Using Image-Based Document Classification and Extraction

Current Situation

- Large organizations continue to process enormous volumes of paper-based information as well as large volumes of relatively unorganized electronic documents and files;
- The cost of processing a paper invoice is ten times that of an invoice handled electronically, and the cost of classification and organization of electronic files runs from $.05 cent to $1.00 per page and even more for extraction of information therefrom;
- It is imperative for large organizations to focus on the conversion of paper-based data into a computer-readable form and, also, to classify and organize them as well as all other electronic files;
- Only now is it possible to significantly decrease the use of human labor for these activities while maintaining data accuracy and speeding the extraction and introduction of data into otherwise automated business systems;
- Goals: lower costs; enhance competitive position; and become more responsive to both customers and vendors.

Capture Systems

- Software systems designed to automate data conversion processes are often referred to as capture systems;
- The typical capture system incorporates a number of subordinate processes that range from document imaging and image enhancement to document interpretation, data extraction, data validation and, finally, publication of the data to business systems;
- Most challenging of these processes is document classification, due to the wide variety of document types, layouts and content present in today’s business environments;
- Capture systems offered by IBM, Kofax, EMC, OpenText, Top Image Systems, Abbyy and others require significant manual activities such as the creation of detailed and complex templates to reflect documents to be analyzed and associating such documents with the template. There is much guess work and the validation process is flawed. Also most systems have Optical Character Recognition (“OCR”) technology at their core and OCR is slow, linear in approach, and often inaccurate.
- BR Visual Similarity technology automatically classifies and clusters documents that are visually similar as they would be seen by the human eye, without human intervention or template making, and whether they had been stored as native files or otherwise.
- In the case of business or project forms such as invoices, the technology automatically identifies the static features of the form and differentiates them from the variable information associated with them. So, in the case of a box for a “name”, it would classify based upon the position and the word “name” and not the actual name contained in the box.

New Automated Capture and Classification

- Fully automated classification of documents to clusters is now available and at a reasonable cost;
BeyondRecognition ("BR") deploys a glyph based engine to interpret and then compare the visual characteristics of a document without requiring that the textual content of the document be obtained through OCR processing.

- A glyph is any non-verbal symbol that imparts information, i.e. in this case discrete groupings of pixels in forms and shapes;
- As such, the engine only understands and uses these forms and shapes and how they relate to each other based upon their Visual Similarity; and
- The engine works independently of language or text.

This approach yields the following very positive results:

- BR is language agnostic…handles any language and this has been verified on 40+ today;
- BR is Symbology agnostic…recognizes and relates non-text elements;
- BR uses Glyph similarity rules … makes decisions independently of text and based upon non-text elements;
- BR clusters to visual similarity…works on all kinds of documents; and
- BR is fast and more than 99.9% accurate.

BR deploys Glyph dictionaries and adds dictionary entries specific to the content being processed. These dictionaries provide for rapid and consistent identification of items for visual similarity. They also provide a basis for universal error correction and other operations outlined below.

This image based approach offers important speed, accuracy and maintainability advantages over approaches that rely solely on OCR-based classification as outlined below.

**Using BR Visual Similarity to Classify Documents**

- Classifying documents according to their visual similarity is a way to normalize documents that may have been created or stored in many different formats.

  - For example, a report could have been created at various times as an Excel spreadsheet, a Word document, or an Acrobat form, yet have almost exactly the same appearance. The spreadsheet or Word document could have been “printed” to PDF or printed to paper and then scanned as a TIF or PDF document.
  - Normalizing documents based on their final appearance lets like-kind documents to be classified or clustered together despite differences in the files in which the documents were stored or communicated.
  - It also enables the development of data extraction or coding rules that are not dependent on accounting for all the different ways that documents can be created, stored or communicated. In the example from the above paragraph, one set of extraction rules works for the visual representation of the form that was created or communicated in the Excel spreadsheet, the Word document, the Acrobat form, the scanned TIF image, or the scanned image-only PDF file.
  - This means that Visual classification works equally well on native files (e.g. Word, Excel), scanned document images or both.

- Visual classification does NOT involve an analysis of the text that can be extracted from documents – it is based on a visual comparison of the way the documents look when printed or displayed as intended by their originators.
BR uses, on average, about 1500 glyph references per page as points of comparison to identify visually similar documents. The process is very reliable and typically results in hundreds of thousands of documents being grouped/classified into hundreds or a few thousand clusters.

Once documents have been visually classified, the different classifications can be routed as appropriate for downstream processing and/or BR’s visual coding can be used to extract different data elements from different clusters or classifications. Visual classifications permit project managers to be fully aware of the types of documents contained in a collection at the very beginning of a project.

Important Practical Advantages over OCR based Classification Systems: Speed, Accuracy, Scope of Application and Ease of use.

Why does OCR fail to meet the higher standard created by BR? Even including OCR based specialized point solutions such as OMR and ICR, OCR methodology disassembles document text and characters to a one dimensional linear string that does not carry forward the spatial and all relational characteristics of a document. BR’s approach is far more complete—producing an average of 1500 glyphs per page and retaining all relative and absolute spatial information. It knows where everything came from on the page and its position relative to everything else on the page. That goes way beyond the capacity of the OCR based strings. With OCR the original spatial relationships and information is lost. Unlike BR, with OCR no positional search is possible and only Boolean or character based delimiters are available. With BR, easily created but more complex delimiters and filters, including those based upon position, are available in addition to Boolean and text logic for more discrete rule creation and ultimately better decision making.

The resulting advantages of BR:

- **Speed.** Image-based page classification can be performed within a small number of milliseconds while full-text OCR typically requires a minimum of several seconds. Speed advantage: over 100-to-1. When millions of pages need to be classified, this large speed advantage creates large cost savings.

- **Accuracy.** OCR-based classification often fails because of OCR errors. The accuracy advantage of image-based classification derives largely from its use of redundant visual features. Thus, the inability to detect some of the image features, perhaps due to scanning imperfections or stray marks on a page, does not prevent correct classification. In contrast, OCR errors commonly do prevent correct classification since the same level of information redundancy does not exist within the textual content as within the visual content.

- **Scope of application.** Image-based classification that does not require text finding or text recognition can be applied to documents written in all of the world’s languages, not just those for which an OCR is available, because in this case visual appearance determines the classification, or grouping, of documents. This breadth of application is frequently of importance in applications relevant to governments and multinational organizations.

- **Ease of use.** OCR-based classification systems typically require the manual construction of many rules that describe how to classify individual pages or documents on the basis of their
textual content. In practice, these rule sets commonly number in the dozens and are difficult to maintain because the rules often interact with each other. There is no such difficulty with image-based classification methods.

- **Example uses of image-based classification:**
  - **High-Speed Scanning.** Modern high-speed scanners typically process a page in several hundred milliseconds. It is common for the processed pages to be physically binned as they are processed. Thus, a decision concerning a page’s destination must be made within that very short timeframe. As a result, full-text OCR classification cannot be used. However, image-based classification, combined with region-based field recognition for greater flexibility, can easily operate within the required time and serve as the basis for the binning function. In the past, high-speed binning was performed using very limited types of bar code, mark sense and field-based recognition. Image-based classification opens the door to a much broader set of high-speed document processing applications.

  - **Forms processing.** The automated processing of both structured and semi-structured forms requires the ability to locate automatically the data to be extracted from the document. The key: automatic determination of the document’s identity. Given the speed, accuracy and flexibility of image-based classification, it is an ideal solution to the document identity problem and, as a result, enables large-scale processing centers to handle a wide variety of form types without requiring manual presorting.

  - **Specialized Data Products.** In many cases today, data must be manually entered because of the nature of the document’s content, its image quality, or its layout complexity. These factors often make accurate conversion using text recognition (OCR) impossible. In these cases, it is common to prepare keying instructions for individual document types and then transmit the instructions and the corresponding set of document images to individual data entry specialists. Image-based classification can be used to automatically group similar document images together, which occurs more rapidly and accurately than if it were performed by hand. This is especially true when the quantity of documents is large.

**Visual Classification and Clustering Enables Accurate and High Speed Extraction of Information from Groups of Similar Documents**

- By and large, data extraction from documents such as business and other forms does not work except for situations where a common set of rules can be applied against a uniform set of data. Classification of documents in existing capture applications is imprecise and yields only limited extraction capabilities. From an actual customer of a competing technology:

  "...had a conversation with [competitor X] about what drove their "automated indexing" and classification... They have a pretty good, inexpensive solution for "header" information that they offer, but they don't touch the line level capture as it takes ~5 minutes per vendor to set up (without guaranteed success) and their average client has thousands of vendors (read: payback on the heavy setup isn't seen as a good ROI when savings vs. manual processing is put up against time to"
- BR’s visual classification creates uniform sets of data automatically – visually similar documents are placed in the same classification or cluster. Once that is done, all the user has to do is drag a cursor over the data elements to be extracted. The function is automatically applied to all documents in the cluster.  

To repeat, this process is as simple as clicking and dragging the cursor to draw a box around the information of interest. No need for separate solutions for each vendor as automated clustering eliminates the need.

- The visual coding associated with the classification/clustering activity examines visual representations of the documents which serve to normalize the information so that individual rules do not need to be created for each type of source file – e.g. Excel, Word, PowerPoint etc. -- let alone each separate file.

- Full-text only approaches to classification and extractions fail similarly as they do not permit (a) sorting by a variety of factors, (b) field-limited searching, (c) report formatting, (d) performing statistical or mathematical operations on certain types of data, or (e) comparing document values against values stored in business support systems. The limitations of full text are most pronounced in scanned paper documents but are still significant in documents stored in native electronic files where the associated metadata may be incomplete or incorrect if present.

- In addition to the positional identification of information to be extracted, BR provides operators with a full set of delimiters to specify what information is to be extracted. Delimiters can be inclusive or exclusive and can denote the beginning and ending of what is to be extracted. Some examples include beginning text strings and the number of lines to include.

- BR also provides a number of filters that can be used to format specific fields or values, e.g. to use commas in numeric values or to round up dollars and cents to just dollars.

- The information identified positionally, with delimiters and/or filters is easily organized and tracked. As soon as the operator specifies what data to extract, BR displays the values that will be extracted from the documents in the cluster by putting them in a column in a spreadsheet-like display where each row represents a document in the cluster.

- BR offers real time feedback by immediately calculating and displaying what the current rule will yield in terms of extracted data; BR enables the user to see right away how the rules operate and avoids problems of rules not operating as expected.

- The output from BR’s visual coding can be validated data in the form of a delimited file or XML tagged to the user designated schema.

Management and Clean Up of Document Sets

- Single-Instance Editing
  - Huge advantage of BeyondRecognition’s glyph clustering approach to text conversion includes single instance editing, correcting or otherwise changing text or other symbols, and having the changes automatically effected or “cascaded” throughout the dataset. Users can leverage BR’s glyph catalog to reduce required editing by orders of magnitude compared to legacy OCR applications.
As an example, consider the following diagram that represents that a number of documents were processed and that many of the glyphs were placed in multiple clusters that were all associated with the text value “a”:

![Diagram of documents, glyph clusters, and text values showing the process of editing from 'a' to 'q'.]

- Assume in this example that the last glyph cluster on the right really should have been associated with the letter “q”. In some systems, that could require going back to each of the text files associated with the underlying documents and editing individual words. Not so in BR. With two keystrokes, the text value associated with the last glyph cluster can be changed from “a” to “q”. Because BR’s global glyph catalog remembers all the places the glyphs from that cluster had appeared, the edits are propagated throughout the collection, as shown below:

![Diagram showing the propagation of the edit from 'a' to 'q'.]

- The “force-multiplier” aspect of single-instance editing is sometimes described as being “cascading” because the one correction flows back to potentially hundreds or thousands or even millions of pages. It is also described as “persistent” because the change in the text value associated with the glyph cluster persists or remains even for documents that are processed in the future.

- With a few hours’ effort, character and word accuracy can be raised well above 99% – and, once again, that editing benefits documents to be processed in the future as well as those that have already been processed.
BR Requires Negative Learning: Dispelling of Myths and Beliefs About How Things Work

- BR destroys widespread implicit assumptions in the market about the way things work; “Knowledge” is so ingrained that it never gets questioned and the person does not even examine new paradigms that are contrary to those implicit assumptions, such as these FALSE ASSUMPTIONS ABOUT THE MARKET:
  - **FALSE ASSUMPTION 1**: 95% or less word accuracy is acceptable when converting images to text;
  - **FALSE ASSUMPTION 2**: The only way to achieve above 99% word accuracy is with extensive manual editing;
  - **FALSE ASSUMPTION 3**: The only way to search through documents is by the use of textual search terms;
  - **FALSE ASSUMPTION 4**: The only way to accurately classify documents is by human review;
  - **FALSE ASSUMPTION 5**: Delays of days are acceptable when processing thousands or tens of thousands of forms a day with <95% accuracy;

**WHEN THE TRUTH IS** **YOU CAN HAVE 99.9% ACCURACY AT 250,000 PAGES PER HOUR ON AN AUTOMATED BASIS.**

**SUMMARY:** As a review, BR performs these primary functions:

- **Visual classification.** BeyondRecognition (“BR”) automatically classifies or clusters documents based on visual similarity. (e.g. Word, Excel, or PDF) or scanned paper copies.
- **Visual data extraction.** With BR, extracting specific data elements from the classifications or clusters can be as easy as drawing a box around relevant portions of the document;
- **Using glyph cataloging,** BR is able to provide textual representations of documents that have initial word and character accuracy levels far exceeding legacy optical character recognition systems.
- The glyph cataloging system enables single instance, cascading, persistent error correction that enables even higher levels of word and character accuracy and makes brute force linear editing a thing of the past.
- Single keystrokes can make ALL instances of a change (correction or edit) in a document, sometimes affecting literally millions of characters in hundreds of thousands or millions of documents.
- BR has automated the logical separation of documents that happened to be copied to a single file by knowing where document boundaries should be i.e., without beginning and ending document tags and, therefore, freeing scanner operators from having to insert slip sheets and minimizing prep time.
- **BR catalogs the coordinates of every glyph,** and those coordinates enable BR to provide highly accurate redactions of words or patterns – “GPS” for documents and content, if you will.
- And, more GPS for documents and content, BR can locate all instances of a given glyph, whether or not it has a textual representation or not.
• Individual forms are automatically placed in the same clusters where BR can identify the unchanging part of the cluster (the form), and the parts that change from form to form (the data). Without human intervention, BR creates tags based on the text that appears in position to box or form labels and extracts the data entries to XML using those tags in XML hierarchy. Users can modify the tags or map them to a user designated XML schema. As such, the XML can be exported directly for use in enterprise applications.