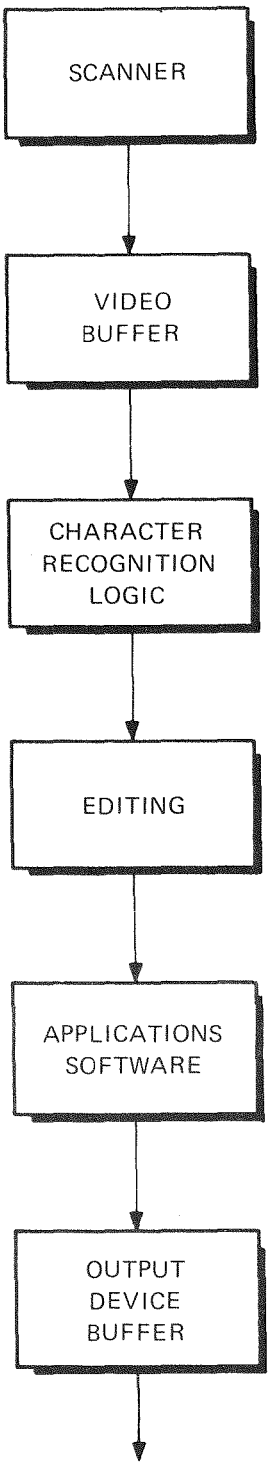

FIGURE 2 BASIC FLOW DIAGRAM

character image tell us where to set the wheels. In figure 3, the first measurement has the value 6, the second value is 2, etc. Each lock is tried to see if it will open. If one does, then we have our answer. If none of the locks open, then we force the locks with just enough pressure to break only one tumbler. In other words, we find an answer that matches in all except one. When all of the characters on a typed line have been recognized, the resulting line buffer is transferred to the Editing process.

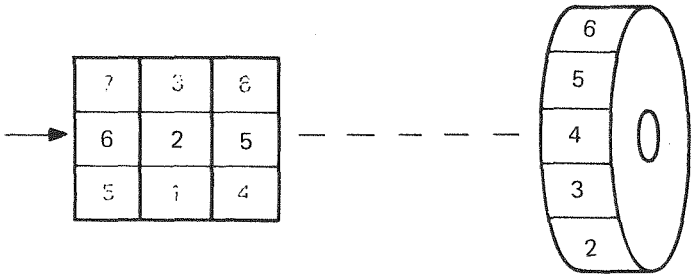

FIGURE 3 COMBINATION LOCK ANALOGY

4.4 Editing. **Editing operations, either character or word delete are performed as well as hand deletions and typewritten insertions. The completely edited line is then operated on by the Applications Software.**

4.5 Applications Software. **One complete line is operated on by the Applications Software. Typically, a line consists of a symbol stream of text characters plus typesetting commands. The Applications Software is divided into three modularized packages. That portion of the software which is common on all Customer/object machines is built into all Autoreaders. Customer dependent, non-machine dependent character and typesetting command definitions are provided in the BASIC-PREP software module. Machine dependent functions may be automatically implemented using the AUTO-PREP and/or AUTO-FORMS software modules. BASIC-PREP and AUTO-PREP recognize pre-defined, symbolic format commands and alphanumeric characters. The format**

commands are expanded as necessary by the Applications Software and the ASCII alphanumeric characters are converted to the appropriate output code (e.g. TTS). The new symbol stream (text and commands) is shifted into an Output Device Buffer and ultimately causes paper tape to be punched or other Customer defined output transmissions to be made.

4.6 Modularized Software. BASIC-PREP provides full alphanumeric conversion plus recognition and expansion of the more significant format commands such as START TAKE, END TAKE, PARAGRAPH START, etc. Upon recognition by the Applications Software, they are expanded to satisfy the object machine criteria. Most other format commands must be typed on the copy in their entirety when using BASIC-PREP. Characters are then output as they are recognized. AUTO-PREP provides recognition and expansion for over 50 common format commands. It simplifies copy preparation through use of mnemonic shorthand for format commands and character strings by allowing the object machine criteria to be replaced on the copy with brief mnemonics such as "BF" for BOLD FACE, "IL" for INDENT LEFT, etc. Recognition of these mnemonics by the Applications Software causes them to be expanded to satisfy object machine criteria. Depending on the current status of the object machine (bold face, light face, hanging indent, etc.) and common syntax built into the software, a format command may be expanded to produce 32 or more unique characters which cause the object machine to perform the required operation. AUTO-FORMS allows predefined formats to be recognized by the Autoreader. This applies to Classified Ad Forms as well as special purpose forms. Portions of a form may be specified as non-scan areas in which case the Autoreader skips over these areas at a faster than normal rate. Forms may be divided into a composition instruction area (CIA) and a text area to simplify copy preparation.

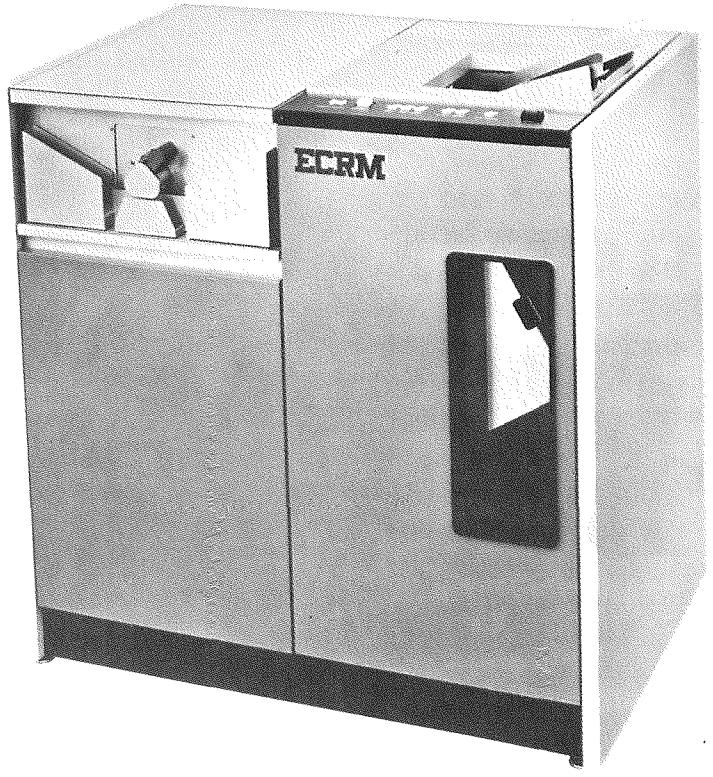

FIGURE 4 5000 SERIES AUTOREADER

INTRODUCTION TO OCR

AND THE

ECRM AUTOREADER

1.0 GENERAL

OCR, or Optical Character Recognition is an automatic means of entering human-readable information directly into a computer. It eliminates the need for keyboards and keyboard operators to translate typewritten information into computer code.

OCR technology is a natural complement to the computer industry's progress. The newer, larger computers of the 1960's had tremendous capacities for handling data. The cost of entering that data was staggering. Industry spent more than $3 billion on keypunch operator salaries in 1968 alone. OCR provided an accurate and cost effective alternative to both small and large companies. Data entry costs have been reduced as much as 90 percent with OCR and companies with as few as three keypunch operators have economically justified conversion to OCR techniques.

When a character is detected, Acquisition Scan is terminated and a Boxing Routine is initiated. This establishes a rectangular box with North, South, East and West co-ordinates. Assuming sufficient lines have been scanned to detect the entire character, the box will circumscribe the character to be identified.

Several measurements are made on the stored character image after it has been boxed. For example, two of these measurements are the height and width of the character. After the measurements are made, a table of allowed values is stored to determine the character which has the best fit. One way to visualize this process is to think in terms of a key and a set of locks. Each measurement corresponds to a tumbler on the lock. Imagine you are at the post office and that each mail box corresponds to one of the possible characters - an A or B or b / A or B or b / etc. For simplicity, assume that the locks on each box are of a type used for bicycles where you set the combination on numbered wheels as indicated in Figure 3 /3./

The measurements made on the character image tell us where to set the wheels. In /f/igure 1, the first measurement has the value 6, the second value is 2, etc. Each lock is tried to see if it will open. If one does, then we have our answer. If none of the locks open, then we force the locks with just enough pressure to break only /ΔIT one ΔLF/ tumbler. In other words, we find an answer that matches in all measurement/s/ except one. When all of the characters on a typed line have been recognized, the resulting line buffer is transferred to the Editing process.

ΔSL

4.4 Editing. /ΔIT / Editing operations, either character or word /ΔLF / delete are performed as well as hand deletions and typewritten insertions. The completely edited line is then operated on by the Applications Software.

ΔSL

4.6 /ΔIT / Modularized Software /ΔBF/ BASIC-PREP provides full alphanumeric conversion plus recognition and expansion of the more significant format commands such as START TAKE, END TAKE, PARAGRAPH START, etc. Upon recognition by the Applications Software, they are expanded to satisfy the object machine criteria. Most other /format / commands must be typed on the copy in their entirety when using BASIC-PREP. Characters are then output as they are recognized. AUTO-PREP provides recognition and expansion for over 50 common format commands. It simplifies copy preparation through the use of mnemonic shorthand for format commands and character strings by allowing the object machine criteria to be replaced on the copy with brief mnemonics such as /"BF"/ for BOLD FACE, /"IL"/ for INDENT LEFT, etc. Recognition of these mnemonics by the Applications Software causes them to be expanded to satisfy object machine criteria. Depending on the current status of the object machine (bold face, light face, hanging indent, etc.) and common syntax built into the software, a format command may be expanded to produce 32 or more unique characters which cause the object machine to perform the required operation. AUTO-FORMS allows pre-defined formats

SAMPLE INPUT PAGES FOR THIS DOCUMENT

(REDUCED)

ACKNOWLEDGEMENT

All copy appearing in this document was prepared and set using an ECRM 5100 Autoreader and a Mergenthaler V-I-P phototypesetter provided by:

CREATIVE COMPOSITION, INC.
ARLINGTON, MASSACHUSETTS