EFDM: Restoration of Single-sided Low-quality Document Images

Reza Farrahi Moghaddam and *Mohamed Cheriet*
Synchromedia Laboratory for Multimedia Communication in Telepresence,
École de Technologie Supérieure, Montréal (QC), H3C 1K3 Canada
reza.farrahi-moghaddam.1@ens.etsmtl.ca, mohamed.cheriet@etsmtl.ca

Abstract

This paper addresses the problem of restoration and enhancement of very old single-sided document images. At first step, a degradation model is developed for the generation of synthesized degraded document images in both double-sided and single-sided formats. Then, we propose a novel method, which is based on the anisotropic diffusion method (ADM), for restoration of the degradations in single-sided document images. Due to local characteristics of ADM, we empower our method with two flow fields to play the role of global classifiers in separating the meaningful pixels. Also, the new method uses an extra diffusion of background information that provides an efficient and accurate restoration of the interference patterns and degraded backgrounds. The performance of the method is tested on both real samples, from Google Book Search dataset and UNESCO's Memory of the World Programme, and synthesized samples provided by our degradation model. The results are promising.

Keywords: PDE-based image processing, Degradation modeling, Nonlinear model, Document enhancement and restoration

1 Introduction

The problem of restoration of bleed-through degradation in document images, which is a major task in analysis of very old documents, is studied from several points of view [6, 8, 9]. In this work, we use diffusion-based methods, which have a great similarity with the physical processes of degradation, to restore the bleed-through and other similar effects.

If we look at the degradation problems in documents form a physical point of view, it appears that many of document degradations are the results of some kind of diffusion processes which occur along the time. Therefore, a good utilization of an inverse diffusion model can result in a very good and direct restoration of these types of physical diffusion. Moreover, the denoising nature of diffusion models provides a clear output without any need for extra denoising postprocessings [3].

The first problem that arises when using a diffusion method is the diffusion of strokes' ink which causes loss of sharpness of strokes. This side effect can be prevented by using small values for the diffusion coefficients. However, on the other hand, the diffusion coefficients must be large enough in order to denoise and remove the background of document. The second problem is that diffusion models are local in the sense of spatial coordinates. In other words, in diffusion models, only the data which are at the neighbors of a pixel can alter the information of that pixel. If a pixel is surrounded by, for example, dark regions, there is no way to insert bright and white information in that pixel. This high spatial dependency of diffusion models limits their ability in information layer (source) separation and removal.

In this work, we introduce a novel restoration method for application to degraded and very old documents. Also, a degradation model is provided in order to generate synthesized degraded document images to test the restoration method.

2 The basics of diffusion modeling

The limits of the diffusion methods can be removed by providing some means of data exchange between different regions of an image. The actual way for providing this data exchange depends on the application. In our method we use two techniques: global classifiers and external diffusions. Global classifiers help the method in preserving of weak and thin connections and edges. The latter technique, which is actually a diffusion-based technique, provides diffusion of information from different sources, for example background information, to all regions of the document image. This additional diffusion process can easily break the spatial barrier in the diffusion models. For a better understanding of the extra diffusion process, we can imagine the sources of information, which create the document image, as some plates that on each of these plates only the information of one source resides. When a document is printed and generated, several sources of information (plates) are usually presented: the main text plate, the background plate, the other-side-of-

the-paper's text plate. Over the time, some other important plates, such as handwriting, may be added. We also consider a common and working plate where the information from the other plates will feed in and, through some interactions, compose the final image. The interaction of each plate with the common plate can be considered as a diffusion process with a specific rate. The schematic diagram in Figure 1 shows the processes that in this modeling results in the final document image. This physical modeling covers the physical degradations which are usually excluded in other degradation models. Here, by physical degradations we mean degradations which persist on the document prior to imaging and scanning processes. The state-of-the-art degradation models [2] deal with degradations which arise from imaging and scanning processes. Our model fills the gap for the physical degradations.

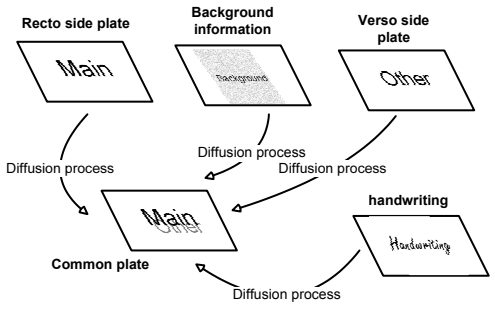


Figure 1. The schematic diagram of degradation model.

3 Methodology

In this section, we first introduce our degradation model that generates the synthesized physically degraded double-sided and single-sided document images. Then, a restoration method will be proposed that is able to remove degraded background and interference patterns from the single-sided document images. The main concepts of the restoration method will be discussed.

3.1 Diffusion-based modeling of physical degradations in very old documents

There are many variations of diffusion-based models in the literature. However, the basic of all of these models is the following governing equation [7]:

$$u_t = \nabla \cdot (c\,(\nabla u)\,\nabla u) = \text{div}\,(c \nabla u) =: \text{DIFF}\,(u, u, c) \tag{1}$$

where c is the diffusion coefficient. Here, we introduce the extended notation of DIFF (u, s, c) to represent the diffusion process of the source s to the target u with the diffusion coefficient c. Usually, equation (1) is referred as the anisotropic diffusion model (ADM). As it was discussed in the previous section, the physical degradations of documents can be modeled as superposition of several physical diffusion processes (see figure 1). Therefore, the corresponding governing equation for the physical processes can be represented, in a similar form to the equation (1), as follows:

$$\begin{aligned} u_t &= \sum_{i \in \text{sources}} \text{DIFF}\,(u, s_i, c_i) \\ &= \text{DIFF}\,(u, u_{\text{recto}}, c_{\text{recto}}) + \text{DIFF}\,(u, s_{\text{bg}}, c_{\text{bg}}) \\ &\quad + \text{DIFF}\,(u, u_{\text{verso}}, c_{\text{verso}}) + \text{etc} \end{aligned} \tag{2}$$

Equation (2) is the governing equation of our degradation model. By providing the recto and verso side images and the background information, the model is able to generate

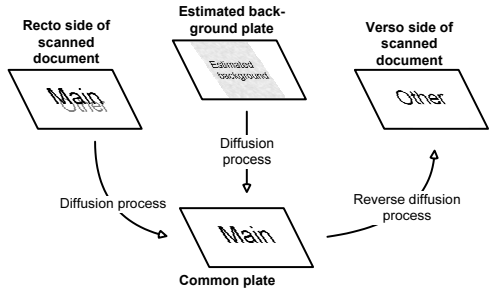


Figure 2. The schematic diagram of the double-sided restoration model formulated in equation (3).

degradations such as aging and bleed-through. The degradation model is used for generating both double-sided and single-sided document images. For example, in Figures 3(c) and 3(d) two sides of a degraded document which is created by the model are shown. The model simulates nonlinear ink seepage and ink spread effects.

The degradation model can be easily changed to a restoration method for double-sided documents. In this method, the backward diffusion of verso-side information is used for removal of interfering patterns on the recto side:

$$\begin{aligned} u_t &= \text{DIFF}\,(u, u_{\text{recto}}, c_{\text{recto}}) + \text{DIFF}\,(u, s_{\text{bg}}, c_{\text{bg}}) \\ &\quad - \text{DIFF}\,(u, u_{\text{verso}}, c_{\text{verso}}) + \text{etc} \end{aligned} \tag{3}$$

The schematic diagram of the method for restoration of double-sided document images is shown in Figure 2. The backward diffusions weaken and remove the unwanted layers and at the same time the diffusions from the background and the image itself fill up the gaps and sharpen the strokes.

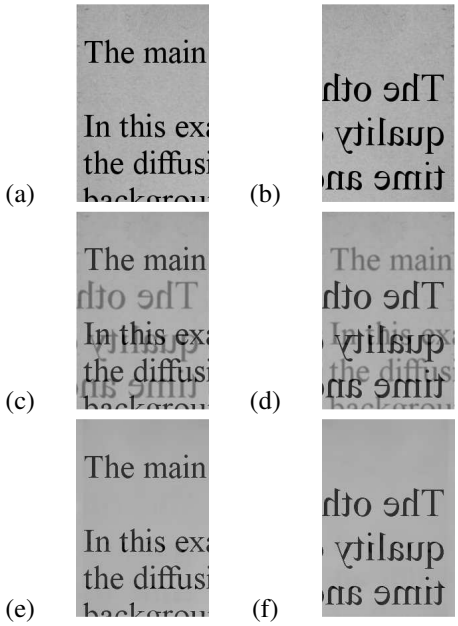


Figure 3. (a) and (b): The original images used as recto and verso sides to generate synthesized degraded document images. (c) and (d): The synthesized degraded double-sided document images which are created using the degradation model. (e) and (f): The restored version of document using the equation (3).

If we apply the method of the equation (3) to the generated input images of Figures 3(c) and 3(d), the interfering patterns of the verso side will be removed (see Figures 3(e) and 3(f)). The outputs are very clear, and unwanted non-text layers are also removed.

We focus from now on the restoration of the single-sided images. Indeed, many archives of the document are only available in single-sided format. Also, in the single-sided cases, registration is not required.

3.2 Restoration of single-sided documents

In many cases, the verso side of the document isn't available. In this section, a diffusion method is presented which can be used for restoration and enhancement of the single-sided documents. The methods based on the ADM are very simple but powerful and they preserve the edges and boundaries. However, because of its local nature, the ADM is unable in preserving very-fine strokes and connections.

The flowchart of the proposed method is illustrated in Figure 4. The input image is placed on the common plate. Also, there are several diffusion processes that take place from other plates to the common plate. Similar to the degradation model, the background information is fed

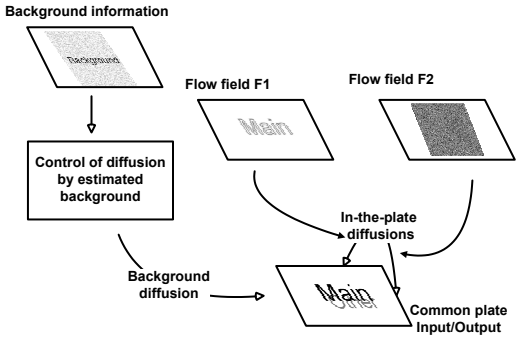


Figure 4. The schematic diagram of proposed method for restoration of single-sided document images.

to the image. For having a uniform background, the background information, s_{bg}, is actually a constant. The rate of the background diffusion is controlled spatially by the estimated background information s_{ebg}. The method of computing the estimated background is a blind method and will be discussed in subsection 3.2.1. The control function, which is actually the diffusion rate for the background diffusion, is as follows:

$$c_{bg} = \left(1 + \tanh \frac{u - s_{ebg}}{\delta_{bg}}\right)\left(1 + \tanh \frac{u - u_{\text{dark}}}{\delta_{\text{dark}}}\right)$$

where u is the document image, s_{ebg} is the estimated background information, u_{dark} is an estimation of the strokes' gray level, and δ_{bg} and δ_{dark} are two bandwidth parameters. The parameters δ_{bg} and δ_{dark} are actually constant, and we use the following values in all computations: $\delta_{bg} = 0.05$ and $\delta_{\text{dark}} = 0.05$. The idea behind the c_{bg} function is that the pixels which are likely to be part of the background could be painted by the uniform background information in order to have a clean and uniform background. This is very important, especially for regions which are surrounded by text loops and curves. In addition to the background diffusion, similar to other diffusion methods, there are some diffusion processes involving the eight neighbors of each pixel. As stated before, our method is based on the ADM. Although, it has many benefits such as simple implementation, ADM is local. This local characteristics lowers its performance especially when there is high degree of noise or spatial variations in the image.

In order to overcome the shortcomings induced by the local aspect of the ADM, we introduce two global classifiers, the flow fields F1 and F2, in our method to improve the performance of the diffusion method when facing very weak and thin strokes or noisy images. The detail of the classifiers will be discussed in the subsections 3.2.2 and

3.2.3. By a flow field we mean a new variable which extends on the entire image region, and hence breaks the locality barrier. The first classifier, the flow field F1, provides membership $[0, 1]$ and states how likely each pixel belongs to the boundaries. Pixels with the flow field zero are boundaries. On the other hand, for the pixels which are within uniform areas, F1 is approximately one. This means the gray value of the pixel has a great correlation to its neighborhood. Moderate values of the flow field F1 actually show the state of the pixel form a global point of view; if a pixel is on a boundary its flow value will be very small. As we move away from the boundaries, the flow value increases and reaches its maximum value in an asymptotic manner. This global variable is very helpful in cases when there are very thin and weak connections and strokes. Even with high levels of diffusion rate, the flow field F1 ensures that thin connections will not receive destructive diffusions from the surrounding large background regions. The second global variable which we use in our method is the flow field F2. The flow field F2 classifies the pixels based on the probability that they are noisy or true data. Thus, we can use F2 to prevent the noisy data to diffuse. Diffusion of noisy data to the surrounding pixels destroys the true data, and this is one of major drawbacks of the averaging techniques. But, if we prevent false data from spreading over nearby regions, the diffusion of true data will remove the noisy data gradually. Also, as a result, the true data will be the main data which is exchanged for enhancement and restoration. For implementation of the two new fields in the diffusion processes, we modify the "div" operator to an extended version. In this extension, the diffusion rate from each neighbor is not only a function of the diffusion coefficient of two pixels, but also it is a function of the flow field F2 and the gradient of the flow field F1. In an implicit form, the governing equation of our method which contains both the background diffusion and the flow-based in-plate diffusion can be written as follows:

$$\frac{\partial u}{\partial t} = \text{DIFF}(u, s_{bg}, c_{bg}) + \text{div}_{F2,\nabla F1}\,(c\nabla u)$$

The first term represents the one-pixel diffusion from the background information s_{bg} to the image plate and is controlled by the diffusion coefficient c_{bg}. The diffusion coefficients c are usual anisotropic ones, and we use the square form of it [7]. Both flow fields compute the relation of the target pixel data to its surrounding pixels. We call the new method for single-sided document images Extended Flow-based Diffusion Method (EFDM).

3.2.1 Estimation of Background

To estimate the background information s_{ebg}, we use a blind method. At the first step an average background is computed with a large sampling window. Then for each pixel, the value of average background is used as the threshold of background, and in an iterative manner a new background value is computed but now with a small sampling window. For the strokes which have small and finite width, the average background will be very brighter than their gray values, and they will be excluded in the subsequent iterations. The algorithm is shown in Algorithm 1. The function BKGDEST(u, l, s) uses a sampling window with size s and based on l, threshold of background in gray level, estimate a background value for each pixel. The parameters ϵ and ϵ' are small values used for stability of the algorithm. The parameters $\text{size}_{\text{large}}$ and $\text{size}_{\text{small}}$ are robust and have small influence on the result. The values $\text{size}_{\text{large}} = 11$ and $\text{size}_{\text{small}} = 3$ are used in computations.

Algorithm 1: Estimation of the background

1 $n = 0$, load image u;
2 Compute the average background:
$s_{ebg} = \text{BKGDEST}(u, \epsilon, \text{size}_{\text{large}})$;
3 **repeat**
4 $n = n + 1$, compute new background:
$s_{ebg} = \text{BKGDEST}(u, s_{ebg} - \epsilon', \text{size}_{\text{small}})$;
5 **until** *steady state* is obtained ;

3.2.2 Data Pixels Classification using Flow Field F1

The flow field F1 computes the structured correlation between a pixel and its neighbors. Therefore, it can easily recognize and find boundary and edge pixels. For boundary pixels, the fraction of neighbors which have similar gray level is very small. Thus, the boundary pixels and sole pixels will appear on the flow field F1 as dark lines and dots. In EFDM, use of F1 saves the boundary pixels from dissolution in the huge background information. Also, at the end of computations, the flow field F1 will provide clear and continuous boundaries and edges of the text. This field can be computed as follows:

$$F1_{i,j} = \sum_{k,l \in N_{i,j,\text{size}_f}} \frac{\exp\left(\frac{-(u_{i,j}-u_{k,l})^2}{\sigma_f^2}\right)}{(i-k)^2 + (j-l)^2}$$

where σ_f is a level parameter for selection of the degree of detail in the flow field computation. The size of the sampling window size_f is selected to be 7. Very low values of size_f cause loss of globality of the flow field and very high values result in unshaped transitions across the boundaries. The values of F1 are normalized to one.

3.2.3 Noise Pixels Classification using Flow field F2

In contrast to the flow field F1, the flow field F2 is a measure for the unstructured correlation between a pixel

and its neighbors. In a similar formulation to F1, we have:

$$F2_{i,j} = s\left(\left\|\sum_{k,l} \frac{R\vec{v}\,\mathrm{sgn}(u_{i,j}-u_{k,l})\exp\left(\frac{-(u_{i,j}-u_{k,l})^2}{\sigma_f^2}\right)}{||v||^2}\right\|\right)$$

where $\vec{v} = (i-k, j-l)$, R is an operator which converts vectors in normalized rays, and s is the saturation function: $s(x) = \tanh(x/\delta_{\text{sat}})$, where δ_{sat} is the saturation parameter and depends on size_f.

4 Experiments and Results

4.1 Experiments setup

We apply the restoration method of the previous section to the real physically degraded samples from the Google Book Search Dataset [5] and Memory of the World Programme of UNESCO [1][1]. Google recent database contains scanned images of books which are flattened and resized. The DVD of the dataset contains data for 68 books, while the hard drive contains 1,000 books, and so something like 300,000 pages are available in the dataset. Also the performance of the method is tested on the synthesized degraded document images that are generated using our degradation method of previous section. In the following some examples of the experiments are presented.

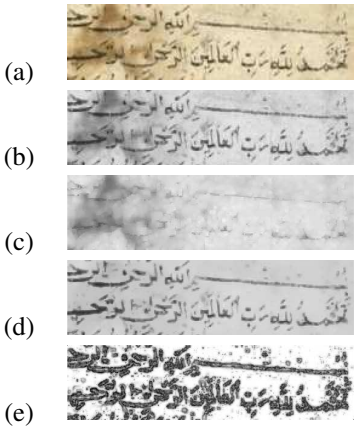


Figure 5. From (a) to (e): The color and gray scale version of a degraded very old document image. The estimated background. The output of the model. The flow field of the output.

4.2 Experiments

As the first example, we use the degraded color document image in Figure 5(a). The image is obtained from the photo database of UNESCO's Memory of the World Programme [1]. The gray scale form of the document is shown in Figure 5(b). The background is nonuniform and in some positions the background's gray level is very close to the main text's one. We start with the estimated background which is shown in Figure 5(c). Figure 5(d) shows the result of application of EFDM. Not only the variable background is removed, but also all the details of the handwriting are preserved. The low value of flow field for the edges' and boundaries' pixels stops the diffusion process across these boundaries even when the gray levels of the two sides of the boundary are very close to each other. This saves the boundary pixels and also enhances the contrast across the boundary. It is good to see that, due to the global control of the flow field F1, the weak strokes and connections on the right part of the image are saved while they have actually very close value of gray levels to the dark background on the left side. On the other hand, the flow field F2 accelerates the diffusion to the noisy pixels. The flow field F1, which is shown in Figure 5(e), provides the boundaries and edges of the image. Although the main output of the model is shown in Figure 5(d), the boundaries and edges in the flow field F1 can be used for other postprocessing tasks. The performance of the method can be easily seen from Figure 6. In this figure, the histograms of the different parts of the input image of figure 5 are plotted with the same scale before and after application of the method. In this example, the left side of the image suffers from a degraded and dark background that can be easily observed from the corresponding histogram. This high intensity and distributed background actually masks all the texts and valuable information. After the application of the method, the background is shrunk and concentrated in a bright and small part of the gray scale, and the low intensity but meaningful

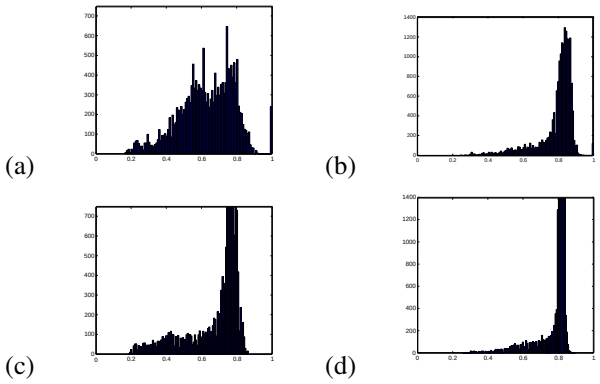
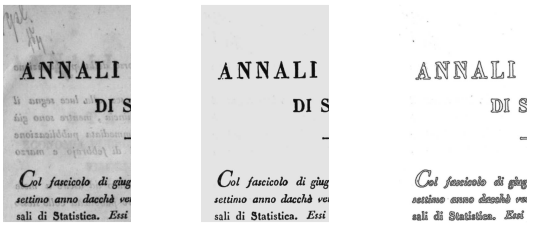


Figure 6. (a) and (b): The gray values histogram of the left and right sides (150×120) of the degraded input image shown in Figure 5(b). (c) and (d): The histograms after restoration and enhancement, computed for output image shown in Figure 5(d).

[1] http://portal.unesco.org/ci/photos/showgallery.php/cat/531

parts can be easily separated from the background. For the right side of the image, background is not very degraded and there is no noticeable interference pattern. However, the method results in better and higher contrast between the background and the text while preserving all the low and weak strokes.



(a) (b) (c)

Figure 7. (a): The gray scale version of a degraded very old document image. (b): The output of the model. (c): The flow field F1: The output boundaries.

(a) (b) (c)

Figure 8. The synthesized degraded single-sided document image. (a): The input in color. (b): Gray level equivalent of the input. (c): The restored output.

In the second example, from the Google Book Search Dataset, we apply the same method to the degraded document shown in Figure 7(a). The main text has several values of the gray level. The background is not clean, and several layers of information such as handwriting are presented. The output, and the flow field F1 are shown respectively in Figures 7(b), and 7(c). Once again, the continuous boundaries are available via the flow field F1. The output is clean without variable background and the strokes are enhanced and sharpened. The distribution of ink within strokes is also more uniform. However, our single-sided restoration method cannot be used for extraction of weak overlays. We are carrying out restoration of document images using their both sides in order to separate and extract weak information layers, such as the handwriting annotations.

Finally, the method is applied to the synthesized image in figure 8(a) (the gray level is shown in Figure 8(b)). The output is shown in Figure 8(c) which provides a clear and binarized document image.

For the sake of comparison with the state of the art methods [4, 9], we noticed that our restoration method gives far better results as it relies on the modeling of various degradations. For objective and goal-oriented evaluation of the model, we are undertaking evaluations on several databases such as Google Book Search dataset and datasets from Jumua Al Majid Center for Culture and Heritage - Dubai.

5 Conclusion and Future Prospects

A restoration method for the single-sided document images (in color or in gray level) is introduced. The method is based on the anisotropic diffusion method and is empowered by addition of two global flow fields. The flow fields determine whether or not the gray level of a pixel belongs to a boundary in the image and reliability of their data respectively. Also, an external diffusion from a background plate, which helps to recover unreachable areas, is included in the method. The method is tested on real very-old document images from Google Book Search dataset and databases from UNESCO's Memory of the World Programme with promising results. Also, to test the method on synthesized degraded documents, a degradation model for very old documents is introduced. The model is based on the diffusion processes and uses the plate concept to include several sources such as recto and verso sides of documents. The model is able to simulate aging and bleed-through degradations in documents. We have shown the effectiveness of our method on single-sided documents; however, one can easily extend it to double-sided cases.

References

[1] A. Abid, *Museum International*, 49(1):40–45, 1997.

[2] H. Baird, *Digital Document Processing*, pp 261–279, 2007.

[3] F. Drira, F. LeBourgeois and H. Emptoz, F. LeBourgeois, editor, *ICDAR 2007 Vol. 2. Ninth International Conference on*, 2007, volume 2, pp 1068–1072.

[4] E. Dubois and P. Dano, *Proc. IS&T Archiving 2005*, April 2005, pp 170–174, Washington DC, USA.

[5] Google, *Book Search Dataset*, Version v edition, 2007.

[6] G. Leedham, S. Varma, A. Patankar and V. Govindaraju, *Proc. Eighth International Workshop on Frontiers in Handwriting Recognition*, 6–8 Aug. 2002, pp 244–249.

[7] J. Monteil and A. Beghdadi, *IEEE Trans. Pattern Analysis and Machine Intelligence*, 21(9):940–946, Sept. 1999.

[8] G. Sharma, *Image Processing, IEEE Transactions on*, 10(5):736–754, 2001.

[9] A. Tonazzini, E. Salerno and L. Bedini, *International Journal on Document Analysis and Recognition*, 10(1):17–25, June 2007.