imerge *Industry Report*

CharacTell Recognizes Hand Print on "Real World" Documents

by Arthur Gingrande, Partner

imerge
CONSULTING

4 Mead Circle
Lexington, MA 02420
www.imergeconsult.com
(781) 258-8181

Executive Summary

Up until relatively recently, unconstrained intelligent character recognition has been more of a pipe dream than a real world application. However, thanks to improvements in hardware technology and ICR software development, it is now possible for everyday users to translate images of unconstrained hand print characters into computer-usable data.

Unlike conventional ICR, which focuses on segmenting words into discrete characters for individual classification, unconstrained hand print recognition does not depend that much on classifying individual characters to do its job. In fact, natural handwriting recognition, as it is sometimes called, presents many more development challenges than does conventional, constrained, hand print ICR.

This is because, when it comes to recognizing unconstrained or freeform hand print, devices for segmenting characters are nonexistent, and context analysis tools are scarce. Therefore, in order to attain acceptable accuracy rates for freeform hand printed characters, it is necessary to use recognition techniques more robust than even those employed by cursive handwriting ICR engines. CharacTell's ICR system, SoftWriting, employs a powerful set of algorithms - unlike any developed for other ICR systems to date - that achieve the ICR accuracy necessary to successfully translate images of unconstrained hand print characters into computer-usable data.

CharacTell's new product, SoftWriting, uses the "small learning set behavior" of its ICR engine, JustICR, to train on user-provided samples. This unique learning behavior of JustICR is used during recognition of the first sample document. After classifying a small fraction of the characters and words in the initial document, all of the recognized words that appear in its dictionary are used as the training set for the remainder. Then, when SoftWriting tries to recognize a second sample, it uses the learning data from the previous one. This patented method can improve the recognition rate from 50% per word to as high as 90% per word on the first document as more samples are recognized.

Moreover, the SoftWriting algorithms specifically adapt to the individual handwriting style of the user. With each successive use of the system, recognition results continue to improve. Once the initial training is completed, the success rate for mixed upper and lower case lettering is about 98% for most writers at the word recognition level; for upper case only, it is around 99%.

SoftWriting is the sole ICR application available on the market today that is designed specifically to recognize hand print data in unconstrained environments. Because SoftWriting can be used in any situation where the user desires to translate his or her writing into computer-usable format, there are numerous applications for SoftWriting technology. Since use of personal computers is a relatively new practice, there are plenty of people who prefer writing instead of a keyboard as their primary mode of data input and record keeping. For example, scanning field diaries and log reports with SoftWriting on a once-a-week basis could save researchers, supervisors, and journalists a significant amount of labor and inconvenience. Another application currently being considered for widespread adoption is the use of SoftWriting in university and high school libraries as a means to expedite note taking by professors and students.

Background and Overview

Intelligent character recognition (ICR) of hand print data made its first appearance as a “serious application” in the world of electronic imaging during the late 1980’s, when the U.S. Post Office funded several large projects involved with reading and sorting the U.S. mail. The IRS followed suit by funding projects that used ICR to intelligently process IRS “Form EZ”. Companies like AEG, ScanOptics, and Recognition Equipment pioneered hand print ICR in the commercial sector as well, recognizing hand-printed characters on mail-in coupons for order entry applications, like Book-of-the-Month Club and Columbia records.

All of the early ICR installations (with the exception of postal) were forms processing applications that had one thing in common: because the ICR engines of the time were incapable of recognizing touching characters, users were required to print their characters within boxes or between tick-marks, so that the characters would remain separated from one another. Using a graphical device to constrain characters in this manner enables an ICR engine to locate and recognize them one-by-one. At first, techniques similar to those used to recognize machine print were used to recognize hand print, but those techniques proved unreliable. The quest for higher accuracy forced developers to seek other, more sophisticated means to recognize hand print data.

Because human hand print is so varied and diverse, ICR is intrinsically more complex than its older cousin, OCR, which has been evolving for over 30 years. While OCR uses simpler technology (i.e., template matching, comparisons against libraries of fonts, etc.) to classify machine print, ICR uses complicated artificial intelligence algorithms, many based upon neural networks, to do its work. If hand print characters are segmented from one another so that an ICR engine can recognize them in isolation, ICR performance can actually emulate human accuracy. Today, users are encouraged to use “ICR-friendly” forms that are printed in drop-out ink, which makes the form itself invisible to the scanner, so that the ICR software sees only the hand print characters when scanned. This practice can produce ICR results that can equal or even exceed human accuracy in some instances, because humans can make errors due to fatigue and machines do not. In practical terms, this can translate into an ICR application that replaces 60% or more of an existing army of data entry operators. The labor savings can be enormous.

ICR State-of-the Art

During the course of its evolution, ICR software has developed an impressive array of capabilities. Standard image processing features include image cleanup, registration, deskew, line removal, and enhancement for dot matrix characters and fax images. Contemporary ICR toolkits offer a variety of context analysis procedures and scripting languages for writing data validation rules and procedures. Speed has improved, from three characters per second (CPS) in 1988, running on plug-in boards and special chips, to over 1000 CPS today running in software-only mode. Most ICR engines can recognize multiple languages - typically the Latin tongues - and some can recognize

Scandinavian languages, even Arabic and Kanji. Others do more than recognize hand print; they recognize machine print, special math and engineering symbols, check boxes, and bar codes. Moreover, virtually any ICR toolkit vendor will create for a customer a "memory" that is tailored to recognize a specific font or character set - for the right price, of course.

Attaining the Highest ICR Accuracy

High per-character accuracy does not necessarily ensure ICR success. In order to provide a practical benefit, an ICR application must be able to recognize data accurately at the *field* level. For example, correctly recognizing 9 out of 10 digits in a telephone number may yield 90% accuracy, but it also produces a wrong number. This standard is known as the *useful work criterion*.

Field level accuracy can be boosted significantly by using context analysis, look-up tables, check digits and other data validation tools that can compare raw ICR results with predefined expectations. An example would be checking recognition results of hand-printed address fields against ZIP codes and vice-versa. By comparing the two data sets, multiple-character and multiple-field errors in addresses can be corrected automatically to 100% accuracy without human intervention.

The size of lookup tables and the range of other validation devices had to be limited in the days of the 386 and 486 chips with their associated memory constraints. However, with today's extraordinary CPU horsepower, mammoth disk space, and the availability of relatively cheap memory, enormous amounts of context analysis and data validation routines can enable an ICR engine to achieve high rates of recognition accuracy. For this reason, many analysts argue that incremental differences in raw hand print ICR accuracy now make little difference because context analysis and validation routines applied during or after recognition overpower the initial ICR errors. This particularly applies to the world of forms, where context is particularly rich and data fields are so tightly interrelated.

When it comes to recognizing unconstrained or freeform hand print, however, devices for segmenting characters are nonexistent, and context analysis tools are scarce. This is because, unlike a form that by its nature specifies a context for each data field, the content of a freely written manuscript is arbitrary. Moreover, without graphical segmentation aids, the width of each character can vary. Moreover, characters are far more likely to touch one another, which greatly increases the ICR degree of difficulty. In order to achieve acceptable accuracy rates for freeform hand printed characters, recognition techniques more robust than those employed by cursive handwriting ICR engines must be used. CharacTell's ICR system, SoftWriting, employs a powerful set of algorithms - unlike any previously developed for other ICR systems –that can classify unconstrained hand print characters with ICR accuracy high enough to meet the *useful work criterion*.

The Challenge of Recognizing Unconstrained Hand Print

Unlike conventional ICR, which focuses on segmenting words into discrete characters for individual classification, unconstrained hand print recognition does not rely as much on identifying individual characters to do its job. In fact, natural handwriting recognition, as it is alternately referred to, presents many more development challenges than does conventional, constrained, hand print ICR.

For example, one of the major problems that freeform ICR must deal with is that of word segmentation - delineating the boundaries of the character string to be identified from those adjacent to it. This is not a problem in today's "ICR friendly" forms processing environment, because the majority of ICR-based, forms processing applications utilize forms that use combs, tick marks or other graphical devices that strictly divide the data fields from each other while segmenting the individual characters. However, as previously mentioned, unconstrained hand print cannot rely on graphics that segment characters and words because these devices simply are not present in the "real world" of letter writing and note taking.

The additional ambiguity created by lack of constraints means that the conventional approach favored in forms processing and other constrained ICR applications - where individual characters are first isolated and recognized, then validated against rules, look-up tables, and spell checkers to obtain the highest accuracy - invariably fall short of the mark when applied to ICR in unconstrained environments. Humans do not read handwriting on a character-by-character basis. Rather, the preferred unit of reading comprehension is a word, phrase or whole sentence. This suggests the use of an approach different from the character-by-character method. Along these lines, several alternative approaches for recognizing unconstrained handwritten characters have been developed by vendors of ICR engines that are designed to recognize cursively written, natural handwriting.

In cursive handwriting recognition, there is no single technique that can produce accurate ICR results all by itself. The major handwriting ICR developers employ manifold recognition techniques on multiple levels simultaneously, in parallel, to classify natural handwriting. By combining classification of morphological models and data "elements" with recognition of individual characters and ascender/descender analysis - while simultaneously comparing the results against application-specific dictionaries - ICR engines can accurately recognize handwritten words. Recognition techniques include the following:

- **Word level** - Each word is recognized according to its morphology, independent of the characters that compose it. Forgiveness for illegibility is high at this level and requires a huge library of word image models, if that method is used. Ascender/descender analysis also can be employed, which involves examining the pattern of upstrokes and downstrokes that occur relative to an imaginary line drawn through the horizontal axis of the word. Using this method, the longer the word, the greater the odds are that the correct word will be matched.

- **Character level** - Individual characters are classified, useless fragments are ignored, and words are assembled from the successful instances of character classification. Compared to the word level, the number of models used is much smaller, and *ad hoc* word construction is easier to facilitate.
- **Fragment level** - At this level, word fragments are identified using a library of geometric fragment models in combination with fragment construction based upon individual character recognition. Linguistic rules can be invoked at this level. Multiple neural nets are used to differentiate between fragments with low confusion tolerance.
- **Handwriting elements:** Human handwriting is analyzed from the paradigm of a series of strokes made by a writing implement. These movements are represented by eight fundamental elements that describe the trajectories found in all cursive letters. The elements form the basis for a hieroglyphic alphabet and special description language. Because cursive handwriting exhibits so many variations, the same character can often be represented by different sets of these elements.
- **Data field objects:** At this level, combinations of words, sentences, phrases or whole paragraphs may be identified, depending upon the application. Classification takes place through grammatical and syntactical rules, coupled with the use of spell checkers and dictionaries. An example of a data field object would be a date field or name and address field.

Handwriting recognition can be effective in numerous forms processing applications, particularly those that rely on the use of "open" fields, where the person filling out the form is encouraged to respond to a question using their own words. However, as a growing number of users become knowledgeable about how to design and employ "ICR-friendly" forms for optimum recognition results, it is probable that the need to use handwriting ICR in forms processing applications will significantly decline.

So far, the most effective handwriting ICR application is check recognition. Legal amount recognition (LAR), by itself does not produce results that are more accurate than those attained by simply employing courtesy amount recognition (CAR) methods on the numeric amount field in a check. Improvement is achieved through a "voting" approach that recognizes both numeric hand print recognition and alphabetic handwriting ICR, by independently recognizing the numeric or courtesy amount on a check and interactively comparing it with the handwritten legal amount. Using this comparative approach, check recognition accuracy can be boosted tremendously - from 40% of fields recognized at 100% accuracy to 72% of fields recognized at 100% accuracy.

Analysis shows that there is no great miracle at work here. Consider that there are only so many ways to express the amount "$1.29" in written form. The following examples immediately spring to mind:

- one dollar and twenty-nine cents;
- one and twenty-nine hundredths dollars;
- one and 29/100 dollars;
- one dollar and 29/100;

...and so on. In this example, there are only a character string variations that denote the amount “$1.29”, each of which produces a unique morphology. Each answer can be applied against a dictionary of possibilities and then compared independently and in parallel with results of recognizing the numeric, hand printed, courtesy amount. This comparative process can boost the recognition rate on handwritten check amounts to as high as 92% of the fields classified at 100% accuracy.

As stated earlier, however, such comparisons are not possible with ICR in unconstrained environments. Acceptable accuracy rates for freeform hand printed characters can only be achieved if recognition techniques more powerful than those used by cursive handwriting ICR engines are brought to bear on the situation. CharacTell has created a powerful set of algorithms, unlike any other recognition algorithms previously developed, to power its SoftWriting ICR solution. These algorithms deliver the ICR accuracy necessary to successfully convert unconstrained, hand print characters into computer-usable data.

Enabling Technologies

The recognition kernels of all the leading ICR technologies use feature extraction from character images. Generally, hundreds or sometimes thousands of features are extracted from each character. These features are than analyzed using various methods such as neural network technology. All of these methods require a relatively large training set. For example, the leading ICR engines train on tens of thousands of samples for each character they learn.

CharacTell’s recognition engine, JustICR, uses a different approach, which bears more similarity to microbiology than it does to conventional intelligent character recognition algorithms. The first step in the recognition process is to create a string from the image of a given character. This string can be characterized metaphorically as a “DNA chain”. The DNA chain is combined from pieces that can be called “genes”. Within this schema, recognizing the image of a character is like finding the father of a child. Each image-derived gene is matched with a database of similarly produced genes that is based upon the learning set, and each gene assigns a weight for each possible recognition result. The number of genes in each DNA chain is always less than or equal to 28.

While this description of the ICR process makes it sound as if character-derived “genes” can be represented as features, it would be wrong to make that equivalence. The number of possible genes is enormous - theoretically infinite - while the number of features in other ICR algorithms is not, because it is fixed by the algorithms. Each gene can exist or not, while the features are generally integers and not Boolean in nature.

The genes do not contain any information other than whether they exist or not, plus their location in the DNA chain.

What is interesting is that the number of samples that is needed to train the JustICR engine is extremely low. In fact, in order to facilitate learning, just *one* sample for each character is sufficient to generate reasonable ICR accuracy. After training on several writing samples of the same user, the JustICR recognition engine can achieve extremely high recognition performance - performance that is superior to the accuracy of a neural network that is similarly trained.

CharacTell’s new product, SoftWriting, uses the “small learning set behavior” of JustICR. When SoftWriting tries to recognize the second document, it uses the learning data from the previous one. The unique learning behavior of JustICR is also used during recognition of the first document. The algorithm is structured so that, after recognizing a small fraction of the characters and words in the initial document, all of the recognized words that appear in the dictionary are used as the training set for the remainder. This patented method can improve the recognition rate from 50% per word to as high as 90% per word on the first document.

SoftWriting uses several proprietary technologies other than those that comprise the recognition kernel. For example, the scanning is done in gray/color bitmap rather than black and white images like most of other ICR engines. A special algorithm converts the gray/color images to black and white images. This algorithm is extremely important because scanning documents written with blue pens in black and white with conventional ICR software generally creates images of poor quality. After doing this conversion, a proprietary algorithm that analyzes the lines, words and connected characters is applied. The recognition kernel uses the dictionary in order to optimize results.

Deploying the SoftWriting Solution

SoftWriting technology translates easily into practice. The procedure for using the system is as follows. First, a document from each writer is transmitted via the Internet to the CharacTell Website. Then the characters are segmented and recognized by one of several ICR engines. The first ICR engine generally yields between a 50-90 percent recognition rate on all the characters. After recognizing a subset of the words using dictionaries, the recognized words are grouped into a learning set for the JustICR engine.

Next, the entire document is intelligently recognized again using the results of the first ICR engine and the results of the JustICR engine taken together. The learning process is iterative until there is no further improvement on the number of recognized words. ICR results can begin with 40% of the words correctly recognized, and end with more than 85% word recognition accuracy after the iterative process is applied to the first three pages of the document.

This primary limitation of the SoftWriting ICR system is that at least 300 words are required for training on the first document in order for JustICR to be effective.

However, given sufficient legibility, in many cases 100-200 words is a good start. Another limitation is that only the words that appear in the English dictionary are relevant for the first training. Therefore, writing the character string "ABC" would not help in training on the first document. Additionally, numbers are not recognized in the initial training on the first document. Numeric characters are recognized after the user has corrected the errors
because SoftWriting uses the learning experience on the user’s corrections for subsequent documents.

In this manner, the SoftWriting algorithms specifically adapt to the individual handwriting style of the user. With each successive use of the system, recognition results continue to improve. Once the initial training is completed, the success rate for mixed upper and lower case lettering is about 98% for most writers at the word recognition level; for upper case only, it is around 99%.

To facilitate optimum recognition accuracy, training may be required not only for the JustICR engine, but for the user as well. This is actually part of the contemporary trend in handwriting ICR applications. In times past, mandates to change established patterns of behavior were met with resistance by end users. Their attitude was to challenge ICR developers to create technology that could deal with idiosyncratic recognition environments that discouraged, rather than promoted, successful ICR results. Contemporary end users have learned that when it is possible to modify their recognition environment to fit the ICR application at hand, they can incur extraordinary savings through the increased recognition accuracy that invariably follows.

Other systems that involve user training require that the user learn additional symbols, as it is with the Graffiti and Palm products. This is not the case with the SoftWriting solution. All the user has to do is remember to be careful to write his or her characters in unconnected fashion when taking notes or writing manuscripts that are intended for SoftWriting recognition.

After ICR verification is completed, SoftWriting automatically exports its results to a designated word processor and launches it - either Word or any other application that accepts text files.

Key Features of the SoftWriting Solution

Apart from the outstanding accuracy on non-cursive characters that JustICR provides, the SoftWriting freeform recognition solution offers a rich array of supporting features:

- *Language support* - SoftWriting supports recognition of non-connected writing in English, Spanish or German.
- *Dynamic OCR training* - The OCR continues to learn individual handwriting as the user work uses the system, which means continuous improvement in recognition rates.
- *Improved recognition using built-in dictionaries* - SoftWriting uses dictionaries against which it matches recognized words to validate questionable characters

and boost recognition rates. SoftWriting also uses extensive mix-and-match capabilities to complete words that were not fully recognized by the OCR engine.

- *User-defined dictionary* - Each SoftWriting user controls a personal dictionary that is continually updated as new words, names, terms and acronyms are encountered.
- *View scanned pages before OCR recognition* - With SoftWriting, the user can preview the scanned image of the document before the start of character recognition.
- *View original document and converted document simultaneously* - SoftWriting provides an on-screen display that shows the handwritten document and the recognized text side by side and highlights any ambiguous words that it encounters, which enables rapid error correction.
- *Capture and preserve of graphic images* - Areas on documents that contain equations, pictures, diagrams or other graphics that should be retained in original form can be defined by the user on a point-and-click basis for inclusion in the electronic version of the document.
- *Save document in format recognized by word processors* - SoftWriting saves converted documents as Word files or standard text files, both of which can be opened in most word processors and other application environments.
- *Broad scanner support - SoftWriting* supports scanners using the TWAIN interface standard supported by most desktop scanners.
- *Remote scanning* - with SoftWriting, documents can be scanned on a remote station and sent to another for centralized processing.

Taken together, these features compose a solution that not only makes SoftWriting easy to use initially, but also continually improves by adapting more precisely to an individual customer’s needs with each repeated use.

Benefits and Applications

SoftWriting is the only ICR software solution that specializes in recognition of unconstrained, non-cursive, handwritten documents. Three other companies on the market - Parascript, Ceresoft, and A2IA - all recognize cursive handwriting and run-on handprint, but they do so only within applications that are constrained in some fashion, such as check processing and “open” fields in forms. SoftWriting is the sole ICR application available on the market today that is designed specifically to recognize hand print data in unconstrained environments.

There are numerous applications for SoftWriting technology. SoftWriting can be used in any situation where the user desires to translate his writing into computer-usable format. Since PC use is a relatively new phenomenon, that means there are plenty of people who prefer writing instead of a keyboard as their primary mode of data input and record keeping, or they find themselves in situations where the use of a computer is awkward. Scanning field diaries and log reports with SoftWriting on a once-a-week basis could save researchers, supervisors, and journalists a large amount of labor and inconvenience.

One application currently being considered for widespread adoption is the use of SoftWriting in university and high school libraries as a means to expedite note taking by professors and students. In this application, several scanning stations are set up in a library in much the same manner as copying machines are currently installed. The researcher would bring his or her notes to the station and scan them, then do the ICR verification and correction on the attached PC. Once the results are completed, the SoftWriting application sends them, along with an attached image of the original manuscript, to an email address designated by the user. That way, the text can be directly copied into a word processing document at the user's home or office the next time that he or she sits down at their computer.

Until very recently, intelligent recognition of unconstrained hand print characters has been a myth: a high-tech curiosity found only in science laboratories. With the advent of CharacTell's groundbreaking innovations in ICR algorithms, however, that myth is fast turning into a reality. SoftWriting takes unconstrained hand print recognition out of the laboratory into the realm of "real world" documents - for everyday use by professionals and laymen alike.

About the Author

Arthur Gingrande is a founding partner of IMERGE Consulting and one of the most published consultants in the field of intelligent character recognition (ICR) From a national perspective, he is considered one of a handful of qualified experts in image-based ICR and forms automation. He wrote the AIIM-published, definitive work on forms processing entitled *Forms Automation: from ICR to e-Forms to the Internet*. He also wrote the TAWPI publication, *Cost-Justifying an ICR Solution*. Prior to becoming a partner of IMERGE, he founded an ICR development company called Symbus technology (now known as Captiva) and worked as director of marketing and business development for Nestor, a neural network-based, ICR software publisher.

Since 1991, over 200 of Mr. Gingrande's articles have been published in various trade periodicals such as KM World, *e*-doc, Business Solutions, Integrated Solutions, Imaging Business, Inform, Imaging and Document Solutions, VAR Business, Service Bureau News, Transform and Imaging World. The topics of these articles have included workflow, document management, electronic imaging, forms automation, ICR/OCR, and electronic forms. He has also written numerous white papers on those subjects, including research reports for Dataquest, Advanced Technology Group, and BIS CAP (now known as GIGA Information Group).

Mr. Gingrande is the former publisher-in-charge of imaging technology at ISIT.com, a Web site and online library dedicated to integrated solutions in information technology. He is also editor and publisher of *Contemplor*, a newsletter dedicated to ICR, forms automation, and document management technology.