Denoising textual images using local/non-local smoothing filters : A comparative study

Fadoua DRIRA
University of Sfax, REGIM-Lab, ENIS
3038 Sfax, Tunisia
fadoua.drira@ieee.org

Franck LEBOURGEOIS
University of Lyon, LIRIS, INSA-Lyon
69621 Villeurbanne Cedex, France
franck.lebourgeois@insa-lyon.fr

***Abstract*—Textual document image denoising is the main issue of this work. Therefore, we introduce a comparative study between two state-of-the-art denoising frameworks : local and non-local smoothing filters. The choice of both of these frameworks is directly related to their ability to deal with local data corruption and to process oriented patterns, a major characteristic of textual documents. Local smoothing filters incorporate anisotropic diffusion approaches where as non-local filters introduce non-local means.**

Experiments conducted on synthetic and real degraded document images illustrate the behaviour of the studied frameworks on the visual quality and even on the optical recognition accuracy rates.

***Keywords*-Anisotropic diffusion; non-local means; document denoising; OCR;**

I. Introduction

Noise removal is a useful pre-processing step for improving the visual quality and allowing further application of image processing and analysis tasks, such as segmentation and character recognition. As a solution, a wide variety of filtering algorithms have been proposed [1]. The most interesting ones are those removing noise while keeping the integrity of relevant image information such as edges. This propriety is very interesting mainly in the case of textual document processing. For instance, characters layout must be treated carefully since any modification could change one letter to another and consequently change the whole meaning. Thus, we will be faced with a meaningless textual document [2]. To state one example, the loss of singularities could transform the the letter "t" to the letter "l" which are completely different.

In this study, we exclude approaches making strong assumptions about the properties of the signal and/or degradation since they lack the generality to be easily applied to any applications. The literature on denoising filters is vast and a complete review is beyond the scope of this paper. Only local/non-local smoothing filters would be deeply investigated in this work.

Local smoothing filters are one of the most fundamental tools used for noise removal in images. Their formulation takes the average of all the pixels under a local filter to estimate the intensity of a pixel in the output image. The averaging process often employs a Gaussian window which gives higher weights to pixels closer to the center pixel. Anisotropic diffusion filters, characterized as local smoothing filters, have shown very promising results. Such filters respect edges by averaging in the direction orthogonal to the local image gradient. Moreover, these methods known as geometry-oriented methods are local since only interactions between neighbouring pixels are involved.

Recently, non-local filtering kernels have attracted lot of attention due to their efficiency in preserving edges in noisy images. Non-local means, proposed by Buades et al., is an example of such kernels. It performs a weighted averaging of similar pixels located on the whole image, rather than in a close neighbourhood of the center pixel. Thus, non-local means takes benefit of the redundancy and self- similarity of the information in the image. This makes it well suited to the treatment of textual document images as they have enough redundancy. For instance, many similar configurations could be found on flat zones or even on characters layout.

L. Likforman et al. evaluate non-local means and total variation as a pre-processing step to document recognition. Their study [3] proves the efficiency of the non-local means algorithm while processing textual documents. Therefore, we focus in this paper to compare the performance of this algorithm on a large set of PDE-based approaches. Evaluation takes advantages from classical measures while denoising images. We furthermore emphasize results by giving numerical values of the OCR accuracy rates.

This paper is organized as follows.

Section 2 gives a brief overview of the non-local means and anisotropic diffusion filters respectively associated to non-local/local smoothing filters. Section 3 presents experiments conducted on synthetic and real degraded document images. We illustrate the behaviour of the studied frameworks on the visual quality and even on the optical recognition accuracy rates. Concluding remarks and future works are discussed in Section 4.

978-0-7695-4774-9/12 $26.00 © 2012 IEEE
DOI 10.1109/ICFHR.2012.198

II. Denoising framework : Local versus non-local smoothing filters

A. Local somoothing filters: Anisotropic diffusion

In this section, we present a brief overview of anisotropic diffusion methods and mainly of tensor-driven ones [4], [5]. These methods tend to control the smoothing effect according to the local geometry of the image; as such treatment was a limitation of linear PDE methods leading to an isotropic smoothing erasing the main image features. A tensor-driven diffusion equation is given as follows:

$$I_t \quad = \quad div\left(D\left(J\right)\nabla I\right) \tag{1}$$

where D is an anisotropic diffusion tensor which depends on the image via the structure tensor J given by $J = J_\rho\left(\nabla I_\sigma\right) = G_\rho\left(\nabla\left(G_\sigma \otimes I\right)\nabla\left(G_\sigma \otimes I\right)^t\right)$.

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 [6], [7].

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:

$$J = \lambda_- \times \Theta_-\Theta_-^t + \lambda_+ \times \Theta_+\Theta_+^t \tag{2}$$

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

$$D = f_+\left(\lambda_+,\lambda_-\right) \times \Theta_-\Theta_-^t + f_-\left(\lambda_+,\lambda_-\right) \times \Theta_+\Theta_+^t \tag{3}$$

$f_+\left(\lambda_+,\lambda_-\right)$ and $f_-\left(\lambda_+,\lambda_-\right)$ are two functions.

Besides to this formulation based on gradient divergence, D. Tschumperlé [8] introduced another formulation based on the computation of the trace operators and the Hessian matrix instead of the divergence.

Our previous work discusses the efficiency of PDE-based approaches in processing old printed documents and mainly in improving the accuracy rates [9]. Thus, we include, in this study, the filters proposed by Weickert, Perona-Malik, beltrami, Tschumperlé and Drira et al.. The values of the functions $f_{+/-}$ corresponding to each approach are given in the table I.

B. Non-local somoothing filters: Non-local means

The non-local means algorithm is introduced by Buades et al.[1]. The main idea behind the development of this denoising algorithm is to take advantage of the redundancy and self- similarity in the image. Rather than performing a weighted averaging in a close neighbourhood of the center pixel, another solution consists in the case of the non-local means to average every instance of similar pixels located on the whole image. For a more detailed analysis on the non-local means algorithm and a more complete comparison, see [1]. This method is expressed by:

$$NL[u](x) = \frac{1}{C(x)}\int_\Omega e^{-\frac{(G_a * |u(x+.)-u(y+.)|^2)(0)}{h^2}}u(y)\,dy; \tag{4}$$

where u is the original image, $x \in \Omega$, G_a is a Gaussian kernel, h acts as a filtering parameter. The function $C(x)$ is a normalizing constant defined by :

$$C(x) = \int_\Omega e^{-\frac{(G_a * |u(x+.)-u(z+.)|^2)(0)}{h^2}}dz. \tag{5}$$

$$\begin{aligned} K(x,y) \quad &= \quad G_a * |u(x+.)-u(y+.)|^2 \\ &= \quad \int G_a(t)[u(x+t)-u(y+t)]^2 dt. \end{aligned} \tag{6}$$

The value $u(y)$ is used to denoise $u(x)$ if the local pattern near $u(y)$ is similar to the local pattern near $u(x)$.

III. Experimental results

A. Document image quality measures

Quality measures, necessary to compare the visual difference between two images, are a good issue in ranking, evaluating and optimizing image restoration algorithms. Two solutions are possible to measure such a difference using either subjective or objective measures. The subjective measure is a good solution since a Man is the ultimate viewer, yet it is very costly. The objective measure is easier to implement and to test but it does not always agree with the subjective one.

In this section, we choose to use the objective measure while testing the most popular distortion measures, such as the Peak Signal-to-Noise-Ratio (PSNR) and the Mean Square Error (MSE). The growth of the assumed amount of measurement error, caused by the presence of several different features between the two images, leads to larger MSE/lower PSNR measure. Certainly, these classical scalar-valued image quality measures are very commonly used in video and image processing; but they don't often remain appropriate to document image processing. Actually, these measures are based on a point-based measurement in which mutual relations between pixels are not taken into account. An adequate measure of document image degradations must consider the neighbourhood of pixels and mainly pixels around informative parts which are difficult to locate. In general, document images are characterized by the presence of a great amount of pixels without any specific/important information (the paper background) that statistically influences the MSE measure or the correlation measure. Consequently, degraded images, which are no more readable, keep almost the same MSE, correlation rate and PSNR compared to the same document which remains readable but strongly degraded for the OCR.

Perna-Malik	$f_{+/-} = \exp - \left(\frac{\|G_\sigma \otimes \nabla I\|}{K}\right)^2$
Weickert	$f_+ \overbrace{=}^{if\ \lambda_+ \neq \lambda_-} \alpha + (1-\alpha)\exp\frac{-C}{(\lambda_+ - \lambda_-)^2}$ $f_- \underbrace{=}_{else} \alpha = 0.001$
Drira	$f_{+/-} = \exp\left(\frac{-\lambda_{+/-}}{K_{+/-}}\right)$ or $f_{+/-} = \frac{1}{1+\left(\frac{\lambda_{+/-}}{K_{+/-}}\right)}$
Beltrami	$f_{+/-} = \sqrt{\frac{1+\lambda_{+/-}}{1+\lambda_{-/+}}}$
Tschumperl*é*	$\begin{cases} f_+ = \frac{1}{\sqrt{1+\lambda_+ + \lambda_-}} \\ f_- = \frac{1}{1+\lambda_+ + \lambda_-} \end{cases}$

Table I
DIFFERENT FUNCTIONS $f_{+/-}$ CORRESPONDING TO DIFFERENT ANISOTROPIC DIFFUSION APPROACHES.

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

Figure 1. A study of a letter denoising. Noisy image (a) and the restored versions repectively by Perona-Malik(b); Weickert(c); non-local means(d); Drira(e); Beltrami(f) and tschumperle(g).

	Perona-Malik	Weickert	Non-Local-Means	Drira	Beltrami	Tschumperl*é*
MSE	1294	1167	1742	900	829	1531
PSNR	17.0	17.4	15.7	18.5	18.9	16.2
Cor	0.905	0.916	0.877	0.935	0.940	0.894

Table II
A COMPARATIVE STUDY WITH AN OBJECTIVE QUALITY MEASURE FOR THE NOISY LETTER A, (MSE:17891, PSNR:5.6DB).

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

Figure 2. Zoom on an extract (a) of the "Gazette of Leyde" database before and after restoration with the filters of respectively Drira et al. (b); Perona-Malik(c), Weickert(d); non-local means(e), Beltrami(f) and Tschumperl*é*(g).

In turn, the optical recognition accuracy rates are also proposed as an additional evaluation criteria. For instance, we give the OCR recognition rates before and after restoration. The number of characters which are not recognized before (Nb) and after (Na) restoration is also given. By comparing the recognition rates, we define the best treatment process. A program well-established to this purpose counts the errors before/after restoration. This program takes benefit from the edit distance metric applied on both of the text lines and between characters located on a same text line. The system retains the best match between the ground truth text lines and the OCR results on the original/restored images. We must notice that we include in this comparative study different local/non-local smoothing filters where each filter needs specific parameters for denoising. Looking for a fair comparison, these parameters are correctly selected that the best results must be reached in each experiment.

B. Case of synthetic document images

An example of a synthetic document image is given in the figure 1. The table II reports the different metrics done on the different images after binarization with an

optimal fixed threshold. The correlation cœfficient *cor* is used to test whether a given test image has been altered. In theory, we will have a value higher than 1 if the image is intact and a value lower than 1 if degradations occur. The figure 1 illustrates that the different denoising filters removes the noise from the background but the degree of efficiency differs from one filter to another. Actually, we reveal the limitation of the Perona-Malik filter since its diffusion process stops too early near the characters' layout. For the Weickert filter, the reinforcement of the coherence tends to modify the characters' topology. According to the given results, the Drira et al. filter is well-suited for document images since it takes benefit from the Weickert coherence reinforcement while stopping this process around singularities. Beltrami, Tschumperl$\acute{e}$ and non-local means restored images are nearly similar to that obtained with Drira et al. but they are more blurred. From table II, we notice that Drira et al. and Beltrami filters outperform the other studied filters.

We clearly notice, from this example and from others omitted here for lack of space, that with low noise level, non-local means and the different diffusion filters restore correctly. Nevertheless, the presence of noise in the background generates a different behaviour statement between all the denoising models and especially when adding a large amount of noise. For instance, in the case of excessive noise, PDE-based approaches remain more efficient compared to the non-local means filter which fails in restoring the visual degradation.

C. Case of real document images

For real document images, we tested binary and color images. Binary images are extracted from the book entitled "Le bourgeois gentilhomme" of Molière, printed in 1671. We selected the most degraded pages (P1 and P6) with overall 2190 characters. These pages presents severe degradations due to the digitization process affecting the image quality with random dithering and replacing lines by isolated dots. Figure 4 details the effect of the restoration on an extract of an image. A crop and a zoom of the different processed images are introduced for a better visualisation. Other experiments have been conducted on a database of 106 color pages of "Gazette of Leyde". The total number of characters is 142340. This database suffers from the problem of ink-bleed through.

After preprocessing, we notice a visual quality assessment. For instance, document images corrupted by damaged characters are of improved qualities. For very degraded characters as it is illustrated for the dotted letter "E" extracted from Molière, tensor-driven diffusion approaches and mainly Beltrami and Drira et al. are more effective than non-local means. This could be explained by the lack of redundancy in the form of such characters in a way that non-local means fails to restore correctly the characters' layout. Diffusion approaches based on the local geometry of the image resolve the problem.

For a quantitative evaluation, we have used the commercial OCR system FineReader 8.0 and the open-source OCR engine Tesseract 3.0 [10], respectively referred to ABBY and Google. The results are summarized in the Tables III and IV. In general, the studied preprocessing methods improve the character recognition. Actually, for the first set of binary images with 2190 characters, before processing the two pages, the OCR system completes the conversion with the accuracy rate of 87,7% with the Finereader but with only 46,3% with the Tesseract OCR. The latter accuracy rate is explained by the fact that Tesseract fails when treating dot images. The total number of errors are respectively 269 and 1175. According to the restored image, the first OCR system had approximately a 94,4% recognition rate when detecting damaged letters with Drira et al. filter where as the second OCR system achieves 85,8% with the same filter. These rates are noticeably higher than the recognition rates calculated on the original non-processed image and even on the other processed images. The accuracy of the OCR system was well-improved and we succeed to decrease the failure rate.

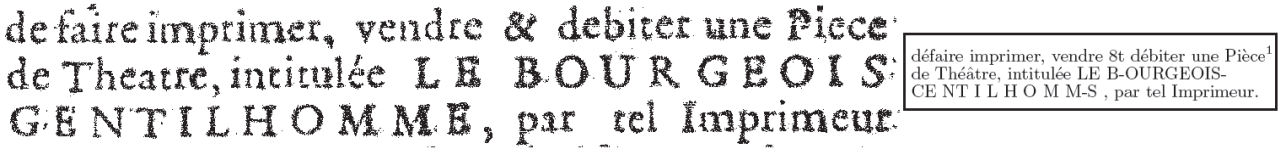
Denoising Methods	Errors After		%Recognition after	
	Tess.	**F.Reader**	**Tess.**	**F.Reader**
Weickert	384	182	82,4%	91,6%
Tschump.	602	211	72,5%	90,3%
Beltrami	646	194	70,5%	91,1%
P.M.	502	182	77,07%	91,6%
NLM	424	169	80,6%	92,2%
Drira	311	122	85,8%	94,4%

Table III
IMPACT OF DIFFERENT DENOISING FILTERS ON THE RECOGNITION RATE: CASE OF TWO PAGES OF A BOOK OF MOLIÈRE.

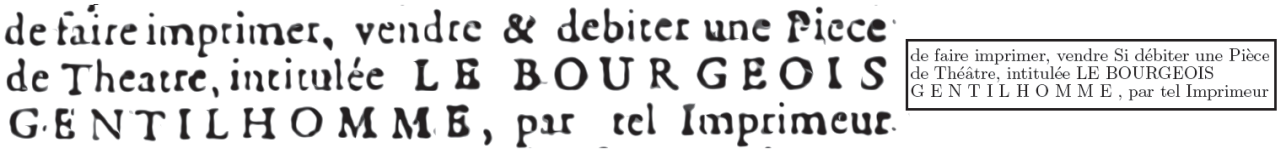
Denoising Methods	Errors After		%Recognition after	
	Tess.	**F.Reader**	**Tess.**	**F.Reader**
Weickert	50680	14199	64,4%	90,02%
Tschump.	40500	9198	71,5%	93,5%
Beltrami	39862	9106	73,7%	93,6%
P.M.	38522	10942	72,0%	92,3%
NLM	34649	10341	75,6%	92,7%
Drira	32245	9282	77,3%	93,4%

Table IV
IMPACT OF DIFFERENT DENOISING FILTERS ON THE RECOGNITION RATE: CASE OF 106 COLOR PAGES OF "GAZETTE OF LEYDE".

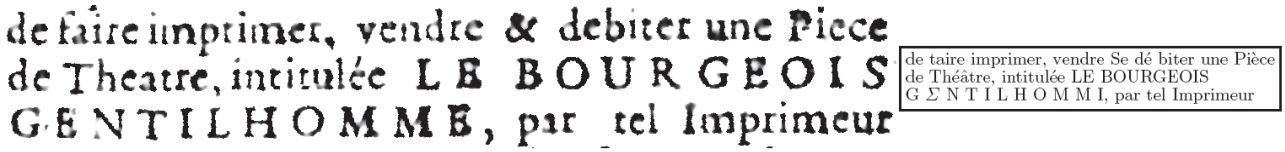
From the second dataset with 142340 characters, we give for lack of space an extract of one letter (Figure 2), the original and the processed versions with the different studied filters. For FineReader, from an OCR rate of 91,6% (11958errors), we reach after restoration the higher accuracy rate with Beltrami filter (93,6%) with 9106 errors. This result is not the same as obtained with Tesseract (with 64744errors) where the recognition rate raises from 54,5% to 77,3% with the Drira et al. filter. These results confirmed



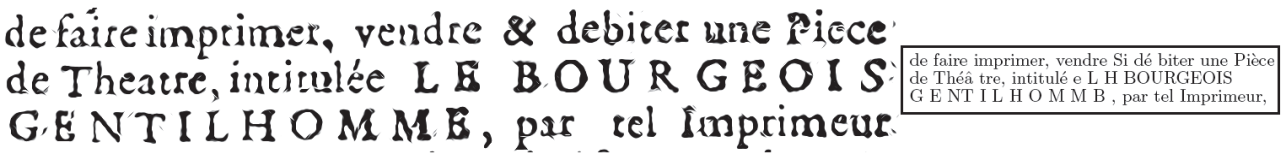
(a) *Original degraded image*



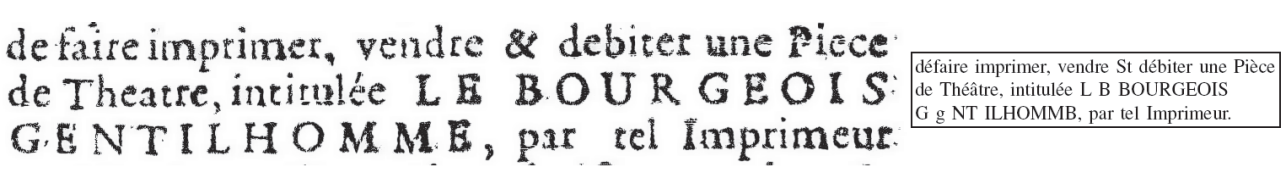
(b) *Image restored with the Drira et al. diffusion filter*



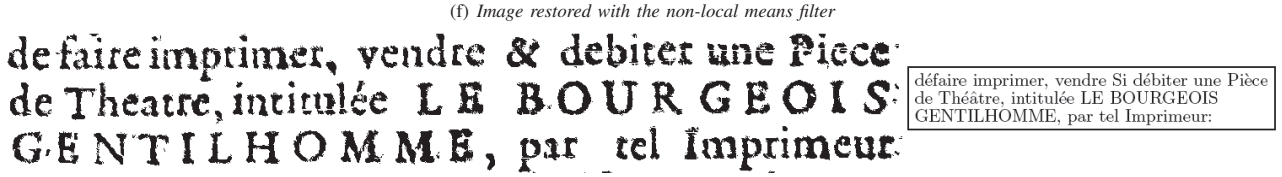
(c) *Image restored with the Perona-Malik-Catté diffusion filter*



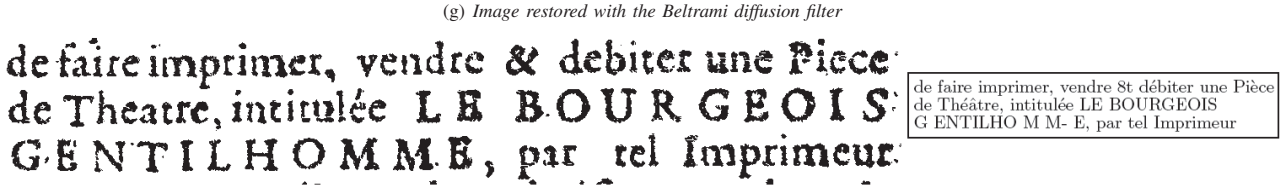
(d) *Image restored with the Weickert diffusion filter*



(f) *Image restored with the non-local means filter*



(g) *Image restored with the Beltrami diffusion filter*



(h) *Image restored with the Tschumperlé diffusion filter*

Figure 3. Details of the image N0070212-TIFF-1-20 (http://gallica2.bnf.fr/ark:/12148/bpt6k70212z) before and after restoration. Small extracts of the ABBY fineReader 8.0 OCR results obtained on the different document images are also given.

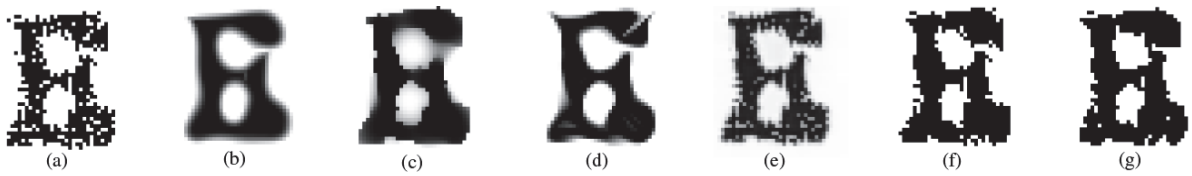


Figure 4. A Zoom on different extracts of document images (Figure 4) : the original image (a) and the results provided respectively by the processes of Drira (b), Perona-Malik(c), Weickert(d), non-local means(e), Beltrami(f) and Tschumperlé(g).

the superiority of local smoothing filters over the non-local means filter which has also very interesting properties to exploit. We conclude that restoring by either local or non-local smoothing approaches leads to a real OCR accuracy improvement. Nevertheless, this improvement is directly related to the importance of the noise level in the textual document.

IV. Conclusions and future works

The first concern of this work is to test the efficiency of local/non-local smoothing approaches in document image denoising. A comparative study is well-established between tensor-driven diffusion approaches and non-local means filters at this aim. Both kinds of filters have interesting properties but we mainly noticed the superiority of diffusion approaches in visual quality assessment and even in improving the accuracy rates in treating very degraded document images. The main idea of the non-local means filters is very interesting as it takes benefit from the self-similarity and the redundancy of the image. Nevertheless, this filter as it is defined could not achieve additional improvement to the exiting state-of-the-art for document image restoration. Further research will investigate the adaptation of the non-local means to textual documents.

References

[1] A. Buades, B. Coll and J. M.Morel, *A review of image denoising algorithms, with a new one*. Multiscale Modeling and Simulation, 4(2):490-530, 2005.

[2] F. Drira *Towards restoring historic documents degraded over time*. In Second IEEE International Conference on Document Image Analysis for Libraries (DIAL'2006), Lyon, France. pp. 350-357. ISBN 0-7695-2531-4.,2006.

[3] L. Likforman-Sulem, J. Darbon and E. Barney Smith, *Pre-processing of degraded printed documents by Non-Local means and Total Variation*, In Proc. of ICDAR, pp.758-762, Barcelona, July 2009.

[4] L. Alvarez, F. Guichard, P. L. Lions, J. M. Morel,*Axioms and Fundamental Equations of Images Processing*, Archive for Rational Mechanics and Analysis, 123 (3), pp. 199-257 (1993).

[5] J. Weickert, *a review of nonlinear diffusion filtering*, In Scale-Space Theory in Computer Vision, volume 1252 of Lecture Notes in Computer Science, Utrecht, the Netherlands, 1997.

[6] P. Perona, J. Malik, *Scale-Space and Edge Detection Using Anisotropic Diffusion*, In IEEE Trans. Pattern. Analysis and Machine Intelligence, pp. 629-639, 1990.

[7] F. Catté, J. M. Morel, P. L. Lions, T. Coll, *Image Selective smoothing and edge detection by nonlinear diffusion*, In SIAM J. Numer. Anal., 29:182-193, 1992.

[8] D. Tschumperlé, R. Deriche, *Vector-Valued image regularisation with PDE's: A common framework for different applications*. IEEE transactions on Pattern Analysis and Machine Intelligence, vol.27, No 4, 2005.

[9] F Drira, F Lebourgeois, H. Emptoz, *Document images restoration by a new tensor based diffusion process: Application to the recognition of old printed documents*, In 10th International Conference on Document Analysis and Recognition (ICDAR), IEEE ed. Spain. pp.321-325, , Barcelona, July 2009.

[10] R. Smith, *An overview of the Tesseract OCR engine*,In Proc. of ICDAR, pp.629-633,Brasil, 2007.