CONSIGLIO NAZIONALE
DE LA RICERCHE
ITALIA

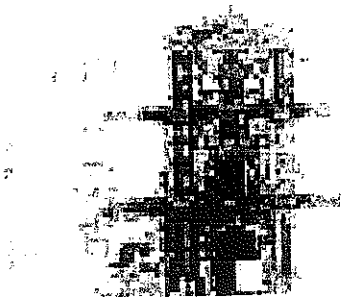
UNIVERSIDAD DE ALCALÁ
ESPAÑA

3rd International Congress on

"Science and Technology for the safeguard of

Cultural Heritage in the Mediterranean Basin"

Alcalá de Henares (Spain)
9 – 14 July 2001



3er Congreso Internacional

"Ciencia y Tecnología Aplicada a la Protección del

Patrimonio Cultural en la Cuenca Mediterránea"

Alcalá de Henares (España)
9 – 14 Julio 2001

OPTIMIZED LASER TECHIQUES FOR THE CONSERVATION OF I156
SCULPTURED ARTWORKS

TLA METHOD AS A COMPARATIVE TOOL TO INVESTIGATE ARTIFICIALLY AND NATURALLY AGED BRONZE SPECIMENS 157

EFFECTIVENESS EVALUATION BY THE TLA OF PROTECTIVE COATING USED FOR OUTDOOR BRONZES 158

COMPARATIVE STUDY OF PROTECTIVE COATING SYSTEMS FOR OUTDOOR BRONZE SCULPTURE 159

SOMBRERO DE INVIERNO PARA LAS ESTATUAS 160

A NEW LOW-COST AND COMPLETE RESTORATION METHOD: A SIMULTANEUS NON-AQUEOUS TREATMENT OF DEACIDIFICATION AND REDUCTION 161

A PAPER-DEGRADING BACTERIAL COMMUNITY: CHARACTERISATION AND ENZYMATIC ACTIVITY OF SOME COMPONENTS 162

DEACIDIFICAZIONE DI MASSA DI BENI ARCHIVISTICI E LIBRARI A MEZZO DI DIAZODERIVATI ALCALINI 163

EFFECT ON THE GROWTH OF A FUNGUS RESPONSIBLE OF PAPER DETERORATION IN ?-RAYS IRRADIATED PAPERS. 164

L'EVOLUZIONE DELLE FORME NELL EUROPA DELLA CARTA 165

INHIBITION OF PAPER BIODETERIORATION BY FUNGI USING SOME NEW TRIAZINE COMPOUNDS 166

NMR CHARACTERIZATION OF PAPER 167

PIXE-? AND MICRO-RAMAN ANALYSIS FOR A NON DESTRUCTIVE CHARACTERISATION OF THE SALERNO EXULTET 168

IMAGE SEGMENTATION AS A PRELIMINARY STEP FOR CHARACTER RECOGNITION IN ANCIENT PRINTED DOCUMENTS 169

RIPRESE AD ALTA RISOLUZIONE PER LA DIAGNOSTICA E LA DOCUMENTAZIONE DI MANOSCRITTI SU PAPIRI ED OSTRAKA 170

ESTRAZIONE DI INFORMAZIONI LINGUISTICHE E TESTUALI DA IMMAGINI DI DOCUMENTI A STAMPA ANTICHI REDATTI IN LINGUA LATINA 171

PRELIMINARY STUDIES FOR TREATMENT OF FOXING ON PAPER: CHEMICAL CHARACTERIZATION 172

OBJECTIVE COMPARISON OF AUDIO RESTORATION METHODS BASED ON SHORT TIME SPECTRAL ATTENUATION 173

ARCHAEOMETALLURGICAL STUDY OF PALEOVENETIAN FIBULAE 175

UNALLOYED COPPER INCLUSIONS IN ANCIENT BRONZE ARTEFACTS 176

IMAGE SEGMENTATION AS A PRELIMINARY STEP FOR CHARACTER RECOGNITION IN ANCIENT PRINTED DOCUMENTS

Luigi Bedini And Anna Tonazzini

Istituto di Elaborazione della Informazione
Area della Ricerca CNR di Pisa
Via G: Moruzzi,1, I-56124 PISA (Italy)
Fax: +39-050-3152810
e-mail: bedini@iei.pi.cnr.it

The development of efficient OCR procedures for very degraded printed documents is still an open issue. On the basis of the results obtained by analyzing and processing several printed ancient documents, we argue that an efficient OCR procedure can be established on the basis of the integration of three processing modules:

- a module for the joint restoration and segmentation of images;
- a module for character recognition, eventually based on neural networks;
- a module for the linguistic analysis of the sequence of characters that have been recognized.

The first module is a fundamental preliminary step for character recognition, which usually relies on isolated characters. This step is particularly critic in the case of ancient printed documents where several degradation processes may cause the characters to touch and merge one another. We propose to integrate techniques of image restoration with techniques of image segmentation, based on Markov Random Field models. The problem is formulated as the minimization of a cost function which accounts for data consistency and expresses both radiometric constraints on the image grey levels and geometrical constraints on the character boundaries. Since the degradation operator, which derives from several and diverse factors, is also unknown, it must be estimated along with the segmented image. Hence, we propose a solution strategy where steps of image estimation iteratively alternate with steps of estimation for the degradation operator. Several results of both simulated and real experiments are shown to validate the method.

IMAGE SEGMENTATION AS A PRELIMINARY STEP FOR CHARACTER RECOGNITION IN ANCIENT PRINTED DOCUMENTS

Luigi Bedini and Anna Tonazzini

Istituto di Elaborazione della Informazione
Area della Ricerca CNR di Pisa
Via G. Moruzzi,1
I-56124 PISA (Italy)
e-mail: surname@iei.pi.cnr.it

Abstract

The development of efficient OCR procedures for very degraded printed documents is still an open issue. On the basis of the results obtained by analyzing and processing several ancient printed documents, we argue that an efficient OCR procedure can be established by means of the integration of the three following processing modules:

-) a module for the joint restoration and segmentation of images;
-) a module for character recognition, eventually based on neural networks;
-) a module for the linguistic analysis of the sequence of characters that have been recognized.

The first module is a fundamental preliminary step for character recognition, which usually relies on isolated characters. This step is particularly critic in the case of ancient printed documents where several degradation processes may cause the characters to touch and merge one another. We propose to integrate techniques of image restoration with techniques of image segmentation, based on Markov Random Field models. The problem is formulated as the minimization of a cost function, which accounts for data consistency and expresses both radiometric constraints on the image gray levels and geometrical constraints on the character boundaries. Since the degradation operator, which derives from several and diverse factors, is also unknown, it must be estimated along with the segmented image. Hence, we propose a solution strategy where steps of image segmentation iteratively alternate with steps of estimation for the degradation operator. Several results of both simulated and real experiments are shown to validate the method.

Introduction

Many Optical Character Recognition software packages are now available to perform automatic recognition of printed characters. Usually character recognition is performed in two steps: in the first step the characters are segmented; in the second step each segmented character is recognized. Unfortunately, these packages cannot be successfully applied to ancient texts where aging of paper, diffusion of ink and other processes have strongly degraded the quality of documents. Thus, the development of efficient OCR procedures for very degraded documents should still be considered an open issue, which presents several difficulties [1]. These are mainly related to non-uniform reduction of the image contrast, non-uniform spacing of the characters, presence of broken, touching and merging characters,

and presence of a strong background noise. All these factors make difficult the correct segmentation of the characters. In [2,3,4] some techniques are proposed to improve the efficiency of the segmentation step. These techniques assume as model of the degradation process a space-variant blur filter and consider the image of the degraded documents as the result of applying this blur filter to the original image and adding noise. The blur filter is considered unknown; thus, segmenting characters means to solve a blind segmentation problem, where blur filter estimation and character segmentation have to be performed simultaneously. Techniques based either on Wiener filter [2] or on Markov Random Fields (MRFs) [3,4] were experimented in segmenting degraded printed documents. These techniques, especially the ones based on MRFs, perform satisfactorily in segmenting characters. However, in presence of strong degradation, they fail and produce a few characters not correctly segmented.

Even in presence of correctly segmented characters, i.e. disjoined and unbroken, the residual degradation and noise still leave a great variability in the character morphology, so that the recognition step, based on standard OCR procedures, performs poorly. Thus many approaches have been proposed, based on the use of Neural Network [5], which are more robust in recognizing degraded characters, when correctly segmented. Global approaches to joint segmentation and recognition have been proposed based on Hidden Markov Models (HMMs) [6]. Such approaches have the advantage that they can be used to recognize a sequence of characters without having first to segment the image in single characters. Moreover, these approaches allow contextual knowledge, expressed in probabilistic form, to be incorporated into the recognition process. In spite of the promising results reported in the literature, when applied to ancient degraded documents, they do not seem to perform satisfactorily, still producing several mistakes, especially when characters are touching or merging.

To overcome the above mentioned difficulties, we argue that the segmentation step and the recognition step have to be integrated in such a way to be able to exploit at best the knowledge we have about the problem. Since a lot of information can be derived from the knowledge we have about the linguistic contents of the text, in [1] we proposed a method based on the integration of the three following processing modules:

-) a pre-processing module for the joint restoration and segmentation of the images
-) a module for character recognition, eventually based on neural networks;
-) a module for the linguistic analysis of the sequence of characters that have been recognized.

The first module, starting with the degraded document, provides a new document restored and segmented into the various characters. The second module, starting from the segmented document, attempts to recognize each character. The zones of the document where the recognition fails are forwarded to the first module, which refines the estimate of the blur mask and the restoration of the zones themselves. The third module takes as input the sequence of characters that have been recognized by the second module and performs a linguistic analysis of the text, by using a reference dictionary. Again, the zones containing those characters that are incorrect from a linguistic point of view are forwarded to the first module, which operates the refining of the blur mask estimate and then produces a new segmentation.

The first module is particularly critic and we have experimented different methods in order to obtain efficient solution to the character segmentation problem. In this paper we describe the method that performed better in segmenting strongly degraded documents. The method tries to integrate techniques of image restoration with techniques of image segmentation, based on Markov Random Field models. The problem is formulated as the minimization of a cost function which accounts for data consistency and expresses both radiometric constraints on the image gray levels and geometrical constraints on the character boundaries. Since the degradation operator, which derives from several and diverse factors, is also unknown, it must be estimated along with the segmented image. Hence, we propose a solution strategy where steps of image segmentation iteratively alternate with steps of estimation for the degradation operator.

Blind character segmentation

The problem of recovering a segmented image and jointly removing the blur is highly ill-posed in that it does not admit a unique solution. To make the problem well-posed we exploit regularization techniques [7,8] based on the definition of suitable mathematical models, and reformulate the problem as an optimization problem to be solved by minimizing a suitable cost function or energy, both with respect to the image f and the blur mask d.

By adopting a Multi-Level Logistic (MLL) model for a 8-neighbors neighborhood system, which is a typical Markov Random Field (MRF) model for piecewise constant images [9,10], we consider a cost function U(f|g,d) which accounts for consistency with the observed image g, smoothness constraints on the gray levels of pixels belonging to homogeneous regions in the image, and geometrical constraints on the character morphology. More 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 formulas, the cost function that we adopt is given by:

$$U(f|g,d) = \|\mathbf{g} - H(d)\,\mathbf{f}\|^2 + \sum_c V_c(f) \qquad (1)$$

where **f** and **g** are the lexicographic 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, where the clique size, shape and orientation depend on the order of the chosen neighborhood system. These potentials can be easily designed to describe both the local smoothness constraint typically enforced by the MLL model and the peculiar features of printed texts above described. With respect to the constraints to be enforced on the blur mask, since we know that it acts as a low-pass filter, we assume that its elements are positive and of unitary sum.

Our blind segmentation problem becomes then:

$$\min_{f,d} \|\mathbf{g} - H(d)\,\mathbf{f}\|^2 + \sum_c V_c(f) \qquad (2)$$

subject to the extra constraints:

$$\sum_{i,j} d_{i,j} = 1 \quad \text{and} \quad d_{i,j} \geq 0 \quad \forall\, i,j \qquad (3)$$

The solution strategy to problem (2)-(3) consists of the alternate execution of steps of image estimation and steps of estimation for the degradation operator, according to the following iterative scheme [11,12,13]:

$$f^{(k)} = \arg\min_f \|\mathbf{g} - H(d^{(k)})\,\mathbf{f}\|^2 + \sum_c V_c(f) \qquad (4a)$$

$$d^{(k)} = \arg\min_d \|\mathbf{g} - H(d)\,\mathbf{f}^{(k)}\|^2 \qquad (4b)$$

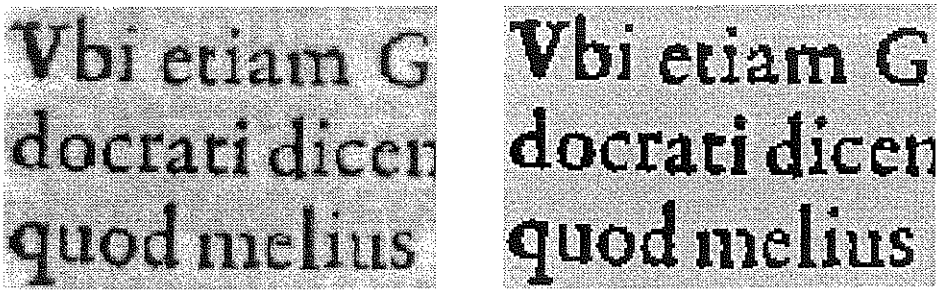
where the constraints (3) are imposed on the solution after each iteration.

In order to perform the minimization (4a), and then to estimate the segmented image, we adopt a Simulated Annealing (SA) algorithm [4,14] with Gibbs sampler [8]. This SA is periodically interrupted to produce a new estimate of the blur mask, via the gradient descent solution of the least-squares problem (4b).

Experimental results

In this Section we show some examples of the performance of the blind restoration method proposed in the paper. From the large set of experiments that we performed, we selected two representative examples, taken from the First Book of the "Opera Omnia" of Cardano. In a first case, we considered a 106x162 portion of a digitized page of the document, taken from a well-preserved microfilm. The resulting image is mildly blurred and noisy. In a second case, we considered a 128x128 portion of a digitized page of the document, taken from an old microfilm. In this case the degradation is much more strong, since it is probably comprehensive of the degradation undertaken by the microfilm support along the time. In all trials convergence to the final values of the parameters and stabilization of the reconstructions were reached in less than 30 iterations of the whole procedure.

Figure 1 (a) shows the available, degraded image of a 106x162 portion of a page of the Cardano's book, taken from the well-preserved microfilm. For purposes of inspection and recognition by the human visual system, the degradation of this image can be considered mild. Nevertheless, it prevents the effective application of standard digital procedures for the separation of the text characters and then a satisfactory application of subsequent OCR functions. Hence, we attempted to improve the quality of the image with blind segmentation, considering a 3x3 blur mask. Considering 25 as gray level for the text characters, and 220 as gray level for the background, we obtained the segmented image shown in Figure 1 (b), and the estimated blur mask of Eq. (5) below.



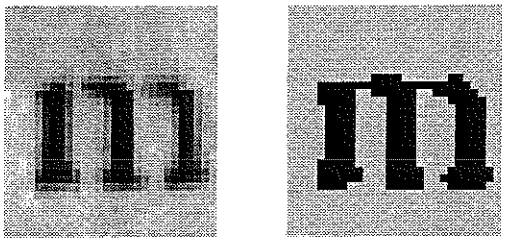
(a) (b)

Fig. 1 - Real degradation: (a) available image; (b) image segmented assuming a 3x3 unknown blur mask.

$$\begin{matrix} 0.095618 & 0.135767 & 0.112601 \\ 0.108409 & 0.158494 & 0.105114 \\ 0.082925 & 0.124810 & 0.092699 \end{matrix} \qquad (5)$$

It is to be noted that the quality of the segmentation is satisfactory, especially in relation to the fact that almost all the characters are correctly separated from each other. Nevertheless, character "m" in the last row of the image appears to be broken. The incorrect separation of some characters, such as the touching of two characters or the fragmentariness of others, could be revealed by means of suitable validation tests to be performed by the neural recognizer or by means of a linguistic analysis of the recognized characters. For those zones the estimation of the blur mask can be refined, in order to further improve the quality of the restoration. For instance, if we process separately the only portion

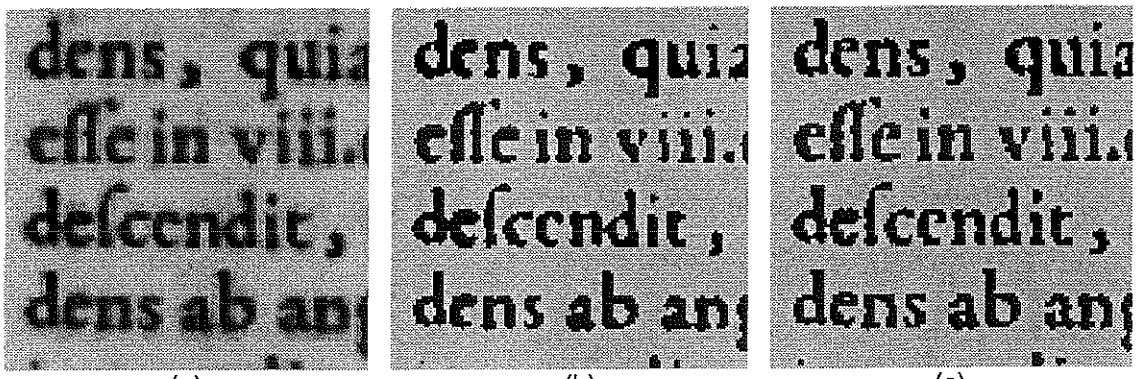
of the image which contains the "m", we obtain the better result shown in Figure 2.



(a) (b)

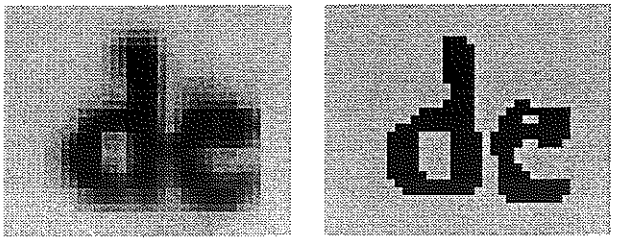
Fig. 2 - Blind segmentation of a portion of the image in Figure 1 (a) containing a single character: (a) original, degraded image; (b) segmented image.

As a second example, we considered the 128x128 image of a region of the same document, this time acquired by an old microfilm (see Figure 3 (a)).



(a) (b) (c)

Fig. 3 - Blind segmentation of a highly blurred image of a portion of an ancient printed document: (a) original, degraded image; (b) image thresholded at grey level 100 (c) segmented image.



(a) (b)

Fig. 4 - Blind segmentation of a portion of the image in Figure 3 (a) containing a couple of merged characters: (a) original, degraded image; (b) segmented image.

Since this time the image is far more blurred, we assumed a 5x5 size for the unknown blur mask. We obtained the segmented image shown in Figure 3 (c). Figure 3 (b) shows the best result we obtained using a simple thresholding procedure, without inverse filtering. It can be noted that in the thresholded image none of the "holes" of the "a" and the "e" have been recovered, and, moreover, it contains several couples of attached characters. The segmented image, although significantly better than the thresholded image, still contains some broken and touching characters, and one "e" without the "hole" inside. These defects confirm the space-variant nature of the degradation which affects the ideal image. Again, the application of the procedure to the only portion of the image containing the incorrectly recovered characters allows for a better estimation of the blur mask, with consequent better

recovering of the characters themselves. As an example, we considered the zone of the image containing the couple of merged characters "d" and "e" in the third row. After processing this zone alone, we obtained the correct separation of the "d" from the "e". Moreover, the "e" presents now a "hole" inside, which allows to distinguish it from the "c". Figure 4 shows the couple "d-e" before and after the blind restoration procedure.

Conclusions

We have proposed a blind image labeling technique, based on MRF image models, to be applied to the segmentation of the text characters of highly degraded ancient printed documents, with the aim to facilitate subsequent phases of recognition and classification of the characters themselves. We formulated our technique as the alternate, iterative minimization of a cost function with respect to the image field and the degradation operator.

Based on the results we have obtained by analyzing and processing several portions of different documents, we conclude that the effectiveness of the proposed technique depends on the severity of the damage affecting the texts. In the case of texts strongly degraded, it is likely to find that some characters are not correctly separated. In that case we found that good results can be obtained by refining the estimate of the blur mask in the zones of the text where the characters were incorrectly segmented. Thus we propose a method based on the integration of the module for text segmentation, herein described, with modules for character recognition and linguistic analysis. In this method, the zones of the text where one of the two latter modules detects an error are forwarded to the module for text segmentation which operates the refining of the blur mask estimate and then produces a new segmentation.

References

1 L. Bedini, A. Tonazzini, "Image restoration oriented to character recognition", *Computer-aided recovery and analysis of damaged text documents*, A. Bozzi Ed., CLUEB, Bologna, 2000, 77-95.

2 L. Bedini, S. Minutoli, A. Tonazzini, "Blind restoration of degraded texts based on Wiener filtering", *Computer-aided recovery and analysis of damaged text documents*, A. Bozzi Ed., CLUEB, Bologna, 2000, 96-120.

3 A. Tonazzini, L. Bedini, S. Minutoli, "Markov Random Field models for blind restoration and labeling of degraded texts", *Computer-aided recovery and analysis of damaged text documents*, A. Bozzi Ed., CLUEB, Bologna, 2000, 121-158.

4 L. Bedini, A. Tonazzini, "Joint blind restoration and segmentation of blurred text characters", Proc. IASTED Int. Conference Signal and Image Processing, 18-21 October 1999, Nassau, Bahamas.

5 H. I. Avi-Itzhak, T. A. Diep and H. Garland, High Accuracy Optical Character Recognition Using Neural Networks with Centroid Dithering, *IEEE Trans. Pattern Anal. Machine Intell.*, 17(2), 1995, 218-224.

6 K. Aas and L. Eikvil, Text page recognition using grey-level features and hidden Markov models, *Pattern Recognition*, 29(6), 1996, 977-985.

7 L. Bedini, I. Gerace, E. Salerno and A. Tonazzini, Models and algorithms for edge-preserving image reconstruction, *Advances in Imaging and Electron Physics*, 97, P.W. Hawkes Ed., 1996, 86-189.

8 S. Geman and D. Geman, Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images, *IEEE Trans. Pattern Anal. Machine Intell..*, 6, 1984, 721-740.

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

10 S. Lakshmanan and H. Derin, Simultaneous parameter estimation and segmentation of Gibbs random fields using simulated annealing, *IEEE Trans. Pattern Anal. Machine Intell.*, 11, 1989, 799-813.

11 G. R. Ayers and J. G. Dainty, Iterative blind deconvolution method and its applications, *Opt. Lett.*, 13, 1988, 547-549.

12 A. Tonazzini, L. Bedini, Using intensity edges to improve parameter estimation in blind image restoration, SPIE's Int. Symposium on Optical Science, Engineering, and Instrumentation, Bayesian Inference for Inverse Problems (SD99), 19-24 July 1998, San Diego, *Proceedings of SPIE*, 3459, 1998, 73-81.

13 Y. You and M. Kaveh, A regularization approach to joint blur identification and image restoration, *IEEE Trans. Image Proc.*, 5, 1996, 416-428.

14 E. Aarts and J. Korst, *Simulated Annealing and Boltzmann Machines*, John Wiley & Sons, 1989.