ABBYY recognition technologies – ideal alternative to manual data entry. Automating processing of exam tests.

Marin Vlada[1], Ivan Babiy[2], Octav Ivanescu[3]

(1) University of Bucharest, vlada[at]fmi.unibuc.ro
(2) ABBYY Ukraine, i.babiy[at]abbyy.ua
(3) Star Storage, Romania, octav.ivanescu[at]star-storage.ro

Abstract

According to statistics, forms has share in 85% of all documents that are used in different economic spheres. Through automate forms processing, company can reduce volume of manual labor in 5 times, increase data quality and speed up documents processing, as result increase effectiveness of company's activity.
ABBYY [1] provides the companies with effective Data Capture solutions which can effectively recognize data from your documents and realize concrete needs for every industry.
ABBYY FlexiCapture transfers paper documents into usable data and offers a full range of state-of-the-art functionalities for document classification, data extraction and indexing.
This easy-to-use and to-deploy yet powerful solution provides a real alternative to manual data entry and other traditional forms of data input.

1. Introduction

For many Romanian commercial and governmental organizations conversion from paper document management to electronic one is the crucial issue.

Automated data capturing technologies have a relatively long history, dating back to when the first optical reading systems were developed to recognize stylized symbols drawn according to templates. Since that time, they have evolved to support a vast industry, utilizing a large set of very different technologies.

The traditional forms processing technologies for fixed (or structured) forms of today are well established. A large choice of systems capable of processing many types of fixed forms is now available.

Today's advanced systems can accurately capture printed and hand-written characters and process thousands of documents per day.

ABBYY FlexiCapture is one of the leading products in the field, capable of handling both printed and hand-printed forms.

2. ABBYY OCR/ICR Technology

ABBYY FlexiCapture is a specialized technology based on ABBYY's experience in recognition and document analysis technologies spanning more than 15 years. It has been in regular use since 1997, and we could probably say that it has served as a platform for many successful projects for 13 years. In fact, since 1997.

Types of documents

Organizations and businesses in different industries have their own features in document processing. ABBYY provides the companies with effective Data Capture solutions that realize concrete needs for every industry.

ABBYY FlexiCapture 9.0 can process one page, multi page documents with any level of complexity, documents with unfixed pages amount as documents with such appendixes as images or texts.

Paper documents can be divided into 3 categories:

- structured documents (fixed forms);
- semi-structured documents;
- unstructured documents

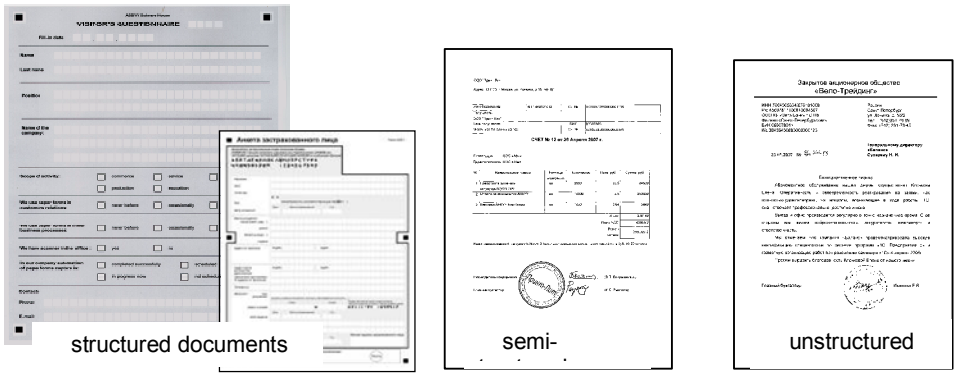

Figure 1. Types of documents

Various questionnaires, forms, examination sheets, reports, inquiry sheets and other similar documents which can be filled either by hand or by means of computer, belong to the structured documents (or fixed forms).

Invoices, payment orders, bills, explanations of benefits, and receipts – semi-structured documents.

Quires, newspaper articles, information from the Internet etc belong to unstructured documents.

Today the automation of structured documents data input is well-mastered. Nonetheless recently the companies have shown great interest in unstructured information input automation. Business of some organizations directly depends on data analysis quality and time and recourses spend for those operations. For example, different educational organizations interested to have data capture solution in order to efficiently automate its current processes. The challenges are to minimize the manual operations associated with questionnaire processing and to leverage data capture in order to increase overall productivity. The solution has to capture the desired data from the questionnaires and export them into usable digital information.

Processing of structured documents (forms) - is a process whereby information entered into data fields should be converted into electronic form:

- data are extracted from their respective fields;
- forms are digitized and saved as images.

In most cases forms processing is completed when the data from all the forms have been extracted, verified and saved. There are only two approaches for data extracting from paper forms: to involve many people in manual data keying in, or to start using automatic forms input system.

Manual data entry requires a lot of time, resources and is troublesome. It implies many problems such as delays in data capture, great amount of operator's misprints, high labor costs, equipment spending, rent-charge, etc. All these costs are avoidable with the help of a data capture solution such as ABBYY FlexiCapture, which enables automated forms processing.

Modern documents processing systems offer comprehensive facilities for automation of these processes and that allows customers to considerably raise overall performance.

Data input stages

Document conversion from paper to electronic type consists of several stages. At the first stage documents are scanned or photographed (given the modern development of digital photography the second method becomes more popular). The next stage is classification during which (for example, incoming letters differ from newspaper articles) is performed.

After scanning (photographing) and classification it is necessary to extract the data and to attribute the electronic document. Practically any document contains data fields: the date, the name of the author, the title, etc. As well as classification, attributing can be performed in the manual, semiautomatic or automatic way, and in semiautomatic mode for accuracy increase various rules are usually involved, having checked with which the system can reduce the number of errors.

ABBYY FlexiCapture interprets machine-print (OCR), isolated handprint (ICR), including alpha and numeric, mark sense (OMR) and barcodes from paper forms gathered from a scanner or a fax machine.

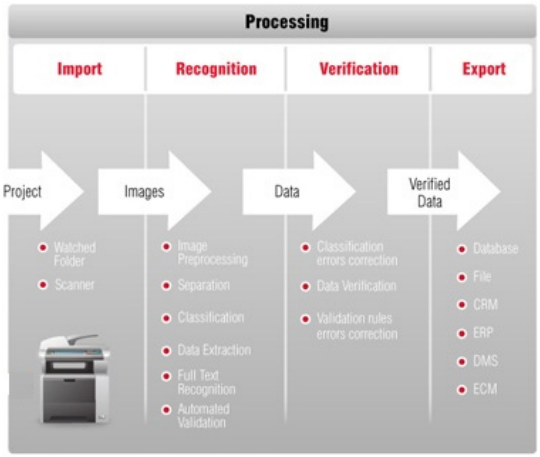

Figure 2. Processing stages

ABBYY FlexiCapture interprets data from paper forms many times faster and immensely more accurately than any professional operator, enabling you to collect data in efficient and secure way. It is noteworthy that the entire process requires only one human operator since all of the stages, except verification, are fully automated.

3. ABBYY FlexiCapture 9.0

ABBYY FlexiCapture 9.0 is supremely intelligent, accurate and scalable data capture and document processing system. It provides a single entry point to automatically transform the stream of different forms and documents of any structure and complexity to usable and accessible data ready to be exported into any business applications and databases.

Historically ABBYY Company developed three directions: document and form input, and applied linguistics. Today in each of these categories the company offers various type products for end users, system integrators and developers. In addition ABBYY integrated products of all listed categories into the uniform solution - FlexiCapture which ensures processing structured and semi-structured documents in a single space.

ABBYY FlexiCapture Software implements a number of processing technologies for checking of the document information relevance. This circumstance has basic value for structured documents processing as this procedure results in databasing. For correct performance of this operation it is necessary to carry out preliminary check of each field in the document on the data type relevance to expected result (for example, whether there is no text in the digital column), lengths of words and other parameters.

Automated paper document input systems are in demand among both governmental institutions and commercial companies.

Figure 3. Processing different kinds of documents

Technology Background

FlexiCapture identifies the document type and assembles one and multi-page documents out of the mix of pages using advanced ABBYY technologies, which allow automatic classification of documents with variable layouts of any complexity including:

- Multi-page documents;
- Documents with variable number of pages;
- Documents containing multi-page tables;
- Documents with image or text attachments.

The ABBYY FlexiCapture enables the recognition system to easily find necessary fields on the semi-structured form. Once located, the data in the fields can be captured using the OCR/ICR/OMR and barcode recognition technology.

FlexiCapture technology is built on powerful and time-tested ABBYY technologies based on the IPA principles (Purposefulness and Adaptability) [2] that imitate the way humans recognize objects.

FlexiCapture accurately extracts data and text from the fields specific for each document type using ABBYY award-winning multi-language recognition technologies. [3]

It offers:

- OCR for more than 180 languages
- ICR for hand-printed text for over 110 languages;
- Checkmark recognition for a wide spectrum of checkboxes;
- Barcode recognition for a variety of 1-d and 2-d barcodes.

Modern OCR technology allows processing a hand-written text as well under the condition of distinct letters writing. These possibilities are in demand among the companies which face the problem of processing of great number of forms and other similar documents filled in handwriting.

4. ABBYY processing examination sheets

Now there are only two technologies for automate processing examination sheets, making it possible to avoid knowledge subjective evaluation. The first one is computer testing. Each entrant

or student answers the question at a separate computer. And in some minutes after completing the test the machine calculates and gives out the result. The method is simple and efficient – however, due to high costs it is suitable only for groups of 10-20 persons.

The second method is more popular. Examination sheets are distributed to entrants where it is necessary to indicate the correct variant of the answer. Other notes - the examination name, a code of the student and a place for the signature. Then works are scanned and converted into the computer which in a minute knows who has passed the test and who has failed it.

Figure 4. Processing of examination sheets

Such technology has been utilized while carrying out “The unified examination” for school leavers in many European Countries for some years so far. In northern countries the youth’s knowledge is supervised by the independent testing centre at the Ministry of Education. On the examination day the centre technologist comes to school with forms package. These people are not acquainted with teachers and get into this or that school on a toss-up. The form with a name, surname of the final-year student and his/her code filled in are processed separately from examination sheet on which the person reduplicates his/her code and marks variants of the answers. There are also mysterious black small squares or other special labels on the form, distinguishing while scanning what examination it is and according to which variant it should be assessed.

All tests are prepared according to certain rules of a special science on questions designing – testology. They can be direct, that is offering only a title or figure or indirect, when the answer needs to be chosen out of four options. All is filled with ordinary pens. But very often due to agitation or indecisiveness peoples (students) put dots instead of "ticks", underline boxes or even fill in the word. ABBYY has developed automatic examination results processing technology, enabling to consider such cases individually, assessing answers in different ways. After scanning and automatic recognition the system assorts each symbol which it is not sure of. Information check is carried out according to special rules and guidebooks, for example, according to reference book confirming each answer.

Fast and qualitative computer check of examination sheets allows to solve three problems at once. Firstly, overcome irregular loads. Since all entrance examinations are held once a year, they need additional expenses on the personnel and teachers. Secondly, avoid health problems. For example, because of a computer-visual syndrome at first eyes get tired, then the weariness goes over the whole body, and the person works slower, his/her attention decreases. Third, and the biggest problem is the reliability of the information. To be assured of data which are entered manually, it is necessary to engage at least two persons and to assign the supervisor comparing their work. It takes a lot of time and is expensive. Automation is useful because all data are distinguished operatively. And only 5-7 percent of the total number of symbols requires the operator’s aid.

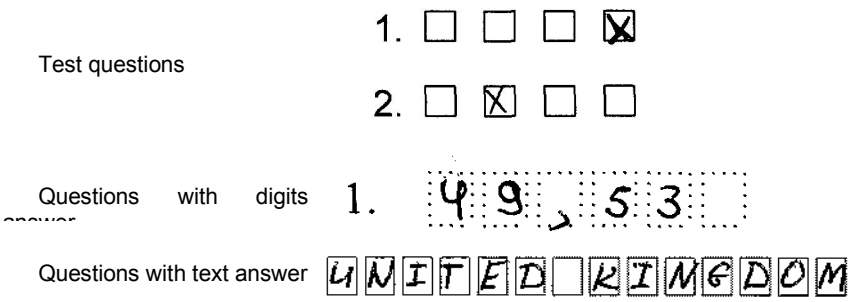

Figure 5. Recognition during processing of examination sheets

But entrance examinations are only a local system task. In fact, it has much greater possibilities: entrants' applications processing which can be recognized in a matter of minutes and databased, teacher feedback forms, data acquisition for plastic ID cards and student's cards etc.

The main concern of the educational organization will be not only to find a short-term, quick and easy solution to serve its current needs, but to further invest in a solution with the potential to meet future needs as well. Investing time and money in a solution should give the possibility to reuse the infrastructure for other data recognition projects and allow future in-house development, according to any project's needs, utilizing the acquired know-how.

5. Conclusions

The introduction of OCR technologies provides organizations with the opportunity to automate routine structured and unstructured data input and processing. The increase of text recognition accuracy, development of handwritten forms processing technologies considerably raises the efficiency of interaction of governmental and commercial institutions with their clients. The automation of these processes provides management with powerful tools to analyze large volumes of information and contributes to taking more exact and prompt decisions, which directly effects business efficiency.

6. References

[1] www.abbyy.com
[2] IPA Principles. ABBYY recognition technologies are built on the principles of Integrity, Purposefulness and Adaptability (IPA). Unlike other recognition technologies, which focus on recognizing patterns, IPA takes recognition a step further by using artificial intelligence to train the computer to analyze documents in the same way that the human brain would analyze them.
[3] ABBYY FlexiCapture 9.0 data and document capture system has been recognized as a Trend-Setting Product of the Year by KMWorld Magazine, the leading information provider serving the Knowledge Management systems market. (August 2010)
[4] www.abbyy.com/CaseStudie/
[5] www.agora.ro/stire/cniv-romania-organizeaza-un-webinar-pe-teme-educationale
[6] www.c3.cniv.ro/?q=2010/webinar