iREAD Forms Product Information (version 4.1)

Summary

CharacTell's iREAD Forms™ is a library of sophisticated character recognition engines and functions (callable as DLLs) that are intended for integration into applications. The primary engine is based on CharacTell's revolutionary Advanced Character Recognition™ (ACR™). Voting in conjunction with other available engines, iREAD Forms supports multiple languages, and is designed for high accuracy, low error rates, and operation under difficult conditions. With our one of a kind ability to train on fonts, special characters, or writing styles (individual and population), very high recognition results are possible.

New in this release, iREAD Forms also offers a wider range of powerful functions that are needed by developers to create robust form processing applications. These include such important functions as form identification and registration during run-time, image cleanup and pre-processing, zone setup and parameter assignment, and interface with the various recognition engines for both output and input.

For users with more specific needs, CharacTell offers JustICR™ and our new CMC7 Reader™. Based on ACR, but not offering voting, JustICR is a lower cost solution to specific recognition needs, primarily reading unique and difficult fonts. Our new CMC7 Reader, also a subset of the iREAD Forms library, offers near 100% success rate for grayscale images at 200 dpi, and can be purchased separately.

Overview

iREAD Forms is CharacTell's all inclusive library of programmable recognition functions (Software Development Kit (SDK)). It also includes JustICR and CMC7 Reader as lower cost subsets that address specific needs. All are available as open API's that allow developers of form processing applications to quickly and easily integrate advanced recognition capabilities (ICR – handprint and OCR – machine print recognition technology) into 32-bit Windows applications using C / C++, Delphi, Visual Basic, or any other development tools supporting DLL components.

iREAD Forms and JustICR are sophisticated recognition engines that – in a single pass – can read machine-print, upper case hand-print and the challenging non-connected, lower case, hand-writing of alpha and numerals. The CMC7 Reader is a specialized engine that is optimized to read and process the CMC7 font that is used in banking applications for check processing.

iREAD Forms – Introduced to the form processing market in the spring of 2002, iREAD Forms is a natural evolution of JustICR. In addition to including CharacTell's advanced technology, iREAD Forms integrates an additional OCR / ICR engine, which works closely in conjunction with the CharacTell engine. The results generated by each engine are combined using a unique voting mechanism, and an advanced algorithm for utilization of context-based information and dictionaries, for even superior and more accurate overall results. The combined engine offers improved recognition in terms of higher correct recognition percentages and substantially reduced errors. iREAD Forms is capable of operating on very low quality characters, training on special fonts and individual hand-writing and is particularly suitable for recognition of critical fields where near 100% accuracy is a must. This further demonstrates CharacTell's ability to offer superior, accurate, and easy-to-use recognition products that are highly competitive in the marketplace.

JustICR – JustICR is a field proven, advanced recognition engine. It has been installed in hundreds of production sites in a wide variety of projects and solutions. JustICR has been applied at FedEx (USA), the Social Security Administration (USA), Swiss Post, Deutsche Post (address change notification project; the largest form-processing center in Europe), numerous country census projects in Brazil, India, Italy, Ireland, Kenya and Cyprus, as well as high volume credit cards voucher processing applications (Visa, MasterCard, American Express and Diner's Club) in Portugal, Israel, and Croatia. Integrators who used JustICR report that without it they would not have been capable of delivering certain demanding solutions, or even successfully bid on certain projects leading to those solutions. Only JustICR offers the combination of character learning ability, very high recognition rates, low error rates and unique capabilities (see below) that were particularly needed for deployment in the field.

CMC7 Reader – Derived directly from ACR's ability to be trained on unique fonts, the CharacTell CMC7 reader offers exceptional performance when it comes to reading CMC7 characters. Processed in grayscale at 200 dpi, the Reader reaches near 100% correct recognition results routinely. Included in iREAD Forms or available separately, the CharacTell CMC7 reader is an excellent interpreter of this widely used font.

All products come complete with documentation, evaluation guide, training and statistics generation tools, and a short demo. The training tool gives developers all they need to conduct a thorough evaluatation without the need to fully integrate the engines, and quickly determine suitability for their needs. This, in most cases, can be done – using your own material – within several hours without writing a single line of code(!) and often without even referring to the user's manual…

Packaging

iREAD Forms offers the convenience of a complete solution to the problem of developing powerful form processing applications. At the same time, developers are offered the flexibility of selecting subsets for specific needs.

The following table summarizes the packaging of each of the products:

Product	Internal engines	Voting	Form Processing Functions	Training Tools	Trained Classifiers	Dictionaries	Evaluation & Demo Tool
iREAD Forms	2	Yes	Yes	Yes	Yes	Yes	Yes
JustICR	1	No	Yes	Yes	No	Yes	Yes
CMC7 Reader	1	No	No	No	No	No	Yes

Why Form Processing Technology from CharacTell?

Where Best to Use CharacTell Engines?

The JustICR and iREAD Forms engines may be used in numerous settings, but are most recommended for image or document processing applications that need accurate and flexible ICR / OCR capabilities, such as:

Processing of machine-readable forms filled in by hand

Archiving and document processing applications

Specialized applications that require searching for and treatment of fonts prior to document recognition

Processing field containing postal address information

What's New in Version 4.1?

iREAD Forms version 4.1 includes several new and important enhancements. The significance of many of the additions to the library is that they expand the scope, and facilitate the process, of application development by making it easier to developers to access the functions they need to build a complete application beyond the core character recognition engines.

The following are new functions and calls:

Form identification: iREAD Forms recognizes the template of a filled form from a library of forms. The identification does not need anchors. The number of possible templates in the library may be very large (even hundreds or thousands of possible templates). The duration of the template recognition process is not dependent on the size of the template library.

Form registration: iREAD Forms returns the registered image in order that the developer will be able to display the exact region to the verifier station.

Field preparation to OCR: iREAD Forms prepares the image for OCR by removing lines or pre-printed text of the empty form.

Optical Mark Sense (OMR): OMR is now a built-in feature of iREAD Forms. The OMR can work in several modes: (a) As is - meaning that the image that should be recognized contains the box (or circle); (b) The box is drooped during the scanning; c) The box is removed by external form removal algorithm.

Expanded file format support: In addition to the ability to read BMP files from memory, support of BMP files is now a built-in feature. Important for banking applications where reading of CMC7 font is needed, iREAD Forms can read CMC7 from grayscale images of checks directly from BMP files.

Easily Trainable Recognition Engine

The key feature of JustICR and iREAD Forms is their ability to train a new font or handwriting style in a very short period and using very small sample sets of only several hundred characters. In most of the form processing applications, there are "problematic" fields that you wish to recognize better than the results that can be obtained from the "off-the-shelf" recognizers. Below are several such cases:

Low quality handwritten fields: in some cases, the quality of the scanning in handwritten fields is poor and characters may be broken or be accompanied by 'dirt'. In these cases the recognition results of the best "off-the-shelf" engines is significantly reduced. Training on these fields with our engine, may yield much improved results. The best way, based on our experience in the field, is to use voting between our engine and other engines that you use normally.

Low quality machine print fields: there are cases that the printed field is of poor quality, such as the case of "stamped field" commonly found in shipping forms (FedEx, Airborne, etc.), credit card vouchers, airline tickets, and many others.

Sometimes a form contains a "sensitive field" (often machine printed), that requires no less than 100% recognition. Teaching the CharacTell engine the specific font of this field accomplishes this goal.

For applications outside US, handwriting styles may be different than the handwriting expected by "off-the-shelf" engines. Again, ACR's ability to be trained on different styles offers a powerful method of addressing this serious problem. This is by numerous accounts the reason that in many national census projects our engine played a significant role both in winning and executing the project.

The same applies when a field contains specific marks. These can also be trained with our engine. One example of the use of this feature comes from the educational market in United Kingdom. The application included tests that were submitted by the students. In these tests the student did have to put a mark on the correct answer, and there were several types of mark signs. The ability to teach these mark signs with JustICR solved this problem immediately.

Searching for a Font in a Form

Many applications involve hundreds of form types. Some are "variants" of the same form, others are different. A common example applies to shipping forms. Often, however, one number stands out in a different font (OCRA in this example), and can be used to identify the form type (different variants of the same form still have the same number). You can train for this font, find it in the form, and ID the form without having or applied any additional knowledge about the form (if such information does exist, it can be used to add the form to a library of templates that can also be used for form ID).

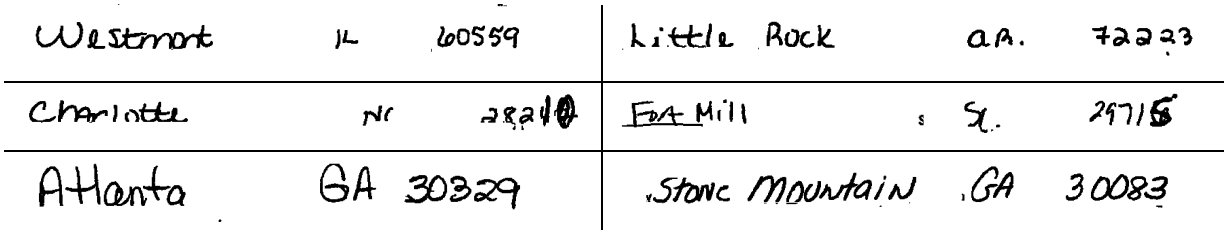
Dictionaries

It is possible to give to JustICR engine dictionaries. The dictionary may be a large dictionary (such as the full English / Spanish / German... vocabulary, which may contain 100,000 to 1,000,000 words). The recognition results are improved significantly even if the dictionary is not full (it does not contain all the possible words). One example can be from an application in Germany in which one of the fields was a First Name (hand print). The recognition rate jumped from 89% per character to more than 97% per character by using a dictionary of first names. Equivalent results achieved with geographic places, or "descriptive fields" in which the field contains a phrase that describes something (such as: occupation, religion, etc.).

City, State, Zip field recognition

iREAD Forms can recognize, verify and return in one operation the City, State, Zip of US-based addresses found in a field or combination of fields. This is a powerful feature for anyone in need of extracting address data from form fields.

The following are examples of City, State, Zip combinations that are successfully recognized by iREAD Forms (notice the difficult conditions present – overlapping characters, noisy images, spelling errors, line interference, broken characters, nearly indiscernible characters, etc.):



In all these cases, the engine returns the correct City, State, Zip combination despite serious deficiencies and complexities in the images.

Feature List

The following are the standard features offered by JustICR and iREAD forms:

1. Image input and support

Supported image formats: TIFF Uncompressed, TIFF Packbits, TIFF Group 3, TIFF Group 3 Modified, TIFF Group 4, PCX

Opening binary images from TIF files

Reading binary images from memory

Support of color and grayscale images from memory

2. Training tool

Capture images from files or from the integrated application

Typing labels of each character

Verifying the suspected characters

Statistics generation tool

3. Image preprocessing

Image deskew

Image despeckle and noise removal

Line removal

Form text removal

Image rotation (degrees): 90, 180, 270

4. Form processing

Fast form identification of filled form based on form attributes from a library of templates

Form registration of identified forms

Easy setup of field processing parameters, choice of engines and desired functions

5. Recognition

Machine-print, handprint (upper-case, lower-case), non-connected handwriting, numerals, Farrington 7, CMC7

Recognition text languages: English, French, German, Italian, Spanish, Portuguese, Dutch

Recognition at field , line, or page levels

Output format: ASCII, confidence levels, coordinates of characters

Special symbols and annotations can be trained and recognized

Support of multi-line fields

Verification of city, state and zip combination for US addresses based on US Postal Service database.

6. Advanced features

Dictionary support

Two engine advanced voting algorithm (iREAD Forms only)

Locating text of specific font in a form

Systems Requirements:

Operating System: Windows 95 / 98 / ME / NT 4.0 or later / 2000 / XP

Runtime environment and Compilers: Dynamic Link Library (DLL) runtime, C / C++ / Delphi / VB

User Interface Language: English

Minimum Hardware Required:

Intel 486 or greater

1X CD-ROM drive (optional)

400 MB free disk space

64 MB RAM

What do people who use JustICR say?

Training of a specific font – Airborne, USA, 2000

“We had a benchmark of a big application in the USA that had an important OCR field of a specific font. This field was so important, that the customer needed to have the highest recognition results with almost no errors. The font was OCRA, but it was of poor quality, because it was not the first copy of the form. The best engines gave us 98% recognition with 1% false positives. After a day of training JustICR, we achieved 99% recognition and no errors at all out of the 50,000 characters in the benchmark.”

Training of low quality writing – Unicre, Portugal, 1998-99

“The project went quite well until we installed the software, and our client faced the recognition quality of his forms. We used three top ICR engines with advanced voting, but the results were so unsatisfactory (82% recognized characters and 5% errors) that the client preferred not to use ICR at all. The writing quality was poor because it was a second copy of the credit card application; some of the writing was too light, and some was too bold. After one year, we decided to try JustICR. We used voting of JustICR with one of the other engines that we had. The results jumped to 90% recognized characters and 2% errors.”

Training of a new writing style and dictionary

The Turkish representative of a big software firm recounts his experience with JustICR: "We decided to work with this software because it was the only engine with the ability to learn the Turkish handwriting style and achieve very good recognition quality."

The Brazilian Census, 1999: "The benchmark recognition requirements of this huge census seemed to be impossible. In order to be qualified, the system needed to achieve more than 90% recognition rate for numeral ICR fields and more than 30% recognition rate for alpha fields. The punishment for each field that was not properly recognized was 30 times bigger than the added points of recognized field.

During our tests, the recognition results for the numerals were above the minimal requirements using voting of 3 engines. However, the situation with the alpha fields was a disaster. We had about 20% recognition rate per field, and 20% errors! It looked impossible to achieve anything close to 30% errors per field with less than 1% errors.

And then we tried JustICR. We designed a form with several fields and gave it to 200 people. After collecting these forms we trained JustICR in one day. We added JustICR into our voting system. We used a smart dictionary option that can handle a partial dictionary with more than 25,000 words. The results were overwhelming: the alpha recognition rate was 57% and 0.3% errors per field, and the numeral field results were improved as well. Needless to say, we were the only competitor in the benchmark who achieved the minimal requirements."

Searching for a font in a form – Mexico Revenue Service, Mexico, 2000

“The application that we needed to process had several hundreds of form variants. Our form identification algorithm could not deal with so many types of forms. However, on each form there was a printed number in a specific font that indicated the form type. We used JustICR's ability to find and recognize this number, and the system worked flawlessly. The competitors used various other approaches: one of them keyed in the information from image, while another spent several hours, manually sorting the forms.”

The same technique was applied in several large applications in the USA, Germany and India.