Kavitha Balakrishnan, Kavitha Sunil, Sreedhanya A.V, K.P Soman / International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622 www.ijera.com
Vol. 2, Issue4, July-August 2012, pp.1529-1534

Effect Of Pre-Processing On Historical Sanskrit Text Documents

Kavitha Balakrishnan, Kavitha Sunil, Sreedhanya A.V, K.P Soman
Centre for Excellence in Computational Engineering and Networking

ABSTRACT

In this paper, the effect of pre-processing on binarization is explored. Here, pre-processing and binarization operations are performed on a historical Sanskrit text document. After scanning, pre-processing operations are applied on the image to remove noise. Pre-processing techniques play an important role in binarization. Newly developed pre-processing techniques are Non Local means and Total Variation methods. Total Variation methods are motivated by the developments in compressive sensing methods like l_1 optimization. Binarization is used as a pre-processor before OCR, because most OCR packages work only on black and white images.

Keywords - Pre-processing filters, Total variation methods, NL means, Binarization, Otsu Binarization

1. INTRODUCTION

There are a number of factors that affect the accuracy of a text document recognized through OCR. Scanned documents often contain noise that arises due to printer, scanner, print quality, age of the document. In the case of historical documents, the quality is usually very low, and the images suffer high degradation. The degradation on the historical document images is mainly due to fading of ink, scanning, paper aging and bleed-through.

The importance of the preprocessing stage of a character recognition system lies in its ability to remedy some of the problems that may occur due to factors presented above. Thus, the use of preprocessing techniques may enhance the document image and prepare it for the next stage in the character recognition system.

The paper is organized as follows. Section 2 shows the steps involved in preprocessing. Section 3 deals with noise removal techniques. It comprises of conventional and new filtering methods. Otsu Binarization algorithm is discussed in Section 4. Section 5 deals with the experimental results and in Section 6 we conclude the discussion.

2. STEPS INVOLVED IN PREPROCESSING

The steps involved in preprocessing is given below Scanning and image digitization

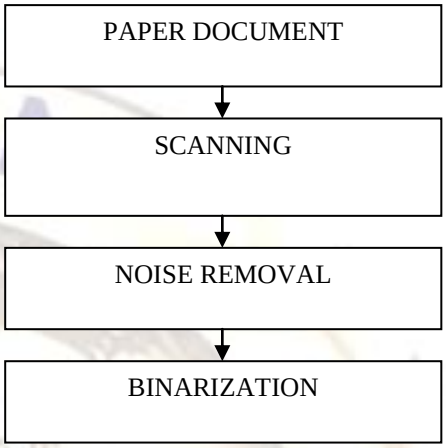

Fig. 1 Overview of basic steps in preprocessing

Before going into the OCR process, one must scan the text document. It is important to do a good quality scanning since if the quality is poor then it will be hard for the OCR software to read the text and to make correct interpretation. The scanned image is then stored in jpeg/bmp format. In order to improve the quality of the image we then have to go for noise removal.

Noise Removal

Scanning of text documents itself can introduce some amount of noise. In order to make it suitable for further processing, a scanned document image is to be freed from any existing noise. This can be achieved by image enhancement. Image enhancement mainly involves noise removal which is done by filtering. Filtering is a neighbourhood operation, in which the value of any given pixel in the output image is determined by applying some algorithm to the values of the pixels in the neighbourhood of the corresponding input pixel. Various methods are applied to reduce noise. The most important reason to reduce noise is to obtain easy way of recognition of documents.

Binarization

Image binarization converts a gray-scale document image into a black and white (0s&1s) format. The choice of binarization algorithm depends on the quality of the image. Hence the binarization algorithms that work best on one image may not be best for another.

3. NOISE REMOVAL METHODS

3.1. Conventional methods

3.1.1. Mean filter

Mean filter is one of the simplest linear filters. It is a sliding window spatial filter, which replaces the center pixel of the window by the average of all the values in the window. The size of the window determines the image contrast. As we increase the size of the window, the image becomes more and more blurred.

3.1.2. Median filter

It is a non linear filter. It is a sliding window spatial filter, which replaces the center pixel of the window by the median of all the values in the window. It is widely used because it will remove the noise while preserving the edges present in the image.

3.1.3. Wiener filter

Wiener filter is an adaptive linear filter, which takes local variance of the image into account. When the variance in an image is large, the Wiener filter results in light local smoothing, while when the variance is small, it gives an improved local smoothing.

Figure 2 shows the effect of filters on text documents. Mean filters smoothen the edges of the character and background. The median filter works better than the mean filter and preserves useful details in the image. Wiener filter produces a fair amount of edge blurring.

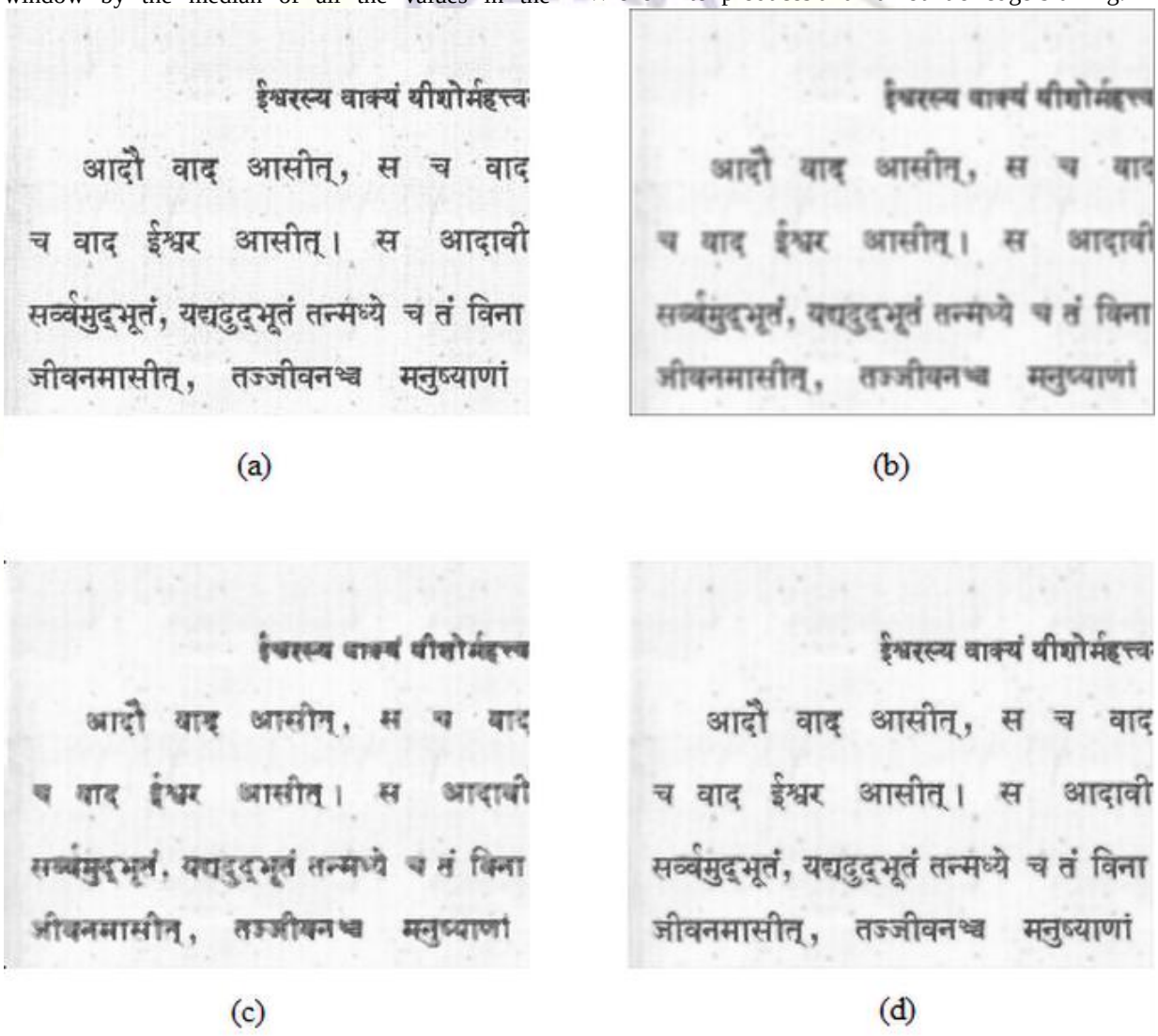

Fig. 2. Samples showing the effect of preprocessing filter (a) The original text image, (b) Mean filter (c) Median filter (d) Wiener filter

3.2 Non Local Means

The conventional filtering methods remove fine details present in the image along with noise. Most denoising algorithms make two assumptions about the noisy image. The first assumption is that the noise contained in the image is white noise. The second assumption is the true image is smooth i.e. it contains only low frequencies. But some images contain fine details and structures which have high frequencies. Filtering removes these high frequency details in addition to the high frequency noise, and these methods do nothing to remove low frequency noise present in the image. These assumptions can

cause blurring and loss of detail in the resulting denoised images.

The Non-local means assumes that the image contains an extensive amount of redundancy and exploits these redundancies to remove the noise present in the image. Adjacent pixels tend to have similar neighborhoods, but non-adjacent pixels can also have similar neighborhoods' as shown in the figure below. The self-similarity assumption is exploited to denoise an image. Pixels with similar neighborhoods can be used to determine the denoised value of a pixel.

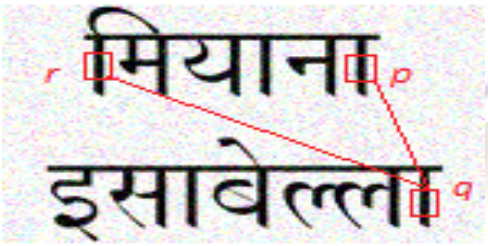

Fig. 3. Example for self-similarity is shown. Non adjacent pixels p, q and r have similar neighborhood in the image.

The non-local means replaces a pixel by the weighted average of other neighborhoods in the image. This method is the best possible denoising method for natural images. First we make a list of all similar neighborhoods in the image. Neighborhood of each pixel is then linearized to form a row in a matrix and L2 norm is computed between each row. Let $N_{x,y}$ and $N_{r,s}$ denotes the neighborhood centered over pixel (x, y) and (r, s) respectively. Let the window size be MxM, where M is the odd. Similarity between the two neighborhoods can be found using L2-norm: $\|N_{x,y} - N_{r,s}\|_2$

This norm then defines a weight to be used in our weighted average

$$w(N_{x,y}, N_{x,y}) = \frac{\exp(-(N_{x,y} - N_{r,s}))^2}{h^2}$$

where h is a parameter that needs to be fine-tuned. Similar neighborhoods give $w = 1$. If the two neighborhoods are very different, $w \approx 0$.

Then each pixel in our new image $g(x, y)$ is a weighted average of the pixels in $f(x, y)$, weighted by how similar the neighborhoods are

$$g(x,y) = \frac{\sum_r \sum_s f(r,s).w\left(N_{x,y}, N_{r,s}\right)}{\sum_r \sum_s w\left(N_{x,y}, N_{r,s}\right)}$$

When historical document image with noise is subjected to NL means algorithm, we get result as shown in figure. Smaller neighborhoods remove noise better.

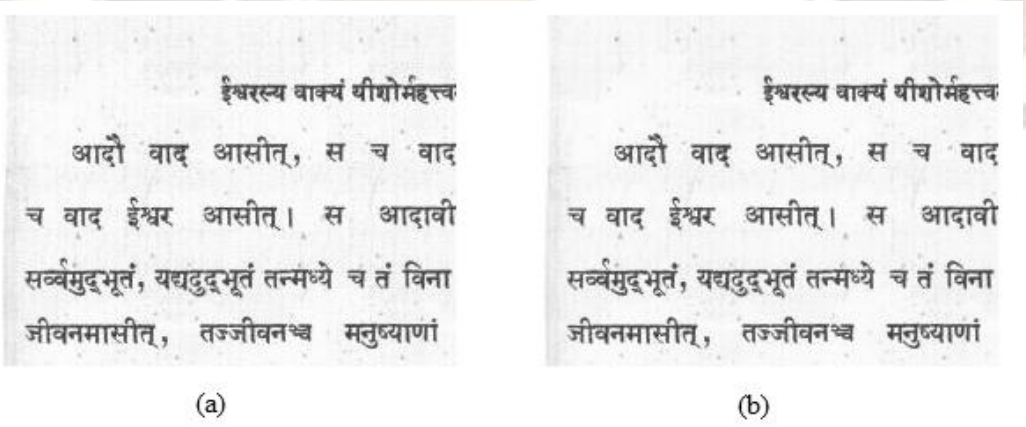

Fig. 4. Effect of NL means algorithm is shown (a) original image, (b) image denoised using NL means with window size=3.

3.3 Total Variation

3.3.1 Tikhnov model

The Total Variation approach is used when characters in the text document are highly degraded. TV Denoising reduces total variation of the image. It filters out noise while preserving the edges. The resulting filtered image should have the same statistical properties as the original image, along with sharp edges and low noise.

Finding the denoised text image can be mathematically described as an optimization problem:

$$u^* = \max_u \{p(u / f)\}$$

where $p(u/f)$ is the posterior probability of a hypothesis *u* (describing our best solution) might be true given by the observation *f*.

The conditional probability $p(u/f)$ can be written as

$$p(u/f) = \frac{p(f/u)p(u)}{p(f)}$$

where p(u) is the prior probability of u and $p(f/u)$ is the conditional describing how well the observed data *f* can be explained by the solution *u*.

The probability of the low noise image with respect to the image f (that is defining our inverse problem):

$$p(u/f) = \frac{p(f,u)p(u)}{p(f)} = \prod_{(x,y)\in D} \frac{1}{4\mu\pi\nu} e^{\frac{-(f(x,y)-u(x,y))^2}{2\mu^2} + \frac{|\nabla u(x,y)|^2}{2\nu^2}}$$

Maximizing this probability is equivalent to minimizing the negative term in the exponent on the set of pixels *D*, that is the overall integral of

$$\frac{(f(x,y)-u(x,y))^2}{2\mu^2} + \frac{|\nabla u(x,y)|^2}{2\nu^2}$$

This leads us to the following variational formulation of our image restoration problem

$$\min_u \{E(u)\} = \frac{1}{2}\int_\Omega |\nabla u(x,y)|^2 d\Omega + \frac{1}{2\lambda}\int_\Omega (f-u)^2 d\Omega$$

This is Tikhonov model. The first term is the regularization term which is derived from prior and the second term as the data fidelity term. The data fidelity measures how far the current solution u is from the observed image f. The parameter λ is a non-negative coefficient that governs the balance between the data fidelity and the regularization term. A large value for λ will produce an image with few details, removing small features, while a small value will yield an image same as u. In Tikhnov model, since we use image prior as a set of smooth images, we obtain blurred images as output.

3.3.2 ROF Model

The ROF model is similar to the Tikhonov model but the regularization term has been changed to the TV norm instead of the quadratic norm. In its original formulation, the ROF model is defined as the constrained optimization problem

$$\min_u \left\{\int_\Omega |\nabla u| d\Omega\right\} s.t \int_\Omega (u-f)^2 d\Omega = \sigma^2$$

where f is the observed image function which is assumed to be corrupted by Gaussian noise of variance σ^2 and u is the unknown denoised image.

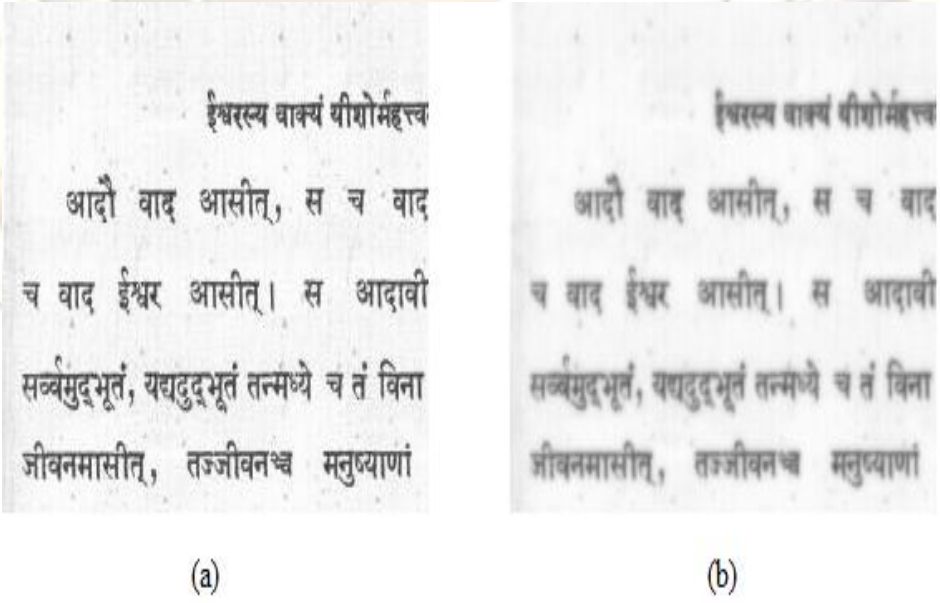

Fig. 5. Result of denoising using Tikhnov model (a) The original image, (b) TV using λ=1.

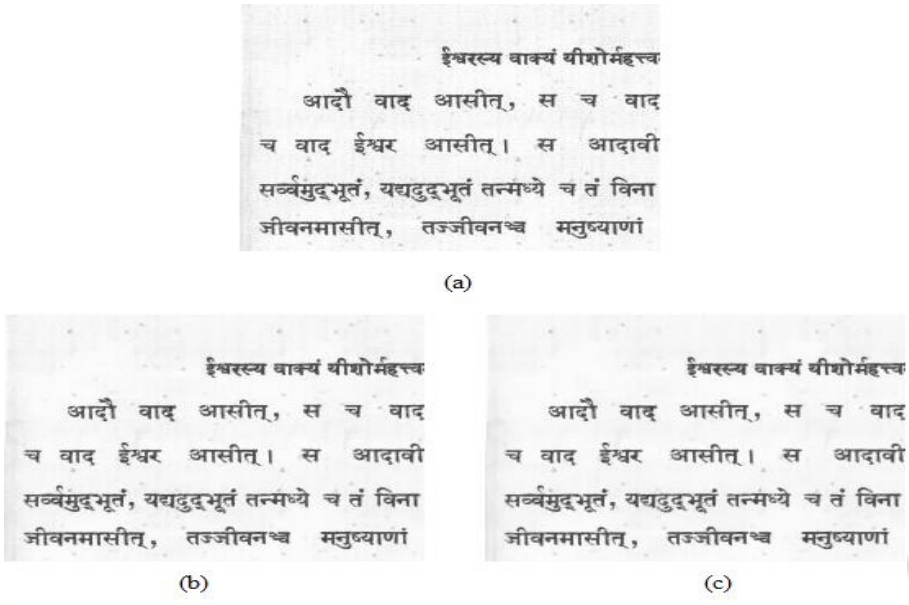

Figure 6. Result of denoising using ROF model (a) The original image, (b) TV using λ=3, (c) TV using λ=5.

The original non-convex ROF model can be turned into a convex problem by replacing the equality constraint

$\int_\Omega (u-f)^2 d\Omega = \sigma^2$ by the inequality constraint $\int_\Omega (u-f)^2 d\Omega \le \sigma^2$

which in turn can further be transformed to the unconstrained (or Lagrangian) model

$$\min_u \left\{ \int_\Omega |\nabla u| \, d\Omega + \frac{1}{2\lambda} \int_\Omega (u-f)^2 d\Omega \right\}.$$

where λ is a Lagrange multiplier.

4. BINARIZATION

Image binarization converts a gray-scale document image into a black and white (0s&1s) format. In binarization, we choose a threshold value to classify the pixels as black and white. If the pixel value is greater than the threshold value, then it is classified as white and if the pixel value is less than the whole document image, but these techniques are not suitable for degraded images. In local binarization technique, the local threshold can be calculated by using different information of the document images, such as the mean and standard variation of pixel values within a local window.

Figure 7. Result of binarization using Otsu algorithm on denoised image using (a) mean filter, (b) median, (c) wiener, (d) Tikhnov, (e) TV ROF using λ=3, (f) TV ROF using λ=5

the threshold value, then it is classified as black. Binarization techniques can be either global or local. In global binarization we choose a single threshold for

4.1 Otsu Binarization Algorithm

Otsu is a global thresholding binarization algorithm .It assumes that the image contains two classes of pixels; one foreground (black) and one background (white). Then it calculates the optimum threshold separating those two classes so that their intra-class variance is minimal. This is equivalent to maximizing the between-class scatter. This establishes an optimum threshold K.

$$I(x,y) = \begin{cases} 1, if\ I_{gray}(x,y) \le K \\ 0, if\ I_{gray}(x,y) > K \end{cases}$$

5. RESULTS

The text image is first filtered by each of the noise removal algorithms described in Section 3. Then each filter output along with the unfiltered original was then binarized by using Otsu binarization algorithm. Two methods are used to do the evaluation measures – MSE and PSNR.

Mean Square Error (MSE) for two $m \times n$ images *I* and *K* is given by,

$$MSE = \frac{1}{mn}\sum_{i=0}^{m-1}\sum_{j=0}^{n-1}[I(i,j) - K(i,j)]^2$$

where one of the images is considered a noisy approximation of the other .

Peak Signal-To-Noise Ratio (PSNR) is generally used to analyze quality of image in dB (decibels). PSNR calculation of two images, one original and an altered image, describes how far two images are equal.

$$PSNR = 10\log_{10}\left(\frac{M_I^2}{MSE}\right)$$

$$= 20\log_{10}\left(\frac{M_I}{\sqrt{MSE}}\right)$$

M_I is the maximum possible pixel value of the image.

These measures were calculated for every image-filter combination. The score obtained using preprocessing algorithms are shown in Tables 1. The image binarized using Otsu algorithms are measured and the scores got are shown in Table 2.

In most cases the best pre-processing filter was NL means. Next best results are shown by Total Variation filter with λ=3. For the conventional filters mean, median and Wiener filter, Wiener gives best results.

Table 1. Performance measure of various preprocessing algorithm

	Mean	Median	Wiener	Tikhonov	ROF λ=5	ROF λ=3	NLmeans
MSE	702.1072	713.0556	50.9573	338.8237	22.8120	13.4386	2.7225
PSNR	19.6668	19.5996	31.0587	22.8311	34.5492	36.8473	43.76

Table 2. Performance measure after Otsu binarization

		Median	Wiener	Tikhonov	ROF λ=5	ROF λ=3	NLmeans
MSE	4476.7	3557.5	325.4874	1923.7	169.1009	111.8863	84.2275
PSNR	11.6212	12.6194	23.0055	15.2895	25.8493	27.6430	28.9978

6. CONCLUSION

This paper presents a system that enhances the readability of an old Sanskrit document, through effective preprocessing methods for binarization. The pre-processing techniques have been successfully tested on a degraded document. Experiments show best results for the NL means algorithm, thereby showing good binarization performance. So the proposed methods can be used for preprocessing of historical machine-printed documents.

7. REFERENCES

[1] Laurence Likforman-Sulem, Jérôme Darbon Elisa H. Barney Smith, "Enhancement of Historical Printed Document Images by Combining Total Variation Regularization and Non-Local Means Filtering", Image and Vision Computing (2011). Vol. 29, Nr. 5, p. 351-363.

[2] Elisa H. Barney Smith, Laurence Likforman-Sulem, Jérôme Darbon, "Effect of Pre-Processing on Binarization", Conference: Document Recognition and Retrieval - DRR , pp. 1-10, 2010

[3] A. Buades, B. Coll, and J Morel. "On Image denoising Methods", *Technical Report 2004-15, CMLA, 2004.*

[4] A. Buades, B. Coll, and J Morel., "A non-local algorithm for image denoising", *IEEE International Conference on Computer Vision and Pattern Recognition*, 2005.

[5] W. Evans. Image denoising with the non-local means algorithm.Available:http://www.cs.wisc.edu .

[6] Chambolle, A. and Lions, "Image recovery via total variation minimization and related problems", Numer. Math., 76:167-188(1997)

[7] Chan, T., Golub, and Mulet, P. (1999), "A nonlinear primal-dual method for total variation-based image restoration", SIAM J. Sci. Comp., 20(6):1964-1977.

[8] Rudin, L., Osher, S., and Fatemi, "Nonlinear total variation based noise removal algorithms". Physica D, 60:259-268. 1992.