Digital Documents Recognition

I

Nicola Barbuti
University of Bari Aldo Moro, Italy

Tommaso Caldarola
D.A.BI.MUS. Ltd – Digitalizzazione di Archivi, BIblioteche e MUSei, Italy

INTRODUCTION

In years, digital document recognition has been one of the most important challenges in digitization fields, and worldwide research institutions and companies have researched, studied and tested a lot of different technologies and produced several systems and software.

The present article concisely both outlines contemporary digital document recognition technologies used to convert scanned images into machine-encoded texts and some digital recognition systems specific for scanned images of ancient handwritten and/or hand printed documents. It outlines too the future research directions by describing a newly created recognition technology.

BACKGROUND

Currently, digital document recognition encompasses the following technologies:

1. Optical Character Recognition
2. Intelligent Word Recognition (IWR)
3. Intelligent Character Recognition (ICR)
4. Pattern Matching.

This paragraph is divided concerning to the above different processing methodologies and specifying the features related to each one.

Digital Technologies Definition for Document Recognition

OCR, ICR and IWR are commonly considered analytical artificial intelligence technologies created to process sequences of fonts (OCR and ICR) or whole words, phrases (IWR), with the aim recognizing the content of digital images.

These technologies make best guesses at fonts or words by analyzing the sequence of lines and curves of the text contained in digital images and with the aid of database lookup tables, in order to associate closely or match the strings of fonts/words that form words/phrases.

To date, ICR has been considered an evolution of OCR because it can recognize digital images of handwritten documents, whereas IWR is considered an evolution of ICR as it can recognize digital images of cursive handwritten.

Optical Character Recognition

OCR technology is currently defined as electronic conversion of scanned images which reproduces printed text into machine-encoded text. OCR is widely used as a form of data entry from original paper source, documents, or any number of printed records. It is the common method for digitizing printed text so that it can be electronically searched, stored, displayed on-line, and used in a number of machine processes (Stokes, 2009).

This technology extracts data from digital document content by selecting and combining the letters to form words, and assembling words into phrases. The OCR recognition allows to access the contents of the original document and to process them. Early versions needed to be programmed with images of each font, and worked on one font at a time.

Advanced OCR works with the following steps:

- Segmentation. The OCR analyzes the structure of the document image, then divides the page into elements such as blocks of text, tables, im-

DOI: 10.4018/978-1-4666-5888-2.ch379

ages, lines, etc.; the text lines are divided into words and these into characters;
- Once all of the font have been distinguished, the program matches them with a sample of font previously selected and makes different assumptions about which letter it might be;
- Based on these assumptions, the OCR analyzes the different variants aiming to split lines into words and words into letters;
- After processing a large number of probability of this kind, finally the OCR can decide and display the recognized text.

The OCR needs to include default fonts and languages sets. It recognizes digital images only when their content is written by the same font the program uses. So, this technology turns out to be closed and does not work fine on images of ancient handwritten or of hand printed documents (Smith & Merali, c1985).

Intelligent Word Recognition

The Intelligent Word Recognition is considered the evolution of OCR and ICR technology, as it can extract and recognize cursive handwritten as well as normal manuscripts. However, as shown later, it differs from ICR because it can be used to index only small database and requires the former creation of thesauri structured with words extracted within the image content.

The IWR process is made of the following steps:

- **Preliminary transcription of the whole image content or of keywords extracted from it:** This step is necessary to achieve correct recognition;
- **Image segmentation in regions containing words:** This step produces an image sequencing that ends when the amount of contrast of each word reaches a fixed threshold; each image is then segmented into regions, and each region contains a word or a set of words;
- **Proper recognition of image content:** All the segments extracted are processed by the recognition algorithm; subsequently, each segment matches the corresponding electronic keyword by means of a pattern-matching process, which varies according to image noise and poor content reading;
- **Storage of the recognized information:** All of the information obtained from the recognition step are developed, classified and stored: words family, words type, full text, as well as standard information like author, title, etc. are necessary information for the data storage.

It may be easily inferred that IWR technology too turns out to be closed and can only be used on small, homographs, homogeneous databases.

Intelligent Character Recognition

According to common definition, the ICR recognizes and indexes the textual content of digital images of handwritten documents. Until today, this technology has been considered an evolution of OCR (von Ahn, Maurer, McMillen, Abraham, & Blum, 2008). In last years, some ICR technologies has been researched with the goal to recognize digital images of ancient handwritten and hand printed documents with interesting results (Feldgajer, 2012).

More recently, an innovative one has been created, able to recognize historical hand printed or handwritten documents by employing a new *training process* (Barbuti & Caldarola, 2012). This methodology uses an approach based on intelligent semi-automatic self-learning recognition of the digital document content.

The new training process is made of the following steps:

- Semi-automatic self-learning of fonts;
- Image segmentation;
- Proper recognition of text or content contained in each segment;
- Storage of the recognized information;

By comparing OCR and IWR with this new ICR technology, it can be easily inferred that there is a fundamental difference between those technologies specifically within the use of the training process. This latter method allows to structure the digital recognition ICR-based systems as open systems. Instead, OCR and IWR are closed technologies because require the creation of font sets (OCR) or structured thesauri of