PARASCRIPT®

Strategic White Paper

Intelligent Document Recognition (IDR) – Advanced Technology for Increased Productivity

Takeaways

- Understanding IDR - data capture, classification, recognition and validation
- Review of business processes, requirements, and output needs to evaluate IDR
- Awareness of more advanced IDR capabilities

What's in this Paper?

6273 Monarch Park Place, Longmont, CO 80503 USA
T: 303.381.3100 | toll free: 888.225.0169
F: 303.381.3101 | info@parascript.com

parascript.com

Introduction

In today's increasingly complex business environment, more companies, government agencies and other organizations process higher volume of forms and documents with increased levels of complexity. The ultimate goal is to streamline document processing and move critical business information into business processes quicker and with greater accuracy. As a result, Intelligent Document Recognition (IDR), or capture technology that includes data capture, classification and recognition is often considered in order to increase productivity and reduce costs. IDR allows organizations to reduce manual data entry; automate data capture from all documents and data types; and employ solutions that are easy to implement, affordable, fast and scalable.

IDR continues to evolve. Just as cell phones have evolved into smart phones with many more features and capabilities than just a few years ago, so has capture technology made significant advancements. This paper describes these many innovations and key factors to understand IDR:

- Basic knowledge of current terminology and technology
- Awareness of more advanced features
- Value of capabilities in relation to current and projected processing requirements

Intelligent Document Recognition

Intelligent Document Recognition (IDR) is a complex solution to accomplish diverse processing needs based on data use, form structure, data type, document classification, business rules workflows and application integration requirements. When selecting an IDR solution, consider business processes, content requirements, and output needs. For those not familiar with IDR or those who have not investigated the technology over the past few years, it is important to understand current capabilities of this software, as they are changing quickly.

IDR is comprised of four major functions:

Capture – Document content is scanned and digitized to 'capture' all the data in documents or forms for processing. Data is identified and extracted from various document types that can include structured, semi-structured, and/or unstructured forms or from selected data fields within these types of documents or forms.

Classification – Classification is based upon a number of technologies and techniques including document layout, presence of images or logos, regions of data, presence of keywords or patterns, and the content itself.

Recognition – Technologies to recognize data include OCR (machine text); ICR (constrained and unconstrained), and handprint, including cursive; 1D/2D barcodes; and OMR. Data can also be read from zones or areas and tables. Unstructured recognition employs advanced location and recognition techniques using dictionaries and pattern rules to identify data and perform extraction. Data output can be exported per the users' requirements that can include to databases, to file systems as text or XML, as well as to content management systems such as SharePoint.

Validation – Data validation workflows can be page centric where all the fields are presented on the page or field centric where only the fields that require review are displayed. Validation can be organized based upon specific need such as accuracy optimization, throughput optimization, or to handle security for sensitive data.

Functionality to Consider – Value to be Gained

During the IDR evaluation process IT professionals responsible for evaluation, selection and implementation should determine the need for a solution's ability to:

- Capture all types of document information without additional costs
- Handle all data structure capture within a single workflow - structured, semi-structured, and unstructured
- Allow for classification and configuration of document types/classes and workflows without having to move through multiple applications or user interfaces
- Provide a comprehensive test capability for data extraction with easy and comprehensive access to results and statistics
- Allow for efficient data validation including data entry and correction
- Support confidential or private information including data redaction and assignment capability to limit access to specific data reviewers

IDR Technology Solutions: Details & Functionality

IDR users require a solution that allows their organization to streamline document processing with results that are faster and more accurate.

Data Types

Recognition of machine print, handwriting including handprint and cursive, barcodes, OMR marks.

OCR – Optical Character Recognition is software that converts printed characters or machine text into digital files and has been in use since the 1960s.

ICR – Intelligent Character Recognition, a step above OCR, reads constrained handprint - printed characters in a form, box or other document limitations.

"Advanced" ICR includes the ability to read unconstrained handprint and any other type of handwriting, including cursive. This is a much more sophisticated and advanced technology because it interprets the patterns of human writing, unique to each individual. Only selected ICR packages offer these capabilities.

IDR users require a solution that allows their organization to streamline document processing results that are faster and more accurate.

Document Structure

Documents for capture can be structured, semi-structured, unstructured or a combination.

Structured documents, most commonly forms, provide data in a uniform manner and tend to be transactional. Many structured forms are designed and implemented by a standards or regulatory body for use within defined industries or for a specific purpose such as health claim applications for Medicare/Medicaid, tax returns and other standardized forms.

Semi-Structured forms have elements of structured forms in that they include uniform data but also have structure variance based on the organization and its needs. Organizations use semi-structured forms to create and manage their own form layout for documents such as checks, invoices, purchase orders, policies and others that require customization. Semi-structured forms also collect and deliver transactional data but due to unique fields, this form type presents challenges in transferring data into processes and systems.

Unstructured documents are not represented in a pre-determined form but have data that must be captured for processing. Examples include customer correspondence by letter or email, contracts, doctors' notes, patient records, manuscripts, court records, video, images and more. These documents may include information relevant to a transaction or process but in a less defined way. Most provide context for a transaction – for example, a customer complaint. The challenge to process unstructured documents is that the software must identify the type of document and then extract the most relevant data.

Dynamic Document Classification

Documents are classified by a variety of methods using data and image analysis on all type of forms and documents:

Data Type – Classifies information by data types; i.e. machine print, handwriting including both print and cursive and combinations of data types.

Image-based – Processes images through comparison to samples; performs layout and graphical analysis to determine document type; and utilizes image comparison algorithms independent of language and format. This is the best method to use for document sorting of standardized layouts.

Region of Interest – Classifies images by examining a specified location or graphical element in an image. This approach can be used independently or in combination with other forms of classification. It is best for classification of variable layouts.

Keyword Analysis – Uses either structured or semi-structured techniques to locate content and classify it by pre-determined keywords.

When evaluating IDR, it is important to identify the type of data and documents to be captured and extracted. Different solutions offer different features and approaches. Users can increase productivity by combining technologies to address specific processing requirements or may benefit from a single solution that can process multiple form types (structured, semi-structured, and field-based data) and all data types (machine print, constrained and unconstrained handprint, and cursive).

	Document Type		
Data Type	**Structured**	**Semi-structured**	**Unstructured**
Tables (fixed, variable, unformatted)	Locate tables and process data. Some solutions can recognize tables automatically	Using rules, automatically locate cells of data regardless of variability	Keyword location and associated data
Machine print (OCR)	Process structured fields	Variable located fields using keywords or patterns	Process full page documents
Constrained hand print (ICR)	Boxes, combs or individual fields	Using machine print as keyword locator, handwriting can be recognized	N/A
Unconstrained hand print and cursive (Advanced ICR)	Fields with multiple handwritten words	Using machine print as keyword locators, handwriting can be recognized	Locate handwriting anywhere on a page and compare to keyword dictionaries

Data Entry and Validation

Most capture solutions employ document-centric reviews. While support can be provided to focus on field-centric views, the entire document is processed and/or displayed, resulting in inefficiencies and potential security risks.

Advanced capture solutions support both document-centric and field-centric validation enabling routing of specific fields to identified users based upon security, validation accuracy, keying speed, and other customized specifications. This enables organizations to increase processing speed and protect the privacy of sensitive information.

Real-time Statistics

Managers and administrators need real-time batch and user statistics including fields successfully recognized verses those that need to be keyed. Also required are user performance, recognition accuracy statistics and other KPIs.

Data Analysis & Testing

Users want an easy to use GUI dashboard format to view results of classification and recognition, set confidence thresholds, compare against truth data, and see immediate results prior to production.

Output Options and Requirements

Expanded API capabilities add new functionality and ensure integration with other systems. Output information should be made available to SharePoint, as well as to databases through common file formats.

Business Rules and Workflow

.NET scripting and full access to recognition context is important to business analysts who want to create document processing rules and workflows leveraging easy-to-use VB.NET or C## syntaxes. Additionally, developers want access to the .NET API to incorporate customized functions for a seamless process.

Conclusion

For those organizations with high document processing requirements, IDR is a practical and cost saving solution. To adequately evaluate IDR and implementation solutions, it is important to assess current and projected processing requirements and evaluate software functionality accordingly.

For more information on Parascript IDR solutions,
visit parascript.com and learn about *Form*Xtra Capture.

About Parascript

Parascript is a leading developer of cursive, handprint, and machine print recognition solutions. Leveraging digital image analysis and advanced pattern recognition, its software enables business automation in forms processing, postal and financial automation, and fraud prevention; and supports cancer screening in medical imaging. Parascript's award-winning technology draws on a proven 15+ year track record and processes billions of document images annually. Fortune 500 companies, postal operators (including the U.S. Postal Service), major government and financial institutions rely on Parascript products, which are distributed through its OEM and Value Added Reseller networks, including partners such as: IBM, EMC, Bell and Howell, Fiserv, Selex Elsag, Lockheed Martin, NCR, Siemens, and Burroughs. Visit Parascript online at http://www.parascript.com.