Document images restoration by a new tensor based diffusion process: Application to the recognition of old printed documents

Fadoua DRIRA
REGIM - ENIS
Route de Soukra km 4
w.3038 Sfax, Tunisie
fadoua.drira@ieee.org

Frank LeBourgeois,Hubert Emptoz
LIRIS - INSA Lyon
20 avenue Albert Einstein
69621 Villeurbanne Cedex, France
frank.lebourgeois, hubert.emptoz@liris.cnrs.fr

Abstract

A modification of the Weickert coherence enhancing diffusion filter is proposed for which new constraints formulated form the Perona-Malik equation are added. The new diffusion filter, driven by local tensors fields, takes benefit from both of these approaches and avoids problems known to affect them. This filter reinforces character discontinuity and eliminates the inherent problem of corner rounding while smoothing. Experiments conducted on degraded document images illustrate the effectiveness of the proposed method compared to another anisotropic diffusion approaches. A visual quality improvement is thus achieved on these images. Such improvement leads to a noticeable improvement of the OCR system's accuracy proven through the comparison of OCR recognition rates before and after the diffusion process.

1. Introduction

Differential Equation (PDE) techniques have been widely applied over the last decade in image processing and they have been mainly devoted in image enhancement. The PDE-based formalism enables to regularize an image through successive iterations. The most interesting PDE-based approaches are those using a non-linear diffusion filtering since it overcomes the limitations of linear methods leading to isotropic smoothing. They could be driven by either a scalar-valued diffusivity [3] or a matrix-valued diffusion tensor [2]. Faced with a wide range of such PDE formalisms, each regularization method has its advantages and its disadvantages.

The tensor driven diffusion process, proposed by Weickert [1], leads to good results in case of anisotropic structure: in particular, we can note its ability, compared to the other processes, to close interrupted structures. However, the Weickert process destroys singularities and leads generally to curved structures. This problem is avoided with a direct application of the scalar diffusion introduced by Perona-Malik. This process has the interesting property to vanish when the value of the local smoothed gradient is important. Looking for an approach that avoids corner rounding and takes into account the coherence-enhancing property, we propose to combine into one approach the best of Perona-Malik and Weickert processes. For instance, it would be desirable to design a filter performant in noise removal with coherence-enhancing property and both of corner and edges preservation. Having all these properties grouped in one filter would be very interesting mainly in the treatment of textual document images with poor quality. In fact, noise removal would be beneficial if the legibility of the text as well as the visual quality of an image are enhanced [4]. Such enhancement is correctly achieved if this noise removal takes into account the reinforcement of broken characters (coherence-enhancing property) and adapts smoothing as much as possible to the geometry of the image. By image geometry, we mean every important feature in the image such as singularities and edges.

This paper is organized as follows: Section 2 presents a brief review of both of the Perona-Malik and the Weickert filters. They are chosen among the different PDE-based approaches since they are respectively known for their performance in singularity preservation and coherence-enhancing properties. These features are combined into a new PDE approach that details represent the motivation of the next section (section 3). Some experiments, done on degraded document images, indicate in section 4 that our proposition provides an efficient tool for improving the visual quality of these images. This leads to a noticeable improvement of the OCR system's accuracy proven through the comparison of OCR recognition rates before and after the diffusion process. A comparative study with results generated with other PDE-based approaches is also given.

978-0-7695-3725-2/09 $25.00 © 2009 IEEE
DOI 10.1109/ICDAR.2009.109

IEEE computer society

2. Perona-Malik/Weickert filter: a brief review

2.1. Case of the Perona-Malik filter

Perona and Malik's work [3] on nonlinear isotropic diffusion has been the first issue in the area. They proposed a selective diffusion that preserves edges; a diffusion coefficient function is chosen to be inversely proportional to the image's gradient. Therefore, the diffusion flow increases within homogeneous regions where the gradient tends to be small and decreases across non-linear discontinuities where the gradient is very high. This is clearly illustrated in the Perona-Malik equation given as:

$$\left\{ \begin{array}{rcc} I_t & = & div\left(C\left(\|\nabla I\|\right) \times \nabla I\right) \\ I_0 & = & I_{Noisy} \end{array} \right. \quad (1)$$

where $\|\nabla I\|$ is the gradient magnitude of the image I and $C\left(\|\nabla I\|\right)$ is an edge stopping function chosen to satisfy $C(0) = 1$ and $lim_{v\to\infty} C(v) = 0$.

From this equation, we notice that the coefficient $C\left(\|\nabla I\|\right)$ serves the purpose of selecting the locations in the image for smoothing. If C is chosen constant at all image locations, the diffusion equation yields isotropic Gaussian smoothing. The function C could be defined by $C\left(\|\nabla I\|\right) = \frac{1}{1+\left(\frac{\|\nabla I\|}{k}\right)^2}$ or $C\left(\|\nabla I\|\right) = \exp\left(-\frac{\|\nabla I\|}{k}\right)^2$

In general, nonlinear isotropic diffusion filter processes like linear diffusion filter inside uniform zones but acts differently near edges where diffusion is inhibited. Consequently, noise at edges cannot be eliminated.

2.2. Case of the Weickert filter

The previous cited work was further extended to another nonlinear process which is anisotropic and driven by an adapted diffusion tensor instead of a scalar diffusivity. A nonlinear anisotropic diffusion equation driven by a diffusion tensor is given as follows:

$$\left\{ \begin{array}{rcc} I_t & = & div\left(D\left(J\right)\nabla I\right) \\ I_0 & = & I_{Noisy} \end{array} \right. \quad (2)$$

where D is an anisotropic diffusion tensor which depends on the image via the structure tensor J defined as $J = J_\rho\left(\nabla I_\sigma\right) = G_\rho\left(\nabla\left(G_\sigma \otimes I\right)\nabla\left(G_\sigma \otimes I\right)^t\right)$. Here G_ρ and G_σ are Gaussian convolution Kernels. $D\left(J\right)$ is performing as an edge stopping function to preserve edges.

For $D = C\left(\|\nabla I\|\right)\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$, we have the nonlinear isotropic diffusion equation of Perona-Malik.

The tensor J could be defined in its orthonormal system of eigenvectors describing the directions where the local contrast is maximal or minimal. This contrast is measured by its eigenvalues. It is defined as follows:

$$J = \lambda_- \times \Theta_-\Theta_-^t + \lambda_+ \times \Theta_+\Theta_+^t \quad (3)$$

where $\lambda_{+/-}$ et $\Theta_{+/-}$ are the eigenvalues and the eigenvectors of the tensor field J above-cited. Therefore, the diffusion tensor is given by:

$$D = f_+\left(\lambda_+, \lambda_-\right) \times \Theta_-\Theta_-^t + f_-\left(\lambda_+, \lambda_-\right) \times \Theta_+\Theta_+^t \quad (4)$$

$f_+\left(\lambda_+, \lambda_-\right)$ and $f_-\left(\lambda_+, \lambda_-\right)$ are two functions satisfying both of these local constraints:

1. Edge preservation: if $\lambda_{+/-} \gg 0$, the smoothing process should be anisotropic and directed along Θ_+.
2. Smoothing within homogenous areas: if $\lambda_{+/-} \approx 0$, the smoothing process should be isotropic.

Among the promising anisotropic PDE-based approaches, we have noticed the Weickert approach [1]. The values of the functions $f_{+/-}$ corresponding to this approach are given by:

$$\left\{ \begin{array}{ccc} f_+\left(\lambda_+, \lambda_-\right) & \overset{if\ \lambda_+ \neq \lambda_-}{=} & \alpha + (1-\alpha)\exp\left(\frac{-C}{(\lambda_+ - \lambda_-)^2}\right) \\ & \overset{else}{=} & \alpha \\ f_- & = & \alpha = 0.001 \end{array} \right. \quad (5)$$

3. Proposition of a new tensor diffusion process

Our proposition takes benefit from the scalar diffusion model of Perona-Malik and the tensor driven diffusion process of weickert. As we have previously noticed, the filter of Perona-Malik is a special case of the Weickert filter. Thus, the main difference between these two approaches relies on the diffusivity coefficient $f_{+/-}$. This coefficient in the case of Weickert filter helps in smoothing along the flow direction. The shape of the diffusivity matrix is thus strongly anisotropic that fails to respect singularities. For Perona-Malik filter, $f_{+/-}\left(\lambda_+, \lambda_-\right) = \exp - \left(\frac{\lambda_+ + \lambda_-}{K}\right)^2$, the diffusivity matrix has no specific orientation but it vanishes near important features (edges and corners). To take advantage of both of the previously compared filters, we propose to maintain in the new filter the Weickert diffusion oriented in the direction of Θ_+ but to change the diffusivity coefficient $f_{+/-}$. This new coefficient must ensure a correct trade-off between complete noise removal and important feature preservation. To this end, we must differentiate between diffusion near a contour where the values of λ_+ and λ_- are large and the diffusion near a corner where λ_+ or λ_- presents high values. The proposed solution to such suggestion is to define separately the values of λ_+ and λ_- in the diffusivity coefficient $f_{+/-}$. After several attempts, we

succeed to design a stable filter that verify all the imposed requirements: Our new filter employs two thresholds K_+ and K_- useful to identify the diffusion along respectively Θ_+ and Θ_- axis. For instance, the quantity of diffusion changes with the encountered feature: contour or singularity. Therefore, the parameters K_+ and K_- are used to define the sensibility threshold necessary to detect a variation in the image.

By comparing both of the values of K_- (useful for the preservation of the edges) and K_+ (useful for the preservation of singularities), we can notice two situations:

- If $K_- \gg K_+$, then the proposed filter processes oriented structures by smoothing along the flow direction. This represents the basic treatment of the Weickert coherence-enhancing diffusion.
- If $K_+ \gg K_-$, then we have a better preservation of image singularities and a significant loss of contours with slight variations. However, with an important value of K_+, the image contours will no longer be preserved.

In its complete form, the equation of the proposed diffusion filter is written as follows:

$$\begin{cases} I_t &= div\left(D^* \times \nabla I\right) \\ I_0 &= I_{Noisy} \end{cases} \quad (6)$$

$$\begin{cases} D^* = \exp\left(\frac{-\lambda_-}{K_-}\right) \times \Theta_+\Theta_+^t + \exp\left(\frac{-\lambda_+}{K_+}\right) \times \Theta_-\Theta_-^t \\ or\ D^* = \frac{1}{1+\left(\frac{\lambda_-}{K_-}\right)} \times \Theta_+\Theta_+^t + \frac{1}{1+\left(\frac{\lambda_+}{K_+}\right)} \times \Theta_-\Theta_-^t \\ \Theta_{+/-}, \lambda_{+/-}\ eigenvectors\ et\ eigenvalues\ of\ T_\sigma^\rho \end{cases}$$

As this EDP is designed to smooth edges, reduce noise in uniform zones and preserve singularities, we define a specific treatment for each of these situations as following (Fig.1):

- Near an edge, we have $\lambda_+ >> \lambda_-$, $\lambda_+ > K_+$ et $\lambda_- < K_-$. In this case, the diffusivity coefficient f_+ (respectively f_-) is close to 0 (respectively to 1) in the direction of θ_- (respectively θ_+). As such treatment similar to that of Weickert, enhance the consistency of features and accentuate smoothing along edges (Fig.1(a)).
- In uniform area with low intensity variation, we have $\lambda_+ < K_+$ et $\lambda_- < K_-$. In this case, the diffusivity matrix D^* becomes totally isotropic as both of diffusivity coefficients f_+ and f_- are close to 1. The diffusion here is equivalent to that done with the heat equation erasing all the small variations (Fig.1(b)).

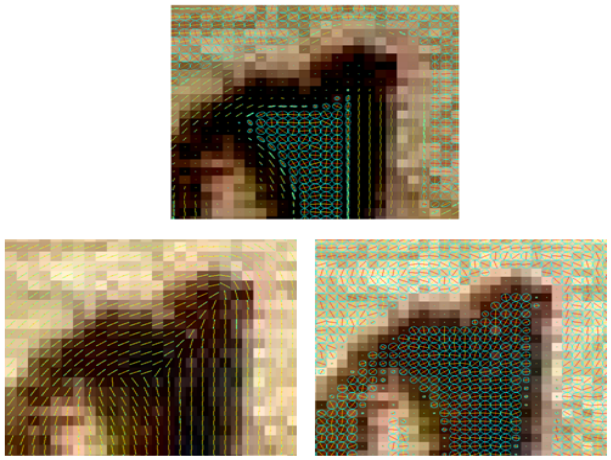


Figure 2. The behaviour statment around contours and corners of (From top to down): the proposed filter, (From left to right) the Weickert filter and the Perona-Malik filter

- Near a corner, we have $\lambda_+ >> 1$, $\lambda_- >> 1$, $\lambda_+ > K_+$ and $\lambda_- > K_-$. In this case, the diffusivity coefficients f_+ et f_- are close to 0. The diffusion process, simulated here to the Perona-Malik process, is stopped in order to preserve right angles and corners (Fig.1(c)).

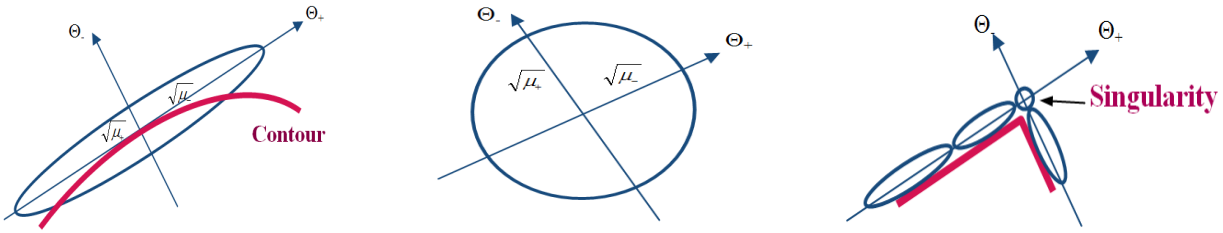
To illustrate these ideas, we give in figure 2, the behaviour statment of the diffusivity matrices D. Each matrix, drown in the case of the image of a letter <<a>>, is caracterized by the shape of an ellipse directed by Θ_+ and its axis length is equal to $\sqrt{f_{+/-}\left(\lambda_+, \lambda_-\right)}$. This figure illustrate that our proposition is a mixture of both of Perona-Malik and Weickert processes.

4. Experimental results : Application on degraded document images characters

4.1. Visual quality improvement

To illustrate the performance of the proposed diffusion filter, we give some experimental results. A first test image is an extract of a degraded character image provided from the database of <<Gazette de Leyde>> (Fig.3). We clearly notice a good visual improvement results of the noisy original image.

Other tests have been done in order to compare the performance of our diffusion filter with its basic approaches applied seperately. For instance, we test these approaches on a degraded character extracted from a document provided by the archive of <<Chatillon-sur-Chalaronne>>. Figure 4 illustrates the image quality before and after processing

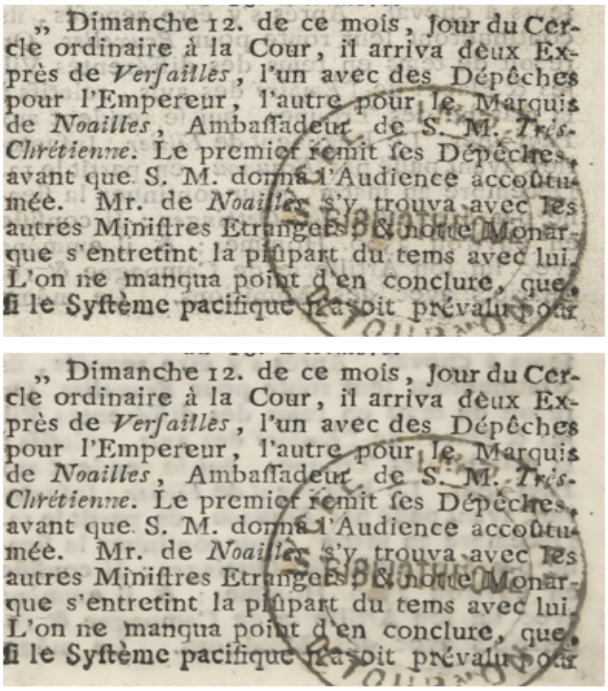


(a) *Diffusivity Matrix D^* in the case of edge processing*

(b) *Diffusivity Matrix D^* in the case of processing areas with slight variations*

(c) *Diffusivity Matrix D^* in the case of singularity processing*

Figure 1. The shape of the diffusivity matrices associated to the proposed filter in the case of three different processed areas



„ Dimanche 12. de ce mois, jour du Cercle ordinaire à la Cour, il arriva deux Exprès de *Verſailles*, l'un avec des Dépêches pour l'Empereur, l'autre pour le Marquis de *Noailles*, Ambaſſadeur de S. M. *Très-Chrétienne*. Le premier remit ſes Dépêches, avant que S. M. donnât l'Audience accoûtumée. Mr. de *Noailles* s'y trouva avec les autres Miniſtres Etrangers, & notre Monarque s'entretint la plûpart du tems avec lui. L'on ne manqua point d'en conclure, que ſi le Syſtème pacifique n'avoit prévalu pour

„ Dimanche 12. de ce mois, jour du Cercle ordinaire à la Cour, il arriva deux Exprès de *Verſailles*, l'un avec des Dépêches pour l'Empereur, l'autre pour le Marquis de *Noailles*, Ambaſſadeur de S. M. *Très-Chrétienne*. Le premier remit ſes Dépêches, avant que S. M. donnât l'Audience accoûtumée. Mr. de *Noailles* s'y trouva avec les autres Miniſtres Etrangers, & notre Monarque s'entretint la plûpart du tems avec lui. L'on ne manqua point d'en conclure, que ſi le Syſtème pacifique n'avoit prévalu pour

Figure 3. Application of the proposed filter on a degraded document image (source <<Gazette de Leyde>>)

Figure 4. From left to right: Original image and restored images with respectively the proposed filter, the Perona-Malik filter and the Weickert filter

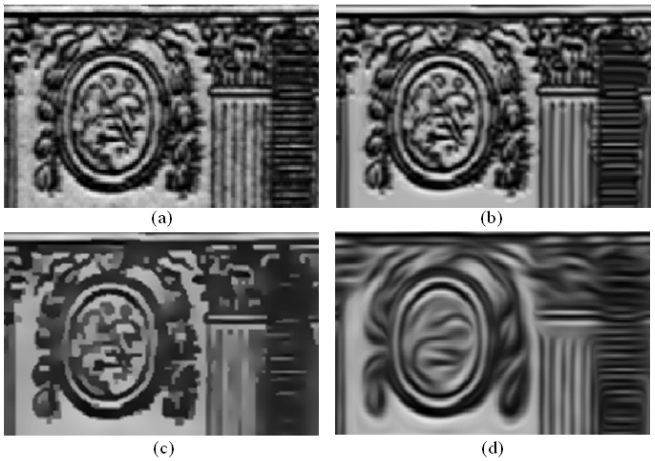


Figure 5. Application on a printed graphical image (a) of The proposed filter (b), the Perona-Malik.filter (c), the Weickert filter (d)

it with respectively the proposed filter, the diffusion filters of Perona-Malik and Weickert. Our proposition generates good visual improvement results compared to the other filters.

Moreover, the proposed filter is not confined to the processing of textual documents; other kinds of images could also be processed. We give for example results on graphical printed images generated after the application of the proposed filter and even after the application of both of Perona-Malik and Weickert filters. Better preservation of image details is achieved with the new tensor diffusion filter (Fig.5).

4.2. OCR accuracy improvement

If the evaluation of natural image quality after a processing step with a PDE filter is possible due to standard measures such as SNR (Signal to Noise Ratio) or MSE (Mean Square Error), this is not the case for degraded textual document images. In fact, these global measurements are more

quantitative rather than qualitative. They don't take into account the importance of information along characters. We can give as an example a small amount of noise located on the part that distinguishes the letter <<t>> from the letter <<l>>. The value of the SNR close to 0 in this case gives the illusion that the image is not degraded but in reality this is not true as this little modification has changed the meaning. Thus, the evaluation of the quality of restored textual document images is not an easy task. There exists no general methodology to evaluate the correctness of a given result. Since the ultimate aim of our proposed PDE-algorithm is to restore damaged characters in noisy document images, we propose to evaluate the result through the study of the optical character recognition accuracy rate. An improved recognition rate is certainly related in this case to an improved quality image. Thus, this section focuses on studying OCR system response before and after restoring damaged characters with PDE-based approaches.

The figure 6 illustrates good visual improvement results of a degraded document image after the application of our filter. As such improvement is accompagned by a good recognition rate proven by the ABBY FineReader 8.0, the OCR software widely known for its high precision character recognition. Actually, before processing the initial document image given in figure 6, the OCR system completes the conversion with the accuracy rate of 89,53%. Thus, a 10,47% failure rate is resulting. According to the restored image, the OCR system had approximately a 93,84% success rate when detecting damaged letters. This rate is noticeably higher than the recognition rates calculated on the original non-processed image. The accuracy of the OCR system was well improved by processing a document page with the proposed filter and we succeed to decrease the failure rate to 6,16%.

Experiments have been done on a database of 106 color pages of <<Gazette de Leyde>>, scanned with 300 dpi. With 144847 as a total number of characters, we obtained 11279 errors before restoration and 8013 errors after restoring with our proposed filter. The OCR system had approximately a 94,46% success rate when detecting restored damaged letters compared to a 92,21% rate calculated on non-processed characters. It is shown that for real document images corrupted by damaged characters, the final restoration images are of improved qualities. We notice that optical character recognition takes great advantage of this improvement for the majority of the tested pages.

5. Conclusions

We have proven in this work the efficiency of a novel PDE-based approach in enhancing the quality of degraded textual document images and even in reducing their OCR errors. This approach doesn't require training nor segmentation steps. It is based on the combination of the Weickert tensor driven diffusion filter and the Perona-Malik scalar diffusion filter and inherits their respective interesting properties: coherence-enhancing and singularities preservation. As such properties are very essential in processing noisy and broken characters. For instance, any degradation affecting these discontinuities might lead to a non-computer-readable format. Hence, it is strongly recommended to process degraded documents with an accurate restoration method in order to improve readability and thus to improve OCR recognition rate. Our experiments show that our proposition outperforms both of Perona-Malik and Weickert filters applied separately.

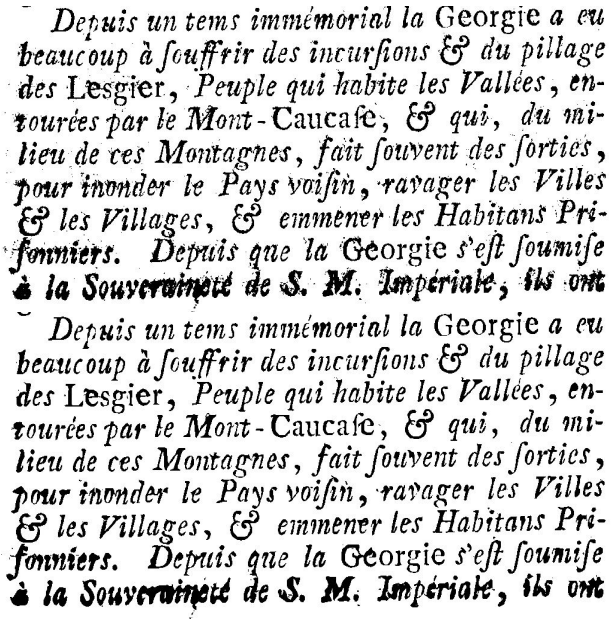


Figure 6. Application of the proposed filter : (From top to down) An extract of a degraded document image and its restored version

References

[1] J. Weickert. Coherence-enhancing diffusion of color images. 7^{th} National Symposium on Pattern Recognition and Image Analysis, Barcelona, Spain, 1997.

[2] J. Weickert. Anisotropic Diffusion in Image Processing. Teubner, Stuttgart, 1998.

[3] P. Perona, J.Malik. Scale-Space and Edge Detection Using Anisotropic Diffusion, PAMI, pp.629-639, 1990.

[4] F. DRIRA, Towards restoring historic documents degraded over time, DIAL'2006,Lyon,pp.350-357,2006.