# An Efficient New PDE-based Characters Reconstruction After Graphics Removal

Louisa Kessi, Frank Le Bourgeois, Christophe Garcia

# An Efficient New PDE-based Characters Reconstruction After Graphics Removal

Louisa Kessi
INSA-Lyon, LIRIS,
UMR5205, F-69621, France
Louisa.kessi@liris.cnrs.fr

Frank Lebourgeois
INSA-Lyon, LIRIS,
UMR5205,F-69621, France
franck.lebourgeois@liris.cnrs.fr

Christophe Garcia
INSA-Lyon, LIRIS,
UMR5205,F-69621, France
christophe.garcia@liris.cnrs.fr

***Abstract*— The separation between texts and graphics when they are overlapped is a challenging problem for digitization companies. In a previous work [1], we presented the first unsupervised fully automatic segmentation system adapted for colour business document with significant colour complexity and dithered background. The system achieves several operations to segment automatically colour images, separate text from noise and graphics and provides colour information about text colour. After split overlapped characters and separates characters from graphics, characters are broken. The OCR system becomes unable to recognize successfully broken characters and its efficiency is thus seriously affected. This paper presents the first Character Reconstruction System through a new PDE (Partial Differential Equation)-based approach. Our approach takes benefit of the combination of the anisotropic morphology proposed by Breuß and the Weickert Coherence enhancing shock filter diffusion. We introduce and present a continuous anisotropic morphology method driven by the main direction of the first order tensors applied in the neighborhood of the missing part left by the separation between text and graphics. It reconstructs the missing part even when the left area is larger than the strokes width. The coherency of the orientation of the tensors around missing parts overcomes the problem of image noises. The application of the ABBY FineReader OCR engine proves an important reduction in OCR errors. Our experiments show that our proposition compared to the existing state of the art requires no training steps and outperforms both of anisotropic morphology and the Weickert Coherence enhancing shock filter diffusion applied separately.**

***Keywords— PDE-based approach; continuous anisotropic morphology; characters reconstruction; OCR accuracy.***

## I. INTRODUCTION

The Optical Character Recognition (OCR) systems accuracy depends considerably upon the quality of the processed images that are often affected by several kinds of damages and degradations. Our previous work [1] presents the first robust data-driven and pixel-based approach suited for color business document with significant color complexity and dithered background that does not need a priori information, training or manual assistance.

Each step of the system is independent from previous steps as the parameters are optimal for a very large range of documents. The system achieves several operations to segment automatically color images, processes inverted and non-inverted text automatically using color morphology, even in cases where there are overlaps between the two, separate text from noise and graphics, and provides color information about text color. After the separation between characters from graphics at pixel level described in [1], characters are broken and cannot be recognized correctly by the OCR (figure 1). Therefore, the OCR system becomes unable to recognize successfully broken characters.

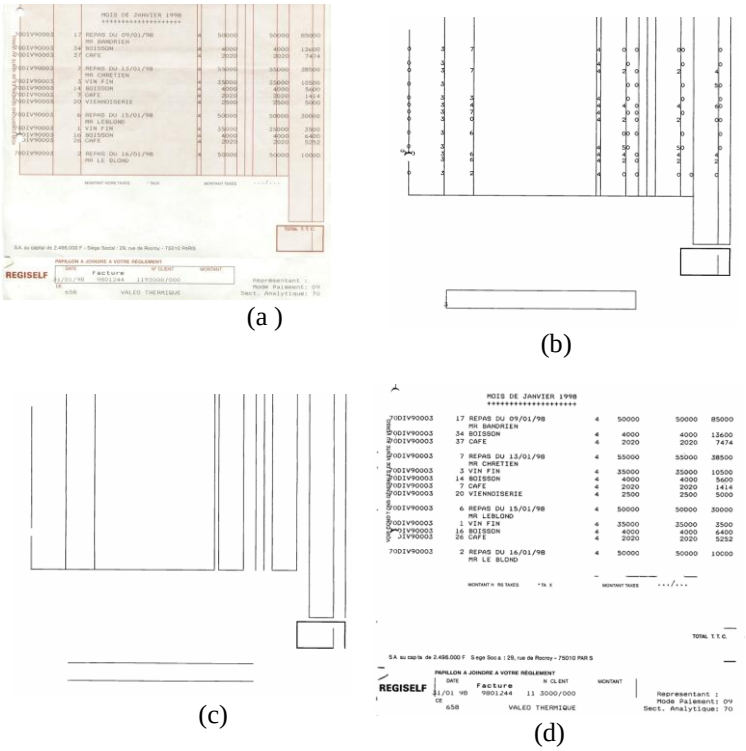



Figure 1: Separation of characters connected to graphics [1]. (a) Original Image, (b) overlapped characters touching graphic are placed in the graphical plan, (c) Graphical plan after separating characters, (d) Textual plan with the characters retrieved from the graphical plan .

To the best of our knowledge, the state of the art of the main works dealing with textual document images quality

assessment and character restoration is very limited [2,3,4,5]. Most of the developed algorithms focused their performances on efficient character segmentation or character recognition which limits their application. Thus, the reconstruction of broken characters is an ill-posed yet challenging problem due to its formulation as a continuous diffusion problem. Hence, the main topic of this work is textual document images, quality assessment and restoration, more precisely, we investigate on the reconstruction of broken characters.

Unlike to the all-discrete methods that dominated image processing in the recent past, continuous models based on partial differential equation (PDEs) appear in a large variety of image processing and computer vision areas and they have been mainly devoted in image enhancement. In this paper we are interested in the PDE-based approaches that using the anisotropic morphology diffusion driven by the direction of the first order tensors matrices. This last kind of diffusion is well adapted to deal successfully with our problem since it is suitable for the processing of oriented features. This paper addresses the most challenging problem of continuous anisotropic morphology, we choose to combine both the Anisotropic Morphology from Breuß [10] and the model of the Weickert Coherence enhancing shock filter diffusion [11] for their proved performance in the coherence-enhancing properties. We use this theory to develop the first character reconstruction system based on anisotropic morphology. This paper is organized as follows: Section 2 presents a brief comparative study of several possible solutions to reconstruct the shape of character. Section 3 describes and details the proposed new PDE approach. Some experiments done on degraded several hundred degraded characters from various fonts overlapped by various lines of different thicknesses. Document images indicate in section 4 that our proposition 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. In Section 5 conclusions are drawn.

## II. RELATED WORK

Based on a study done on image restoration techniques for restored degraded characters, to reconstruct the shape of the characters, we focus our study on continuous approaches which can process the discrete image at the sub-pixel level by Partial Derivatives Equations (PDE) in the scale-space to retrieve the general orientation of the strokes and reconstruct by diffusion or by continuous morphology. Three alternatives have been presented:

### *A. Image inpainting:*

Image inpainting consists to diffuse the color information around the missing parts in the direction of the isophotes (lines of the same intensities orthogonal to the gradient vector). The work proposed by [6] based on diffusion in the orthogonal orientation of the gradient cannot reconstruct curved strokes or large missing parts when it is greater than the width of the strokes. These problems have been corrected with the Curvature-driven Diffusion (CDD) proposed by [7] describe by (1). When the missing part is too large, even the CDD cannot reconstruct correctly the shapes of characters. The curvature Kappa equals to the divergence of the normalized gradient. The parameters *a* and *b* are the main parameters of the PDE. For *a*=1 and *b*=0, we get the Total Variation model and for *a*=0 and *b*=1 we retrieve the first version of the CDD inpainting model entirely driven by the curvature. This model also generalises the Beltramio inpainting model [6].

$$I_t = div\left(\left(a + b\kappa^2\right)\frac{\nabla I}{\|\nabla I\|}\right) \quad \kappa = div\left(\frac{\nabla I}{\|\nabla I\|}\right) \qquad (1)$$

### *B. Anisotropic based diffusion approaches:*

The Coherence-enhancing diffusion of Weickert [8] can be used to reinforce the continuities of strokes and fill the missing parts (2). It has been used by [9] for contour completion. If *D* equals the identity matrix *I*, (2) becomes the classical heat equation.

$$I_t = div\left(D\nabla I\right) \qquad (2)$$

Then *D* is the 2x2 diffusivity matrix which drives the heat equation. Weickert uses the tensors of the first order derivatives to build the diffusion matrix *D*. The model of Weickert reinforces the coherence of strokes and can be used to reconstruct characters if we diffuse into the empty parts left by the graphical deleted elements. But this model is a diffusion and not a morphology, the result on binary image provides blurred images.We search for an anisotropic morphology which is better suited to the reconstruction of binary patterns.

### *C. Anisotropic morphology based diffusion*:

To process fine strokes at the sub-pixel level, we studied continuous morphology approaches based on PDE (3) which dilates or erodes the image in function of the sign of the coefficient of diffusivity *d*. The PDE proposed in (3) achieves the basic operation of dilation or erosion by adding or removing a fraction of the contours of the objects expressed by the magnitude of the gradients $\|\nabla I\|$ . The isotropic morphology described in the equation (3) does not dilate or erode in specific directions. The anisotropic morphology described in the next section will provide the solution.

$$I_t = \pm d \times \|\nabla I\| \qquad (3)$$

Breuß et al [10] proposed to dilate the extremities of the missing parts by an anisotropic continuous morphology in the direction of the strokes (4).The 2x2 matrix *D* describes the orientation of the dilation or erosion. *D* is calculated from the main orientation Θ+ of the structure tensor *T* like in the Weickert model (4).

$$I_t = \pm d \times \|D\nabla I\| \quad (4)$$

$d$ is the coefficient of diffusivity which must be in the direction of the strokes. We can use the sign of the main tensor orientation defined in the work of Weickert [11]. The main idea of a shock filter consists in using a dilation process along the direction of the isophotes. The Weickert coherence enhancing shock filter is directed by the sign of the second derivatives $I_{\Theta_+\Theta_+}$ from the Hessian rotated in the direction $\Theta_+$ of the main orientation of the tensors. This filter is given as follows:

$$I_t = -sign(I_{\Theta_+\Theta_+})\|\nabla I\| \quad (5)$$

$I_{\Theta_+\Theta_+}$ is the second derivative of the image I in the direction Θ+ of of the tensors T .

This morphology both enhances the coherency of the strokes and sharpens the image. In reality it is a morphology of dilation of strokes in the direction $\Theta_+$ of tensors.

## III. Proposition of a new PDE-based Characters Reconstruction process

To reconstruct character shapes, we propose the PDE equation (6) which combines (4) and (5) in the same scheme. We propose a continuous anisotropic morphology method driven by first order tensors applied in the neighborhood of the missing parts M.

$$I_t = -sign(I_{\Theta_+\Theta_+})\|D\nabla I\|_{(x,\,y)\in M} \quad (6)$$

We first detect all the missing parts to reconstruct by finding the characters which overlap graphical lines and then compute a binary image M of all zones to reconstruct, enlarged by a discrete dilation of radius r=4 in order to get enough information about the structure tensors in the neighborhood of the missing parts. The tensors $T$ is computed for pixels inside M by (7).

$$T_\sigma^\rho = G_\rho \otimes (\nabla(G_\sigma \otimes I)\nabla(G_\sigma \otimes I)^T) \quad (7)$$

$$T_\sigma^\rho = G_\rho \otimes \begin{pmatrix} (I_x^\sigma)^2 & I_x^\sigma I_y^\sigma \\ I_x^\sigma I_y^\sigma & (I_y^\sigma)^2 \end{pmatrix} \quad (8)$$

Here $G_\rho$ and $G_\sigma$ are Gaussian convolution Kernels. The pre-smoothed original image by a Gaussian kernel with variance σ allows a correct calculation of the gradients. The outer product of the gradients vectors generates a 2x2 symmetric and diagonalizable matrix (8).

The convolution of the tensors field by a Gaussian kernel with standard deviation ρ reinforces the coherency of the orientation of the tensors around missing parts and overcomes the problem of image noises. To compute the gradient vectors and the second derivatives, we use the 5x5 stencil described in [12] which is very stable numerically. The second derivative in the direction of the strokes $I_{\Theta_+\Theta_+}$ is calculated by the rotation of the Hessian matrix $H$ in the direction of the eigenvector $\Theta_+$ associated to the largest eigenvalue $\lambda_+$ of the tensor $T_\sigma^\rho$ [11].

$$I_{\Theta_+\Theta_+} = \Theta_+^t H \Theta_+ = \begin{pmatrix} u & v \end{pmatrix} \begin{pmatrix} Ixx & Ixy \\ Iyx & Iyy \end{pmatrix} \begin{pmatrix} u \\ v \end{pmatrix} \quad (9)$$

The matrix $D$ is computed with the Weickert model described in [8]. The integration scale $\rho$ is generally set to ρ=3×σ. In our case, the integration scale $\rho$ is important because it defines the maximal size of the missing part to reconstruct. For too large values of ρ, curved strokes cannot be reconstructed correctly. If $\rho$ is too small, the completion is not achieved, because the main orientations of strokes are not guessed in the missing part. During experiments, we set $\sigma$=1.0 and $\rho$=3.0 for 300 dpi images and text height of 24 pixels in average. Figure 2 shows the differences between an anisotropic diffusion (2) and an anisotropic morphological dilation (6). It illustrates the difference between an anisotropic diffusion and an anisotropic morphology.

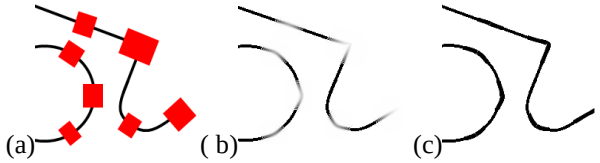
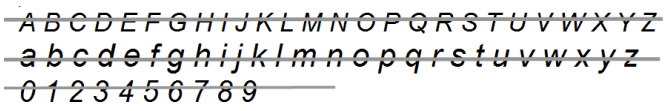



Figure 2: (a) Results of the reconstruction of missing part materialized in red by ,(b) anisotropic diffusion (2) , (c) anisotropic morphology (6).

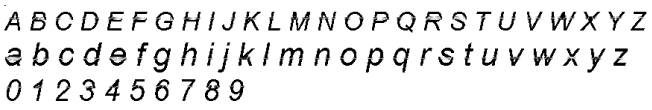
## III. Experimental result and discussion

### *A. Visual quality improvement:*

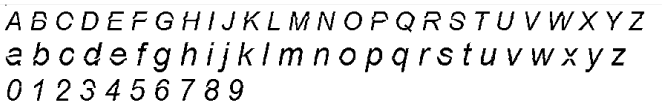
To illustrate the performance of the proposed continuous morphology diffusion method, we give some experimental results. We have carried out tests on a dataset of a dozen damaged characters with a visual evaluation. This permits us to check if this method was relevant. We clearly notice good visual improvement results of the damaged original image. Other tests have been done in order to compare the performance of our diffusion process with its basic approaches applied separately. 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 synthetic images generated after the application of the proposed filter and even after the application of both of the Anisotropic Continuous-Scale Morphology and the Weickert Coherence enhancing shock filter diffusion. Better results has achieved with the new tensor diffusion filter (figure 3).




Truncated characters by horizontal lines




reconstruction by using the anisotropic dilation alone (4)




Reconstruction by using our scheme (6)

Figure 3: Comparison between the anisotropic dilation alone (4) and the anisotropic dilation guided by the tensors of Weickert (6).

*B.OCR accuracy improvement:*

In the case of natural image quality after a processing step with a PDE process, the evaluation can be done with measures such as SNR (Signal to Noise Ratio) or MSE (Mean Square Error), but in the case for damaged textual document images, these kinds of measures do not take into account the importance of information along characters. Thus, the evaluation of the quality of the reconstruction of the shape of damaged characters and the restoration of textual document images is not an easy task. Since the ultimate aim of our proposed PDE-algorithm is to reconstruct damaged characters after splitting them from a graphic line which is detailed in previous work [1], we propose to evaluate the result through the study of the optical character recognition accuracy rate.

Therefore, this section focuses on studying OCR system response before and after reconstruct damaged characters with our PDE-based approaches. We have tested the reconstruction of several hundred characters from various fonts overlapped by various lines of different thicknesses randomly placed. The reconstruction depends on the information in the neighborhood of the missing part. If the lines do not delete important information like intersections, junctions or extremities, the reconstruction performs well as expected, especially on curved strokes. The performance depends on the position of the missing parts on the characters. On 680 characters randomly broken by lines, the OCR recognizes only 36 characters (5,3% of recognition). After reconstruction by our anisotropic morphology, the OCR recognizes 674 characters (99,1% of recognition). Figure 5 and 6 show a sample of the results of the OCR Abby Finereader 11. As expected, the OCR doesn't recognize any broken characters. Despite the wrong reconstruction of some characters like B,G,a,d,x, the OCR recognizes almost all reconstructed characters.

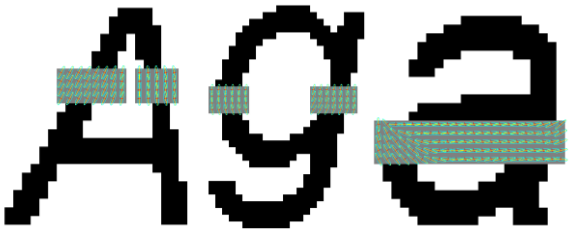



Figure 4 shows the tensors on well reconstructed characters. (A,g) and wrong reconstructed character (a)

Figure 4 displays the tensor fields in the missing parts M. For the character 'a' not correctly reconstructed, we can see that the main flow of the tensors field follows the horizontal stroke and not the two vertical strokes. It is due to the fact that the graphic line deletes important information, like the vertical junctions, necessary for a correct estimation of the tensors orientation like for 'a' and 'g'.

## V. CONCLUSION

We have proven in this work the efficiency of a continuous anisotropic morphology in the quality of degraded textual document images. We have proposed a new PDE-based diffusion. It is based on the combination of both the Anisotropic Continuous-Scale Morphology and the Weickert Coherence enhancing shock filter diffusion for their proved performance in the preservation of the singularity and coherence-enhancing properties. We use this theory to develop the first character reconstruction system. The quality of the restored damaged characters is visually improved. The application of the ABBY FineReader engine 11.0 OCR proves an important reduction in OCR errors. Our experiments show that our proposition outperforms both of Anisotropic Continuous-Scale Morphology and the Weickert Coherence enhancing shock filter diffusion applied separately.

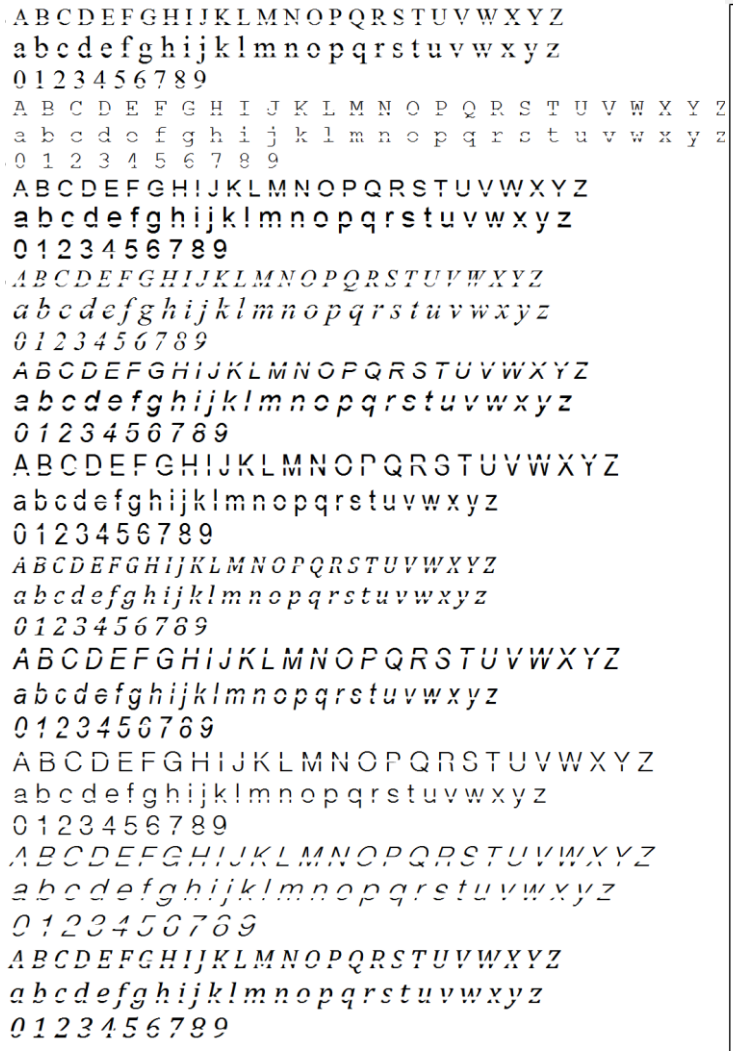


Figure 5a : Images of the broken characters by straight lines before the reconstruction used in the experiments.

Figure 5b : 5,3% of OCR recognition on image fig 2a (errors are in red color)

Figure 6a : Reconstructed images by using (6)

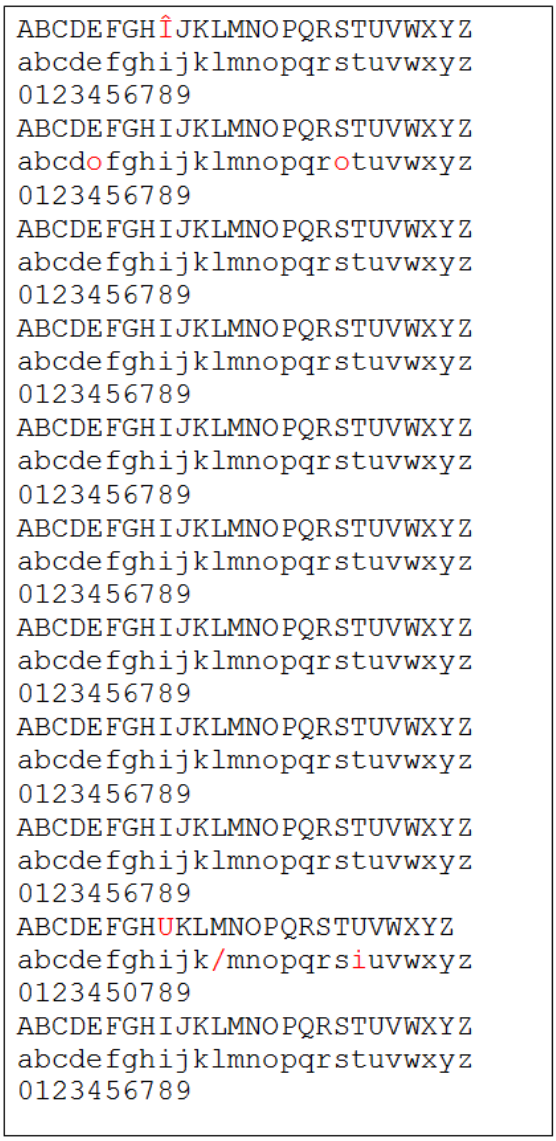


Figure 6b : 99,1% of OCR recognition on image fig 3a after character reconstruction. (errors are in red color)

## REFERENCES


[1] L. Kessi, F. Le Bourgeois, C. Garcia, J. Duong "AColDPS: Robust and Unsupervised Automatic Color Document Processing System", accepted for VISAPP'15.

[2] F. Drira, Towards restoring historic documents degraded over time. In IEEE DIAL.06, Lyon, pp. 350-357, 2006.

[3] John D. Hobby Tin Kam Ho, Enhancing Degraded Document Images via Bitmap Clustering and Averaging, ICDAR.97, pp. 394--400, 1997.

[4] John D. Hobby, Henry S. Baird, Qigong Zheng and Tapas Kanungo, Morphological degradation models and their use in document image restoration. ICIP.01, Greece, pp. 193-196, 2001.

[5] J. D. Hobby, H.S. Baird, Degraded Character Image Restoration, In Proc. of Document Analysis and Information Retrieval, Las Vegas, 1996.

[6] M. Bertalmio, G. Sapiro, V. Caselles, C. Ballester. Image inpainting, Proceedings of the SIGGRAPH, 2000.

[7] T.Chan, J.Shen, Mathematical models for local non texture inpaintings, SIAM J.Appl. Math. 62 (3) pp. 1019-1043, 2002.

[8] J. Weickert. Anisotropic Diffusion in Image Processing, PhD, University of Kaiserslautern, Germany, 1996.

[9] D. Gil, P. Radeva, F. Vilariño Anisotropic contour completion, Proc. IEEE ICIP03, vol 1, pp 869-872.

[10] M. Breuß, B. Burgeth, J. Weickert, Anisotropic Continuous-Scale Morphology, LNCS Vol. 4478, 2007, pp 515-522.

[11] J. Weickert. Coherence enhancing shock filters. LNCS, Vol. 2781, Springer, Proceedings of DAGM 2003, pp. 1-8, Germany, 2003.

[12] J. Weickert, H. Scharr, A Scheme for Coherence-Enhancing Diffusion Filtering with Optimized Rotation Invariance, JVCIR, Vol. 13, Issues 1–2, 2002, pp. 103–118.

[13] J. Weickert. Anisotropic Diffusion in Image Processing, PhD, University of Kaiserslautern, Germany, 1996.