IJDAR (2009) 11:183–201
DOI 10.1007/s10032-008-0076-2

ORIGINAL PAPER

Low quality document image modeling and enhancement

Reza Farrahi Moghaddam · Mohamed Cheriet

Received: 22 February 2008 / Revised: 29 November 2008 / Accepted: 18 December 2008 / Published online: 28 February 2009

Abstract In order to tackle problems such as shadow-through and bleed-through, a novel defect model is developed which generates physically damaged document images. This model addresses physical degradation, such as aging and ink seepage. Based on the diffusive nature of the physical defects, the model is designed using virtual diffusion processes. Then, based on this degradation model, a restoration method is proposed and used to fix the bleed-through effect in double-sided document images using the reverse diffusion process. Subjective and objective evaluations are performed on both the degradation model and the restoration method. The experiments show promising results on both real and generated data.

Keywords Defect modeling · Document enhancement · Document image processing · Anisotropic diffusion method

1 Introduction

In the past, primitive printing and imaging technologies were usually responsible for generating low-quality document images. By virtue of incorporating advanced optical technology and accurate modeling of the scanning process and its corresponding defects [3] into hardware and software in the printing and imaging industries, low-quality document images are now only produced by human error. However, there are many cases where dealing with such images is unavoidable. Very old documents and ancient texts, for example, have suffered from physical degradation of several types as a result of their age. The quality of many types of printed media, such as newspapers and magazines, is also poor because of the often very thin or inferior papers used to produce them. Even modern books and texts which have not been kept in an appropriate environment may not be easy to read. As many of these physically damaged documents contain important information, and there is a large number of them, fast, accurate methods are needed to restore them [1,5,10,13,20,38,39]. Defects in document images arise mainly from two sources: printing-imaging processes and physical phenomena.

For document images which have suffered degradation as a result of process defects, there are very well-developed models which can be used to generate datasets of text in the form of a single character, a single word, or even a whole page [2–4,15–18,28,29,43,47].

Much of the degradation suffered by document images that stems from physical processes is the result of some type of diffusion which occurs over a period of time. Such defects are common in very old documents, as well as in media and documents made of poor-quality materials, and they are highly destructive [22]. Despite state-of-the-art advancements in defect models for printing and imaging degradation, there is, to our knowledge, only one model [46,47] which covers the defects of document images that have external and physical origins. In that model, the shadow-through effect is modeled using blurring and transformation operators.

To better illustrate stroke seepage, a real example of ink bleed-through[1] is shown in Fig. 1 [11]. The nonlinear nature of ink seepage can easily be seen from the interference patterns of the verso side that show through onto the recto side.

R. Farrahi Moghaddam (✉) · M. Cheriet
Synchromedia Laboratory for Multimedia Communication in Telepresence, École de Technologie Supérieure, Montréal, QC H3C 1K3, Canada
e-mail: reza.farrahi-moghaddam.1@ens.etsmtl.ca; imriss@yahoo.com

M. Cheriet
e-mail: mohamed.cheriet@etsmtl.ca

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

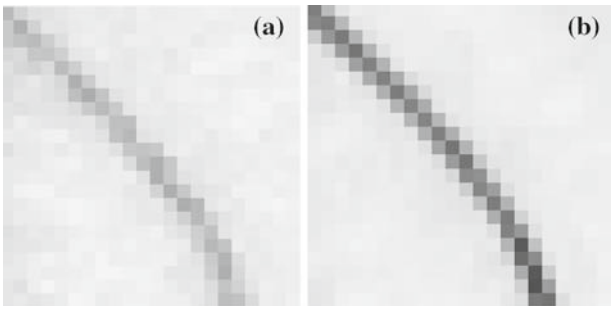


Fig. 1 **a** Recto side of a document image showing a *thin stroke* on the verso side and its *nonlinear pattern* on the recto side. **b** Verso side of the same document

This nonlinear behavior only can be addressed if we use proper nonlinear modeling of the phenomenon. Seepage of ink through a paper is a complex physical phenomenon in which many parameters, such as thickness, the characteristics of the paper, the distribution of the paper fibers, and ink quality, are involved. From the fluid mechanics point of view, seepage is the flow of liquid through a porous medium that actually is a collection of numerous of tiny cubages [42]. The presence of cellulose fibers in particular results in a very complex fluid dynamics. Although at micro scale, almost all fluid-related phenomena in porous media are known [44], performing numerical computations at such a scale is impossible because of the high computational cost. Consequently, an equivalent model at macroscopic scale is usually used, representing the average behavior of micro-scale phenomena. The details of this equation depend on the nature of the paper and the ink used in the document under study. For each type of paper using one of the micro-scale methods, such as molecular dynamics, the macro-scale behavior can be estimated and modeled as functions and parameters in the governing equation. There are different models for studying liquid seepage in porous media [42]. However, despite the use of various coefficients, the final macro-scale equation is usually diffusion-like. This is a common feature of many other systems as well, such as soil–water [7,30] and soil–oil [40] systems.

After developing the degradation model, we use the diffusion-based methods, which are very similar to the physical processes of degradation, to restore a document suffering from the bleed-through effect. The problem of the restoration of double-sided documents is of great interest, and it has been studied from various points of view [12,21,33,38]. The restoration methods designed to remove interference include the smart binarization methods [21,24,34], statistical methods, such as independent component analysis (ICA) and blind source separation (BSS) [32,38], neural networks [45], and methods which are a combination of several techniques such as segmentation and inpainting [12,35,36].

The structure of the paper is as follows: In Sect. 2, a defect model for producing defects of physical origin into document images is developed. This model is based on the diffusion processes and is very similar to the natural processes that result in physical defects such as aging and ink seepage. Then, in Sect. 3, a restoration method is discussed which is capable of recovering degraded double-sided document images. This method emerges from our defect model and is based on artificial reverse diffusion processes which remove interference patterns. The nonlinear nature of our methods and models

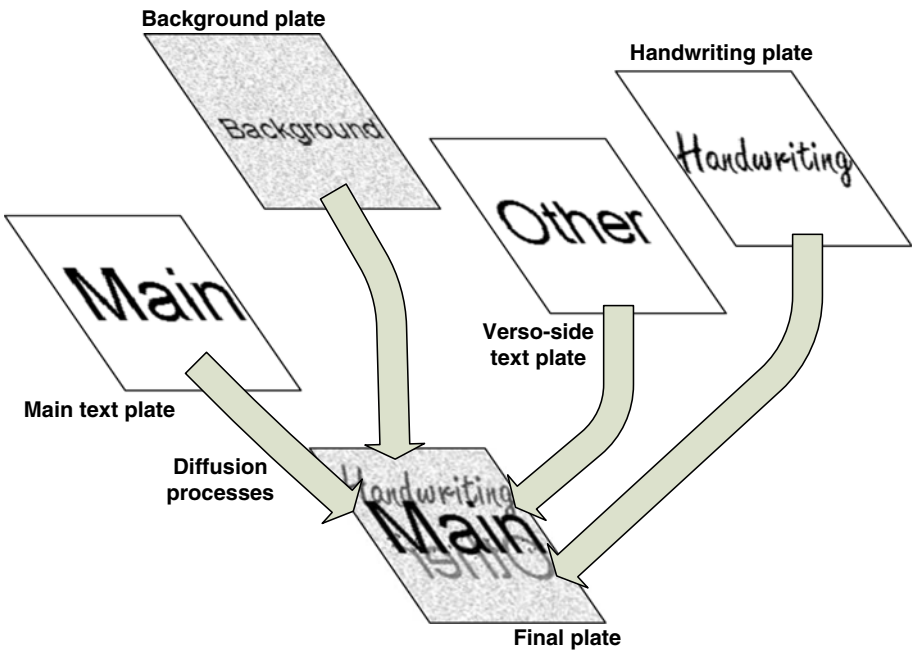


Fig. 2 A schematic diagram of the defect model. The handwriting plate is an example of the possibility of having multiple sources for the recto side

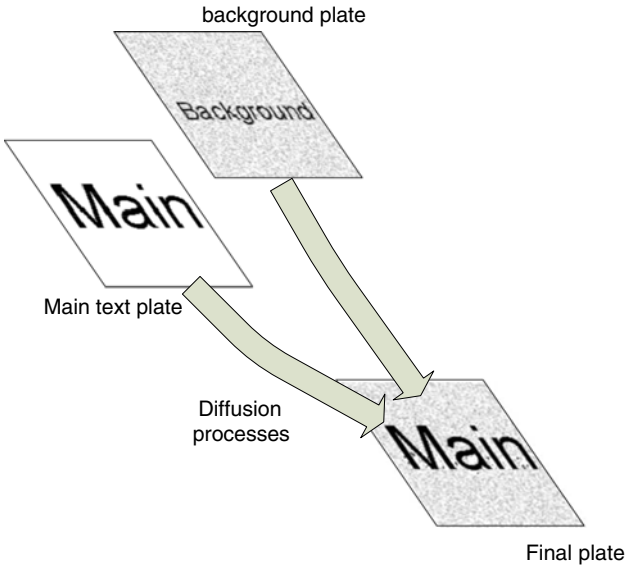


Fig. 3 A schematic diagram of the ideal history of a document image

makes them perform remarkably well in terms of generating or eliminating physical defects. Finally, in Sect. 4, application of the restoration method to real samples is described, and examples obtained using the degradation model are provided. A detailed study of the parameters of the degradation model is also presented.

2 Diffusion-based modeling of degradation in very old documents

In this section, a diffusion-based degradation model for physical effects such as bleed-through is developed. Then, using this model, any enhancement or restoration method for low-quality documents can be evaluated on a sufficiently large set of degraded, double-sided document images.

For the sake of notation abstraction, and before discussing our defect model, we introduce the general diffusion operator, DIFF (u, s, c), which represents a diffusion process from the source s to the target u with the diffusion coefficient c. In terms of the diffusion operator, the anisotropic diffusion method (ADM) can be rewritten as follows [27]:

$$\frac{\partial u}{\partial t} = \nabla \cdot (c\,(\nabla u)\,\nabla u) =: \mathrm{DIFF}\,(u, u, c)\,, \tag{1}$$

where $u(x, y, t)$ represents the gray value of the image pixels at virtual time t. At $t = 0$ the ADM starts with the input document image as $u(x, y, 0)$. The explicit relation for the diffusion coefficient c can be written as follows [23]:

$$c = \frac{1}{1 + (\nabla u/\sigma)^2}. \tag{2}$$

We select the square form because of its simplicity and lower computational cost. Usually, the value of σ is set at every time step based on the estimated noise variation [41]. This parameter also has a scale factor, σ_{scale}, which is set based on the desired level of fine image details to be kept. This scaling factor is set at the beginning of the computations. In denoising applications, σ_{scale} is usually 1.0. For document images, σ_{scale} may be different, depending on the goal which could be to remove all the interference patterns and save very weak features. To be accurate, σ_{scale} is chosen in such a way that the value of σ is less than the minimum of the gradient between the main text and the background.

Using the concept of the diffusion operator, we can introduce, and use, diffusion processes which are not restricted to the image itself. It is worth noting that a simpler form of the diffusion operator has already been used in the variational methods where a relaxation term between the current image and the input image is added to the right-hand side of the governing equation [9,25,31]. However, DIFF is general and can be used for the information exchange between every source and every target. It should be noted that, in

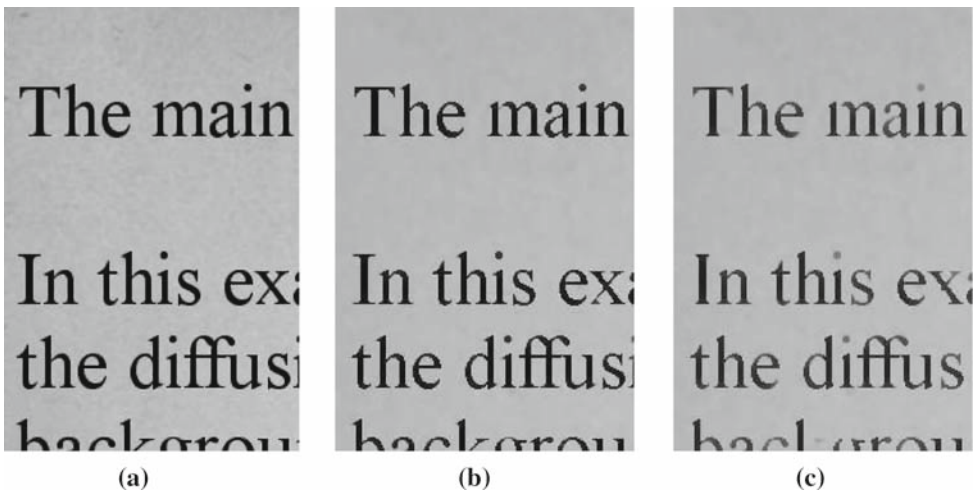


Fig. 4 An example of the aging effect. The input image is shown in **a**. The result after 300 iterations is shown in **b**. The result after 600 iterations is shown in **c**

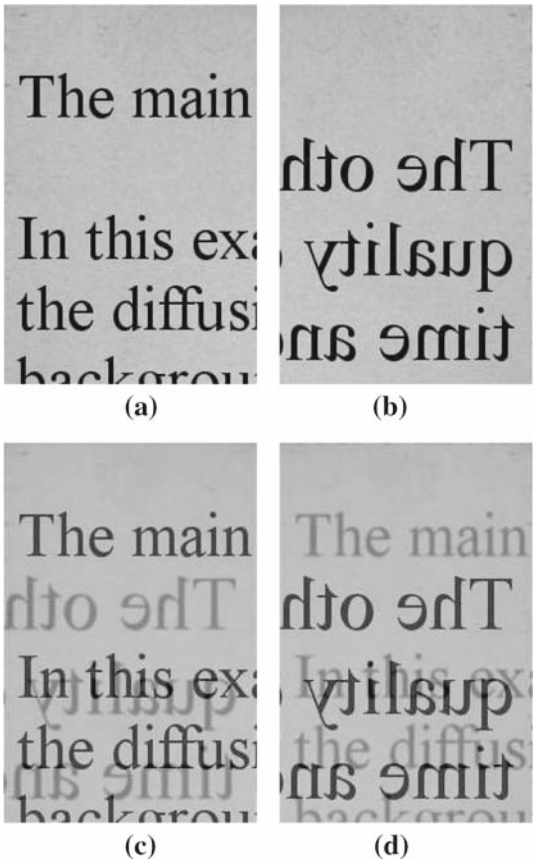


Fig. 5 An example of the bleed-through effect. The input images are shown in **a** and **b**. The results are shown in **c** and **d**. The following parameter values are used: $1/d = 6.0$ and $\sigma_b = 100.0$, and the number of iterations is 20

this work, we do not allow DIFF to exhibit any sort of tensor behavior. But, in the future, and by properly defining its diffusion coefficient, DIFF can simulate many directional phenomena.

If we examine the physical processes that result in many of the defect types, we may conclude that a considerable number of them can be attributed to diffusion. Examples would be degradation of the cellulose structure of the paper or persistent external particle accumulation on a document which occurs over time. The latter is usually considered to be a defect of external origin, in this case from a noisy source to the document image. Another would be ink on the document surface which has spread to nearby regions, and which can be modeled as a process of diffusion from one part of a document image to another. A third would be the seepage of ink from the verso side of a piece of paper to the recto side. This and many other effects can also be modeled as a diffusion process from the source (the document image of the verso side) to the target (the document image of the recto side). What this means is that the physical degradation of documents can be modeled as the superposition of several physical diffusion processes (see Fig. 2). Using the DIFF notation, we can model the physical processes associated with the defects by the following governing equation:

$$\frac{\partial u}{\partial t} = \text{DIFF}\,(u, s_{\text{recto}}, c_{\text{recto}}) + \text{DIFF}\,\big(u, s_{\text{bg}}, c_{\text{bg}}\big) + \text{DIFF}\,(u, s_{\text{verso}}, c_{\text{verso}}), \tag{3}$$

where s_{recto} is the recto-side image, s_{bg} is the background information, and s_{verso} is the verso side image. The first term represents the diffusion on the document image from its own data (spread of the ink). The second term takes into account dust and also changes to the characteristics of the paper. The seepage of the ink through the paper is modeled with the third term. Other possible sources can be added to the right-hand side of the equation. In this work, the problem of generating a realistic background is not addressed. Instead, we use scans of real papers without text as s_{bg}. The coefficient c_{recto} has the same formula as c in Eq. (2), and c_{verso} will be discussed below. The diffusion coefficient has the form

$$c_{\text{bg}} = d_{\text{bg}}\left(1 + \tanh(u - s_{\text{bg}} - \delta_{\text{bg}})/\sigma_{\text{bg}}\right)$$

to ensure that the text is not changed by this effect; for text regions, the difference between u and s_{bg} will be very high, and therefore the background diffusion stops. The parameters δ_{bg} and σ_{bg} are positive and have a value of less than one. In the experiments, the following values are used: $\delta_{\text{bg}} = 0.2$ and $\sigma_{\text{bg}} = 0.3$. The parameter d_{bg} is a positive number with a value of less than one, and determines the ratio of the background diffusion to the normal diffusion on the recto side. Equation (3) is written only for the recto side. The corresponding equation for the verso side can be written easily by considering the symmetric nature of the problem.

Equation (3), which models some physical phenomena, can be used to develop a model in which the defects are generated artificially on the document image. This defect model makes it possible to generate an unlimited number of physically defective images which can be regenerated and adapted to particular applications. The governing equation of the model can be written as

$$\frac{\partial u}{\partial t} = \sum_{i \in \text{sources}} \text{DIFF}\,(u, s_i, c_i). \tag{4}$$

The schematic diagram of the defect model is the same as the one in Fig. 2, except that the processes are performed in the virtual simulation environment.

The defect model can be arranged for several working states. For example, for modeling the phenomenon of document aging, we can include only the defect-free input document image and the noise information (the background information). The schematic diagram of this state is shown in Fig. 3. The information from the two sources flows to the result over time, and thus the degraded image for the desired

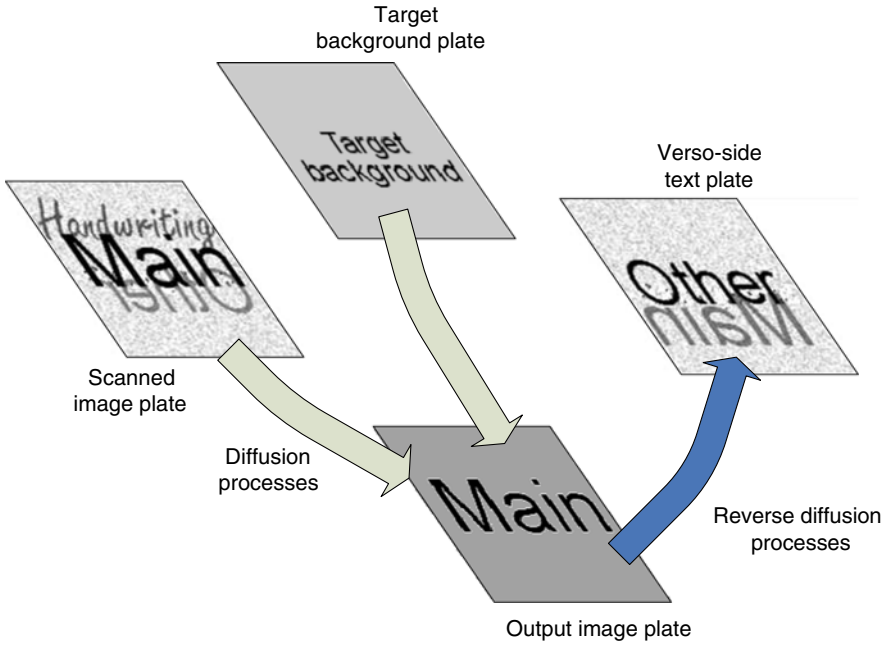


Fig. 6 A schematic diagram of the doubled-sided restoration method. The reverse diffusion of the verso side image helps in the removal of interference patterns

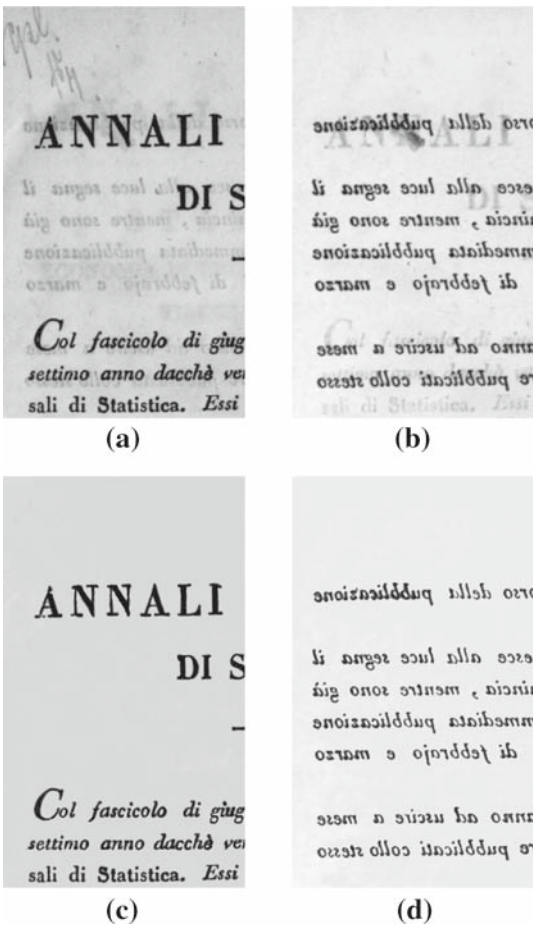


Fig. 7 **a**, **b** Input images from the Google Book Search dataset [14]. **c**, **d** Outputs obtained after restoration

aging period can be obtained. As time passes, the spreading of the ink, the degradation of the paper, and the degradation in ink quality change the document image. An example of these processes is shown in Fig. 4. The original image and two results at different times are shown.

Another possible configuration for the defect model is possible for the ink seepage through the paper. In this case, we have three sources: defect-free document images of both sides of the paper and the background. For the diffusion from the verso side to the recto side, we use the coefficient (when the equation is discretized, there is a diffusion coefficient from each neighboring pixel on the verso side to the target pixel)

$$c_{\text{verso}} = \frac{d}{1+(s-u)^2/\sigma_b^2}\frac{1}{1+s^2/\sigma_{\text{ink}}^2}, \tag{5}$$

where s is the gray value of the verso side on each neighboring pixel and u is the gray value of the target pixel on the recto side. The parameter d is the ratio of the verso diffusion to the normal diffusion on the recto side. The parameter σ_b has the same role as σ in the diffusion process and controls the degree of ink seepage through the paper, and σ_{ink} is a general parameter which restricts diffusion to the ink only. In other words, only the ink on the verso side can penetrate the paper to the recto side. Therefore, the diffusion coefficient for the white or background information on the verso side must be very small. The second term in Eq. (5) controls the rate of diffusion based on the gray level of the source information pixels. As the gray levels of the ink and the background are usually distributed on the interval [0, 1], we set σ_{ink} to 0.2 for all the computations. An example of the bleed-through

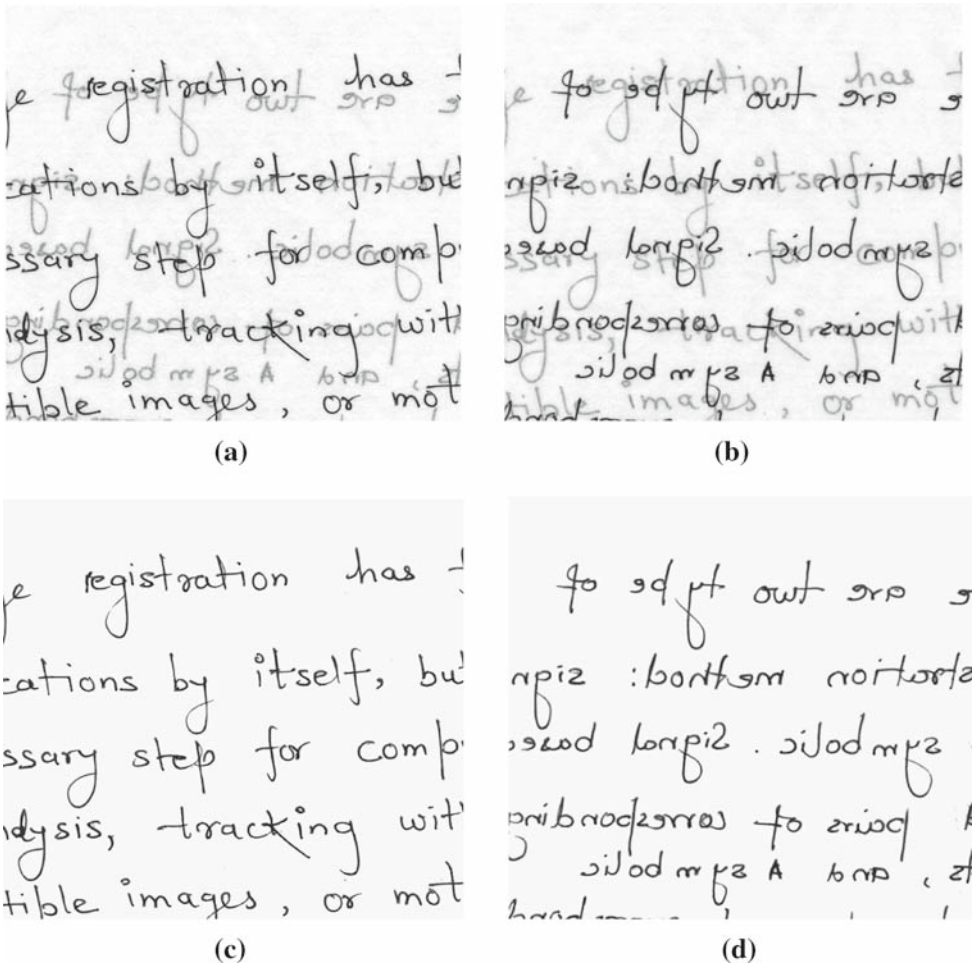


Fig. 8 Performance of the restoration method in Sect. 3 on real samples. **a**, **b** Input images [12]. **c**, **d** Outputs obtained after restoration

defect is shown in Fig. 5. The source images are shown in Fig. 5a, b, and the results, which are the two sides of a page bearing the bleed-through defect, are shown in Fig. 5c, d. The intensity of the interference patterns depends on the ratio of the diffusion coefficient of recto-to-recto diffusion to the diffusion coefficient of verso-to-recto diffusion. Also, the time interval has an important role to play in setting the degree of severity of the defects. It is worth noting that Fig. 5 represents the situation when the quality of the paper is high or the document is not very old. In these cases, the ink on the verso side cannot reach the recto side, and there is no smudging of interfering ink on that side. But, as the paper is not completely opaque, the ink that seeps into the paper is visible from the recto side. These cases of bleed trough are very similar to the shadow-through effect, but with a lower degree of blur.

3 A restoration method for very old double-sided documents

Not only does the appropriate use of diffusion models result in a very good and direct tool for the restoration of documents suffering from diffusion of some kind, but the inherent denoising ability of these models provides clear output without the need for additional denoising. However, PDE-based diffusion models are not suitable in their conventional form for application to problems in which several sources of information are involved. In particular, the problem of recovering very old documents is a complicated one for diffusion models. In very old documents, not only are there several layers of information, but these layers are very close to each other on the gray-level scale. Therefore, they can easily diffuse into one another, and this results in a mixture of layers which are then inseparable. This makes it is more difficult to apply diffusion models to bleed-through problems, as these models act very locally. In other words, with these models, only the data of a pixel's neighbor can alter the information of that pixel. If, for example, the pixel is surrounded by dark regions, then there is no way to insert light or white information into that pixel. This high spatial dependency of diffusion models limits their ability to separate and remove layers, although this limitation can be remedied by providing some means of data exchange between different regions of an image. The way to provide for this data exchange depends on the application, but one of the most common techniques, which is in fact also

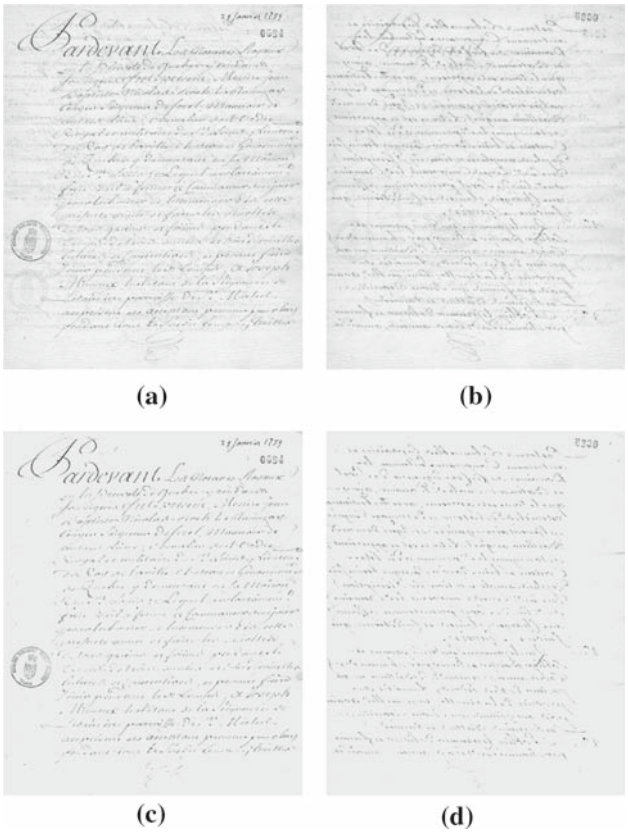


Fig. 9 Another sample from [12]. **a**, **b** Input images. **c**, **d** Outputs obtained after restoration

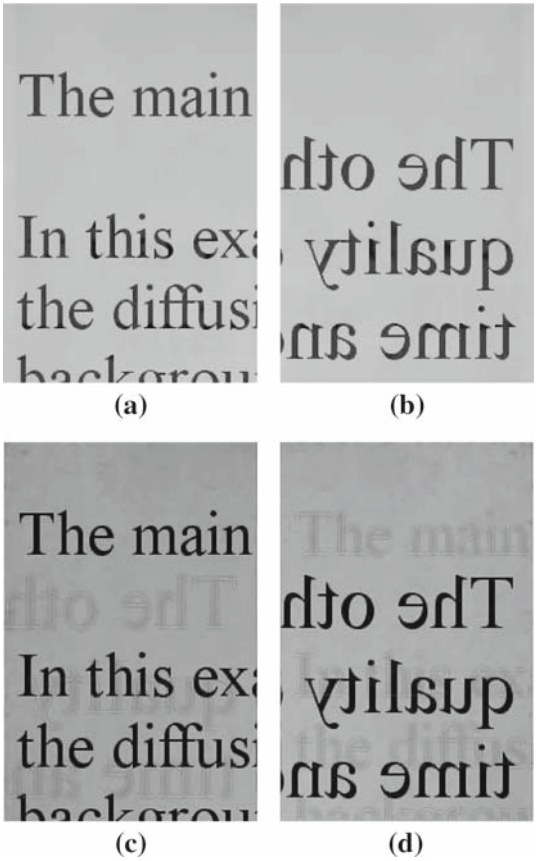


Fig. 10 An example of the restoration of bleed-through degradation. **a**, **b** Output results. The images in Fig. 5c, d are used as input images, the parameters of the restoration method are as follows: $\sigma_{\text{scale}} = 0.36$, $1/d_i = 6.0$, $\sigma_{i,b} = 0.1$, and the number of iterations is 200; **c**, **d** results of applying the ICA method

a diffusion-based technique, is to provide a diffusion process from "target" background information. This additional diffusion process can easily break the spatial barrier in diffusion models. Using the target background, the unwanted information layers can be removed and replaced by the background information, while the other layers are only sharpened and enhanced via in-the-image diffusion. Information can also be exchanged between other sources of information in a document, such as the main text (and perhaps other sources on the recto side, such as handwritten text) and the verso-side text by means of similar diffusion processes.

In the previous section, a defect model was introduced which can be used to generate document images which are totally preregistered and also suffer from these types of degradation. Using this model, we can propose a restoration method which is based on the diffusion processes, but at the same time performs the reverse phenomenon on the damaged document image. Again, because of the symmetric nature of the problem, only the recto side will be discussed here. In practice, the recto and verso sides are processed simultaneously. Consider the proposed schematic diagram in Fig. 6. In this figure, the information relative to the verso side of the document is used in a reverse diffusion process to eliminate the interference on the recto side. Although we do not actually have access to the true information relative to the verso side, we can, in a modification, use the degraded image on the verso side for this purpose. This degraded verso side image contains the required information for an adequate reverse diffusion process, which can be used to recover the true data on the recto side. This process should not be mistaken with backward/inverse diffusion which is an enhancement method for increasing the gradient of the image for certain gradient values which are known to be features of the image. The reverse diffusion weakens and removes the unwanted information layers, and, at the same time, the diffusion from a "target" background, $s_{i,\text{bg}}$, and the diffusion from the image itself fill the gaps and sharpen the strokes. Usually, a constant gray value is used as the target background. The governing equation of the restoration method is as follows:

$$\frac{\partial u_r}{\partial t} = \text{DIFF}\left(u_r, u_r, c_{i,\text{recto}}\right) + \text{DIFF}\left(u_r, s_{i,\text{bg}}, c_{i,\text{bg}}\right) - \text{DIFF}\left(u_r, u_v, c_{i,\text{verso}}\right). \quad (6)$$

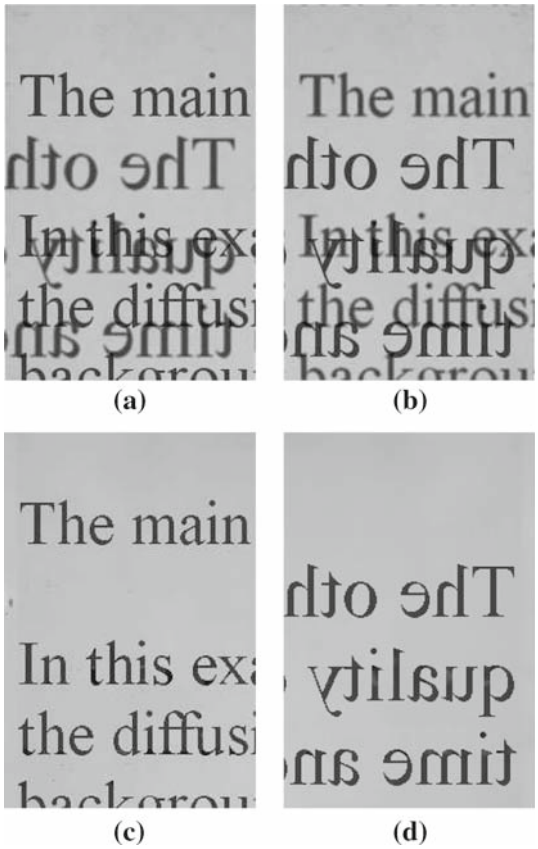


Fig. 11 Another example of bleed-through restoration. **a**, **b** Input images created using the degradation model in Sect. 2 from the original images in Fig. 5**a**–**b**. Output results are shown in **c**, **d**

To differentiate between the notation of the degradation model and that of the restoration method, all the symbols in the restoration method have an extra subscript i. u_r and u_v represent the recto and verso sides, and the final output will be obtained at the steady state. At $t = 0$, we have $u_r(x, y, 0) = u_{i,\text{recto}}$ and $u_v(x, y, 0) = u_{i,\text{verso}}$, where the degraded input images are denoted by $u_{i,\text{recto}}$ and $u_{i,\text{verso}}$. The main difference is the diffusion rate of the reverse diffusion $c_{i,\text{verso}}$, which now has a simpler form:

$$c_{i,\text{verso}} = \frac{d_i}{1 + (s - u_r)^2/\sigma_{i,b}^2}.$$

It is computed on all the neighboring pixels on the verso side u_v. The effect of the restoration method is to eliminate the unwanted diffusion on the document image. However, it is worth noting that the bleed-through effect is not mathematically reversible. In other words, there is no one-to-one relation between the degraded document images and their original sources. However, the restoration method introduced by Eq. (6) is capable of removing all the data that appears to constitute interference patterns. The method uses the basic physical idea of the degradation model to differentiate among patterns. Nevertheless, because it is a diffusion model, it will also change the true data. This is the reason why in Sect. 4 and for the least degraded examples the peak signal-to-noise ratio (PSNR) measure of the outputs of the restoration method will be very much lower than the degraded inputs. The ability of the method to address real samples is evaluated in the next section by applying the method to real and synthesized samples.

4 Experimental results and discussion

For all numerical calculations, a finite-difference scheme on an 8-pixel neighborhood is used. As the value of c is less than one, the upper limit of time steps is the reciprocal sum of all the distances from neighboring pixels to the target pixel, i.e. $1/(4+2\sqrt{2})$. Also, in all the restoration experiments we use $1/d_{i,\text{bg}} = 6.0$, $1/d_i = 6.0$, $\sigma_{i,b} = 0.1$, $\sigma_{i,\text{bg}} = 0.3$, and $\delta_{i,\text{bg}} = -0.01$. Although these values are not optimal, the parameters show small variations (as discussed previously). This fact also shows the robustness of the restoration method.

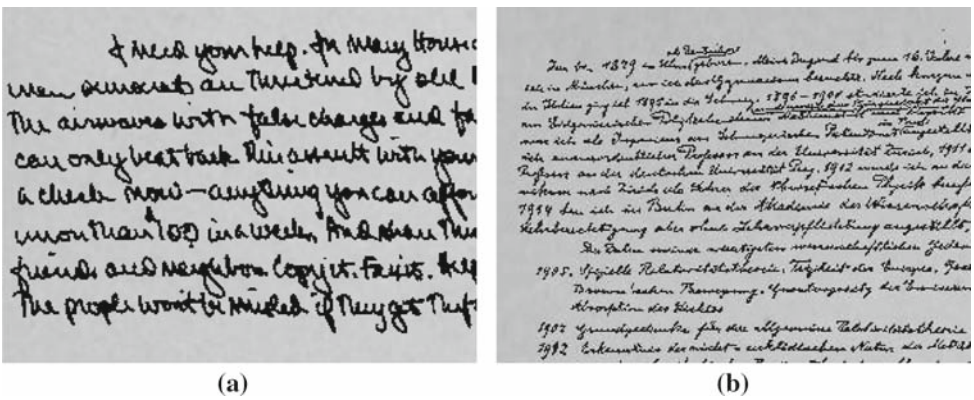


Fig. 12 The original images used for generating document images with the bleed-through defect

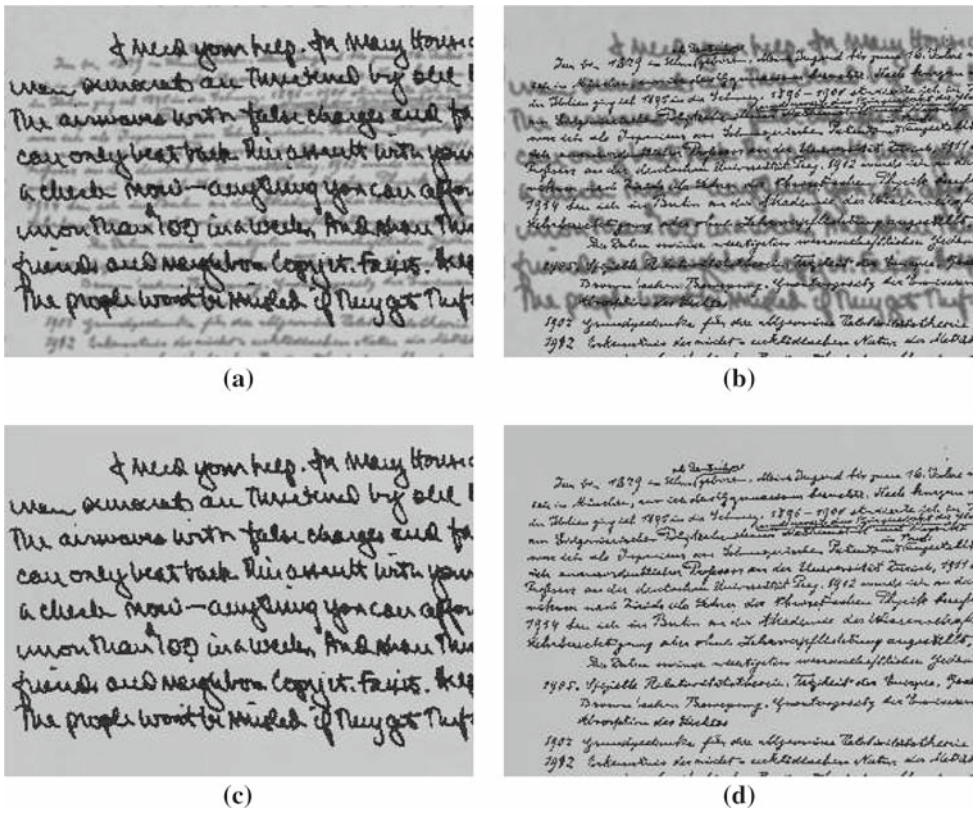


Fig. 13 An example of the restoration of bleed-through degradation. **a**, **b** Input images created using the degradation model in Sect. 2 from the original images in Fig. 12. Output results are shown in **c**, **d**

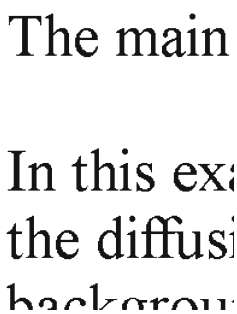


Fig. 14 The diffusion coefficient of the background diffusion process of the output images in Fig. 10. This function can be used as the binarized version of the document

4.1 Enhancement of real samples

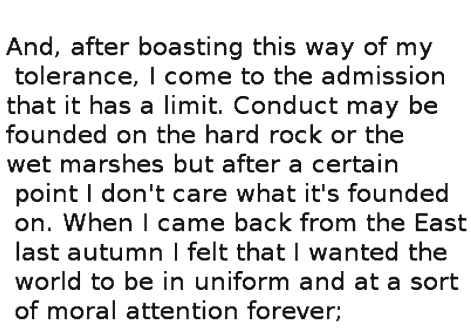
We applied the new restoration model to samples from the Google Book Search dataset [14]. An example is shown in Figure 7. Each sample was selected from one page of a book. The verso side image was inverted horizontally. These two images were then registered using the standard spatial transform of the Matlab Image Processing Toolbox: *cp2tform* [37] (a set of registered double-sided images is shown in Fig. 7a, b). After applying the restoration method (Eq. 6) to the input images in Fig. 7, the interference patterns are removed (outputs shown in Figs. 7c, d). The outputs are very clear, and unwanted non-textual layers are removed as well.

Figure 8 provides a real sample from another dataset. The double-sided document images in this dataset[2] were preregistered by the author of the dataset [12]. Again, the outputs are very clean, and many thin connections and overlapping crosses have been saved. The same promising results for another sample from the same dataset are shown in Fig. 9. As was stated in Sect. 3, we assume that the input images are preregistered. In restoring double-sided document images, in contrast to other problems, the amount of correlation between the two sides is very limited and actually depends on the degree of bleed-through degradation. However, in many cases where an affine transformation (a rotation followed by a shift) is sufficient for registration, especially where the paper has not been deformed, this amount of correlation is enough to register the recto- and verso-side images in an automatic and simple way [6,12].

4.2 Enhancement of synthesized samples

Our first example involves the degraded document images that were obtained in Sect. 2 with the defect model. We use them as input images to the restoration method (the images in Fig. 5c, d). As the images were preregistered, there is no need for registration at this point. The results of application

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



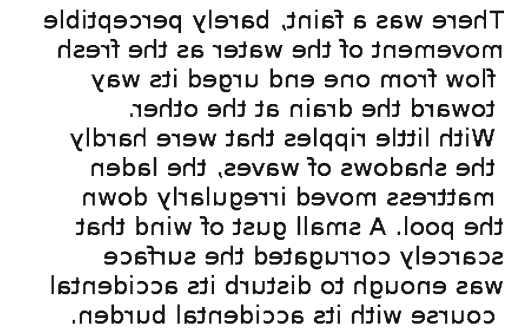


Fig. 15 The original recto and verso sides used for generating degraded images in Figs. 16, 17, 18 and 19

Fig. 16 The effect of n on the output of the degradation model. *First column* The recto side of the output. *Second column* The restored recto side obtained using the method in Sect. 3. *Third column* The restored recto side obtained using the ICA method. From *top* to *bottom*, n is 1, 21, 51, and 71, respectively. The other parameters are given in the text

of the restoration method to the degraded images in Fig. 5 are shown in Figs. 10a, b). The parameters are stated in the figure caption. The outputs are very clear, and actually all the interference has been removed. For the sake of comparison, the results of applying the ICA method [8,26] to the same degraded images are shown in Fig. 10c, d. The sources are well separated, but some weak patterns have remained near the edges of the interference patterns. This is due to

Fig. 17 The effect of d on the output of the degradation model. *First column* The recto side of the output. *Second column* The restored recto side obtained using the method in Sect. 3. *Third column* The restored recto side obtained using the ICA method. From *top* to *bottom*, $1/d$ is 1.5, 3.375, 11.39, and 25.63, respectively. The other parameters are given in the text

the one-to-one the nature of the ICA method, which limits its ability to deal with the nonlinear seepage of ink around patterns.

In a more challenging example, we set the parameters of the defect model to the following values: $1/d = 6.0$ and $\sigma_b = 100.0$, and the number of iterations to 30. The original images in Fig. 5 are used as the inputs to the defect model. The resulting degraded document images are shown in Fig. 11a, b. The restored images, shown in Fig. 11c, d, are obtained using the same parameters as those in Fig. 10, with the governing equation being (6). The method effectively recovers the bleed-through defects. It also shows its effectiveness with respect to the nonlinear nature of the defect (in this case, ink spread over the surface of the paper, which is noticeable in Fig. 11a, b). In Figs. 12 and 13, another example of bleed-through effect is presented, but with handwritten texts.

The diffusion coefficient of the background process can provide more useful information. Figure 14 shows the diffusion coefficient of the background process at the end of computation. The figure actually represents the binarization of the recovered text in a clear form. We can therefore state that the restoration model is capable of providing additional information about the document which is usually obtained by applying other operators to the results. Consequently, our method not only inherently integrates these tasks, but it also eliminates the possibility of introducing further error because of extra processing of the data. These errors are addressed conventionally by other processing tasks, such as smoothing

Fig. 18 The effect of σ_b on the output of the degradation model. *First column* The recto side of the output. *Second column* The restored recto side obtained using the method in Sect. 3. *Third column* The restored recto side obtained using the ICA method. From *top* to *bottom*, σ_b is 0.005, 0.125, 15.625, and 390.625, respectively. The other parameters are given in the text

and gap recovery. By contrast, this information can be integrated in the restoration method to increase its performance. However, in this work, where the objective is to provide a simple and clear introduction, the restoration method has been only briefly described, since it is not the main focus of the paper.

4.3 Role of the various degradation model parameters

The degradation model in Sect. 2 has four main parameters: (1) the number of iterations n that plays the role of time; (2) the scaling factor d; (3) the verso diffusion parameter σ_b; and (4) σ_{ink}. In this subsection, the role of each of these parameters is studied in several experiments, in which the constant parameters are $n = 30$, $1/d = 5.$, $\sigma_b = 100.0$, and $\sigma_{\text{ink}} = 0.2$. The fact that the restoration method works based on a balance between several terms in Eq. (6) results in a robust and insensitive behavior of the method with respect to its parameters. This robustness enable us to choose fixed, rather than optimized, values for the parameters of the restoration method: $\sigma_{\text{scale}} = 0.12$, $1/d_{i,\text{bg}} = 6.0$, $1/d_i = 6.0$ $\sigma_{i,b} = 0.1$, $\sigma_{i,\text{bg}} = 0.1$, and $\delta_{i,\text{bg}} = -0.01$. The original recto and verso sides are shown in Fig. 15. We start with n. Figure 16 shows a series of degraded document images for different values of n. The first column contains the recto side of the degraded document images (the verso sides are omitted to save space), the second column contains the restored recto sides obtained using the method in Sect. 3 and the third

Fig. 19 The effect of σ_{ink} on the output of the degradation model. *First column* The recto side of the output. *Second column* The restored recto side obtained using the method in Sect. 3. *Third column* The restored recto side obtained using the ICA method. From *top* to *bottom*, σ_{ink} is 0.0025, 0.0156, 0.244, and 1.53, respectively. The other parameters are given in the text

column contains the restored recto sides obtained using the ICA method as a reference. The parameter n varies from 1 to 51 in this experiment. As n increases, the degree of bleed-through increases, and, at some point the restoration method actually captures the interference patterns. As the bleed-through effect is nonlinear and there is no one-to-one relationship between the verso side and its interference patterns on the recto side, the ICA method captures the deformed edges and boundaries as an independent source instead of original information for major manifestations of this effect.

The second parameter considered is d. Figure 17 is the same as Fig. 16, except for variations in d. In this figure, $1/d$ varies from 2.25 to 25.63. Actually, as can be seen from the results, the parameters d and n have similar effects. However, in situations where the diffusion of the main text on the recto side is also high, for example in a very humid environment, d determines the relation between seepage diffusion and recto-side diffusion. The next parameter, σ_b, is the key parameter of the degradation model. This parameter represents the thickness of the paper. For low values of this parameter, a small amount of the ink can pass through the paper and reach the recto side. The effect of this parameter on the output of the degradation model is shown in Fig. 18. The variation ranges from 0.005 to 390.625. Again, for high seepage values, the ICA method suffers from residuals around the edges and boundaries. It is also evident from the PSNR curve of the degraded image that σ_b must be greater than 10 in order for there to be a noticeable degree of degradation. The last parameter studied is σ_{ink}. This parameter reflects the fact that only the ink can flow and diffuse to other places. In other words,

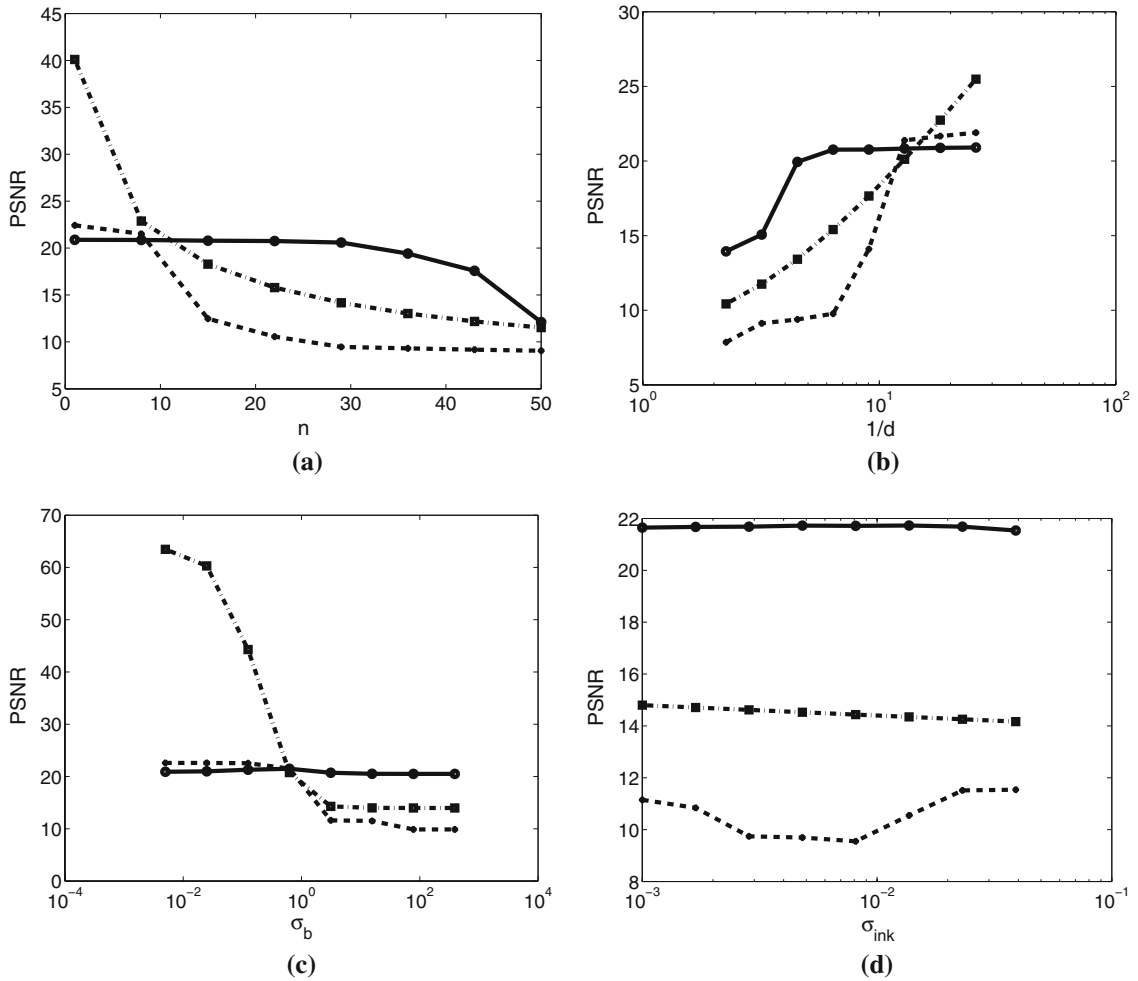


Fig. 20 Comparison of the PSNR performance of the restoration method with the ICA method for variations of several parameters. From **a** to **d**, the figures correspond to n, $1/d$, σ_b, and σ_{ink}, respectively. The *continuous line with circles*, the *dashed line with crosses*, and the *dashed-dotted line with squares* correspond to our method output, the ICA method output, and the degraded input, respectively. The images are shown in Figs. 16, 17, 18 and 19

the white information and clean regions on the verso side cannot produce any effect on the recto side of the document image. However, for the sake of completeness, an experiment on the variations of this parameter is performed, and shown in Fig. 19. The variation ranges from 0.001 to 0.039. For large values of σ_{ink}, the degraded images are actually a linear combination of recto and verso texts. This is why the ICA method has a good performance for this region. For real applications, a fixed value of 0.02 can be used for this parameter.

Two quantitative evaluations of the restoration method are performed based on the results in Figs. 16, 17,18, and 19. The first measure is the PSNR [19]. The variation of the PSNR for each parameter is shown in Fig. 20. The PSNR of the recto side of the original degraded image, the result of the method in Sect. 3, and the result of the ICA method are compared. For n, when the number of iterations is very small and therefore the degree of degradation in the degraded image is unnoticeable, the restored image has a lower PSNR than the degraded one. This is because the diffusion method actually introduces some changes into images. For very high values of n, the phenomenon of capturing the other side of the document is evident from the PSNR curves. It is worth noting that, for the ICA method, because of normalizing and change of gray levels, the curve is usually lower than the degraded one. Therefore, to arrive at a better measure, the OCR recognition rate has also been measured and will be discussed later. For the second parameter d, it can be observed from the curves that the performance of the restoration method drops for high values of this parameter, although it is still good. For the third parameter, σ_b, the restoration method shows a high degree of robustness. Although, the value of the PSNR is not very high for almost clean cases, it is nearly constant over whole the range. Finally, for σ_{ink} the restoration method shows its superiority over all values.

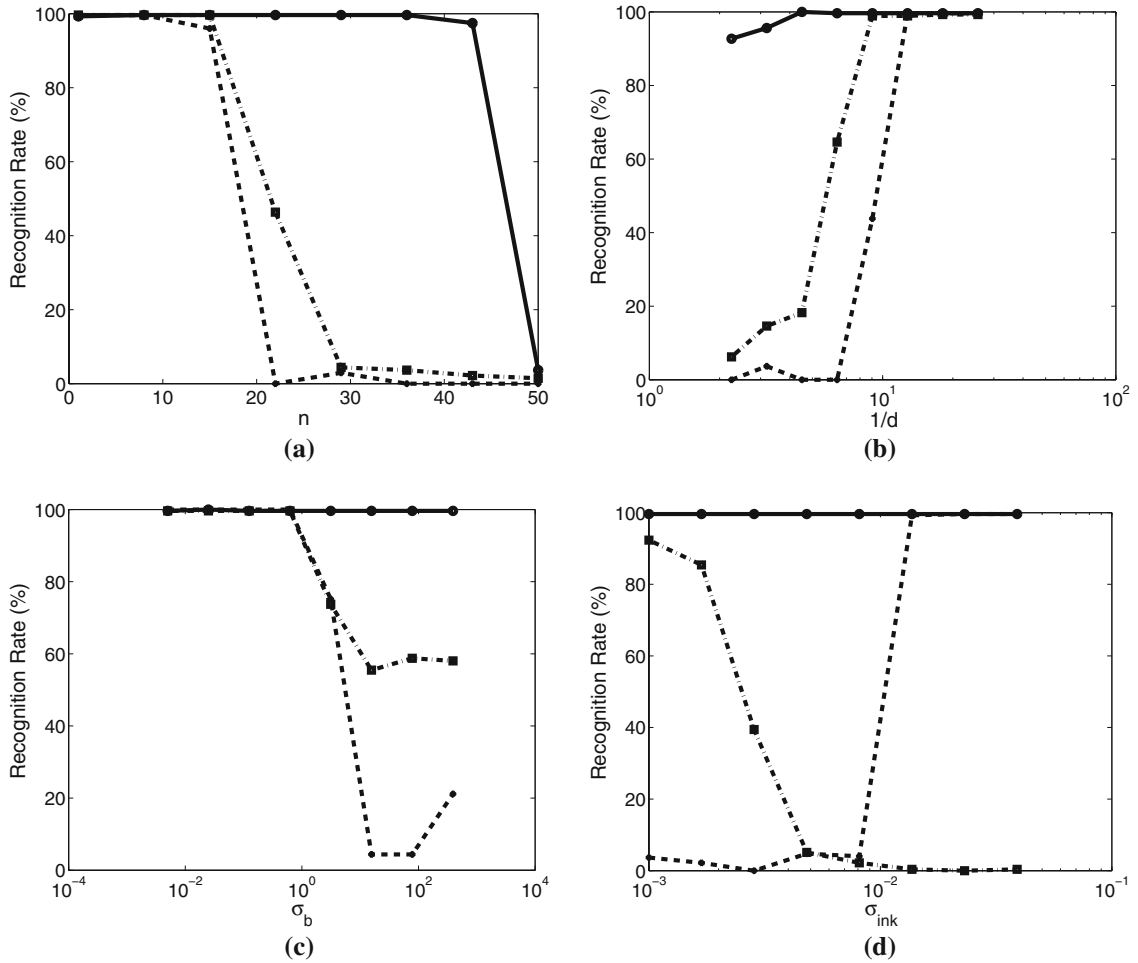


Fig. 21 Comparison of the OCR recognition rate of the restoration method with that of the ICA method for variations of several parameters. From **a** to **d**, the figures correspond to n, $1/d$, σ_b, and σ_{ink}, respectively. The *continuous line with circles*, the *dashed line with crosses*, and the *dashed-dotted line with squares* correspond to our method output, the ICA method output, and the degraded input, respectively. The images are shown in Figs. 16, 17, 18 and 19

The result of measuring the OCR recognition rate in the same experiments is also presented in Fig. 21. A commercial OCR software, FineReader version 9.0, is used. This measure is more robust with respect to normalization of gray values. The rates correspond to recto side of document images. As the outputs of the ICA method do not have any one-to-one relationship with the input images, the two outputs have been compared, and the highest rate chosen. Again, in all cases, the restoration method in Sect. 3 outperforms the ICA method. For n, our method has good performance up to $n = 40$, but the ICA method stops at $n = 20$. For two parameters, d and σ_b, the precision of our method is almost one hundred percent that shows high degree of robustness. Finally for σ_{ink}, the performance of our method is higher. However, it can be seen for low-degraded cases, such as for low values of n, d and σ_b, the performance of the ICA method is also good. The images and recognized texts are available on Internet.[3]

In many cases, the behavior of the OCR curves is very similar to that of the PSNR curves, and this reveals the consistency that exists between the two measures. It is worth noting that, in terms of the OCR recognition measure, the performance of methods is almost binary; full recognition or recognizing a little amount of text.

Another measure for OCR evaluation is presented in Fig. 22. In this figure, the percentage of wrong recognized characters of the OCR output is presented. Usually, the percentage of the incorrect recognized characters is the complement of the recognition rate. However, in some cases such as $1/d = 3.375$ or $\sigma_{\text{ink}} = 0.0156$, the ICA method output is so complex that the OCR software doesn't recognize any text.

Another measure for comparing the performance of the methods is computational complexity. The computational cost of the degradation model is not very high, usually requiring about 30 iterations to generate a double-sided document image. For a double-sided 512×512 pixel image, the computational cost of every ten iterations is about one second.

[3] http://www.synchromedia.ca/web/ets/expreswgf7.

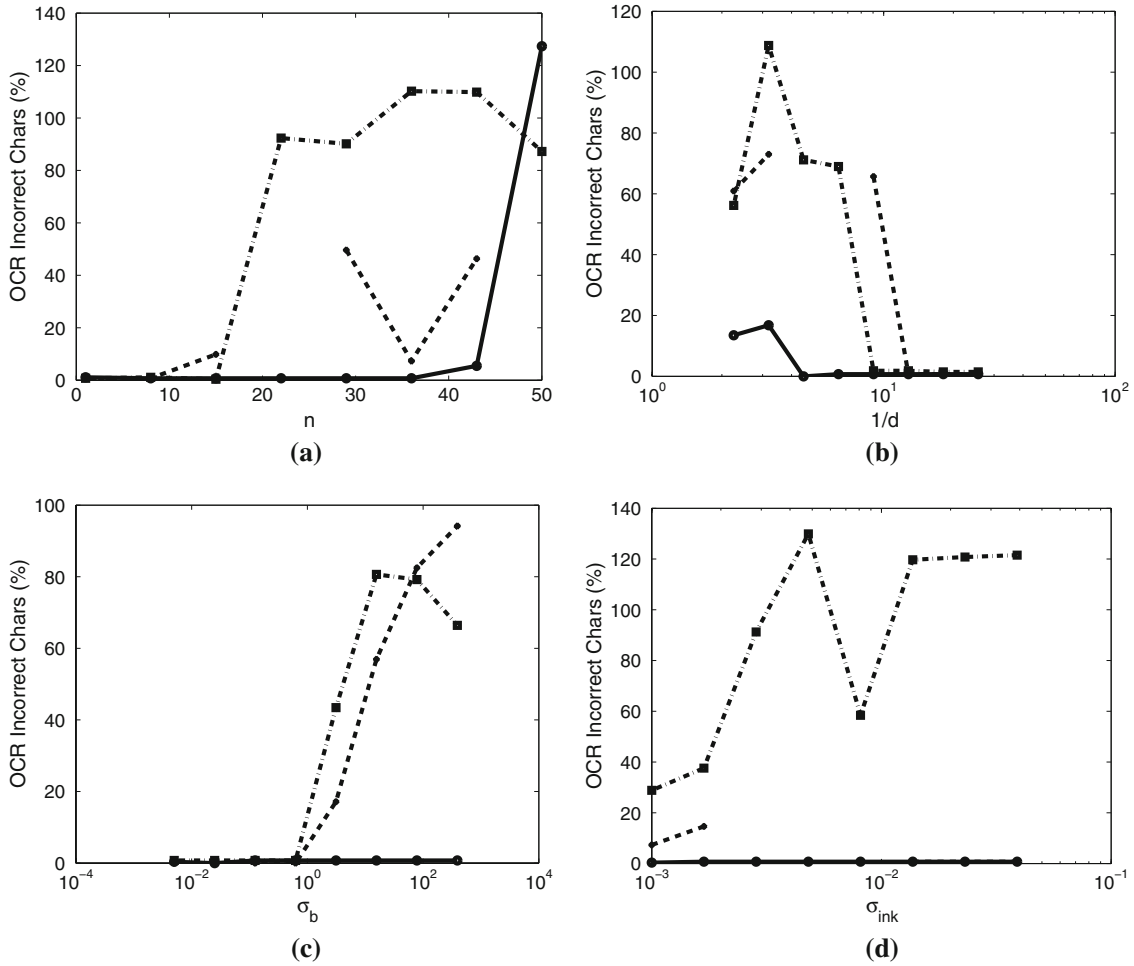


Fig. 22 Comparison of the OCR incorrectly recognized character rate of the restoration method with that of the ICA method for variations of several parameters. From **a** to **d**, the figures correspond to n, $1/d$, σ_b, and σ_{ink}, respectively. The *continuous line with circles*, the *dashed line with crosses*, and the *dashed-dotted line with squares* correspond to our method output, the ICA method output, and the degraded input, respectively. The images are shown in Figs. 16, 17, 18 and 19

Because the degradation model and the restoration method have a similar structure, the computational cost per iteration of the restoration method is also the same. However, to reach the steady state, the restoration method needs about 100 iterations. In comparison, the ICA method needs two seconds for a double-sided, 512×512 pixel document image.

It is worth noting that the parameters of the degradation model can easily be non-stationary and non-uniform over the image domain. As an illustration, Fig. 23a, b shows the recto and verso sides of a degraded document in which σ_b has a synthetic spatial dependency. In this case, the maximum value of σ_b is 100.0. The other parameters of the degradation model are $1/d = 5.0$, $n = 30$, and $\sigma_{\text{ink}} = 0.1$. It is interesting that, despite the non-uniform degradation in this example, the restoration method is able to recover the original text (see Fig. 23c, d). The parameters of the restoration method are the same as of those stated at the beginning of this subsection, and are constant on the images. For comparison purposes, the outputs of the ICA method are presented in Fig. 23e, f.

As stated at the beginning of this subsection, the parameters of the restoration method are robust to the degradation level. Figure 24 illustrates the PSNR curve of the restored image on a large interval of the d_i values. As can be seen from the figure, the curve is almost constant, and this shows the robustness of the method with respect to the rate of the reverse diffusion d_i.

Considering all the above mentioned results, it can be concluded from OCR evaluation that the low-cost ICA method is a good choice for recognition applications for low-degraded cases. On the other hand, if the goal of processing is visual enhancement of document images for archiving, the method of Sect. 3 is the best choice based on the PSNR measure (see Fig. 20). Finally, for highly degraded images, the ICA method fails and our method must be used.

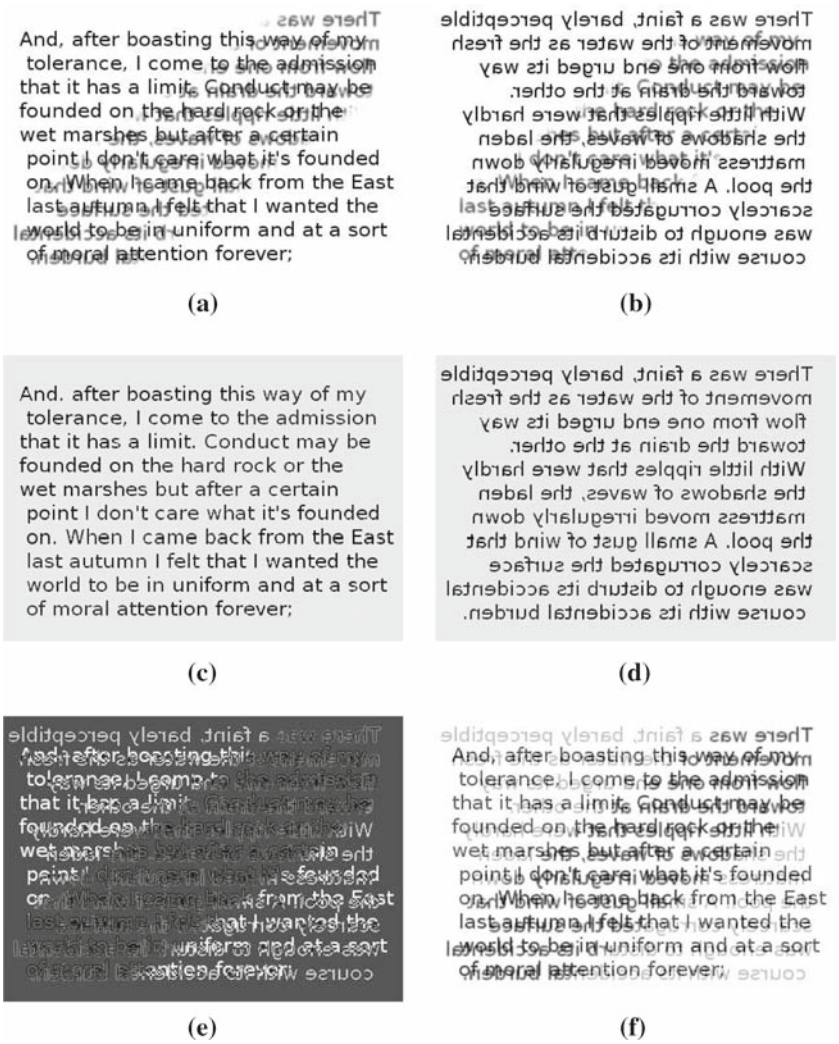


Fig. 23 An example of the restoration of non-uniform bleed-through degradation. **a**, **b** Degraded input images; **c**, **d** results of applying the restoration method; **e**, **f** results of applying the ICA method. The parameters are given in the text

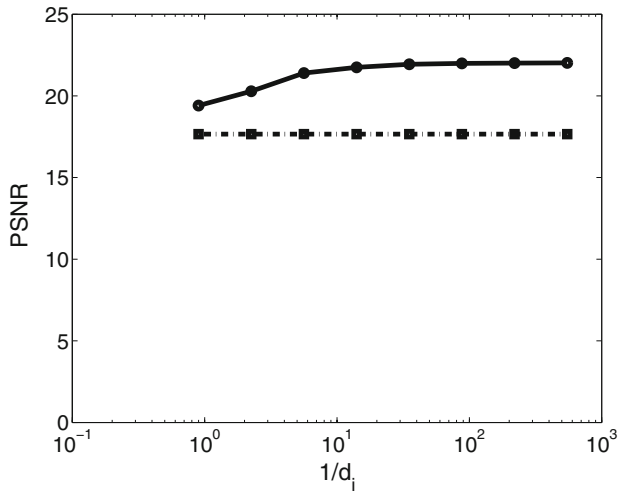


Fig. 24 The PSNR curve of the restored recto-side image versus d_i. The degraded input corresponds to the second row of Fig. 17. The *continuous line with circles* and the *dashed-dotted line with squares* correspond to our method output and the degraded input, respectively

5 Conclusion and future prospects

The problem of enhancement and restoration of low quality document images is studied from the diffusion point of view. To address the problem of double-sided degraded document images, and based on the idea of general diffusion originating from several sources, a defect model for generating the physical defects in the documents is proposed. The model can be used for inserting defects, such as aging and ink seepage. Using this defect model, the datasets of physically degraded document images can be generated to develop, train, compare, and evaluate restoration methods. Also, the model can be cascaded with other well-developed defect models which simulate defects originating from imaging processes.

A method derived from the defect model is then introduced to remove the interference patterns from double-sided documents. The results of application of the restoration model to generated degraded document images, and also to real documents, are promising. In addition to recovering the main text,

the restoration method can provide other useful information, such as simultaneous outputs. For example, the coefficient of diffusion of background information can be used as a binarization version of the document, which would obviate the need for other processing tasks, such as binarization. Finally, it is worth noting that, although the registration of double-sided documents is not an easy task, the introduction of new types of document images, such as multi-spectral images which are automatically preregistered, has resulted in a demand for multi-source degradation models and restoration methods. Ours and other similar work also provides a basic framework for such cases.

Currently, we are working on methods which combine the benefits of several approaches, such as the ICA method and the diffusion method. Also, as a prospect for the future, we propose to build a physical simulation environment for the rapid aging and degrading of documents. Using the samples that will be obtained through this process, we will able to model the macro-scale parameters for a specific type of paper and condition of preservation. The degradation model can then be adapted to those types of paper, and an unlimited number of degraded double-sided document images for any specific type of paper can be generated for the calibration or evaluation of restoration methods on those types of paper.

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

References

1. Al-Khatib, W.G., Shahab, S., Mahmoud, S.A.: Digital library framework for arabic manuscripts. In: Shahab, S. (ed.) Computer Systems and Applications, 2007. AICCSA '07. IEEE/ACS International Conference on, pp. 458–465. Amman, Jordan (2007). doi:10.1109/AICCSA.2007.370922
2. Baird, H.: Document image defect models. In: Proceedings of IAPR Workshop Synthetic and Structural Pattern Recognition. Murray Hill, NJ (1990). (Reprinted in L. O'Gorman & R. Kasturi (eds.), Document Image Analysis. IEEE Computer Society Press, Washington, pp. 315–325, 1995
3. Baird, H.: Digital Document Processing: Major Directions and Recent Advances. The State of the Art of Document Image Degradation Modelling, pp. 261–279. Springer, Berlin (2007). doi:10.1007/978-1-84628-726-8_12
4. Barney Smith, E.H.: Estimating scanning characteristics from corners in bilevel images. In: Proceedings of SPIE, vol. 4307. Document Recognition and Retrieval VIII, pp. 176–183. San Jose (2001). doi:10.1117/12.410835
5. Boussellaa, W., Zahour, A., Taconet, B., Alimi, A., Benabdelhafid, A.: Praad: Preprocessing and analysis tool for arabic ancient documents. In: Zahour, A (ed.) Document Analysis and Recognition, 2007. ICDAR 2007 Vol. 2. Ninth International Conference on, vol. 2, pp. 1058–1062. Curitiba, Parana (2007). doi:10.1109/ICDAR.2007.4377077
6. Castro, P., Almeida, R., Pinto, J.: Restoration of Double-Sided Ancient Music Documents with Bleed-Through, vol. 4756/2008, pp. 940–949. Springer, Berlin (2007). doi:10.1007/978-3-540-76725-1
7. Chen, L., Zhu, J., Young, M., Susfalk, R.: Modeling polyacrylamide transport in water delivery canals. In: ASA-CSSA-SSSA International Annual Meetings, pp. 294–296. Indianapolis (2006)
8. Cichocki, A., S, A., K, S., Tanaka, T., Phan, A.H., Zdunek, R.: Icalab—matlab toolbox ver. 3 for signal processing (2007)
9. Deriche, R., Faugeras, O.: Les edp en traitement des images et vision par ordinateur. Tech. Rep. 2697, INRIA (1996)
10. Drira, F.: Contribution à la restauration des images de documents anciens. Ph.D. thesis, École Doctorale Informatique et Information pour la Société (EDIIS), LIRIS, UMR 5205 CNRS (2007)
11. Dubois, E., Dano, P.: Joint compression and restoration of documents with bleed-through. In: Proceedings of IS&T Archiving 2005, pp. 170–174. Washington DC, USA (2005)
12. Dubois, E., Pathak, A.: Reduction of bleed-through in scanned manuscript documents. In: Proceedings of IS&T Image Processing, Image Quality, Image Capture Systems Conference (PICS2001), pp. 177–180. Montreal, Canada (2001)
13. Fathalla, R., Sonbaty, Y.E., Ismail, M.: Extraction of arabic words from complex color image. In: Sonbaty, Y.E. (ed.) Document Analysis and Recognition, 2007. ICDAR 2007 Vol. 2. Ninth International Conference on, vol. 2, pp. 1223–1227. Curitiba, Parana (2007). doi:10.1109/ICDAR.2007.4377110
14. Google: Book Search Dataset, version V edn. (2007)
15. Kanungo, T., Haralick, R., Baird, H., Stuetzle, W., Madigan, D.: Document degradation models: parameter estimation and model validation. In: Proceedings of International Workshop on Machine Vision Applications, pp. 552–557. Kawasaki, Japan (1994)
16. Kanungo, T., Haralick, R., Baird, H., Stuezle, W., Madigan, D.: A statistical, nonparametric methodology for document degradation model validation. Trans. Pattern Anal. Machine Intell. **22**(11), 1209–1223 (2000). doi:10.1109/34.888707
17. Kanungo, T., Haralick, R.M., Phillips, I.: Nonlinear local and global document degradation models. Int. J. Imaging Syst. Technol. **5**, 220–230 (1994). doi:10.1002/ima.1850050305
18. Kanungo, T., Zheng, Q.: Estimating degradation model parameters using neighborhood pattern distributions: an optimization approach. Trans. Pattern Anal. Machine Intell. **26**(4), 520–524 (2004). doi:10.1109/TPAMI.2004.1265867
19. Kim, B.J., Pearlman, W.: An embedded wavelet video coder using three-dimensional set partitioning in hierarchical trees (spiht). In: Data Compression Conference, 1997. DCC '97. Proceedings, pp. 251–260. Snowbird, USA (1997). doi:10.1109/DCC.1997.582048
20. Klijn, E., Bibliotheek, K.: The current state-of-art in newspaper digitization: a market perspective. D-Lib Mag. (2008). doi:10.1045/january2008-klijn
21. 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: Proceedings of Eighth International Workshop on Frontiers in Handwriting Recognition, pp. 244–249 (2002). doi:10.1109/IWFHR.2002.1030917
22. Lesk, M.: Substituting images for books: the economics for libraries. In: Symposium Document Analysis and Information Retieval, pp. 1–6. Las Vegas, Nevada (1996)
23. Monteil, J., Beghdadi, A.: A new interpretation and improvement of the nonlinear anisotropic diffusion for image enhancement. IEEE Trans. Pattern Anal. Machine Intell. **21**(9), 940–946 (1999). doi:10.1109/34.790435
24. Nishida, H., Suzuki, T.: Correcting show-through effects on document images by multiscale analysis. In: Suzuki, T. (ed.) Pattern Recognition, 2002. Proceedings of 16th International Conference on, vol. 3, pp. 65–68 (2002)
25. Nordström, N.: Biased anisotropic diffusion–unified regularization and diffusion approach to edge detection. Computer Vision – ECCV 90 pp. 18–27 (1990). doi:10.1007/BFb0014846

26. Oja, E., Yuan, Z.: The fastica algorithm revisited: Convergence analysis. IEEE Trans. Neural Netw. **17**(6), 1370–1381 (2006). doi:10.1109/TNN.2006.880980
27. Perona, P., Malik, J.: Scale-space and edge detection using anisotropic diffusion. IEEE Trans. Pattern Anal. Machine Intell. **12**(7), 629–639 (1990). doi:10.1109/34.56205
28. Rice, S., Jenkins, F., Nartker, T.: The fifth test of ocr accuracy. Tech. Rep. TR-96-01, ISRI, Univ. Nevada Las Vegas, Las Vegas (1996)
29. Rice, S., Kanai, J., Nartker, T.: A report on the accuracy of ocr devices. Tech. Rep. TR-92-02, Univ. Nevada Las Vegas, Las Vegas (1992)
30. Roth, K.: Scaling of water flow through porous media and soils. Eur. J. Soil Sci. **59**(1), 125–130 (2008). doi:10.1111/j.1365-2389.2007.00986.x
31. Rudin, L.I., Osher, S., Fatemi, E.: Nonlinear total variation based noise removal algorithms. Phys. D **60**(1–4), 259–268 (1992). doi:10.1016/0167-2789(92)90242-F
32. Salerno, E., Tonazzini, A., Bedini, L.: Digital image analysis to enhance underwritten text in the archimedes palimpsest. Int. J. Doc. Analy. Recogn. **9**(2), 79–87 (2007). doi:10.1007/s10032-006-0028-7
33. Sharma, G.: Show-through cancellation in scans of duplex printed documents. IEEE Trans. Image Process. **10**(5), 736–754 (2001). doi:10.1109/83.918567
34. da Silva, J.M.M., Lins, R.D., Martins, F.M.J., Wachenchauzer, R.: A new and efficient algorithm to binarize document images removing back-to-front interference. J. Univ. Comput. Sci. **14**(2), 299–313 (2008)
35. Tan, C.L., Cao, R., Shen, P.: Restoration of archival documents using a wavelet technique. IEEE Trans. Pattern Anal. Machine Intell. **24**(10), 1399–1404 (2002). doi:10.1109/TPAMI.2002.1039211
36. Tan, C.L., Cao, R., Shen, P., Wang, Q., Chee, J., Chang, J.: Removal of interfering strokes in double-sided document images. In: Applications of Computer Vision, 2000, Fifth IEEE Workshop on., pp. 16–21. Palm Springs, USA (2000). doi:10.1109/WACV.2000.895397
37. The Mathworks Inc., Natick: MATLAB Version 7.5.0
38. Tonazzini, A., Salerno, E., Bedini, L.: Fast correction of bleed-through distortion in grayscale documents by a blind source separation technique. Int. J. Doc. Anal. Recogn. **10**(1), 17–25 (2007). doi:10.1007/s10032-006-0015-z
39. Tonazzini, A., Salerno, E., Mochi, M., Bedini, L.: Blind source separation techniques for detecting hidden texts and textures in document images. Image Anal. Recogn. 241–248 (2004). doi:10.1007/b100438
40. Vaziri, H.H., Xiao, Y., Islam, R., Nouri, A.: Numerical modeling of seepage-induced sand production in oil and gas reservoirs. J. Petrol. Sci. Eng. **36**(1-2), 71–86 (2002). doi:10.1016/S0920-4105(02)00264-4
41. Voci, F., Eiho, S., Sugimoto, N., Sekibuchi, H.: Estimating the gradient in the Perona-Malik equation. Signal Process. Mag. IEEE **21**(3), 39–65 (2004). doi:10.1109/MSP.2004.1296541
42. 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.) Image and Graphics, 2007. ICIG 2007. Fourth International Conference on, pp. 880–885 (2007). doi:10.1109/ICIG.2007.179
43. Yam, H.S., Barney Smith, E.: Estimating degradation model parameters from character images. In: Barney Smith, E. (ed.) Document Analysis and Recognition, 2003. Proceedings of Seventh International Conference on, vol. 2, pp. 710–714. Edinburgh, Scotland (2003)
44. Zhang, D.: Stochastic Methods for Flow in Porous Media: Coping with Uncertainties. ISBN 0-12-779621-5. Academic Press, San Diego (2002)
45. Zhang, X., Lu, J., Yahagi, T.: Blind separation methods for image show-through problem. In: Lu, J. (ed.) Information Technology Applications in Biomedicine, 2007. ITAB 2007. Sixth International Special Topic Conference on, pp. 255–258 (2007). doi:10.1109/ITAB.2007.4407395
46. Zi, G.: Groundtruth generation and document image degradation. Technol. Rep. LAMP-TR-121/CAR-TR-1008/CS-TR-4699/UMIACS-TR-2005-08, University of Maryland, College Park (2005). http://lampsrv01.umiacs.umd.edu/pubs/TechReports
47. Zi, G., Doermann, D.: Document image ground truth generation from electronic text. In: Doermann, D. (ed.) Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on, vol. 2, pp. 663–666 (2004). doi:10.1109/ICPR.2004.1334346