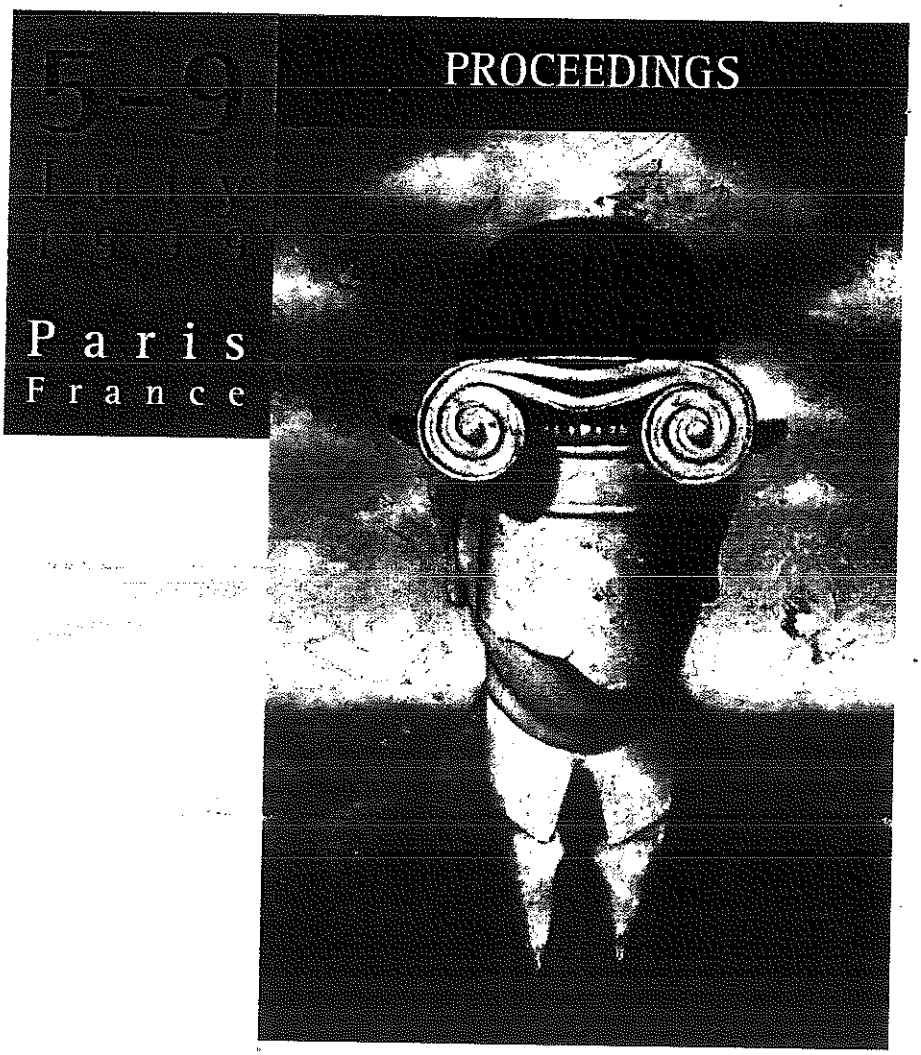
2nd international congress on

"science and technology for the safeguard of cultural heritage in the mediterranean basin"



CNRS
CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE
France

CONSIGLIO NAZIONALE DELLE RICERCHE
Italia

Vol. 2

ELSEVIER

Science and Technology for the Safeguard of Cultural Heritage in the Mediterranean Basin

Proceedings
Vol. 2

Editor: A. Guarino

Paris, Amsterdam, New York, Oxford, Shannon, Tokyo
23, rue Linois, 75724 Paris cedex 15
http://www.elsevier.fr

ELSEVIER

DIGITAL TECHNIQUES FOR CHARACTER RECOGNITION IN OLD PRINTED BOOKS AND IN MODERN DAMAGED DOCUMENTS

Luigi Bedini[1], Andrea Bozzi[2], Anna Tonazzini[1]

[1]Istituto di Elaborazione della Informazione - Consiglio Nazionale delle Ricerche
Via S. Maria, 46 - 56126 Pisa, Italy
[2]Istituto di Linguistica Computazionale - Consiglio Nazionale delle Ricerche
Via della Faggiola, 32 - 56126 Pisa, Italy

Keywords: Document Analysis, Blind Image Restoration, Optical Character Recognition, Computational Philology

1. INTRODUCTION

A central objective of the CNR Special Project "Safeguard of Cultural Heritage" is the development of computerized tools to retrieve and restore textual information contained in ancient printed documents, accessed as digital images. Within this objective, we propose an integrated system that improves the quality of the images and, at the same time, activates optical character recognition (OCR) functions. The aim is to implement a computer-assisted workstation that is not limited to providing catalogue information, but offers direct access to the textual data contained in the ancient books[1].

To enhance the quality of degraded images we adopt blind image restoration techniques, based on regularization theory and constraints derived from known features of the ideal image[2,3,4]. These constraints are enforced via Markov Random Field (MRF) models, and the solution is obtained by means of the optimization, with respect to both the image and the unknown degradation operator, of a cost function which expresses data consistency and fidelity of the image to the adopted model. The enhanced images are then segmented to locate each single character of the text. Each character is recognized and classified, using techniques based on a neural network model that has a good classification capability along with a fast training stage. The workstation graphical interface shows the interpreted text and the word by word segmented image in two separate windows on the screen so that the operator can check any mistakes. The "word in image" and "word in text" concordance is therefore recorded and from this moment onwards a number of queries can be made, for linguistic and philological analysis.

2. IMAGE ENHANCEMENT

The various degradation factors affecting the images of ancient or damaged documents globally act as an unknown space-variant blur operator that, especially when strong, makes the various characters to spread and overlap one another. Standard commercial OCR systems and even those based on Hidden Markov models and/or neural networks[5,6] can fail in these situations, since the blur may cause the segmentation step to produce joined and/or broken characters[7]. Hence, enhancement techniques are necessary, in order to remove the blur and, at the same time, to separate each character from the others and from the background. Thus, we propose to integrate techniques of blind image restoration with techniques of image segmentation, so as to recover the ideal undegraded image in an already segmented form. As a sub-product, we also obtain an estimate of the blur mask[8]. To simplify the problem, we assume that the images can be partitioned into sub-images, where the space-variant blur can be approximated as a space-invariant blur. We then propose a recursive procedure that, starting with the estimation of a single blur mask for the whole image, refines the estimate, and hence improves segmentation, in those zones where suitable validation tests, based on a linguistic analysis, reveal errors on the result of the subsequent OCR process.

Within regularization techniques, we adopt a Multi-Level Logistic (MLL) model, which is a typical MRF model for piecewise constant images[9], and consider a cost function accounting for consistency with the observed image g, smoothness constraints on the gray levels of pixels belonging to homogeneous regions in the image f, and geometrical constraints on the character morphology. More

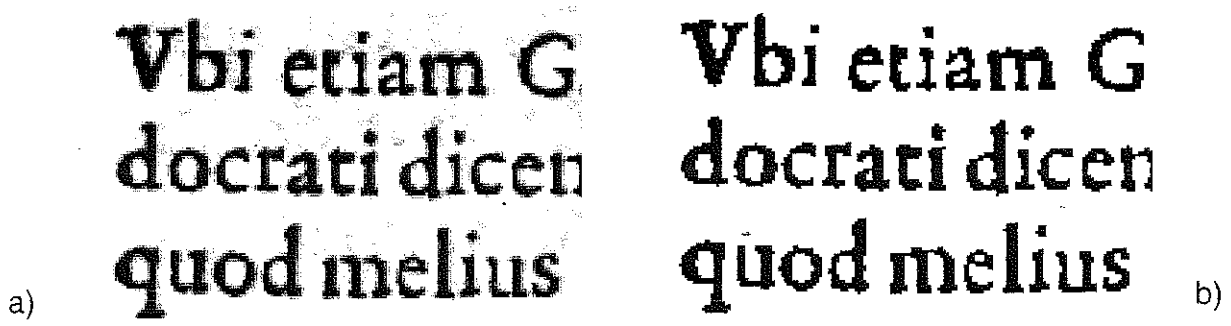


Fig. 1 - Blind segmentation of a 110x165 portion of a page of the First Book of the Opera Omnia of Cardano: (a) original, degraded image; (b) segmented image.

specifically, our models exploit some relevant characteristics of printed texts, such as the fact that the ideal image is essentially a two-level image, one level corresponding to the background and the other to the characters, and the fact that the characters present sharp and regular boundaries. In addition, since the blur mask acts as a low-pass filter, we exploit the constraint that its elements must be positive and of unitary sum. Our problem becomes then:

$$\min_{f,d} \| g - H(d)f \|^2 + \sum_{c \in C} V_c(f) \qquad (1)$$

subject to the extra constraints:

$$\sum_{i,j} d_{i,j} = 1 \quad \text{and} \quad d_{i,j} \geq 0 \quad \forall i,j \qquad (2)$$

where **f** and **g** are the vector form for f and g, respectively, H(d) is a block Toeplitz matrix, whose elements derive, according to a known rule, from the blur mask d, and $V_c(f)$ are potential functions that enforce the interaction of cliques c of adjacent pixels. The solution strategy to problem (1) consists of the alternate execution of steps of image estimation and steps of estimation for the degradation operator[4,8], according to the following iterative scheme:

$$f^{(k)} = \arg\min_f \| g - H(d^{(k)})f \|^2 + \sum_{c \in C} V_c(f) \qquad (3a)$$

$$d^{(k)} = \arg\min_d \| g - H(d)f^{(k)} \|^2 \qquad (3b)$$

where the constraints (2) are imposed on the solution after each iteration. In practice, we adopt a simulated annealing (SA) algorithm with Gibbs sampler[2] for the estimation (3a) of the segmented image, periodically interrupted to produce a new estimate of the blur mask, via the gradient descent solution of the least-squares problem (3b).

To test the performance of the procedure we applied it to several zones that are different for size and position in the considered documents. Fig.1 shows an example of the typical results obtained. The marked "m" was correctly restored by using the refining procedure based on the re-processing of those zones of the image where the OCR or a subsequent linguistic analysis have revealed an error.

3. CHARACTER RECOGNITION

The restored image constitutes the input to the module for character recognition, based on neural network. At present, this module is available in prototypical form and is named OCRLab. It performs three main tasks: the segmentation of the characters within the restored image; the learning phase for training a neural network, based on Self Organizing Maps (SOMs)[10]; the recognition of the segmented characters by means of the trained neural network. Herein we report the main procedures executed by OCRLab.

3.1 Manual Transcription Procedure

The operator selects an area in the image to be transcribed; which appears already segmented as shown in fig.2. Therefore, the operator starts the manual transcription of the text. As this phase goes on, the system creates a correspondence map between the typed character and its image, to be used in the learning phase. It also produces statistics about the number of the examples collected for each character, which can be shown in the so-called "set document window". This makes the operator able to decide whether to stop or not the manual transcription (note 1).

3.2 Network Training Procedure

This procedure uses the examples collected in the previous phase to train the neural network. It is activated by means of the button on the toolbar. Once the training is finished (generally it takes about two minutes on a Pentium 300), the automatic interpretation phase can begin by selecting the part of the document to be recognized.

3.3 Automatic Translation Procedure

The automatic translation procedure produces:
- the file with the interpreted text;

- the check of the interpretation by a linguistic-statistical module and a very large machine dictionary;
- other information like the alternatives of an interpretation and the dependability rate of each. For example: Char 533 535 550 559 indicates the bounding box parameters in the image of a character which has been interpreted like: a (a, 32) (o, 2) (c, 2) (e, 12) (?, 2).

3.4 Exportation of the recognized text

The recognized text is exported as .rtf file in a Digital Library Workstation which is able to access images and texts in a very innovative way.

4. THE DIGITAL LIBRARY WORKSTATION

The workstation supports all the facilities for displaying documents in the form of both images and transcribed texts, for correcting these latter and for performing different types of analysis. In particular the workstation makes available all the functions that are necessary for linguistic and philological analysis. To this purpose, the software examines the text transcription produced by the OCRLab system or manually by the philologist in a document memorized with a commercial Word Processor. It is able to produce automatically the relationship between each word in the image and each word in the transcription. The information is stored in a database and from this moment onwards a number of queries can be made:

a) by selecting a word on the transcription (or in the *index locorum*), a window appears showing the word evidenced in the image. The system supplies both the reference to the selected *locus*, and all the *loci* in which that word can be read. The selection of any other reference implies immediate visualization of the corresponding parts of text and image;

b) on the other hand, by selecting a word in the image, a window appears which shows the transcription of the word in the text. Even in this case, the system suggests the references of the other *loci* in which it appears (see fig.3)[11].

The data relevant to the texts and the images are encoded in different ways: original encoding envisions RTF for text representation and JPEG for the representation of compressed images. The database with image and text documents segmented have been modeled according to HyTime.

5. CONCLUSIONS

The activities connected with the research program described above are positioned in at least two areas of interest: in the first place that of computerized technology applications for libraries. Another application area concerns more directly the possibility of full-text interrogation of the books' contents. In fact, the program described herein represents an indispensable premise for the design of a workstation assisted by a computer that does not limit itself, as is currently the case with SBN (the Italian National Library System), to providing catalogue information, but, thanks to the OCR method specialized in recognizing the characters present in ancient books, offers direct access to textual data contained therein. The potential users of the system are represented by librarians, philologists, archivists, epigraphists, papyrologists and, in general, the students of medieval source documents.

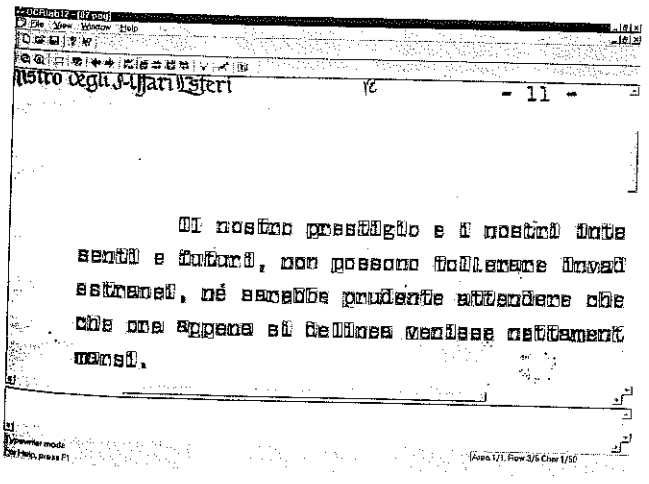


Fig. 2 - The character segmentation procedure.

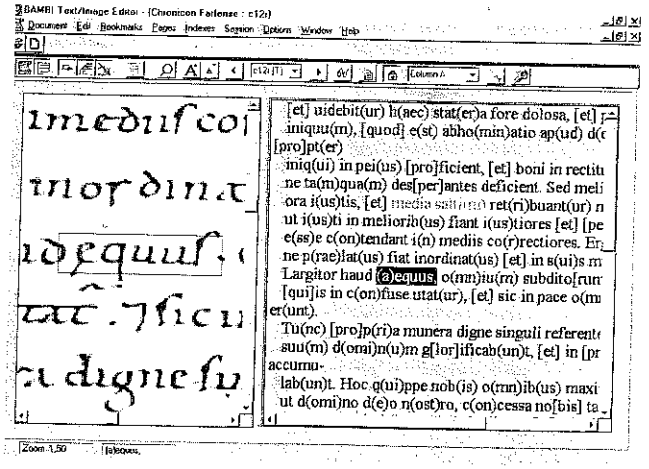


Fig. 3 - Text-image automatic link.

NOTE

1 If there is a lot of text, 2 pages are sufficient. If, however, the amount of text is scarce, 3 or 4 pages are required. Some Italian Libraries made available the documents for the experimental phases: Istituto per il Catalogo Unico delle Biblioteche Italiane, Roma; Biblioteca Nazionale Centrale, Roma; Biblioteca Nazionale, Napoli; Biblioteca Nazionale, Firenze; Biblioteca Medicea Laurenziana, Firenze; Gabinetto Storico Letterario G.P. Vieusseux, Firenze; Fondazione Primo Conti, Fiesole; Istituto di Studi Giuridici sulla Comunità Internazionaleand the Italian Ministero degli Affari Esteri. The software development has been partially realized by MacsTech S.r.l., Pisa.

REFERENCES

[1] Bedini L., Bozzi A., Tonazzini A., 1997, *ERCIM, News* 28, 24.

[2] Geman S., Geman D., 1984, *IEEE Trans. Pattern Anal. Machine Intell.*, 6, 721-740.

[3] Bedini L., Gerace I., Salerno E., Tonazzini A., 1996, in *Advances in Imaging and Electron Physics*, 97, P.W. Hawkes ed., Academic Press, San Diego, 86-189.

[4] You Y., Kaveh M., 1996, *IEEE Trans. Image Processing*, 5, 416-428.

[5] Avi-Itzhak H.I., Diep T.A., Garland H., 1995, *IEEE Trans. Pattern Anal. Machine Intell.*, 17(2), 218-224.

[6] Aas K., Eikvil L., 1996, *Pattern Recognition*, 29 (6), 977-985.

[7] Taxt T., Flynn P.J., Jain A.K., 1989, *IEEE Trans Pattern Anal. Machine Intell.*, 11, 1322-1329.

[8] Tonazzini A., Bedini L., 1998, *Proceedings of SPIE*, 3459, 73-81.

[9] Li S.Z., 1995, *Markov Random Field Modeling in Computer Vision.*, Springer-Verlag, Tokyo.

[10] Hynninen J., Kangas J., Laaksonen J., Kohonen T., 1995, *Technical Report A31*, Helsinki University of Technology, Laboratory of Computer and Information Science.

[11] Bozzi A., 1997, in *Better Access to Manuscripts and Browsing of Images*, A. Bozzi ed., CLUEB, Bologna, 69.