Cognitive Forms
the ultimate document capture system

www.cognitiveforms.com
Introduction

Today tens of millions of similar documents, including postcards, money orders, customs entries or tax return forms, bank checks, ballots and all sorts of application forms are added to Information Management Systems on a daily basis. Hundreds of thousands of operators perform monotonous routine sequence of actions which are to review next document page promptly, find and read the filled in data and retype it rapidly. A professional typist makes several hundred keystrokes per minute, but usually an operator can input about 10,000 words or numbers for a shift.

As an alternative to manual data input, Automated Data Collection technologies are used that have several distinct advantages: modern scanners are able to process up to 200 pages per minute, Optical Character Recognition software (OCR software) could «read» a few hundred characters per second and with no break for lunch needed.

Beside data input cost savings and higher quality, document scanning and recognition solutions have another significant advantage which is a correctly identified document flow, including recognized information and graphic images. The one may form the basis for an electronic filing system where you can search, retrieve, transfer and print documents (with a quality similar to document photocopying). The developing global computer networks providing remote access to those filing systems gradually highlight the technological advantage.

Document Types

There are different types of paper documents that we use in various life situations: business forms, application forms, invoices, delivery notes and payment orders; insurance polices, contracts, letters, identification documents, etc.

Documents can be distinguished by style and formatting and therefore by a level of data processing complexity:

• Structured documents with fixed field location and dimensions. The documents form is strictly determined.
• Flexibly structured documents, or «flexible documents». The documents have the same horizontal and vertical structure and sequence order of their parts, but parts may vary in size or scope.

A document can have several parts with different formatting styles (e.g. invoice). Those can also be a single and multi-page documents, hard page break and «flow» documents. Documents can have a colored background. It could be useful for ease of work when document contains a number of various fields. If an automated data input system is used such background you should better remove for more comfortable work.
The System Development Approach

There could many different reasons you will need to use Automated Document Capture systems. However, one of the main reasons to use Automated Document Capture systems is to increase quality of data input process.

For structured documents, the Cognitive Forms Automated Data Input system can solve a number of problems which are often quite complex and with no universal and effective solution. Those also could be problems of document graphical image scanning and processing (binarization, slope definition, graphic primitive selection), structured document simulating and identifying by a set of predefined structure templates, recognition of printed and hand printed documents, contextual analysis and logical control, etc. We will not go into details telling of problem solving issues while describing the system, but only show how the solutions combined into functional modules provide a set of tools based on which conveyor flows of structured document large volume data input can be created.

To solve the above problems a set of principles was formulated for the system while developing the system.

1. Hierarchical organization - each system level is isolated and runs over the data came from the lower level, using its own prior knowledge and techniques in the context established by the higher level.
2. «Fuzzy» recognition - all methods of detection are not clear classifiers, each one does not give one reply but a list of alternatives and their evaluations, sorted in rating descending order. Therefore the recognition system is more error resistant and provides possibility to methods of syntactic and structural pattern recognition make the most difficult final decision, which have more field prior information. In addition, it allows us to create a more accurate error determination mechanism.
3. Modularity and Standardization - modules that solve a particular problem and hierarchically separated from other one, the module inputs and outputs are standardized and strictly regulate any changes.
4. Adaptability - modules settings are moved to parameters when possible. While developing algorithm, considerable attention is paid to repeated application.
5. Sustainability - modules of the system and the whole system can adequately handle cases which are not initially provided for. For example, if an unfamiliar symbol is received by recognition system input, the system should refuse it.
6. Continuity - the whole system and all modules are designed to operate according to 7 / 24 formula. Automated processes will not require constant human control.
7. Scalable and Distributed Computing - task that can effectively performed in parallel way and those must be solved taking into account the parallel execution. When solving problems you should avoid solutions that do not scale well. The system implements a cluster parallelism type technology.
8. Quality and Efficiency - formulated method estimating system quality and efficiency. There are tools developed assessing and optimizing the system parameters to achieve an optimal quality / efficiency balance of each specific task. Assessment methodology, factors that affect the quality and efficiency, and quality and efficiency improvement methods will be discussed later herein.
Fig. 1. Document Capture Process Main Stages
The System Summary

The Cognitive Forms System is a software package designed to organize process flow of large scale capture of document templates. System modules are installed on computers connected in a local area network, and those while interacting with each other line up a data processing pipeline that allows you to enter up to 70,000 and plus pages per day. The system includes 12 executable files (main and auxiliary) and 149 dynamically loadable libraries. The base system is implemented in C++, the total package consists of about 3 million rows.

All the Cognitive Forms work is divided into two stages: Document flow preparation and processing. The first stage is supported by a Cognitive Forms Designer module. This module describe document template. The program allows you to create a document template from an existing one or develop a new document template.

General scheme of document processing include such process stages as documents flow scanning, recognition, verification and result export to an external information system (Fig. 1).

The basic stages are as follows:

Scanning
At the scanning stage filled documents templates are loaded into the scanner resulting with a set of image files (usually in TIFF format, compressed to CCITT Group 4 standard).

Sorting and Packaging
At the stage of document batch sorting, combined in one package to improve scanner productivity by separation special templates which are on the first document pages. Key fields are identified to control document completeness. If a document is complete so it will be passed to recognition stage otherwise to sorting adjustment stage.

Sorting Stage Result Correction
At the stage of sorting result correction, an operator is notified about document batches which do not meet completeness criteria or if any structure disruption (for example, some pages stuck together while being scanned). The operator either corrects the error and sends the document once again for recognition or scans/exclude it.

Recognition
At an identification stage, Recognition Module determines the type of the next symbol pages, selects input fields, set in automatically selected template, recognizes and stores in a specialized database information on document type, field values and graphical arrangement. The database is transferred to a verification stage.
Recognition Result Verification
At verification stage, an operator is notified about document fields where system failed with text recognition, or has no confidence in identification reliability. The operator is also notified about fields that do not meet document completion rules. The operator may either discard a document sheet or as a whole or export it to an External Information System (database, banking system, etc.)

Export Data
Export processing stage completes the document processing cycle when a document is converted to a standard format (DBF, XML,) and sent for processing to an External Information System.

The System Main Components
The Cognitive Forms system consists of several modules, each of which performs work at a certain stage of data input automation, including document preparation and scanning from one side to interaction with external information systems and all input data process administration from the other. Thanks to a modular architecture an each input stage (scanning, image processing, recognition, verification and exporting) can be easily scaled and increased in its performance by increasing module number. The main process flow is supported by:

- Scanning modules: CF ScanPack batch scanning module and CF AutoScan automated page scanning module;
- CF Processor automated template identification and detection module;
- CF Editor recognition results verification module.

Including some other software programs designed to work with a document package rejected by automated processing stages.
Documents Scanning Modules

Scanning modules link up document graphic image flow for automated processing. Document electronic image of a paper document is received by scanning document sheets using either a conventional scanner or a special scanner with automatic paper feed. To implement this function, the system provides a special set of software scanning modules i.e. CF ScanPack, CF AutoScan, which serves as the communicators with scanner, save image and documents flow formation.

CF ScanPack module is designed to scan documents in batches (Fig. 2). The package passed through the scanner. The image of each scanned page is a separate file. All package files are written to disk in one directory. The name of the package formed in accordance with the configuration settings. Friendly user interface will let you configure the scan mode, resolution, scanned document size and other input parameters.

The module implements the CF AutoScan page scanning (Fig. 3). After launching the application, scanning process starts automatically as soon as a document sheet gets to a scanner tray. After scanning the page, TIFF file is created. File is placed to the directory specified in the module settings. This application allows you to configure scanning mode, resolution, scan area size and other operating parameters.

The process overall functional diagram is as follows: Once the module is launched, an operator selects the module configuration settings. An Operator then puts subsequent document batch / one document to a scanner and if a scanner has a tray filling detector it automatically starts the scanning process and if such detector is not provided, then the operator should press button initiating scanning. Once a subsequent page scanning is completed special image processing algorithms start to improve document image quality. The image is then saved in a specified format and once the package scanning is finished, it goes to the next stage. As it is a semi-automatic process and an operator work is required (for example, to load paper into the scanner), scan process parrelling is possible within local area network. The phase is resulted in a document image portion or package which goes for further processing.

Fig. 2. An example of the batch scanning implementation

Fig. 3. An example of the page-by page scanning implementation
Image Processing Module

Recognition Server Module (CF Processor) is designed for automated processing of scanned documents. Document type, data input field and content are identified while processing is performed. All server work is logged. It is important to note that the system automatically identifies document with no manual document sorting, identify types of documents in the stack during scanning. It does not require application of reference elements in the document. Module:

- Preprocesses a graphic image (binarization, inclination angle definition, cleaning) and graphic primitive selection i.e. text strings, numbering lines, checkboxes and other standard elements (Fig. 4);

- Automatically identifies document template that is most suitable for processed image file by associating a document template and selected set of image file primitives;

- Identifies and recognizes input fields that control document completeness;

- Keeps track of document pages and compliance of the document structure to template syntax that defines possible variants of the page sequence one after another.

Fig. 4. An example of document image processing
If a document is incomplete, it is sent to a special stage of processing or re-scanned. Then, while document is recognized, CF Processor:

- Identifies document input field (printed text (OCR), handprinted text (ICR), label / checkboxes (OMR), one-and two-dimensional bar codes);

- Recognizes typed or hand printed text, check boxes, bar codes, monitors signatures and stamp presence, cuts file image fragments which correspond to photography fields (Fig. 4);

- Perform recognition result contextual analyses and logical control, choosing most plausible file image field interpretation version (taking into account «geometric» descriptor and «logic» assessments, text logic context appropriateness (Fig. 5).

Fig. 5. CF Processor Module Main Window during Sequent Document Automatic Processing

The module ultimately sends recognized information for verification phase, Module supported CF Editor.
Recognition Result Verification Module

Verification Module (CF Editor) is intended for recognition result processing (Fig. 6). CF Editor Module is used for sampling or adjustment of the document field values for those cases where a correct data input can not be performed automatically. For example, if the field is filled in violation of the document completion rules or graphic image quality does not guaranty uniqueness of the recognized value.
The verification process goes as follows. The next package is loaded; an operator sees an image and an electronic template with recognized data and fields that have not been monitoring the reliability and logical consistency are highlighted in color to attract operator’s attention. The operator while moving the focus between e-form fields see an error diagnostics and either fixes the bug or if an error is not corrected decides to transfer the document to «bad» document processing stage. When moving through fields, the module automatically highlights image field frame. To increase operator efficiency, there are two specific input modes: Only field passing mode not passed control, and blank field skipping mode. In addition, if there is a dictionary containing field valid values, you can specify in interaction model description section to inform an Operator about the dictionary and allow the operator to insert a field value from the dictionary (Fig. 7). Figure 8 shows a CF Editor main window screenshot during barcode verification process. You can see template field requiring confirmation or correction at the bottom and corresponding section of image document at upper part. Depending on your current field properties, it can be associated with vocabulary, formatting methods, such as «date» or «phone» or other special editing and monitoring tools.

Once verification process is completed an operator will be asked to either postpone a proposed package or pass it to the data export stage. The complete script for operator and verification module interaction is specified by the document interaction model.
Data Export Module

Data export process provides converting interaction model data to content model data. Those are then verified as to correctness in accordance with content model. If data tested is completed, it is transferred in accordance with description specified by the content model. Image transfer is carried out in a similar way. If data is not considered valid, relevant sections are marked «not valid» and a package is returned for additional verification. Document package containing images of documents and corrected documents, including a variety of templates are considered to be input ones. Automated process carried out jointly with verification module and may require operator interaction. The module is implemented as a set of dynamic libraries. As per document flow stage description, paper documents should be kept only till the end of sorting, all the other stages go with no paper needed which can significantly improve efficiency.

Electronic Document Form Design Module

CF Designer module is intended to create electronic document forms. The electronic form contains a full list of document elements. The list indicates each element properties, including type, location and dimensions. For elements such as checkboxes and input fields, their completion rules are optionally specified. The form can be created basing on a blank document form. Elements one after another are drawn on a template or copied from previous form. The other option is to create an electronic template coping paper document image. Template creation based on a scanned document includes automated template processing i.e. binarization, inclination angle definition and removal, graphic primitive selection, including numbering lines, text strings, and checkboxes. Identified primitives can be added to the template and modified. The templates cover the basic properties of the document and its elements i.e rigid / non-rigid format, layout, alphabet and style and other properties used during further document processing. When you save a template as a pdf file it can be used as a replication template. When you save template in the Cognitive Forms Internal Data Format the one is used as a template during document recognition process.

Integration with external systems

The document flow is aimed at template transformation to electronic format for further data processing by a corporate information system. ERP, CRM, DMS, EAS and other storage and processing systems are usually used as information systems. The Cognitive Forms system can be integrated with information systems in two ways: Input module connection on directory import level and input data export; Direct integration using API user interface.
Export Module Installation

You can export data using export modules that are part of the basic product package to the following formats i.e. XML, DBF, CSV, TXT or use ODBC connection. Specialized export modules could be developed on request.

You should import directory file to provide document primary control at data input stage. The directory file examples could be customer lists, bank account numbers, etc. The directory files could be uploaded in text format or DBF.

Installation via API

Programming interface (API) allows you to connect document scanning and recognition functions, access document field values and graphic image. The API also provides access to Cognitive Forms Internal Data Format. The API includes a set of dynamic link libraries (DLL) and ActiveX components. The Software Development Kit includes connection examples in the following environments: MS Visual C / C++, MS Visual Studio.NET (C#), Borland Builder, Delphi.
Our company

Cognitive Technologies is the Russia's leading software developer based in Moscow, Russian Federation. Founded in the late 1960's by a team of revolutionary scientists of Institute for Systems Analysis of Russian Academy of Sciences, it is the one and only Russian company having over 40 years of software development experience. Currently we employ over 1 000 people at our wide office branch in 67 regions of Russian Federation.

Cognitive Technologies focuses on software development in the following areas:

- Enterprise Content Management systems
- Document Management systems
- Electronic Archiving systems
- Enterprise Resource Planning systems
- Procurement Management systems
- Information Analysis Systems
- Master Data Management systems

Our mission

At Cognitive Technologies we rely on fundamental science which provides us with truly priceless competitive advantage. Scientific approach to us is innovative theory applied to business, complex developments converted to economic growth and profitable inventions that make life more comfortable.
Why Cognitive Technologies?

Quality
Over 500,000 companies have chosen Cognitive Technologies’ products all over the world. Continuous investment in R&D allows us to guarantee our products’ consistent high-grade performance.

Productivity
Cognitive Technologies’ software products have been designed to provide the maximum power and scalability vital for large enterprises to operate in geographically distributed environment.

Opportunity
We offer flexible system architecture and strive for excellence in satisfying our customers’ specific requirements. Your ideas, our resources – great product and new opportunities for your business.

Stability
Sustainable growth, innovations and progressive approach to the conduct of business throughout our 40 years of operation have secured our reputation of a reliable business partner.

Technology
Our software takes advantage of the latest technologies to let you quickly capture, manage, store, archive and recall your documents and data.

Efficiency
Our target is to create software products that are most productively applied to the real-life business cases of your company. Efficiency to us is your economic growth.

Innovation
Curious and committed minds, talented scientists and managers are our company’s foundation that enables breakthrough inventions and contributes to your business day by day.