

## Context Is King

### Improving Automatic Character Recognition Using Data

By Brian Ball

ICR (intelligent character recognition) has been around for a while and OCR (optical character recognition) even longer. Recent developments in character recognition enhancements promise to yield further cost savings from these advanced technologies.

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 and so forth. Recognition software takes advantage of choice narrowing in the form of ICR.

But this is not news. What is news is the opportunity to take 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. 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*, up to 40 percent 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. 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, date of birth or gender. This database can be used for reference during recognition.

A nearly 63 percent 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, 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.

This significant reduction in keystrokes and decrease in operational time was realized without introducing new errors.

The new performance results were realized from the use of dynamic dictionaries that allow the system to accumulate valid answers over time. The possibility 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 of just last names associated with that zip code.

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 and so forth). 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 percent 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. Still improvements in effectiveness are 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.

*Brian Ball is vice president of Forms Processing at Parascript. He can be reached at 303-381-4126 or [brian.ball@parascript.com](mailto:brian.ball@parascript.com).*

## Methods

### 1. Blind double-key verify to reduce keystrokes and increase accuracy

- a. Matching manual data entry with recognition answer reduces keystrokes, facilitates faster data entry, & enhances accuracy.
- b. Low confidence recognition answers serve as the first pass of keying and can be compared to multiple passes of keying.

### 2. Accumulate dictionaries for repeat occurrences

- a. On many forms, several data fields will have answers repeated throughout the life of a project.
- b. Field data can be stored as a record in a database for each unique instance (name, address, phone, DOB, gender).
- c. The database is then used for reference during recognition.

### 3. Use dynamic dictionaries for recognition

- a. Dynamic database lookups further refine the possible list of answers, boosting recognition rates and accuracy.
- b. Any field, such as zip code or phone number, can be used to dynamically create other databases on-the-fly.
- c. A zip code answer can be used to create a dynamic dictionary of one answer for last name and so forth.

### 4. Pre-fill city, state and zip for address data entry

- a. For unrecognized full address answers, partial answers (city, state and zip) are frequently available and correct.
- b. City, state and zip can be pre-filled for data entry, requiring data entry for the street address line only.

### 5. General demographic table

- a. Address or phone gives unique index for records related to a specific field such as name, address, DOB, phone, gender.
- b. Commercially available marketing/demographic databases can increase context and can be used to create lookups in real-time.