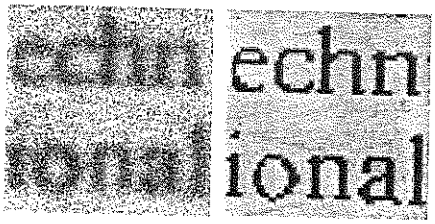
Digital Techniques for Character Recognition in Old Documents

by Luigi Bedini, Andrea Bozzi and Anna Tonazzini

Two research projects have recently been activated by the Italian National Research Council (CNR) with the objective of designing and developing computerised tools to retrieve and restore textual information contained in ancient documents, accessed as digital images. The first, LAperLA, concerns old printed books, while the second, carried out in the framework of a CNR special project that aims at preserving our cultural heritage, regards old manuscripts. Both projects are aimed at implementing an integrated system that improves the quality of the images and, at the same time, activates optical character recognition (OCR) functions. Three CNR Institutes are involved in this activity: the Istituto di Linguistica Computazionale (ILC-Pisa), the Istituto di Elaborazione della Informazione (IEI-Pisa) and the Istituto di Applicazioni del Calcolo (IAC-Rome).

The first step is to convert the documents into digital form. This is not only a valid method for conservation but also makes it possible to operate on the graphic characteristics of the document at various levels. The image is first segmented automatically, identifying and separating those zones containing illustrations, graphs, musical scores, formulae, drawings, etc. These objects are virtually detached from textual zones that only contain alphanumeric characters. This stage, which includes classifying the illustrations and storing them in a database, is performed by IAC.

IEI is responsible for the second step which concerns the implementation of procedures to enhance the quality of the images, especially unfocused images or those obtained from damaged microfilm. Specific image restoration algorithms have been designed for this purpose; they are based on the definition of suitable mathematical models, and the reformulation of the restoration problem as the optimisation of a function that expresses a measure of the fidelity of the recovered image to the degraded image and to known features of the original image. Within a Bayesian framework,



Example of restoration of a section of a degraded printed document.

Markov Random Field models are used to describe local properties of brightness and of the boundaries of the text characters. A generic model for piece-wise constant images is augmented by specific constraints inferred from the peculiar features of printed and/or handwritten characters (eg quasi-binary intensity, variance with respect to a finite set of samples of undegraded characters, length and connection of the boundaries, etc.). Particular attention is also paid to the estimation of the best model for the underlying degradation operator, which derives from a number of often unknown factors (ageing of the documents and/or the microfilms, deterioration due to usage, aberrations of the optical imaging system, low resolution of the digital acquisition). The function thus defined is optimised using algorithms that should ensure a good solution under the enforced constraints and acceptable computational costs. The performances of deterministic descent algorithms, such as the graduated non-convexity and the iterated conditional modes, are thus investigated, together with their possible parallel implementations on dedicated hardware.

The zones of the images that contain written text, enhanced when necessary using the procedures described above, are then processed by a system for segmentation and interpretation developed by ILC. For each word-zone, identified by means of a preliminary phase which roughly segments the page, a finer segmentation is performed which locates each single character within a word or a numeral. Each character is then normalised, ie transformed into a standard form, varying its size and orientation. Finally, the characters are recognised and classified. This step is performed using techniques based on a Neural Network model whose architecture is similar to that of the well known Radial Basis Function Network; it has a good classification capability along with a fast training stage. The number of neurons needed is proportional to the number of characters which are present in the alphabet under consideration, and to the resolution used to acquire the documents. The goal is therefore to implement a neural module whose size permits effective simulations on normal-performance machines, possibly assisted by accelerated parallel cards.

The activities of the research projects described above are mainly of interest in two application areas. The first concerns the use of digital retrieval techniques for the consultation of catalogues of images and texts. Thanks to digital imaging, direct visualisation of the document as an object of study on the computer screen is possible: the access keys for this form of consultation are derived from information usually found in traditional catalogues (author, title, date, language, etc.). Another application area regards the full-text interrogation of the contents of the books. Our aim is to implement a computer-assisted work station that is not limited to providing catalogue information, as is currently the case with SBN (Sistema Bibliografico Nazionale), but, thanks to the OCR method that has been developed, offers direct access to the textual data contained in the ancient books. The potential users of the system are represented by librarians, philologists, archivists, epigraphists, papyrologists and, in general, the students of medieval source documents.

■

Please contact:
Andrea Bozzi – ILC-CNR
Tel: +39 50 560481
E-mail: bozzi@ilc.pi.cnr.it

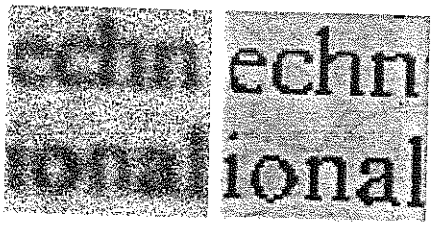
Digital Techniques for Character Recognition in Old Documents

by Luigi Bedini, Andrea Bozzi and Anna Tonazzini

Two research projects have recently been activated by the Italian National Research Council (CNR) with the objective of designing and developing computerised tools to retrieve and restore textual information contained in ancient documents, accessed as digital images. The first, LAperLA, concerns old printed books, while the second, carried out in the framework of a CNR special project that aims at preserving our cultural heritage, regards old manuscripts. Both projects are aimed at implementing an integrated system that improves the quality of the images and, at the same time, activates optical character recognition (OCR) functions. Three CNR Institutes are involved in this activity: the Istituto di Linguistica Computazionale (ILC-Pisa), the Istituto di Elaborazione della Informazione (IEI-Pisa) and the Istituto di Applicazioni del Calcolo (IAC-Rome).

The first step is to convert the documents into digital form. This is not only a valid method for conservation but also makes it possible to operate on the graphic characteristics of the document at various levels. The image is first segmented automatically, identifying and separating those zones containing illustrations, graphs, musical scores, formulae, drawings, etc. These objects are virtually detached from textual zones that only contain alphanumeric characters. This stage, which includes classifying the illustrations and storing them in a database, is performed by IAC.

IEI is responsible for the second step which concerns the implementation of procedures to enhance the quality of the images, especially unfocused images or those obtained from damaged microfilm. Specific image restoration algorithms have been designed for this purpose; they are based on the definition of suitable mathematical models, and the reformulation of the restoration problem as the optimisation of a function that expresses a measure of the fidelity of the recovered image to the degraded image and to known features of the original image. Within a Bayesian framework,



Example of restoration of a section of a degraded printed document.

Markov Random Field models are used to describe local properties of brightness and of the boundaries of the text characters. A generic model for piece-wise constant images is augmented by specific constraints inferred from the peculiar features of printed and/or handwritten characters (eg quasi-binary intensity, variance with respect to a finite set of samples of undegraded characters, length and connection of the boundaries, etc.). Particular attention is also paid to the estimation of the best model for the underlying degradation operator, which derives from a number of often unknown factors (ageing of the documents and/or the microfilms, deterioration due to usage, aberrations of the optical imaging system, low resolution of the digital acquisition). The function thus defined is optimised using algorithms that should ensure a good solution under the enforced constraints and acceptable computational costs. The performances of deterministic descent algorithms, such as the graduated non-convexity and the iterated conditional modes, are thus investigated, together with their possible parallel implementations on dedicated hardware.

The zones of the images that contain written text, enhanced when necessary using the procedures described above, are then processed by a system for segmentation and interpretation developed by ILC. For each word-zone, identified by means of a preliminary phase which roughly segments the page, a finer segmentation is performed which locates each single character within a word or a numeral. Each character is then normalised, ie transformed into a standard form, varying its size and orientation. Finally, the characters are recognised and classified. This step is performed using techniques based on a Neural Network model whose architecture is similar to that of the well known Radial Basis Function Network; it has a good classification capability along with a fast training stage. The number of neurons needed is proportional to the number of characters which are present in the alphabet under consideration, and to the resolution used to acquire the documents. The goal is therefore to implement a neural module whose size permits effective simulations on normal-performance machines, possibly assisted by accelerated parallel cards.

The activities of the research projects described above are mainly of interest in two application areas. The first concerns the use of digital retrieval techniques for the consultation of catalogues of images and texts. Thanks to digital imaging, direct visualisation of the document as an object of study on the computer screen is possible: the access keys for this form of consultation are derived from information usually found in traditional catalogues (author, title, date, language, etc.). Another application area regards the full-text interrogation of the contents of the books. Our aim is to implement a computer-assisted work station that is not limited to providing catalogue information, as is currently the case with SBN (Sistema Bibliografico Nazionale), but, thanks to the OCR method that has been developed, offers direct access to the textual data contained in the ancient books. The potential users of the system are represented by librarians, philologists, archivists, epigraphists, papyrologists and, in general, the students of medieval source documents.

■

Please contact:
Andrea Bozzi – ILC-CNR
Tel: +39 50 560481
E-mail: bozzi@ilc.pi.cnr.it