An Innovative Character Recognition for Ancient Book and Archival Materials: A Segmentation and Self-learning Based Approach

Nicola Barbuti[1] and Tommaso Caldarola[2]

[1] Department of Classical and Late Antiquity Studies, University of Bari Aldo Moro
n.barbuti@ateneo.uniba.it
[2] D.A.BI.MUS. L.L.C., Spin Off of University of Bari Aldo Moro, Italy
t.caldarola@dabimus.com

Abstract. The paper illustrates the invention of a method and an apparatus able to recognize the text in a set of digital images referring to pages of ancient manuscripts or printed books. It includes the following macro steps: identifying and connecting in sequence regions containing words in a subset of the images; structuring a thesaurus of fonts used in those regions; performing the character recognition of one or more images belonging to the set, associating to this recognition a first value of efficiency. The prototype is patent pending (National Pat. Pend. n. BA2011A000038 – Intern. Pat. Pend. n. I116-PCT).

Keywords: Intelligent Character Recognition (ICR), Manuscripts, Ancient printed Books, Digital Library, Digital Database of ancient Heritage.

1 Introduction

Existing digital libraries, containing digital collections of ancient and valuable handwritten and printed documents and books dating up to the second half of XIXth century, show a level of interactivity still extremely low. For these specific digital contents, indeed, has not been yet possible to develop optical-digital recognition systems and/or text recognition of virtual pages, able to provide an efficient indexing of databases content either already accessible or to constitute over the web 2.0.

None of the latest and most important projects of digital libraries currently available on the web 2.0 (Europeana, World Digital Library, The European Library, etc.) has accessibility and usability features that allow users to see the text content of the reproduced digital objects without having to scroll them through in full. Excluding common cataloguing research (author, title, release notes), in these databases it is not possible to develop any indexing that allows in-depth studies based on the analysis of the recurrence of words, inference about different texts, etc.

This difficulty arises from the nature of the artifacts in question. The complexity and divergence of ancient manuscript spellings, even those paleographic more linear and regular; the kind of old inks used; the obsolescence of the materials, in most cases with damages caused by biological or biochemical factors, mechanical accidents and

M. Agosti et al. (Eds.): IRCDL 2012, CCIS 354, pp. 261–270, 2013.

human carelessness: all these factors have so far prevented all attempts to go beyond the simple reproduction of these digital artifacts.

Neither the currents OCR, ICR and IWR available on the market can be applied to solve the problem of text recognition in ancient documents[1].

If this situation seems almost obvious in the case of manuscripts, because of their nature, it should be less understandable for the printed books. Instead, even for this kind of artifact, in particular for books produced by handprinting (and therefore the entire print production from 1456 up to 1850), the situation is very similar to that of the manuscripts.

The problems, in fact, are not different, even if they affect to a lesser extent. The techniques of composition of the printing plates, the inks used, the alignment of stamps within words, and of the words within lines, the different graphic fonts representative of certain letters, as compared to those commonly used (eg., the "*s*" represented by a printing font very similar to "*f*"), different linguistic conventions, the various noise of the images (background noise caused by the press on the reverse side page, smudges and breakage of stamps, ink stains and some other varied cause due to time and men) are all factors that, today, frustrate any attempt to index the contents of digital images of ancient materials through application of recognition systems with satisfying results.

2 State of the Art

Research. Concurrently with the research and prior to the implementation of the prototype, a survey was carried out both in research on intelligent recognition systems, and among international patents relating to existing applications for recognition of the content of digital images.

It became apparent that the research on intelligent recognition systems, which is able to operate effectively on images of handwritten or printed ancient materials, especially before 19th century, has not yet produced significant results despite the efforts made for several years.

Shape prior model – Ben-Gurion University. In our opinion, the only interesting research about recognition was carried out by a team of Ben-Gurion University in Israel, whose first results were published in 2008[2].

The paper describes a method of segmentation and recognition of characters which aren't perfectly legible in damaged ancient manuscripts. The process is based on the manual construction of *shape models* representing the possible variability of the characters previously segmented from images of damaged manuscripts. On these is performed a training set that, by matching with the reference models, progressively reduces them to a core, generating for each segmented character a *shape prior* which is the essential reference for the reconstruction of damaged characters and not legible to human eye.

The system, which works on grayscale images, implies a preliminary long and laborious step of manual construction of models of reference, and requires more

progressive training set. Despite the complex laboriousness of the process, the result is certainly interesting.

Even so, the *ratio* of the system has completely different requirements than those of the application object of this work, which will be discussed in the following paragraphs.

Patents. The survey among international patents produced no most important results. Faced with an astounding amount of existing applications, it's easy to detect the almost complete identity of functions and applications they exhibit in the output. There are very few exceptions, however always conditioned by elaborative processes that require high manual skills, thus making not very user friendly. We describe briefly some of those that seem to be a useful paradigm to better illustrate the newness of the process that we developed.

reCAPTCHA[3]. An interesting patent is that developed in 2008 by researchers from Carnegie Mellon University, USA. They have revised the existing CAPTCHA systems, enabling them to interpret the doubtful words identified by OCR programs, according to a simple, but efficient, system.

When two OCR systems identify differently a word, this word is associated with a known word and sent to a user who has to pass a CAPTCHA test to access a service. It is assumed that if a user is able to identify the known word correctly, then there is an high probability he/she also identify the unknown word. When three users give the same answer, the system stores the word as correct.

In September 2009 the project has been acquired by Google, who uses it to correct errors resulting from OCR scanning of texts. It should however be noted that, for images of books printed prior to the second half of the 19^{th} century, the results are not at the level of expectations created at the moment of the discovery and distribution of the system. In fact, the rates of return are still quite low, as it oscillates between 30% and 60% for ancient printed documents, with the highest percentage obtained exclusively on printed texts from the late nineteenth century, while for manuscripts the system did not show any noteworthy working.

Multifont Optical Character Recognition Using a Box Connectivity Approach (EP 0649113 A2)[4]. The approach of the system is based on a pattern recognition obtained setting a minimal bounding rectangle around the pattern, sharing out the pattern into a grid and comparing a partitioned vector derived from this grid with other vectors obtained in a similar way starting from known pattern.

Finally, you choose a set of pattern according to Pareto and select one of the patterns thus obtained. The process is laborious, and it is not able to operate effectively on images of ancient documents.

Document Digitization (Fr 2768825 A1)[5]. The system is based on the digitization of generic documents, acquiring the image with a scanner connected to one of two computers linked to a network. Scanned images are stored in a high capacity data storage system.

This storage system is also used to save text files "searchable" produced from the second computer with an OCR process on document images. The application does not present any significant functionality able to operate on images of ancient documents.

Method and Apparatus for Isolating an Area Corresponding to a Character or Word (Us 5144682 A)[6]. This system works in order to isolate an area corresponding to a character or a word in an OCR device. The main technical problem that must be solved is to recognize characters and words disposed on lines considerably inclined.

The method works on black and white images. Although it is designed to solve a noise level, which is one of the most problematic factors of the images of ancient documents, the system is not able to operate effectively on this kind of images.

Technique for Correcting Character-Recognition Errors (Gb 2463577 A)[7]. The system is structured on a method for identifying and correcting failures in the information extracted from images using character-recognition software like OCR or ICR. However, the level of operation on which it is effectively able to act is strictly limited to current documentation concerning financial transactions.

A2iA's Proprietary IWR, Intelligent Word Recognition[8]. Some systems, while using more sophisticated methods than aforesaid, base the recognition on the segmentation into words of the text regions.

Such approach is used by the *A2iA Proprietary IWR, Intelligent Word Recognition,* developed by the A2iA, USA. Although this IWR has been successfully used in projects for recognition of handwritten documents, the system is able to operate only if interfaced with specific semantic thesauri structured prior to recognition phase, otherwise it is not capable to make any refund of text. Once again, it is assumed the necessity of a preliminary laborious manual work.

Some Remarks. As can be easily inferred from what we have above outlined, nowadays there is not a method or system able to recognize and index images of ancient documents in either automatic or semi-automatic way.

The models and the systems able to work on such kind of document have in common test on ideographic script. Some questions remain about working on alphabetical script. Furthermore, in order to work satisfactorily on such documents, all of them require either a complex manual transcription in electronic format of the content of documents to index, or to structure specific semantic thesauri on which to match the images, followed by an equally laborious training process.

They need too the planning of complex algorithm to extract information to use as models for the matching of digital images, but the output is incomplete and unsatisfactory. And all the scientific and research papers about the digital recognition have the same limit: they purpose not systems, but pure models without sufficient certainty about their working on digital database of paleographic materials.

We consider the reason why the existing systems for optical and/or intelligent recognition of digital images don't work on ancient documents is the methodological approach used in the structuring of such systems[9-15].

This approach maps the words on the scanned images of document pages by associating them in their entirety either to an electronic text inserted manually by an operator, or to thesauri of reference preliminarily structured still manually. This approach, just as it requires a long and complex preliminary manual work, seriously limits the possibility to electronically recognize a large amount of historical texts.

In addition, if this method works fine on certain more recent materials (from the late 19th century onwards) because of the linearity of the typographic and graphics composition of the pages, it cannot be the solution to the problem of opening to scholars and mankind an interactive access to the enormous amount of older works both "more" and "minor" still unknown, stored in thousands of historical libraries in the world, whose reproduction presents graphical, typographical, and noise complex and unsolved issues.

3 The Method and Apparatus to Recognize Text in Digital Images Reproducing Pages of an Ancient Document (Pat. Pend. Nat. n. BA2011A000038 – Intern. n. I166-PCT).

3.1 A New System for Recognizing Text in Digital Database of Ancient Manuscripts and Printed Books

Considering the above, the purpose of the following research has been to set up a method and an apparatus able to recognize and to transcribe full text a percentage rate greater than or equal to 50% of content in a set of digital images, each of which depicts a page of an ancient manuscript or printed document, without requiring a laborious and long preliminary manual work.

The methodological approach used has been different from those previously used for similar systems, as it has had its own premises in the characteristics of discrepancies and noise peculiar to the digital reproductions of ancient artifacts.

The process aims no longer the regions/words of text (regions that contain a word), but the regions/fonts (regions that contain a font), each of which is associated with a sample of corresponding electronic font transcribed manually by an operator.

3.2 Training Stages

The process has been tested on samples of images of printed and manuscripts documents, different in dating, typographic and graphic characteristics and noise index, calculated over the entire of intrinsic and extrinsic factors of each sample (*intrinsic*: printed books: stamps set used, cleaning of pages, presence of spots or dirt, handwritten gloss, etc.; manuscripts: handwriting readable to naked eye, non homogeneous graphic sign, irregular text lines; *extrinsic*: image quality, resolution, background noise, etc.).

The amount of images to be used for the training set has been calculated as a sample of 100 for the printed documents, 30 for manuscripts, selected according to the following characteristics:

- *ancient printed books*:
  - 16th century; font: italic; noise rating: 80% (very high)
  - 17th century; character: round; noise rating: 60% (high)
  - 18th century; character: round; noise rating: 30% (average)
- *ancient manuscripts*:
  - letters: handwriting cursive; noise rating: 90% (significantly high)
  - census: handwriting chancellery cursive; noise rating: 75% (very high).

Before running the training, it has been calculated the threshold of iteration of handwriting/typographical fonts used in the selected image set, that is the threshold beyond which the fonts set used to compile the document begin to be iterative and equal to the previous.

Therefore, the image percentage settled as exhaustive of the whole fonts set has been used as sample for the training. This percentage never exceeds 10% of images for each sample, and often the threshold has been reached already with very low percentage (2%-5%).

Then, has been performed in electronic the manual transcription of the content of each percentage of images, to use it as text to matching recognizing font. Obviously, the greater the amount of content transcribed at this stage and reconnected to the regions extracted from the images, the greater the precision in the return of correct semantic structures.

The whole training has been divided into following steps: a) document scanning; b) self-learning of fonts of digital document; c) image segmentation either in regions/words (each region matching one word) or in regions/handwritten or printed fonts (each region matching one handwritten or printed font) varying according to the image noise and to to the hard reading of the content; d) proper recognition of text contained in each segment; e) storage of the recognition information; f) application; g) facility.

a) Document scanning. This step has carried out the manual scanning of the document which has been submitted to recognition. It has been used a planetary scanner with trilinear CCD 3 X 10.000 pixels rgb real (not interpolated), with real resolutions of 400 dpi up to 2xA2, 600 dpi up to 2xA3, 800 dpi up to 2xA4, 1200 dpi up to 2x5A.

b) Self-learning of fonts. For each digital document or whole databases has been assumed the existence either of a set or of multiple sets of fonts that the system should learn to recognize assimilating them permanently.

The learning of the fonts has been the key step of the system, on which the whole process is based: if there are errors or flaws at this stage everything that follows may result inaccurate.

This process is *iterative* and *incremental*. *Iterative* because it is based on a number *n* of iterations, *incremental* because at each iteration a new information is added to the set of recognized font, namely *extension of the set of the characters*. In fact, the font for the system can be provided not necessarily by the number of characters provided

for the single alphabet, but it's open, unlimited in relation to the possible variants (graphics, typographical, coloring, etc.) that each character brings with when digitalised.

The receipt of the font for the system occurs either when the index of noise in the iteration is lower than a given threshold ξ defined *noise rating*, or when during the iteration the noise rating to the *i-th* step concurs with that in the next step, i.e.:

$$\xi_{i+1} = \xi_i$$

that is, the system has finished the learning.

If at the end of the process the results are out of threshold, it may need human intervention which will analyze the noise to manually classify it and train the font to recognize the non-recognized character/s.

c) Document segmentation. When all the fonts of a document have been known to the system, the training proceeded with the *segmentation*. This is even an iterative process. Through a multiple temporary and in image memory processing, the segmentation produces a series of image processing that ends when the amount of contrast of character reaches a fixed threshold.

At the end of this step, depending on the settled segmentation, a series of segments has been selected, each of which contains the character set recognized. The setting of the segmentation can be variable according to character, word, line, etc., and must be defined referring to functionality that will be applied.

During this stage, through subsequent proceedings, it has been evaluated the functionality of the system by analyzing the steps of testing and acting on each step to refine the results recorded different from those contemplated.

For this specific step have been assumed different eligibility criteria of the segmentation process, possibly based on statistical values that self-refine as the number of processes increases.

This step has been closed when the segmentation of a relatively large number of digital documents matches to a very low percentage of noise.

d) Recognition of document content. After the segmentation step, all the segments produced by the recognition of the sets of characters contained in each segment have been processed. Before switching to the storage of information concerning the recognition, a further process has been performed in order to permanently remove any residual noise due to not properly recognized characters for many reasons (e.g., graphic rendering other than the modern, etc.).

e) Storage of the information carried out from the recognition. Once completed the recognition step, the storage and classification of information started. All information obtained from step b) have been developed, classified and stored: the font family, font type, text, noise ratio, as well as standard information like author, title, number of pages, segments per page, different kinds of fonts for the document, etc. All these

information are necessary for the data warehouse in order to make available to users more features as possible even through the use of facilities.

f) Application. During this step has been tested the functionality of the system by subjecting it to different test cases performed on different sample set of digital images. In case of further problems detection, it has been tried to refine and correct any step that came into play during the test cases. The testing stage was completed with check of full and effective functioning of the system.

g) Facility. The system has been meant to be as open, so it can be implemented with various and diversified facilities. The facilities are extensions to the system that allow to exhibit additional features to relate, classify, map information and then allow the end user to enjoy a richer information.

3.3 Percentages of Font Recognition

The basic algorithm has been used in an univocal way on each sample of images. Then it has been calibrated in relation to the feedback obtained from the stage c) and d) of the training step. In particular, for manuscripts it has been necessary to set up different modes of segmentation of the regions, per character on the census (functionality ICR), per word on the letters (functionality IWR), due to the high heterogeneity of the graphic sign in the documents of this latter sample.

At the end of the training, the percentages of fonts and text properly recognized resulted more than satisfactory, although with some differences between different samples of documents:

- *ancient printed books*:
  - 16^{th} century:
    - fonts: 87% exactly recognized, 13% error
    - words: 65% exactly recognized, 35% error
  - 17^{th} century:
    - fonts: 84% exactly recognized, 16% error
    - words: 57% exactly recognized, 43% error
  - 19^{th} century:
    - fonts: 98% exactly recognized, 2% error
    - words: 89% exactly recognized, 11% error
- *ancient manuscripts*:
  - letters[1]:
    - words: 57% exactly recognized, 43% error
  - census:
    - fonts: 36% exactly recognized, 64% error
    - words: 42% exactly recognized, 58% error

[1] As specified previously, it has been tested some segmentation processes variable depending on the kind of function that will be applied, and in the case of the letters we have chosen to test the function IWR instead of the ICR, due to low legibility of the documents used to the naked eye too.

The above percentages refer, of course, to execution of the process in the first and only solution, without further refinements and calibrations. Additional manual calibrations can be done to correct and eliminate the inevitable noise, achieving a recognition with a high index of accuracy.

As it should be noticed, for nearly all the samples the system was able to carry out an accurate recognition with percentages >50%. The only exception have been the samples related to census, but, whereas among the manuscripts they constituted the highest dating (first half of the 18th century), in this case too the percentage of refund can be considered fully satisfactory, especially if parameterized with the current state of the art outlined above.

Moreover, we must not forget that the presented results refer to a training first and only performed on representative samples not numerically significant. It follows that as much information the system receives at this stage, that is to say as many are the images on which carries out the training by having a minimal portion of extracted text as a reference base, the greater the percentage of information correctly identified, and consequently the less the noise, which would then further reduced and, plausibly at least for printed documents, almost entirely phased out in subsequent steps of manual correction, of course, also iterative.

4 Conclusions

This paper describes a new prototype of *Intelligent Character/Word Recognition* able to recognizing text in digital images of ancient manuscripts and printed documents. The system currently is patent-pending (Pat. Pend. Nat. n. BA2011A000038 – Intern. n. I166-PCT) and is named *ICRPad*.

It has been tested with features ICR, IWR and OCR on sample sets of digital images of ancient manuscripts and printed documents with positive feedback, such as to sustain right now that further trials, which is currently undergoing, will open possibilities for research, study and interactive use of digital libraries of cultural archival and book heritage, and perhaps not only, both differentiated and with a high level of innovativeness.

The algorithmic structure developed for this application will allow a level of accessibility to the digital documents that to date has not yet reached by any similar system. In fact, it allows two levels of usability applicable contextually.

The first allows the user to search through the document without the need to indexing the content: this procedure, however, would require time, because the segmentation would be contextual to the research step, so it would work effectively for the user only on documents of small capacity.

The other involves the launch in batch of the application on the entire document prior to its overflow into the database, with the consequent indexing of the textual content recognized, so that, once the document has been input in the database with keyword search options, the user can do all the searches he wants with an immediate return.

References

1. Feldgajer, O.: Universal Character Section for Multifont (EP 0369761 (A2)), http://worldwide.espacenet.com/publicationDetails/biblio?CC=EP&NR=0369761&KC=&FT=E&locale=en_EP
2. Bar-Yosef, I., Mokeichev, A., Kedem, K., Dinstein, I.: Adaptive shape prior for recognition and variational segmentation of degraded historical characters. Pattern Recognition 42(12), 3348–3354 (2008)
3. von Ahn, L., Maurer, B., McMillen, C., Abraham, D., Blum, M.: reCAPTCHA: Human-Based Character Recognition via Web Security Measures. Science 321(5895), 1465–1468 (2008)
4. Krtolica, R.V., Malitsky, S.: Multifont Optical Character Recognition Using a Box Connectivity Approach (EP0649113A2), http://worldwide.espacenet.com/publicationDetails/biblio?CC=EP&NR=0649113&KC=&FT=E&locale=en_EP
5. Blondy, A.: Document Digitization (Fr 2768825 A1), http://patent.ipexl.com/FR/FR2768825.html
6. Nakamura, M.: Method and Apparatus for Isolating an Area Corresponding to a Character or Word (Us 5144682 A), http://www.patentbuddy.com/Patent/5144682
7. Masami, M.: Technique for Correcting Character-Recognition Errors (Gb 2463577), http://worldwide.espacenet.com/publicationDetails/biblio?CC=GB&NR=2463577&KC=&FT=E&locale=en_EP
8. http://www.a2ia.com
9. Eynard, L., Leydier, Y., Emptoz, H.: Particular Words Mining and Article Spotting in Old French Gazettes. In: Proceedings of MLDM Posters, pp. 176–188 (2009)
10. Gordo, A., Llorenz, D., Marzal, A., Prat, F., Vilar, J.M.: State: A Multimodal Assisted Text-Transcription System for Ancient Documents. In: DAS 2008. Proceedings of 8th IAPR International Workshop on Document Analysis Systems, pp. 135–142 (2008)
11. Le Bourgeois, F., Emptoz, H.: DEBORA: Digital AccEss to BOoks of the RenaissAnce. IJDAR 9(2-4), 193–221 (2007)
12. Leydier, Y., Le Bourgeois, F., Emptoz, H.: Textual Indexation of Ancient Documents. In: Proceedings of the 2005 ACM Symposium on Document Engineering, pp. 111–117 (2005)
13. Leydier, Y., Le Bourgeois, F., Emptoz, H.: Towards an Omnilingual Word Retrieval System for Ancient Manuscripts. Pattern Recognition 42(9), 2089–2105 (2009)
14. Rawat, S., Kumar, K.S.S., Meshesha, M., Sikdar, I.D., Balasubramanian, A., Jawahar, C.V.: A Semi-automatic Adaptive OCR for Digital Libraries. In: Bunke, H., Spitz, A.L. (eds.) DAS 2006. LNCS, vol. 3872, pp. 13–24. Springer, Heidelberg (2006)
15. Toselli, A.H., Romero, V., Pastor, M., Vidal, E.: Multimodal Interactive Transcription of Text Images. Pattern Recognition 43(5), 1814–1825 (2010)