PARASCRIPT®
World's Best Image Recognition Engine – *Getting Better!*

WHITE PAPER

CHALLENGES

Lost efficiency due to confusion about recognition technology and best processes in the commercial arena

TAKEAWAYS

The Benchmarks for Accuracy

How Capabilities Like Contextual Reading and Smart Dictionaries Drive Efficiency Gains and Cost Savings

The Technologies That Can Make a Difference in Your Business

A Look Ahead: Advanced Capabilities that Will Drive Bigger Gains

Best Practices in Image and Character Recognition: Technologies that Can Transform Your Organization

Defining OCR, ICR and NHR:

Exploring the technology and methods that are driving today's biggest efficiency gains in the commercial arena

Part 2 of a 2-Paper Series

In Part 2, we look at the accuracy imperative for automated character recognition, explore some of the methodologies that drive the best OCR and ICR processes, and look ahead at advanced capabilities that can help your company achieve cost savings and efficiency gains.

What's in this Paper?

The Data Accuracy Puzzle: What Are the Benchmarks?

Accuracy levels for automated recognition have been under criticism since the technology was first introduced. Performance has been determined by the amount of information that is 100% correctly processed at an 'x' percentage of the time. Manual keying, which in itself is not error-proof, is the baseline for comparison.

From the standpoint of measuring accuracy, humans are inaccurate from 2% to 6% of the time. This accounts for factors such as boredom, repetition and fatigue that affect an individual's overall performance. Recognition technology does not encounter these problems, but performance varies depending upon the application, quality of the images, and other factors. It is important to fully understand the application and its limitations in order to understand overall performance.

For their specific applications, OCR and ICR demonstrate high levels of accuracy when working with constrained text (i.e., lines, boxes and combs). Working with high-quality machine print, OCR provides nearly 99.9 % recognition accuracy — high enough to be acceptable without additional controls for most OCR applications. With different machine fonts, this high level of accuracy can vary.

When the need for unconstrained handprint or cursive writing recognition is required, the need for more complex analysis arises. Standard read accuracy rates become more an art than a science and depend heavily on the types of form data, the instructions typically included (such as the instructions "please print" or "enter date as MM/DD/YYYY"), the cost of human intervention, and desired processing throughput rate. Again, accuracy can be in the 80% to mid-90% range depending on a number of variables and tolerance thresholds.

Enhancing Recognition Using the Latest Techniques and Technology

A myriad of variables impact recognition accuracy, including image quality, character and word clarity, font identification, word recognition, language, and dictionary. The way that recognition rates and accuracy are derived differs with each of the recognition technologies and with the specific application.

Form Design and Preparation

When forms are not optimized, a certain amount of processing may be required to prepare the image prior to recognition. Preparation can include form registration, removal of lines from the image, and elimination of any “speckles” caused by poor print quality or reproduction.

Form registration ensures that the image is aligned for optimal recognition, and fixes any images that are skewed or slanted due to poor alignment or registration when the form was scanned.

Preprocessing removes any form boxes or lines that separate characters or fields in constrained images. It also removes any “speckles” caused by poor print quality or reproduction of the form.

Context Sensitivity

Context plays a significant role in the recognition process. When humans read handwriting or print they look at entire words — and even the entire document — to correctly identify what is written.

Knowing a range of probable meanings makes the task of reading much easier. This is why recognition engines use context as an effective and flexible tool to compensate for the inherent ambiguity of handwriting and to improve recognition accuracy.

Examples of context can include “alpha” vs. “numeric,” “date” vs. “number,” and “address” vs. “name.” While recognition engines bring many powerful techniques to bear in order to read, recognize, and respond to data extraction processes, any human-based context provided to the engine can significantly improve not only accuracy but performance throughput of the recognition process.

ICR “Narrowing”, Contextual Reading and Smart Dictionaries

The “Intelligence” in Intelligent Character Recognition comes from the context of the data attempting to be captured. “Choices” in answers narrow significantly from having an idea of what data should be in a given field, just as people automatically narrow context when reading information on a form. For example, date fields convey specific meaning with a small number of characters, as do phone numbers, addresses, names, account numbers, etc. Recognition software takes advantage of choice narrowing in the form of ICR.

The most potent technology takes advantage of the repetitive nature of some data streams. By using databases containing records of information previously captured, more contextual information is available for data lookup, increasing recognition rates. For example, in the case of medical claims, there are frequent concentrations of claims by individuals. Much of a person's lifetime health care costs are incurred during the last years of life. In fact, according to the New England Journal of Medicine[i], up to 40% of a person's last year health expenses are incurred in the last month alone. Also, U.S. women are more than twice as likely to visit the doctor as men[ii]. That equates to 150 million more checkups for women versus men.

As a result, "frequency" plays an additional role in the contextual equation. If a repeat patient files a claim, lookups can be used for combined fields like patient ID, name, address and phone number. Similarly, the same providers are regularly filing claims. Provider names, addresses and phone numbers can be captured once and used repeatedly for future data extraction.

The net effect for improved ICR performance is cost savings. With the current focus on the economy and cost reduction, companies are relying heavily on technology to help improve the bottom line. Utilizing existing data to enhance automated recognition is one way to achieve that goal.

Modern forms processing software has the capability to apply such technology today. In the case of medical forms, several data fields will have answers frequently repeated from form to form. Field data can be stored as a record in a database for each unique instance such as name, address, phone, DOB or gender. This database can be used for reference during recognition.

A nearly 63% reduction in keystrokes was realized from recent trials using data accumulation on several decks of multi-thousand Health Care Financing Administration (HCFA) forms. The form sets consisted of complex single- and multi-part forms averaging about 375 output characters per form. Using existing standard OCR/ICR methods in combination with true double-blind/verify keying (i.e. neither keyer knows the input of the other keyer), the typical manual effort was 400 keystrokes per document. With the addition of "data accumulation," but still performing double-key/blind manual correction, that keystroke average went down to about 150. ***The corresponding average time to process these forms went from an average of 3.8 minutes to 1.4 minutes per form***.

Most significantly, this reduction in keystrokes and decrease in operational time was realized without introducing new errors.

The new performance results were enabled by dynamic dictionaries that allowed the system to accumulate valid answers over time. The ability to use this data dynamically further refines the possible list of answers, boosting recognition rates and accuracy. For example, a ZIP code is a relatively easy field to read automatically. Once the ZIP code is read, it can be used to create a dynamic dictionary comprised only of last names associated with that ZIP code.

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.

Additionally, cross-validation also improves performance by providing context for other fields. For example, suppose a form has a complete address field including ZIP code. A recognition engine would extract and recognize the ZIP code, which would use a database validation to identify appropriate street addresses and cities. Armed with this additional context, the engine would then use this information to recognize the other address fields in a quicker manner by reducing the number of potential matches it needs to validate.

Manual Data Correction: The Rewards of Getting it Right

Improvements have also been seen in manual data correction. Automatically recognizing a full address field is a complex task. There are existing ways to use postal addresses for context. And, read rates can be very high with existing technology. Still, parts of the address are read more easily than others: when a ZIP code is read, then the state is a given and so is, usually, the city. The street address is much more difficult given the high variability of how a person might write it and all the “extras” that can be added (like direction indicators, apartment numbers, rural routes, etc.). Traditionally, when an address field is not fully recognized the entire set of sub-fields is sent to keying: street number and name plus city/state/ZIP.

Without using any software technology, the average number of keystrokes required to finalize an address is 32 per single-key and 64 for double. Using advanced methods of OCR/ICR in combination with intelligent keying, this keystroke penalty can come down to about 4 keystrokes for first pass and 6 for double-key. This 90% keystroke savings is more than significant considering that virtually every form has at least one address.

ICR and OCR are tried and true technologies used to improve data capture. But improvements in effectiveness are still possible with new uses of contextual data. Using some methods indicated above, recognition technology improvements are available that significantly increase automatic read rates and drastically reduce expensive, error-prone keystrokes.

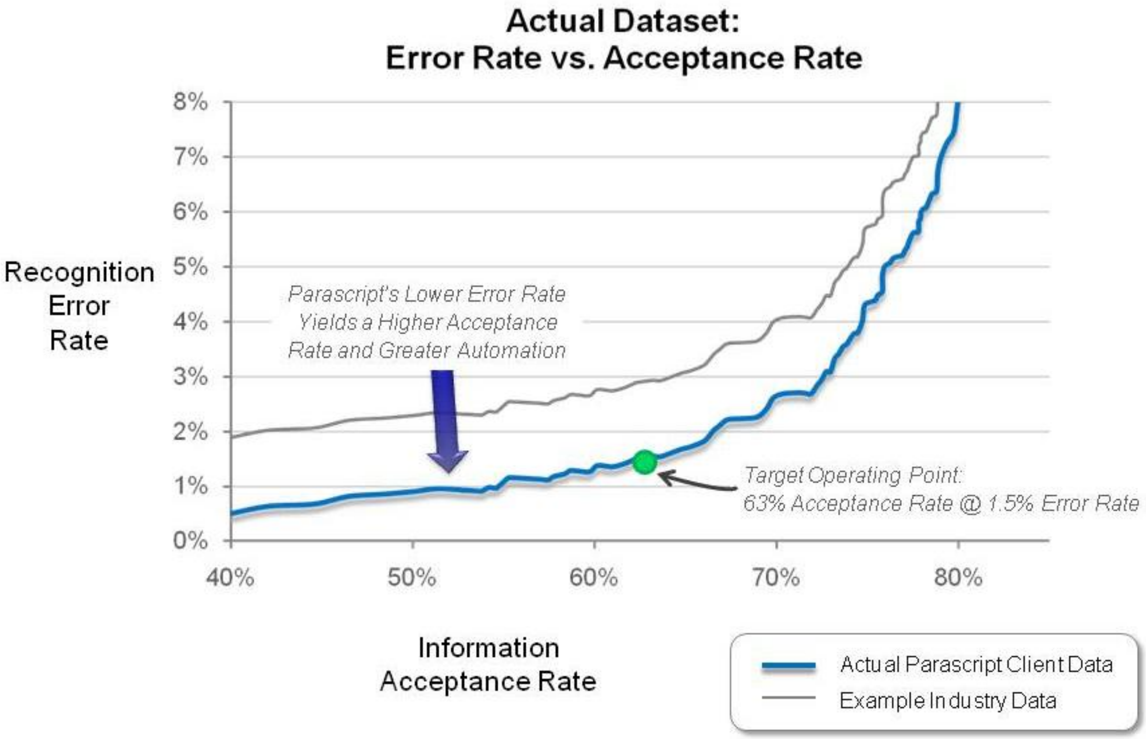
Measuring and Tuning Results

Just as each company has unique objectives and processes that will determine how automated recognition technology is deployed, so will each company experience varying degrees of cost reduction. With the technology in place, gains can be measured and new benchmarks set through a tuning process that takes into account the accepted error rate balanced by desired throughput. Companies will be able to fine-tune their automated recognition processes and establish their own accuracy/cost “curve” that serves as a quantifiable basis for discovering what the outcomes are, and how to get even better. Continuous tuning based on feedback will drive an organization closer to optimal read/accuracy rates.

A Performance & Tuning Example

The following chart presents actual recognition data showing the direct relationship between the recognition error rate and the associated information acceptance rate. A higher acceptance ratio means that less manual intervention is required with the recognized information. A very low recognition error rate results in a lower acceptance rate. Of course, the operational goal is to move the curve down to the lower left as shown to achieve a better cost benefit result.

In this actual example, the target performance was a 63 percent acceptance rate at a 1.5 percent recognition error rate. Thus, 63 percent of recognized data did not require any human intervention or correction. This resulted in the optimal cost benefit to the operation.



Recognition Error Rate: The recognition error rate represents false data recognition. The allowed error rate will vary depending on the type of input data. For example, the error rate in reading a Social Security number may need to be extremely low, whereas the allowed error rate for a first name may be less restrictive.

Information Acceptance Rate: On the other hand, the information acceptance rate is the reduction of manual intervention through recognition. Information is accepted when the confidence rate in a particular recognition process meets or exceeds a defined threshold and thus does not require manual intervention. So a higher acceptance rate results in greater automation. In general, the higher the cost of manual keying or intervention, the higher the allowable acceptance rate.

Accuracy requirements vary according to the business operation or recognition task. For example, some applications require an error rate of no more than 0.1 percent; other applications accept an error rate of 2.5 percent. A one to two percent error rate is common for a human operator, so that rate is often the benchmark for automating recognition — but it is only one data point to be considered.

Technologies that Can Make a Difference: The Future of Total Recognition Solutions

Online Recognition Services

As the market for recognition technology continues to grow, companies will seek solutions that merge document processing with applications such as customer relationship management (CRM) and ecommerce.

Parascript partners with these companies to create a seamless information network that fuels new lines of revenue and better avenues for customer service. Online services offer complete data capture and recognition solutions for customers who do not have the infrastructure or capital to invest in high-cost hardware and software. This enables users who are on a budget or without access to sufficient internal resources to outsource their recognition processes while focusing on their core competencies.

Enhancing In-House Forms Processing

Some businesses will choose to keep their forms processing and data capture systems in-house. Fortunately, automated data capture can benefit different types of government, financial services and banking, healthcare, and other high-volume enterprise form applications such as tax returns, insurance claims, credit applications, health claims, and fulfillment. Total recognition technology captures, interprets and transforms the mass of disconnected forms data – cursive, handprint, machine print, constrained and unconstrained – into a business asset that delivers fast, accurate information to business applications and databases.

Intelligent recognition technology can be used to read any text style – in any combination – and recognize all form field types, including name, address, dollar amount, and check boxes. Intelligent Recognition combines character-based engines with engines that read entire words or phrases. Application-specific context such as vocabularies, templates, alias tables, postal databases, and other contextual information increases recognition speed and accuracy and reduces costly errors.

Applying this technology to your business processes can deliver huge benefits:

- **Cost Reduction and Time Savings** – Automated forms processing reduces tedious and time-consuming manual entry, reducing fully loaded cost per manual data entry operator by up to $40,000.
- **Improved Accuracy** – With the potential for human error reduced, machine print (OCR), handprint (ICR), natural handwriting (NHR®), check marks (OMR) and barcodes can be interpreted with unparalleled speed and accuracy.

- **High Quality** – Identify top performing operators and increase the level of verification for less efficient operators, maintaining high accuracy and quality while minimizing forms processing costs.

- **Complete Workflow** – A complete forms processing workflow scalable to increasing volumes of forms.

- **Enhanced Security** – Form snippets – rather than complete form data – enforce security at the field level. Information can be sent to multiple sites, potentially eliminating security breaches or data misuse. This method also lends itself to optimized keying because operators don't have to scan full pages and actually can improve performance by keying the same field type repeatedly.

Customized Solutions

Parascript recognizes that its customers do not want a one-size-fits-all answer for their problems. Along with technology partners and integration partners, Parascript offers custom solutions that provide the best recognition capabilities at the right cost point for your specific application. In addition to technology integration, Parascript offers form processing expertise that can help improve your performance and overall productivity at critical checkpoints in the application process.

About Parascript

Parascript delivers the world's best Image Recognition Engine. Employing patented digital image analysis, handwriting analysis and advanced pattern recognition technologies, Parascript improves important business operations in areas like forms processing, medical imaging, postal automation, signature verification and fraud detection. The powerful Parascript engine processes over 100 billion imaged documents per year. Fortune 500 companies, postal operators, major government, and financial institutions rely on Parascript products, including the U.S. Postal Service, Bell and Howell, LLC , Fiserv, Elsag, IBM, Lockheed Martin, NCR, Siemens, Xerox and Burroughs. Visit Parascript online at http://www.parascript.com.

For more information, please visit www.parascript.com or call 1-888-PSCRIPT (1-888-772-7478) or (303) 381-3100

[i] http://content.nejm.org/cgi/content/full/330/8/540

[ii] http://www.thenewsenterprise.com/cgi-bin/c2.cgi?053+article+SpecialSection.HealthyLiving+20090409112243053019