Technology White Paper

Understanding Optical Character Recognition

Overview of OCR and Its Applications

MICROSCAN®

Understanding Optical Character Recognition

Optical Character Recognition, commonly known as OCR, is distinct from linear and 2D symbologies in that it is simultaneously machine-readable and human-readable. OCR does not replace the more robust, secure options of linear and 2D symbologies. For traceability applications, 2D symbols offer the highest level of data security and reliability, and linear symbols offer an intermediate level of security and reliability. OCR alone offers the least. OCR is most effective when used to complement linear and 2D symbols. Topics of this white paper include:

- *A History of Optical Character Recognition Technology*
- *OCR vs. Bar Code Technology*
- *OCR Templates vs. "Teachable" OCR Systems and Optical Character Verification*

Microscan Systems, Inc.

A History of Optical Character Recognition Technology

Optical Character Recognition technology has been used extensively in commercial applications since the 1970s. In the early 1970s, a company in Dallas, Texas, called Recognition Equipment, Inc., developed a high-speed system for reading credit card receipts from gasoline purchases. At the time of the transaction, a receipt would be imprinted with the customer's account number (typically embossed on credit cards in OCR-A font). The merchant copy of the receipt would then be sent to a processing center where equipment provided by Recognition Equipment, Inc. would read the OCR-A account numbers on the receipts at document speeds of 45 to 55 feet per second.

By the late 1970s, OCR-B was being used on payment stubs for automated payment processing. Some utility companies, such as Southern California Gas, still use this system. Also in the late 1970s, Recognition Equipment, Inc. released a handheld OCR reader. The product was developed in response to the retail community's desire to switch from punch-hole price tags to price tags with OCR strings. Shortly after Recognition Equipment, Inc. introduced the first handheld OCR reader, Robert Noyce (co-founder of Fairchild Semiconductor and Intel) founded the Caere Corporation. Caere (later acquired by ScanSoft, Inc.) introduced its own handheld OCR reader in 1977. As a result of the new handheld OCR technology, several large retailers, including Sears, JCPenney, and Kmart converted to OCR-A price tags between 1980 and 1983. Sears and JCPenney alone purchased 50,000 readers. OCR-A was used on price tags until 1987, when the retail community selected UPC as their standard.

Since the mid-1980s, OCR technology has been adopted in a wide variety of applications, including remittance processing, passport processing, semiconductor manufacturing, automotive and aerospace manufacturing, secure document processing (checks, financial documents, bills), document handling, postal tracking, publishing, food packaging and consumer goods packaging (batch codes, lot codes, expiration dates), and clinical applications.

In remittance processing applications, individual payment stubs are printed with an OCR string, which is usually decoded by a fixed-mount reader in an automated environment. In passport processing, traveler information is encoded in two lines of OCR text. The use of OCR on passports was introduced in 1983 as part of an international convention. In 1984, Caere Corporation developed the first passport scanner for the U.S. State Department, and some are still in use today. The use of OCR on passports may diminish as biochips and other identification technologies gain greater currency in the coming decades.

Figure 1: *Examples of OCR used in a variety of applications*

Wafers and lead frames in semiconductor manufacturing applications are often marked with OCR strings. In automotive and aerospace manufacturing, OCR strings are placed on parts and sub-assemblies as direct part marks (laser etch, chemical etch, dot peen, etc.) In remittance processing applications, individual payment stubs are printed with an OCR string, which is usually decoded by a fixed-mount reader in an automated environment. In passport processing, traveler information is encoded in two lines of OCR text. The use of OCR on passports was introduced in 1983 as part of an international convention. In 1984, Caere Corporation developed the first passport scanner for the U.S. State Department, and some are still in use today. The use of OCR on passports may diminish as bio-chips and other identification technologies gain greater currency in the coming decades.

Wafers and lead frames in semiconductor manufacturing applications are often marked with OCR strings. In automotive and aerospace manufacturing, OCR strings are placed on parts and sub-assemblies as direct part marks (laser etch, chemical etch, dot peen, etc.)

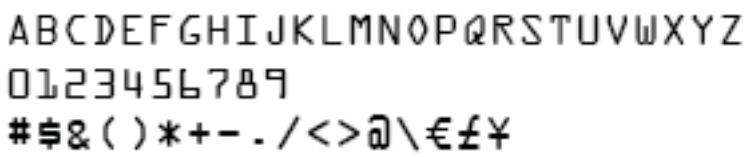
OCR vs. Bar Code Technology

OCR and bar code technology are both data capture methodologies, and each has advantages and disadvantages. The primary advantage of OCR is that it encodes information in a format that is simultaneously machine-readable and human-readable, while linear and 2D symbols are only machine-readable. Data encoded in an OCR string does not require a secondary machine-readable symbol. Data encoded in linear and 2D symbols is considerably more reliable, however. OCR has an inherently high rate of character substitution (particularly the OCR-A and OCR-B fonts)—not typically a concern when using linear and 2D symbols, which offer greater data integrity. Check characters are often embedded in OCR data fields and then calculated by OCR readers or vision systems to avoid substitution errors in data output. Many OCR readers have the ability to re-try the decode process a predetermined number of times, since substitution rates of as many as one of every 3,000 characters are expected in OCR applications. (See Using Checksums to Reinforce OCR Data Integrity.)

OCR Templates vs. "Teachable" OCR Systems and Optical Character Verification

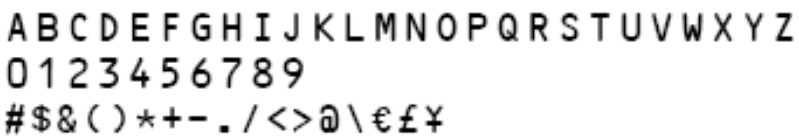
There are different ways to integrate OCR into an application, and different systems for processing OCR-encoded data. OCR templates and OCR fonts are the simplest and most reliable option. Examples of some common OCR fonts are shown below.

OCR-A



OCR-A is a relatively reliable font that supports an alphanumeric character set, along with some additional ASCII characters. It complies with the character shape, size, and printing position requirements for the ANSI INCITS 17-1981 (R2002) standard, which can be purchased from the ANSI website: http://webstore.ansi.org.

OCR-B



OCR-B is less reliable than OCR-A, but its less angular characters are generally considered to be more aesthetically pleasing. It complies with the character shape, size, and printing position requirements for the ANSI INCITS 49-1975 (R2002) standard, which can be purchased from the ANSI website: http://webstore.ansi.org.

MICR E-13B

MICR E-13B is used primarily in the banking industries of the U.S., Canada, Puerto Rico, the UK, and Panama. It is most commonly seen at the bottom of personal checks, where account information is encoded using magnetic ink (MICR is an abbreviation of "Magnetic Ink Character Recognition"). MICR E-13B complies with the character shape, size, and printing position requirements for the ANSI X9.27-2000 standard, which can be purchased from the ANSI website: http://webstore.ansi.org.

MICR CMC-7

MICR CMC-7 is used primarily in the banking industries of France, Spain, and most Latin American countries. It is most commonly seen at the bottom of personal checks, where account information is encoded using magnetic ink. MICR CMC-7 complies with the character shape, size, and printing position requirements for the ISO 1004:1995 standard, which can be purchased from the ISO website: http://www.iso.org.

SEMI M12

SEMI (Semiconductor Equipment and Materials International) is used for wafer and lead frame marking in the semi-conductor manufacturing industry. It complies with the character shape, size, and printing position requirements for the SEMI M12-0706 standard, which can be purchased from the SEMI website: http://www.semi.org.

OCR Templates

OCR templates define several parameters, including the OCR font that is used, layout of OCR text (in a row, in a column, etc.), the number of characters in a row of OCR text, the total number of rows, and the total number of characters in all the rows. Each character position in a row is specified as an ASCII value, a group of ASCII values, a wildcard character, or a combination of known ASCII values and wildcard characters. Limiting the variables of character type and character position as much as possible improves the reliability and efficiency of OCR applications.

Software-Configurable OCR Parameters

The task of optimizing OCR reliability and efficiency is vastly simplified by the current generation of configuration software. Many current providers of OCR technology, such as Microscan Systems, Inc., offer intuitive software interfaces to assist users in setting up OCR applications. See Figures 2 and 3 on the following page.

Teachable/Trainable OCR Systems and Optical Character Verification (OCV)

In contrast to the relative simplicity and reliability of OCR templates, “teachable” or “trainable” OCR systems are a technologically impressive but potentially unreliable option for OCR applications. Typically a feature of higher-end machine vision, teachable OCR systems can be trained to recognize characters in any user-defined font—not just fonts that are created specifically for optical character recognition (OCR-A, OCR-B, MICR, SEMI). OCR systems can be taught to recognize a full character set in any font created for any language. The disadvantages of this type of OCR system are the labor-intensive integration process and the decrease in reliability when using fonts that are not created specifically for OCR applications.

Optical Character Verification (OCV) is one way to address the problem of reliability in teachable OCR systems. Once the desired specifications have been taught to an OCR reader, OCV software can verify that characters are printed to match the user-defined specifications, can ensure that data is encoded correctly, and can guarantee that labels are placed in the correct orientations on the correct items.

Using Checksums to Reinforce OCR Data Integrity

The purpose of using checksums with OCR is to reduce the likelihood of character substitution errors. OCR systems commonly provide users with various checksum options. Checksum types (row or block), weight schemes, and modulo values are all ways of ensuring the correctness of data encoded in OCR strings.

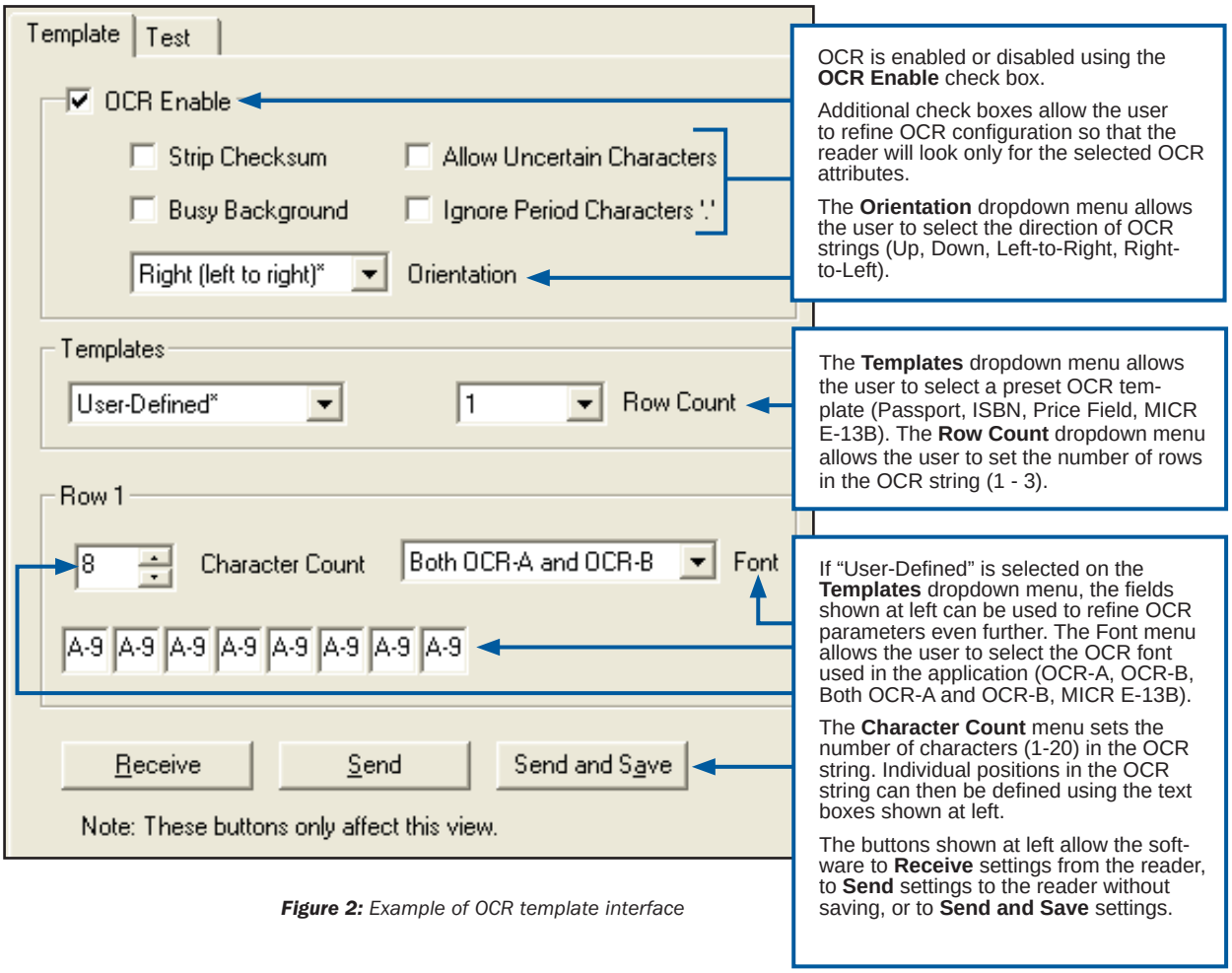


Figure 2: *Example of OCR template interface*

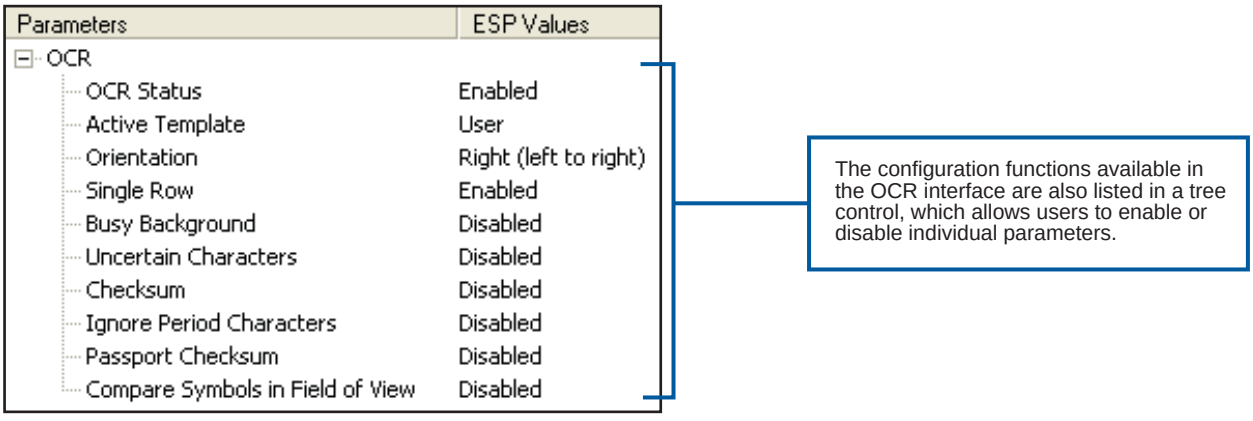


Figure 3: *Example of OCR configuration functions*

Conclusion

Although OCR was originally developed decades ago, it continues to be used in a broad range of application environments, and continues to be supported by a wide variety of products and systems—from high-end machine vision to more compact, easier-to-integrate solutions. Some OCR applications, particularly those using templates, can be fully supported by a lower cost OCR solution. Other applications may have a greater number of variables, requiring a more complex vision system. Thorough evaluation of all the attributes of the target application is necessary before choosing an OCR solution.

MICROSCAN®

www.microscan.com

North America (Corporate Headquarters)
Email: info@microscan.com

Europe
Email: emea@microscan.com

Asia Pacific
Email: asia@microscan.com