DIAR: Advances in Degradation Modeling and Processing

Conference Paper · June 2008

DOI: 10.1007/978-3-540-69812-8_1 · Source: DBLP

CITATIONS
6

READS
90

2 authors:

Mohamed Cheriet
École de Technologie Supérieure
478 PUBLICATIONS **6,449** CITATIONS

SEE PROFILE

Reza Farrahi Moghaddam
École de Technologie Supérieure
86 PUBLICATIONS **1,215** CITATIONS

SEE PROFILE

DIAR: Advances in Degradation Modeling and Processing

Mohamed Cheriet and Reza Farrahi Moghaddam

Synchromedia Laboratory for Multimedia Communication in Telepresence,
École de Technologie Supérieure, Montréal, QC H3C 1K3 Canada
mohamed.cheriet@etsmtl.ca

Abstract. State-of-the-art OCR/ICR algorithms and software are the result of large-scale experiments on the accuracy of OCR systems and proper selection of the size and distribution of training sets. The key factor in improving OCR technology is the degradation models. While it is a leading-edge tool for processing conventional printed materials, the degradation model now faces additional challenges as a result of the appearance in recent years of new imaging media, new definitions of text information, and the need to process low quality document images. In addition to discussing these challenges in this paper, we present well-developed degradation models and suggest some directions for further study. Particular attention is paid to restoration and enhancement of degraded single-sided or multi-sided document images which suffer from bleed-through or shadow-through.

1 Introduction

Despite enormous advances in DIAR and the development of state-of-the-art technologies in document recognition and interpretation, the basics of DIAR have not changed. Almost all OCR systems consist of a multi-parameter decision algorithm and a training set. If a large enough training set that is also well-distributed is available, the error rate will be very small. But, in real applications, the results are not usually very promising. The main problem is that the document Images (DIs) that users feed to the trained systems are not exactly the same as the source images stored in the computer. Even the print process itself may introduce minor changes to the document (these changes are intentional modifications which are made to make the document easier to read). It has therefore been suggested, according to the results of several tests [1,2,3], that the training sets used be as real as possible. For a number of small and specific applications, this strategy, with manual preparation of the training sets, will solve the problem. However, in many applications, manually generating large training sets is actually impossible, as well as being subject to errors. Also, for huge applications, such as universal OCR systems, very general datasets with controlled distribution are needed. For these reasons, among others, degradation modeling (DM) in DIAR has been developed. In the next section, DM in general will be discussed from several points of view. Special attention will be

A. Campilho and M. Kamel (Eds.): ICIAR 2008, LNCS 5112, pp. 1–10, 2008.

given to the modeling of physical degradations, which are common in very old handwritten documents. In section 3, some methods are presented for enhancing and restoring physically degraded DIs.

2 Degradation Modeling

A degradation model is, by definition [3], an algorithm which is able to generate, based on some user-specific parameters and distributions, a set containing an unlimited number of DIs (usually single characters) which suffer from some sort of defect, and substitute a real dataset for it (if one exists). DMs not only reduce the need to compile a real dataset and consequently the labor and costs associated with projects, but they also provide (or, more accurately, are built on) a basic understanding of the defect and degradation phenomena, and can therefore be used in the development of enhancement and restoration techniques to address such degradations. The most important impact of DMs is their ability to provide frameworks for comparing different and competitive recognition and restoration algorithms in an unlimited number of repeatable benchmark tests. Although the idea behind them is very simple, there are some fundamental questions to be addressed, which is why it has taken so long to develop state-of-the-art DMs. Briefly, DMs must be:

1. Capable to be calibrated: It should be possible for the model to be expressed by a set of numerical parameters. For a specific application, the model can be calibrated by adjusting its parameters.
2. Resistant to over-training: The model should behave somewhat randomly, and its effects should have that characteristic as well. This intrinsic behavior ensures that the training sets will be independent of one another, even if they are produced with the same parameter values.
3. Able to differentiate parameter distributions: For any set of defective DIs, it should be possible to deduce which distributions of model parameters may result in the same set of DIs.
4. Able to repeat results: As with any engineering task, the model should generate the same set of defective documents if the same parameters and seeds are used.

Below, we divide DMs into two subcategories. The first contains DMs which have been developed for printed documents and are state-of-the-art. DMs in this category mainly focus on the defects that arise in the printing and imaging phases. The second consists of DMs which represent the defects that are the result of some external phenomena and that persist in the document itself. This type of defect is very common in very old documents, and also in printed media in which low-quality ink and paper have been used.

2.1 Document Image Degradation Modeling (DIDM)

Document Image Degradation Modeling (DIDM) has a long history and has been recognized since the beginning of DIAR [4,5]. The main objective of DIDM is to

model imaging defects such as coarsing, thinning, thickening, geometry deformation, etc. These and other similar defects usually appear when the document is either imaged or printed. Some are due to human error (which will be discussed in more detail in subsection 2.3) and many are the result of the nonlinear and variable nature of the material and equipment used in imaging devices. Human error can be avoided by training users (or by changing the definition of degradation [6]). However, for the nonlinear properties of imaging devices, calibration is the only solution. Practically every month, a new material in optics or some other field of imaging technology is introduced, and, without a general DM which is capable of estimation, calibration is actually impossible. There are many DMs in the DIDM category [3,7,8] (for a review see [3]). The most sophisticated model is presented in [5,3]. It has several parameters, such as output sampling rate, rotation, scaling factors for horizontal and vertical directions, translation offsets, jitter, defocussing, sensitivity, and threshold [3]. Among these many parameters, the two most important are defocussing and threshold [9]. By applying parameter estimation, the proper parameter value can be estimated for any application [10,11,12]. Work on validating DMs has been carried out in [13,14].

2.2 Document Degradation Modeling (DDM)

The analysis of handwritten and very old documents has introduced new DIAR requirements, as this type of document is usually suffering from severe degradations prior to the imaging stage. Although the degradations are very different, they fall into two distinct classes: those that have an external source, and those originating in the document itself. External degradations consist of unrecoverable defects (such as the loss of some part(s) of the document paper itself) and recoverable problems (such as small cracks and thin overlays). The defects originating in the document itself are more common, and reduce the documents readability and recognition rate very seriously. For example, bleed-through and shadow-through are two challenging problems in double-sided documents of poor-quality materials [15].

In Figure 1, a real[1] double-sided DI which suffers from the bleed-through problem is presented [16]. In Figures 1(a) and 1(b), the best estimations of the source document are presented (obtained using the restoration method proposed in subsection 3.2). The input images are shown in Figures 1(c) and 1(d). To observe the nonlinear and complex nature of the bleed-through problem, the ICA method [17,18,19] is applied to the input images, and the results are shown in Figures 1(e) and 1(f), and normalized for better visualization. It can easily be seen from the outputs that some minor interference problems remain in the images, especially associated with the boundaries of the strokes. This effect can be related to the nonlinear nature of the seepage phenomenon, which results in the spreading of ink and the smoothness of the stroke edges. Another reason for the remaining interference patterns is weak registration of the two sides of the document. The high level of sensitivity of the linear methods, such as ICA, to registration limits their application to the restoration of bleed-through

[1] http://www.site.uottawa.ca/~edubois/documents/

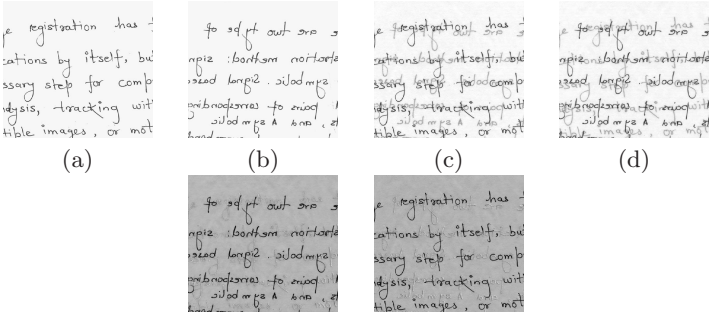


Fig. 1. Performance of the ICA method in a real case. a) and b) show the source images of the recto and verso sides of the document; c) and d) show the degraded images, which are linear combinations of the source images; e) and f) display the results of applying the ICA method.

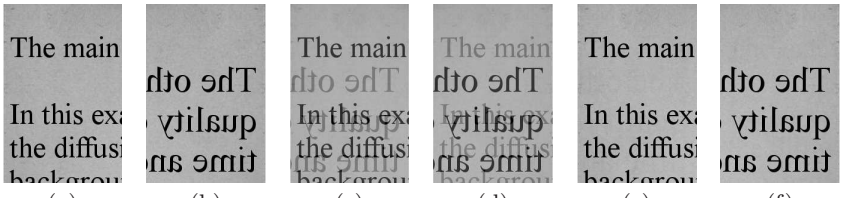


Fig. 2. Performance of the ICA method in a linear case. a) and b) show the source images of the recto and verso sides of the document; c) and d) show the degraded images, which are linear combinations of the source images; e) and f) display the results of applying the ICA method.

interference patterns. In section 3, restoration methods which work on the same basis (and based on the same physics) as the bleed-through phenomenon are proposed. These methods are less sensitive to registration and to the nonlinear nature of the phenomenon. A ground-truth test is presented in Figure 2, in the form of a linear seepage case, in which ICA is able to exactly separate the main text from the interference patterns. This test not only confirms the accuracy of ICA codes, but also proves that the failure of ICA in Figure 1 is due to the nonlinear nature of the phenomenon in the real example. The nonlinear nature of the bleed-through effect has been observed in many studies. For example, in [20], it was found that the edges of the interference patterns are very weak and smooth. This smoothing effect is due to smearing of ink through the paper. The nonlinear behavior of defects can only be addressed if we use proper nonlinear modeling of the phenomenon. If we look at the degradation problems in documents from a physical point of view, many of the degradations are the result of seepage processes which occur over time. The seepage of ink

through paper is a very complex phenomenon, and several parameters, such as the thickness of the paper, the distribution of the paper fibers, and ink quality, all play an important role. Seepage is actually the flow of ink through the porous medium of paper [21]. From this point of view, the paper can be considered as a collection of many small units which can contain a fraction of the ink. At the same time, each unit is able to transfer the ink to other units based on its saturation state and other nonlinear parameters [21]. Many similar phenomena occur in the physical world, such as in water-soil [22,23] and oil-soil [24]systems, in which seepage and containment of the soil bulk are of great importance. Several models for the ink-paper system have been developed, such as Brownian motion [25,26], cellular automaton-based simulation [27], and the balance method [28,29,30]. Almost all these models are based on diffusion processes of some kind, because these processes are closely related to the physics of seepage phenomena.

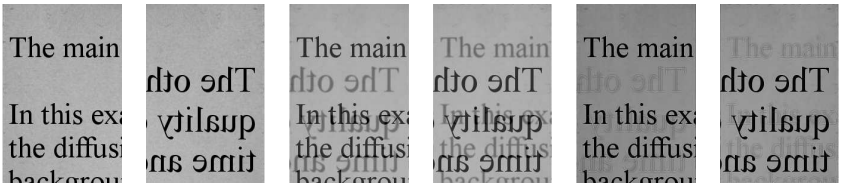
We propose to construct a DM based on some diffusion processes which provide an exchange of information between several sources of information, such as the recto-side image and the verso-side image, as well as additional background information (which represents the surface of the paper after a long period of time). To our knowledge, there is only one other DDM. In that model, the shadow-through effect is modeled using blurring and transformation operators [31]. Our model is based on a physical understanding of degradation. In mathematical terms, our DM can be written in the form of the following governing equation [32]:

$$\frac{\partial u}{\partial t} = \sum_{i \text{ counts on the sources}} \text{DIFF}\,(u, s_i, \cdots) \tag{1}$$

where u is the DI and s_i is the i^{th} source (for example, the verso side of document). Every process of information exchange is formulated by a diffusion process DIFF, which also depends on certain parameters. Any DM of this kind should have at least the following parameters for controlling various aspects of the model:

1. Time period: the aging of the document.
2. Diffusion parameter for the ink and paper: the extent to which ink can become smeared on the paper.
3. Interference growth parameter: the thickness and quality of the paper.
4. External overlay growth parameter: the parameter takes into account the effects of the environment of the document on the quality of the text (aging of the paper is also affected by this parameter).

As a test, in Figure 3, a synthesized double-sided document image which suffers from the bleed-through problem is presented. The degraded images are obtained using our DM, equation (1). After applying the ICA method, the remaining patterns are very similar to the real case in Figure 1, which shows that the DM can be used to generate large-scale datasets which suffer from bleed-through or similar problems.



(a) (b) (c) (d) (e) (f)

Fig. 3. Performance of the ICA method in a simulated case. a) and b) show the source images of the recto and verso sides of the document; c) and d) show the degraded images, which are linear combinations of the source images; e) and f) display the results of applying the ICA method.

2.3 Human Originated Degradations

"O Lord, help me not to despise or oppose what I do not understand. William Penn"

Many state-of-the-art OCR systems are trained and developed under specific and controlled conditions and assumptions. Although these conditions and assumptions can easily be applied to many cases of DIAR, because of their focus on achieving better scanned output, the users of OCR systems usually prevent the fulfillment of these requirements. Indeed, this focus has a major impact on the degradation level of DIs [6]. It has been found, for example, that the threshold and edge spread play a significant role in personal preferences. Studies of this kind help us to better understand readability and user viewing preferences, with the result that OCR systems can now be trained and optimized based on the same parameters that are compatible with user-created DIs.

The printing stage is another area where human preferences play a significant role in changes that have been made to DIs. Based on the limitations of printing technology and computer graphics, and the differences between them, printed material is usually modified for better readability and clarity. The methods used to change printed material are very different, and have been developed and exploited by the manufacturers of printers and copiers [33]. Again, DMs can be used to study and model print-stage defects. A better solution for defects of human origin is to change the principles underlying DIAR systems and make them more perception-oriented [34].

3 Processing of Degraded Document Images

Restoration of defects not only results in improved readability, but also has a significant impact on the recognition rate of OCR systems. Once again, DMs have a major role to play in DI restoration. These models can provide a detailed understanding of defects, and therefore a proper restoration technique and values can be adapted to address them. In this section, we focus on the defects that have a physical origin and which persist on the document. One of most important and

challenging problems for both printed media (newspapers and magazines) and very old documents is the bleed-through (or shadow-though) problem. Restoration of this type of defect has often been studied. In [35,36,37], for example, some transformations have been used for the recovery of the recto and verso sides of double-sided scanned documents. Also, some restoration methods have been designed based on the smart binarization methods [38,39]. Statistical methods such as Independent Component Analysis (ICA) and Blind Source Separation (BSS) have also been used for double-sided documents [40,41,42]. Neural networks have been used for separation and modeling in this field as well [43]. Finally, methods which are a combination of several techniques, such as segmentation and inpainting [44,20], have been suggested and used to obtain interference-free outputs.

In the following subsections, some restoration methods based on the DDM discussed in the previous section (subsection 2.2) will be proposed. These methods are basically diffusion-based, and are very similar to the physical phenomena involved in the degradations.

3.1 Single-Sided Document Images

In single-sided DIs, the main problem and defect is the degraded, complex, and variable background. Also, interference from other sources of information may reduce the readability of the document. With the basics of DM in mind (subsection 2.2, we can propose a diffusion-based restoration method which consists of two sources of information: the input image and the estimated background. The method uses the diffusion process on the input image for enhancement and to sharpen edges and boundaries. At the same time, another diffusion process based on the estimated background information will clean up the degraded background and other interference patterns on the input image. The estimated background will limit the background diffusion to the regions without text and strokes, and will therefore preserve the main strokes of the input DI.

3.2 Double-Sided Document Images

In double-sided DIs, complex degradations, such as the bleed-through effect, can be restored using extra information from the verso side of the DI. By including the information of both the recto and verso sides of a DI, a proposed restoration method is able to remove the interference patterns, even when the intensity of the interference patterns is higher than that of the main strokes. The method is the same as the single-sided method (see subsection 3.1) except that it has an extra inverse diffusion process from the verso-side information to the recto-side information. The effect of this extra diffusion process is to convert the interference patterns of the verso side (which actually constitute meaningful information) to the background information, by pushing them to higher gray-level values. This discrimination of interference patterns will be followed by the clean-up operation of the background diffusion, which will fill up and remove the replaced interference. In actual processing, all three diffusion processes (regular diffusion on the recto side, background diffusion, and inverse diffusion from verso to the recto side) will perform simultaneously.

4 Conclusion

The problem of degradation modeling is discussed. A new category in degradation modeling is introduced which covers the physical and persistent defects. The basic parameters and requirements of this class of models are presented. Then, the human factor in the production and introduction of defects is discussed. Finally, based on the degradation model, some restoration methods are proposed which can be used to recover physical degradations such as bleed-through and shadow-through.

Acknowledgments

The authors would like to thank the NSERC of Canada for their financial support.

References

1. Rice, S., Kanai, J., Nartker, T.: A report on the accuracy of ocr devices. Technical Report TR-92-02, Univ. Nevada Las Vegas, Las Vegas, Nevada (1992)
2. Rice, S., Jenkins, F., Nartker, T.: The fifth test of ocr accuracy. Technical Report TR-96-01, ISRI, Univ. Nevada Las Vegas, Las Vegas, Nevada (April 1996)
3. Baird, H.: The State of the Art of Document Image Degradation Modelling. In: Digital Document Processing: Major Directions and Recent Advances, pp. 261–279. Springer, Heidelberg (2007)
4. Baird, H.: Document image defect models. In: Proc. IAPR Workshop Synthetic and Structural Pattern Recognition. Murray Hill, NJ, June 13–15 (1990)
5. Baird, H.: The state of the art of document image degradation modeling. In: Proc. of 4 th IAPR International Workshop on Document Analysis Systems, Rio de Janeiro, Brazil, pp. 1–16 (2000)
6. Hale, C., Barney-Smith, E.: Human image preference and document degradation models. In: Barney-Smith, E. (ed.) Ninth International Conference on Document Analysis and Recognition, 2007. ICDAR 2007, vol. 1, pp. 257–261 (2007)
7. Kanungo, T., Haralick, R.M., Phillips, I.: Nonlinear local and global document degradation models. Int. Journal of Imaging Systems and Technology 5, 220–230 (1994)
8. Zi, G., Doermann, D.: Document image ground truth generation from electronic text. In: Doermann, D. (ed.) ICPR 2004. Proceedings of the 17th International Conference on Pattern Recognition, 2004, vol. 2, pp. 663–666 (2004)
9. Ho, T.K., Baird, H.: Large-scale simulation studies in image pattern recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence 19(10), 1067–1079 (1997)
10. Kanungo, T., Zheng, Q.: Estimating degradation model parameters using neighborhood pattern distributions: an optimization approach. Transactions on Pattern Analysis and Machine Intelligence 26(4), 520–524 (2004)
11. Barney-Smith, E.H.: Estimating scanning characteristics from corners in bilevel images. In: Proceedings of SPIE. Document Recognition and Retrieval VIII, San Jose, CA, January 21-26, vol. 4307, pp. 176–183 (2001)

12. Yam, H.S., Barney Smith, E.: Estimating degradation model parameters from character images. In: Barney Smith, E. (ed.) Proceedings. Seventh International Conference on Document Analysis and Recognition, 2003, Edinburgh, Scotland, August 3-6, vol. 2, pp. 710–714 (2003)
13. Kanungo, T., Haralick, R., Baird, H., Stuezle, W., Madigan, D.: A statistical, nonparametric methodology for document degradation model validation. Transactions on Pattern Analysis and Machine Intelligence 22(11), 1209–1223 (2000)
14. Kanungo, T., Haralick, R., Baird, H., Stuetzle, W., Madigan, D.: Document degradation models: Parameter estimation and model validation. In: Proc. of Int. Workshop on Machine Vision Applications, Kawasaki, Japan, December 1994, pp. 552–557 (1994)
15. Lesk, M.: Substituting images for books: The economics for libraries. In: Symposium Document Analysis and Information Retieval, pp. 1–6 (1996)
16. Dubois, E., Dano, P.: Joint compression and restoration of documents with bleedthrough. In: Proc. IS&T Archiving 2005, Washington DC, USA, April 2005, pp. 170–174 (2005)
17. Hyvarinen, A., Oja, E.: Independent component analysis: algorithms and applications. Neural Networks 13(4-5), 411–430 (2000)
18. Oja, E., Yuan, Z.: The fastica algorithm revisited: Convergence analysis. IEEE Transactions on Neural Networks 17(6), 1370–1381 (2006)
19. Cichocki, A., Amari, S., Siwek, K., Tanaka, T., Phan, A.H., Zdunek, R.: Icalab matlab toolbox ver. 3 for signal processing (2007)
20. Tan, C.L., Cao, R., Shen, P., Wang, Q., Chee, J., Chang, J.: Removal of interfering strokes in double-sided document images. In: Cao, R. (ed.) IEEE Workshop on Applications of Computer Vision 2000, pp. 16–21 (2000)
21. Wang, X., Sun, J.: The researching about water and ink motion model based on soil-water dynamics in simulating for the chinese painting. In: Sun, J. (ed.) Fourth International Conference on Image and Graphics, 2007. ICIG 2007, pp. 880–885 (2007)
22. Chen, L., Zhu, J., Young, M., Susfalk, R.: Modeling polyacrylamide transport in water delivery canals. In: ASA-CSSA-SSSA International Annual Meetings, Indianapolis, IN, November 12-16, pp. 294–6 (2006)
23. Roth, K.: Scaling of water flow through porous media and soils. European Journal of Soil Science 59(1), 125–130 (2008)
24. Vaziri, H.H., Xiao, Y., Islam, R., Nouri, A.: Numerical modeling of seepage-induced sand production in oil and gas reservoirs. Journal of Petroleum Science and Engineering 36(1), 71–86 (2002)
25. Huang, S.W., Way, D.L., Shih, Z.C.: Physical-based model of ink diffusion in chinese ink paintings. Journal of WSCG 10(3), 520–527 (2003)
26. Yongxin, S., Jizhou, S., Haijiang, Z.: Graphical simulation algorithm for chinese ink wash drawing by particle system (chinese). Journal of Computer-Aided Design & Computer Graphics 15(6), 667–672 (2003)
27. Zhang, Q., Sato, Y., Takahashi, J.Y., Muraoka, K., Chiba, N.: Simple cellular automaton-based simulation of ink behaviour and its application to suibokuga-like 3d rendering of trees. The Journal of Visualization and Computer Animation 10(1), 27–37 (1999)
28. Xiujin, W., Jingshan, J., Jizhou, S.: Graphical simulator for chinese ink-wash drawing. Transactions Of Tianjin University 8(1), 1–7 (2002)
29. Mei-jun, S., Ji-zhou, S., Bin, Y.: Physical modeling of "xuan" paper in the simulation of chinese ink-wash drawing. In: Ji-zhou, S. (ed.) International Conference on Computer Graphics, Imaging and Vision: New Trends, 2005, pp. 317–322 (2005)

30. Yu, Y., Lee, D., Lee, Y., Cho, H.: Interactive rendering technique for realistic oriental painting. Journal of WSCG 11(1), 538–545 (2003)
31. Zi, G.: Groundtruth generation and document image degradation. Technical Report LAMP-TR-121/CAR-TR-1008/CS-TR-4699/UMIACS-TR-2005-08, University of Maryland, College Park (2005)
32. Cheriet, M., Farrahi Moghaddam, R.: Degradation modeling and enhancement of low quality documents. In: WOSPA 2008, Sharjah, UAE (to appear, 2008)
33. Lee, J.H., Allebach, J.: Inkjet printer model-based halftoning. IEEE Transactions on Image Processing 14(5), 674–689 (2005)
34. Saund, E., Fleet, D., Mahoney, J., Lamer, D.: Rough and degraded document interpretation by perceptual organization. In: Doermann, D. (ed.) Proceedings 5^{th} Symposium on Document Image Understanding Technology (SDIUT), UMD (2003)
35. Sharma, G.: Show-through cancellation in scans of duplex printed documents. IEEE Transactions on Image Processing 10(5), 736–754 (2001)
36. Knox, K.T., Rochester, N.: Show-through correction for two-sided documents (July 1997)
37. Tan, C.L., Cao, R., Shen, P.: Restoration of archival documents using a wavelet technique. Transactions on Pattern Analysis and Machine Intelligence 24(10), 1399–1404 (2002)
38. Leedham, G., Varma, S., Patankar, A., Govindaraju, V.: Separating text and background in degraded document images - a comparison of global thresholding techniques for multi-stage thresholding. In: Proc. Eighth International Workshop on Frontiers in Handwriting Recognition, August 6-8, pp. 244–249 (2002)
39. Nishida, H., Suzuki, T.: Correcting show-through effects on document images by multiscale analysis. In: Suzuki, T. (ed.) Proceedings 16th International Conference on Pattern Recognition, 2002, vol. 3, pp. 65–68 (2002)
40. Gerace, I., Cricco, F., Tonazzini, A.: An extended maximum likelihood approach for the robust blind separation of autocorrelated images from noisy mixtures. Independent Component Analysis and Blind Signal Separation, 954–961 (2004)
41. Tonazzini, A., Salerno, E., Bedini, L.: Fast correction of bleed-through distortion in grayscale documents by a blind source separation technique. International Journal on Document Analysis and Recognition 10(1), 17–25 (2007)
42. Salerno, E., Tonazzini, A., Bedini, L.: Digital image analysis to enhance underwritten text in the archimedes palimpsest. International Journal on Document Analysis and Recognition 9(2), 79–87 (2007)
43. Zhang, X., Lu, J., Yahagi, T.: Blind separation methods for image show-through problem. In: Lu, J. (ed.) 6th International Special Topic Conference on Information Technology Applications in Biomedicine, 2007. ITAB 2007, November 8-11, pp. 255–258 (2007)
44. Dubois, E., Pathak, A.: Reduction of bleed-through in scanned manuscript documents. In: Proc. IS&T Image Processing, Image Quality, Image Capture Systems Conference (PICS 2001), Montreal, Canada, April 2001, pp. 177–180 (2001)