PARASCRIPT®

Strategic White Paper

Automated Handwriting Recognition – Not All ICR Software is Created Equal

Takeaways

In this paper you'll learn:

- How recognition works, accuracy, applications and benefits
- Differences between earlier ICR programs and more sophisticated, advanced ICR software that can read any type of handwriting, including unconstrained handprint and cursive
- About increased productivity and cost savings that result from a comprehensive automated recognition solution that includes the best attributes of OCR and ICR

Summary

OCR and ICR technologies deliver increased productivity and reduced costs for businesses, organizations and government agencies that process massive quantities of documents. OCR is an industry standard technology used to digitize machine text located in document images. ICR is a newer technology programmed to think like a human in order to process handwritten documents. This paper presents the technology and functionality of advanced ICR and its associated applications and benefits.

What's in this Paper?

6273 Monarch Park Place, Longmont, CO 80503 USA
T: 303.381.3100 | toll free: 888.225.0169
F: 303.381.3101 | info@parascript.com

Introduction

Most enterprises process huge volumes of forms, applications, checks and other documents. This data is submitted for electronic processing and the content is directly captured into a business system. Information on these documents can be completed by providing a mark, or fill-in-the-blank with machine print, handprint, handwriting (including cursive), or a combination of them all. Conversion to digital data is then required. The system of accurately capturing and recognizing machine text and handwritten data is known as recognition technology.

OCR – Optical Character Recognition software converts printed characters or machine text into digital files and has been in use since the 1960s. ICR - Intelligent Character Recognition (ICR) is a step above OCR. Standard ICR is best known as having the ability to read constrained handprint - printed characters in a form, box or other document limitations. What is not widely known is that ICR systems have evolved to read unconstrained handprint (constrained fields are no longer necessary), and any other type of handwriting. Attempts have been made to redefine this type of technology, but practically speaking people generally refer to all handwriting recognition as ICR. For the sake of this paper, we'll consider this as advanced ICR, which is far more complicated because it interprets the patterns of human writing and no two people ever write identical characters.

As noted standard ICR has limitations that have been significantly expanded by the broader capabilities of advanced ICR. It should also be noted that not all ICR software offer this next level of advanced sophisticated technology.

ICR offers substantial time and cost savings over current manual processes. Understanding how the process works, its many applications and level of accuracy provides confidence to adopt this technology. Document processors can increase their productivity by employing a comprehensive automated recognition software package that incorporates the best of OCR and ICR in a single solution.

Recognition Techniques

Manual Data Entry

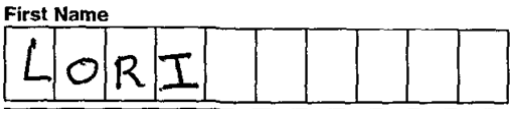
The most widely used form of recognition is still manual entry performed by data entry specialists. People read information from forms and enter it into computers. Manual data entry is still the most versatile of recognition techniques because it uses human brainpower to interpret hand printed and handwritten information. Data entry operators can interpret almost all normal irregularities, including characters that are either stylized or go outside of boxes in a form. Because of this operator adaptability, manual data entry is nearly as commonplace today as it has been over the last century.

The human brain is the reason why manual data entry is effective, but the human body is what leads to its shortcomings. Human accuracy levels are impacted by the extent of expertise, fatigue, mood, boredom, working conditions, and even the time of day. Productivity often declines and errors increase at the end of a day.

Automated Recognition Software

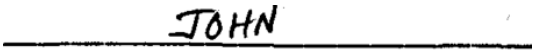
When considering the reality of increased labor costs coupled with increases in the volume of data to be processed, manual data entry has become cost prohibitive. As a result, OCR that reads machine text is an industry standard, yet has only been adopted by 32% of organizations[1]. Example:

ICR has been less widely adopted. ICR that is most familiar and in use now can only read constrained handprint - handwriting in a box or other defined area in a form. ICR for constrained hanprint is used by about 12% of organizations. Example:



However, ICR capabilities have improved dramatically over the past few years. In addition to constrained handprint, more recent advanced ICR programs include the ability to read additional handwriting variations. Advanced ICR has been adopted by only 6% of organizations.

Example 1 - Unconstrained handprint - handwriting not confined to a predetermined area in a document:



ICR Capabilities have improved dramatically over the past few years. Some advanced ICR programs can read uncontrained hanprint, including cursive.

1 AIIM White Paper. “Forms Processing - User experiences of text and handwriting recognition (OCR/ICR).” June, 2012. http://www.aiim.org/Research-and-Publications/Research/AIIM-White-Papers/Forms-Processing

Example 2 - Cursive handwriting:

Additionally, advanced ICR can read machine print when OCR fails, or in applications where a certain recognition rate is required (examples are checks, addresses, invoices, etc.)

When the capabilities and accuracy of this technology are fully understood, advanced ICR is a viable and cost effective alternative to manual operations.

Challenges of Advanced ICR

The challenge is to create advanced ICR software that reproduces a human's ability to instantly apply context and judgment in a mechanized process to:

- Interpret all forms of handwritten and printed documents
- Convert handwritten content into data records with a high level of accuracy
- Address issues of non standard elements such as writing outside form boxes
- Accommodate all of the irregularities of handwriting with a minimal amount of human intervention
- Meet needs by offering the flexibility of processing the work in-house or outsourcing it

Understanding Recognition Accuracy

Variables that impact recognition accuracy include image quality, character and word clarity, document design, and availability and type of context. The way that recognition rates and accuracy are derived differs with each of the recognition technologies and with the specific application of the technology.

Early automated recognition systems faced challenges with levels of accuracy. Performance success is based on the volume of information that is 100% correctly processed during a defined percentage of allotted time. Manual keying is the baseline for accuracy comparison. Human errors are caused by level of experience or expertise, fatigue, motivation, working conditions, quality and legibility of data and other elements that impact people, generally estimated at a low of 2% to a high of 6%.

While automated recognition technology avoids issues inherent to human operators, digital accuracy can also vary. Automation errors can be the result of an application design, quality of the images (writing can be too light, smudged, etc.), unconstrained print (word or characters that are out of defined area such as a box), noise, and other factors. It is important to fully understand the application and any limitations in order to understand overall performance.

Optical Character Recognition (OCR) has a long history of high levels of accuracy reading machine print Working with high-quality machine print, OCR provides nearly

99.9 % recognition accuracy. As a result, additional controls are not typically required for most OCR applications. However, it should be noted different machine fonts and image quality (i.e. faxes) could impact this high level of accuracy.

Intelligent Character Recognition (ICR) performs recognition for constrained handprint – printed characters in a form, box or other limiters that discourage the user from connecting characters and support character definition for the recognition engine. Standard ICR can provide reasonably good accuracy rates, however forces the end user into awkward information entry and is limited in its ability to refer to whole-word context.

Advanced ICR software can perform automated recognition on difficult machine print that may be encountered on forms, unconstrained handprint and cursive handwriting. Due to the virtually unlimited number of writing styles, it should be noted that ICR accuracy levels could be impacted by content quality:

ICR accuracy rates can be achieved up to and in excess of 90%.

- Words or characters located outside boxes or lines
- High noise levels and intrusions in the image
- Content that is illegible
- Character ambiguity – question regarding identity of a letter

Accuracy levels for more advanced handwriting recognition software have improved significantly due to recent technological improvements. Rates vary depending on the application, image quality, field types and context availability, among other factors but accuracy rates can be achieved up to and in excess of 90%.

How It Works - Understanding the Recognition Process

Benefits of ICR related to reduced time and costs are clear. Understanding how advanced ICR technology reads unconstrained text and cursive writing actually works may still present some questions.

The major challenge is the infinite variations in the way people make letters and numbers. In order to recognize unconstrained text and cursive writing, advanced ICR technologies have been programmed to behave more like humans and evaluate data based on its context. Recognition software reads a field from a document and provides a recognition result based on the interpretation of its individual characters, an entire word or a phrase. Similar to the way a person reads a document; the recognition engine creates a hypothesis about a word or number series, evaluates it against a database of possible answers (name, address, date of birth, social security number, etc.), and proves the answer that exists in the database, or adheres to certain standards (numeric, 9 digits in the case of a social security number).

Advanced ICR technology works by taking a “holistic” approach to recognition that combines highly sophisticated mathematical algorithms, neural network applications, image interpretation using a special description language, and content validation within context that happens during the recognition process.

The software identifies, coordinates and consolidates many variables in the process of accurately reading handwriting.

Field and Character Learning

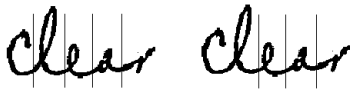
Because of its like-human recognition capabilities, advanced ICR identifies character shape definition. This capability is not inherent in OCR and standard ICR that reads only handprint. The handwriting recognition engine has the ability to:

- Employ advanced methods of single character recognition
- Recognize an entire field and accept or reject the entire field
- Use sophisticated algorithms to cross-validate results

The result is that advanced ICR technology can recognize poor quality or unreadable text that cannot be read by OCR or standard ICR technologies.

Character Identity

Letters A and B below are different from each other and are different from machine-printed letters. Advanced ICR software allows a computer to learn to interpret and evaluate written symbols similar to the human eye and brain.



Handwriting can have characters that are not clear and can result in confusion regarding the identity and meaning of a word. Characters can be ambiguous. In the words below, it is not obvious whether the first symbol is a “d” or a combination of a “c” and an “l.” Depending on the results of segmentation, the recognition result produced by ICR technology may be different, returning either “clear” or “dear.”

Standard ICRs rely on constraining elements that serve as templates or receptacles for hand-printed text. For this technology, print must be written in boxes to show a high degree of accuracy in their recognition of characters. Accuracy is also affected by the ability of humans to write within these constrained areas. These limitations offer even greater challenges when processing cursive handwriting and handprint without constraints.

Context Sensitivity

Context plays a significant role in the recognition process. Humans read handwriting or handprint by viewing entire words and sentences to correctly identify the content. Characters and words are read by understanding a range of probable meanings. Recognition engines use context as an effective and flexible tool to compensate for the inherent ambiguity of handwriting and to improve recognition accuracy.

This character could be considered the number "7" or the cursive letter "T." Through the use of context information such as identifying whether the field type is alpha or numeric, the recognition engine can accurately determine its correct identity. Context information can be specific such as a range of numbers, dates, or values; or as broad as a field type. Proper use of context information helps determine the correct result with a greater amount of confidence.

Recognizing Multiple Field Types

Address recognition is one example of the types of fields identified on a document or form. Depending on the software application, recognition engines can recognize a variety of field types. The most common field types are alpha, numeric, alphanumeric, amount, date-number, telephone number, telephone area code, unit of measurement, first name, last name, middle name, month, day, year, age, city-state-zip, number-street-apartment, and check box.

The recognition engine improves performance with identification of field properties. Fields can be configured to specify the number of digits, range of values, or specific vocabularies. A date field can be configured to specify a date format, a most frequently used date range, and a less frequently used date range.

The unique distinction of including the field context validation as an integral part of the recognition process makes the entire process more accurate, as well as more efficient and economical.

Database Cross-Validation

Database cross-validation also enhances accuracy and increases read rates. Common uses include matching ZIP codes with appropriate mailing addresses for address recognition, or verifying the numeric amount (i.e., $108.35) on a check with the alphanumeric amount (i.e., One hundred eight and 35/100) for check processing applications.

Application-Specific Databases and Customized Dictionary

Context vocabularies are used to customize words and phrases that are relevant to an industry, company, or customer base. This process increases the speed and accuracy of the recognition process since the vocabulary is more defined and the content volume is reduced.

Example of Using the Recognition Process

The scope and power of recognition technology is effectively illustrated in how it is used in an application process. A typical application can be seen in recognition of an address field on a standard warranty form. Like all recognition processes, address recognition occurs in several distinct stages:

1. The form document is scanned to generate a standard digital image format file – such as TIFF, JPEG, or PDF file.
2. The scanned document image is routed through form processing software that

identifies, de-skews and cleans up the form image.

3. The document image is analyzed to locate the appropriate field containing the address block information.
4. The address block is routed to the recognition engine for processing.
5. During the recognition process, the field is divided into a city-state-zip line and a street address line. Based on the interpretation rules for this specific application, the recognition engine processes the city, state, and zip code first.
6. This information is then used to verify the recognition of the street address.
7. During the initial recognition, the system identifies all of the possible combinations of city, state, and zip code and cross-validates them against the U.S. Postal database.
8. A confidence value is assigned to the highest answer. The confidence level is determined by two factors: the score of the second closest alternative and the distance between the two similarity scores. The same process occurs for street address recognition, with additional verification against the final city, state, and zip code.

Conclusion

Recent handwriting recognition technology advancements expand ICR functionality – with the ability to recognize unconstrained and cursive handwriting. When compared to manual data entry and early-generation, character-dependent recognition systems, it offers significant cost savings, greater recognition versatility and the ability to adapt to any style of form. Advanced ICR also delivers high read rates and accuracy levels. This delivery is based on a holistic approach in reading handwriting, advanced methods of software training, context use, and many other technological advances. Organizations charged with moderate to high volume forms processing can ensure a competitive advantage from a comprehensive automated recognition system that includes the best attributes of OCR and ICR.

About Parascript

Parascript is a leading developer of cursive, handprint, and machine print recognition solutions. Leveraging digital image analysis and advanced pattern recognition, its software enables business automation in forms processing, postal and financial automation, and fraud prevention; and supports cancer screening in medical imaging. Parascript's award-winning technology draws on a proven 15+ year track record and processes billions of document images annually. Fortune 500 companies, postal operators (including the U.S. Postal Service), major government and financial institutions rely on Parascript products, which are distributed through its OEM and Value Added Reseller networks, including partners such as: IBM, EMC, Bell and Howell, Fiserv, Selex Elsag, Lockheed Martin, NCR, Siemens, and Burroughs.
Visit Parascript online at http://www.parascript.com.

6273 Monarch Park Place, Longmont, CO 80503 USA
T: 303.381.3100 I toll free: 888.225.0169
F: 303.381.3101 I info@parascript.com