

1-26-2011

A Mask-Based Enhancement Method for Historical Documents

Elisa H. Barney Smith
Boise State University

Jerôme Darbon
CMLA, ENS Cachan, CNRS, PRES UniverSud

Laurence Likforman-Sulem
Telecom ParisTech

A Mask-based enhancement method for historical documents

Elisa H. Barney Smith^a and Jérôme Darbon^b and Laurence Likforman-Sulem^c

^aBoise State University, Boise, Idaho, USA

^bCMLA, ENS Cachan, CNRS, PRES UniverSud, France

^cTelecom ParisTech, Paris, France

ABSTRACT

This paper proposes a novel method for document enhancement. The method is based on the combination of two state-of-the-art filters through the construction of a mask. The mask is applied to a TV (Total Variation) - regularized image where background noise has been reduced. The masked image is then filtered by NLmeans (Non Local Means) which reduces the noise in the text areas located by the mask. The document images to be enhanced are real historical documents from several periods which include several defects in their background. These defects result from scanning, paper aging and bleed-through. We observe the improvement of this enhancement method through OCR accuracy.

Keywords: Document image enhancement, Image processing, Variational approach, Non-local Means, Historical documents, Character recognition

1. INTRODUCTION

A large number of document images are available for consulting, exchange and distant access purposes. These images have been scanned from collections of historical documents in libraries or archives thanks to digitization projects.¹⁻³ Accessing the content of document images is fully enhanced when textual transcriptions are attached to them; this allows users to index and search images through textual queries. For establishing such transcriptions, automatic tools such as OCR systems (Optical Character Recognition) are used to convert document images into text lines and words in ASCII format. However OCR systems are very sensitive; when facing noise, they perform poorly for both segmentation and recognition tasks. Historical documents include many defects due to aging and human manipulations. These defects include bleed-through ink, folding marks, ink fading, holes and spots. The grain of the paper can also produce texture in the background. The digitization process can produce uneven illumination in the image and modify character edges.⁴ Other degradations may be present such as streams on images obtained from microfilms. Thus, reducing or removing noise in document images is an important issue for improving OCR recognition. The main goal of this paper is to show that two recent powerful image restoration techniques can be combined to drastically improve the recognition performance.

Several approaches have been proposed for enhancing document images. Leung et al.⁵ enhance contrast with the POSHE method based on sub-block histogram equalization resulting in improvement of the readability of very low-contrast images. Sattar and Tay⁶ deblur noisy document images using fuzzy logic and multi-resolution approaches. Pan et al.⁷ correct uneven illumination and remove wood grain and shading on images of text incised on wood tablets by filtering allowing handwritten strokes to be more easily extracted. Removing shades in document margins produced while scanning thick documents with a region growing method has been studied by Fan et al.⁸ Shading in the background can also be detected by morphological operations and lightened for removal.⁹ The water flow model of Kim et al.¹⁰ can extract the different background layers of a document while binarizing it. Restoring character edges by PDE-based (Partial Differential Equations) approaches has also been proposed.¹¹ This approach regularizes a document image using anisotropic diffusion filtering through an iterative process. Approches exist to remove bleed-through pixels^{12,13} but they require co-registration of the recto and verso images. Source separation is another framework to address document enhancement.¹⁴ It assumes that each pixel results from the mixture of different sources (background, foreground and, in the case of palimpsests,

Further author information: EBS: ebarneysmith@boisestate.edu, LLS: likforman@telecom-paristech.fr, JD: darbon@cmla.ens-cachan.fr

another writing layer). With such an approach, smoothing, noise removal and thresholding are performed jointly. It contrasts with our filtering-based approach which is fast thanks to efficient algorithms and does not assume any structure of the document. Document image enhancement can be the first step of a binarization task. In Gatos et al.¹⁵ noise reduction by Wiener filtering is performed prior to the adaptive binarization of the document image. Markov Random Fields (MRFs) approaches¹⁶⁻¹⁸ are also suitable for document enhancement. Such approaches include in a single model both the data (the spatial local context of a pixel) and a degradation model.

Our approach is based on regularization and filtering and aims at reducing the noise level in the background and on character edges of document images. The background noise comes from ink bleeding from the verso or from defects of the recto. The existence of such noise makes document segmentation and recognition difficult. Additional pixels may fill the inter-line and inter-word spaces or create confusing character shapes. The proposed method combines two restoration steps based respectively on the Total Variation regularization approach (TV) and Non-local Means (NLmeans) filtering and combines them through the application of a mask. The present approach differs from our previous works.^{19,20} In¹⁹ the two recent filtering approaches, TV and NLmeans are tested in isolation and their comparison is based on an OCR task. The two approaches are compared in isolation in²⁰ but for a binarization task. In the present work, the two approaches are combined in an original single method and we show that this increases the performance of the OCR task.

Fig. 1-a shows the flowchart of the proposed method. The original image is pre-processed by TV in order to reduce background noise. The mask is constructed from this TV-processed image through binarization and dilation operations. The resulting binary image is applied as a mask on the TV-processed image. This results in a grey level image with low background noise and smoothed edges. The NLmeans filter is then applied to enhance character details. In Fig. 1-b, the mask is applied to the image pre-processed by NLmeans rather than TV. The decision whether to apply the combination scheme shown in Fig. 1-a or its variant in Fig. 1-b (referred to as the type-A and type-B combination respectively) is related to character dimensions and document contrast.

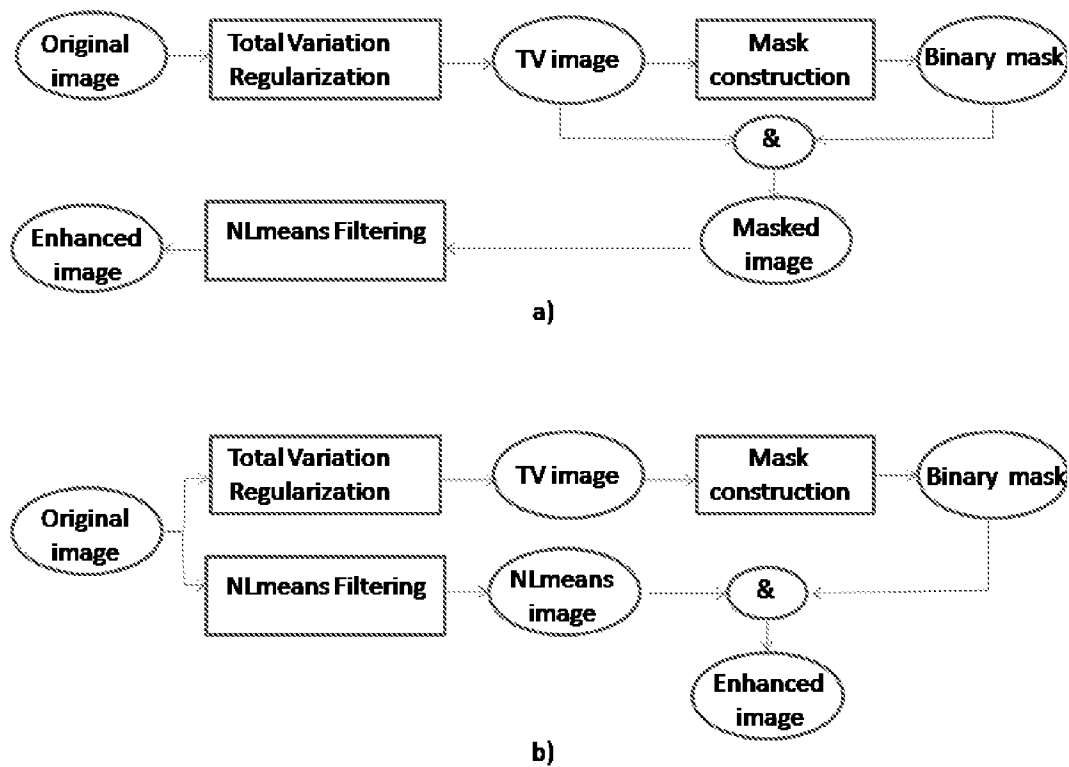


Figure 1. Flowcharts of the two proposed enhancement methods combining TV regularization and NLmeans filtering. a) type-A combination b) type-B combination.

The paper is organized as follows. In Section 2.1, the mask construction based on Total Variation is presented. Section 2.2 presents the foreground noise reduction based on NLmeans. In Section 3, we validate the proposed method on documents from various periods showing its robustness to a range of degradations. For printed documents, our evaluation is based on the recognition rate of an open source OCR, at the character level. Section 4 concludes the paper.

2. OUR APPROACH

2.1 Mask Construction with Total Variation

The documents we consider are historical documents which include a variety of noise such as background noise (bleedthrough, microfilm lines). Background objects in document images reduce the ability of a recognition system to localize text lines and words. A mask is used to reduce background noise. The image is filtered using a state-of-the-art filter based on the Total Variation. With TV grey levels are flattened and small objects may vanish. This is a desirable property for removing or reducing the effect of background objects, but can cause character detail to be lost.

Textual areas in the TV filtered image are located based on a conservative threshold chosen by the Otsu algorithm to locate the set of pixels which in majority belong to text. Second, the thresholded image is dilated by a square structuring element with 9 pixel sides. This results in a binary mask where regions around characters are set to 0 and the remaining ones to 1.

The mask is superimposed on the initial image to construct the masked image. The masked image consists of saturating at value 255 the image values corresponding to mask value 1, and recovering the initial pixel values corresponding to mask values 0. Noise pixels which have their grey levels flattened by TV are thus not included in the mask.

The mask can be applied to any image and we choose to apply it either to the TV-regularized image which has been created during mask construction or to the NLmeans filtered image (see also Section 2.2). This corresponds to the type-A and type-B combination methods respectively. The choice of the image to be masked depends on the character size, since spaces within small-sized characters processed by TV are more likely to be filled (see Section 3.2). There is no such effect with NLmeans. Fig. 2 shows the TV regularized image obtained from an original image. For this document, the TV image has been masked and the masked image (Fig. 2-d) is different from the TV image (Fig. 2-b) since background pixels distant from character pixels have been saturated.

TV has one main parameter β which determines the balance between the data fidelity term and the TV term.¹⁹ We use a default value $\beta = 20$.

2.2 Non-Local Means filtering

The NLmeans filtering is used in our combined approach in two ways (see Section 1). The filtering is applied either to the TV-regularized after being masked as an additional filtering step (type-A), or to the original image prior to being masked (type-B). NLmeans is a non-local filter which can smooth character parts from neighboring data. NLmeans averages neighboring parts of the central pixel but the averaging weights depend on the similarities between a small patch around the pixel and the neighboring patches within a search window.²¹ NLmeans tends to regularize the background more than characters. For background pixels similar patches can be found within the search window. Character pixels are smoothed less since fewer similar patches can be found within the search window. Thus character pixels are preserved which is desirable for fading characters. Default parameters $K = 4$ and $P = 3$ for the NLmeans approach are widely used for a number of applications. We also use the default parameters in our proposed enhancement method. Fig. 3 shows the effect of the NLmeans filtering on a sample word and characters on a document from set XXa. We observe that the noisy background has been strongly smoothed since similar patches can be found for background pixels. The fading pixels of character 'n' have been preserved. They have been smoothed less since fewer similar patches have been found. We also observe a small improvement on the edges of the binarized character 'j' since similar patches can be found along the long vertical stroke of this character. It can be noted that edges are also regularized by TV, but small objects such as fading pixels may vanish with TV.

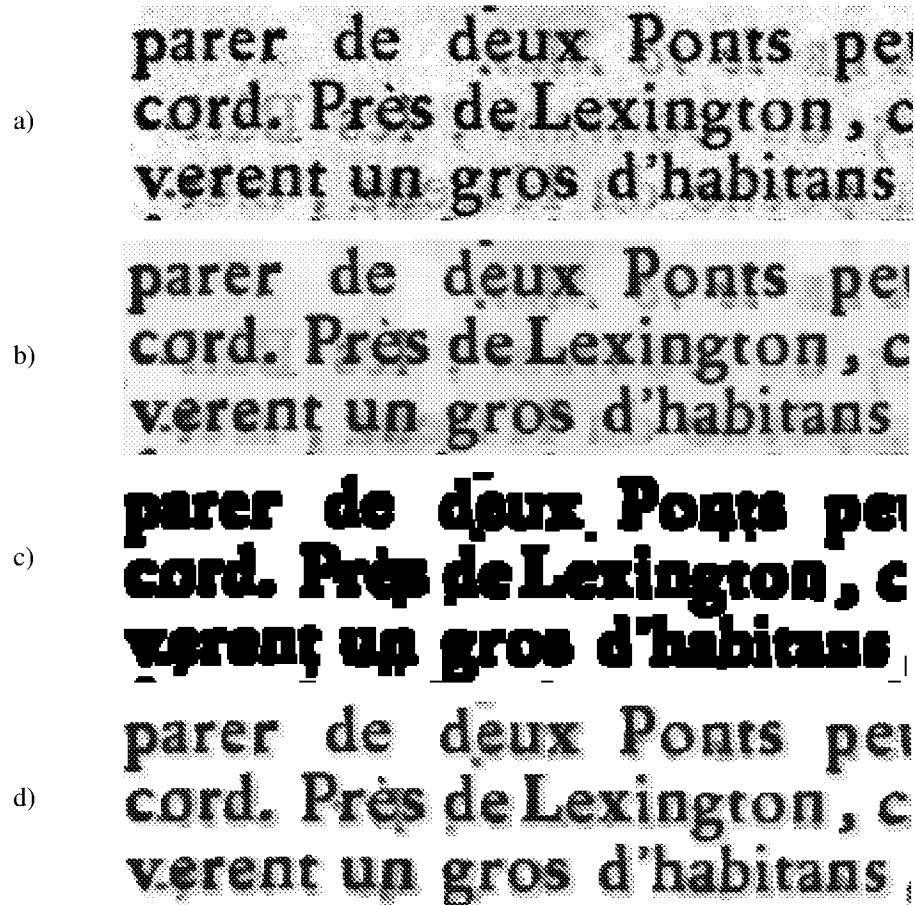


Figure 2. a) Original image b) TV-regularized image c) Mask d) Enhanced image resulting from the application of the mask on the TV image. Pixel values outside character regions are saturated to 255.

3. EXPERIMENTS

Experiments are conducted to evaluate the proposed approach on documents of various periods.

3.1 Data sets

Three sets of real printed degraded documents are used in these experiments. The sets are built and named according to the period in which the documents were created: the XVII, XVIII or XX century (see Fig. 6). Each set currently includes text images from two document collections, so that each set can be separated into set-a and set-b. Set XVII includes 1,457 characters from the electronic collection of the British library.²² Set XVII-a comes from a Hamlet theater piece, while set XVII-b is a festival book in French. Set XVIII includes 4,560 characters of French Gazettes, newspapers from the 18th century.²³ Set XVIII-a (Gazette d'Avignon) is less degraded, while set XVIII-b (Gazette de Leyde) includes more degraded characters. Set XX includes 496,836 characters of twentieth century documents. Sample images from a French journal whose publishing period is around 1930 are used to form set XX-a²⁴ and the whole set News.3G provided by ISRI forms set XX-b.²⁵ Table 1 shows how the size of x-characters varies according to the sets. The smallest characters are found in set XVIIa.

3.2 Evaluation through recognition

The proposed methods are evaluated through recognition performance on printed documents of various periods described in Section 3.1. To evaluate the performance of the proposed approach, we pass the enhanced images through the OCR Tesseract.²⁶ This OCR was originally developed by HP and obtained good results at the

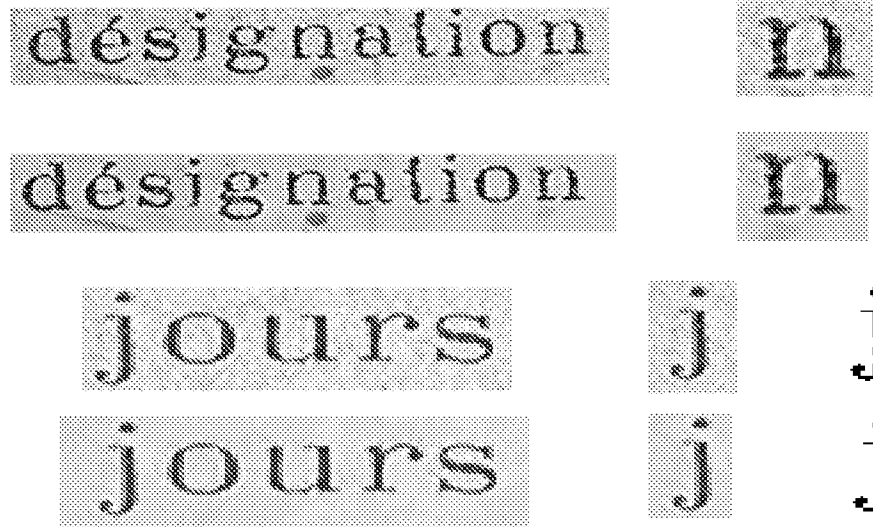


Figure 3. From top to bottom. An original sample word and enlarged character. The same word and character enhanced by NLmeans. Enlarged character and binarized character. Same character enhanced by NLmeans. NLmeans filtering reduces background noise and preserves character details.

test set	x-dimensions	image degradation	#characters
XVII-a	10x10	low contrast	730
XVII-b	23x23	bleed through	733
XVIII-a	16x16	bleed through, textured background	1,689
XVIII-b	37x37	bleed through, textured background, scanning streams	2,871
XX-a	21x21	low contrast, folding marks	4,756
XX-b	17x20 to 100x100	bleed through, low contrast	492,080

Table 1. Dataset image specifications. Size of x characters and qualitative description of the primary noise effects.

UNLV accuracy test in 1995.²⁵ It is now available open source through Google. The set XX-b is one of the sets tested in the UNLV evaluation as set News.3G. The OCR engine is a means for evaluating the improvements brought by the enhancement method we are proposing.

We consider one specific tuning for the OCR. This tuning consists of reducing the influence of dictionaries in order to test the improvement brought by the proposed enhancement approaches at the pattern recognition level, character by character. This is done by setting the Tesseract configuration variables *ok_word*, *good_word*, *non_word*, and *garbage* to one. This setting is suggested by²⁷ and allows us to run the OCR without dictionary-based corrections.

We provide to Tesseract the grey level document images, original or enhanced. The OCR uses the Otsu thresholding algorithm to binarize the images. However in our combined approach, there are two kinds of background pixels: saturated and not saturated (see Section 2.1). Consequently, we have modified the Otsu algorithm within the OCR system to take into account the specificity of these images which include three modes. The modification consists of removing the saturated mode from the histogram. Similarly, the modification of the Otsu thresholding algorithm is necessary for set XX-b (News.3G), since XX-b images have large white background zones around the clipped news articles whose paper intensities are moderately dark gray. It can be noted that preliminary experiments on set XX-b¹⁹ used a common global threshold for XX-b images, the value (75) being chosen as suggested in²⁵ for this data set. However for the sake of comparison, all sets are binarized in our experiments according to the same Otsu-based framework.

3.2.1 Mask-based enhancement

We evaluate the proposed method on the set of real printed documents. The method relies on two types of combination. The first type (type-A) method applies the mask to the TV-regularized image, while the

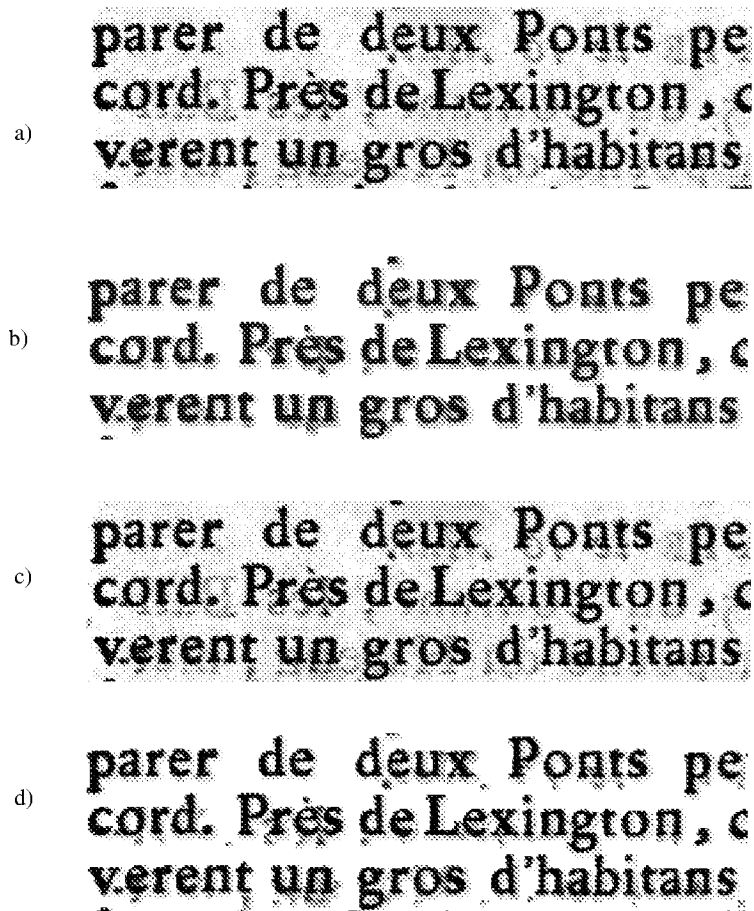


Figure 4. a) Original image b) Proposed method: Combination of TV and NLmeans (type A) c) Wiener filter d) Combination method with TV replaced by Wiener.

second type (type-B) applies it to the NLmeans filtered image. The type-B combination is more appropriate to documents which include small-sized characters since small characters may be filled or vanish with TV. Thus the type-A combination is applied to sets XVIIIb, XVIIIa, XVIIIb, XXa while the type-B combination is applied to sets XVIIa and XXb (see Table 1).

In Fig. 5, we compare the proposed method over no-enhancement. The improvement brought by our method is very high for sets XVIII-a and b: the increase reaches 20% in absolute value for set XVIII-b. This is due to the efficiency of the background noise reduction step on these highly degraded sets. Additional results are provided in Fig. 6.

To show the advantage of using TV in the proposed method, we replace TV by another enhancement method. We choose the Wiener filter since it is a popular filter for document enhancement¹⁵ which, like TV, has the ability to reduce background noise. The mask is either applied to the Wiener-filtered image or to the NLmeans-filtered image depending on the combination type (type-A or type-B). In both cases the mask is constructed using the Wiener filter. Using the Wiener filter in a combined scheme as shown in Fig. 4-d provides a quite different result than when it is used as a single enhancement technique as in Fig. 4-c. Results shown in Fig. 5 show that regularization and mask construction is more effective with TV than with Wiener.

4. CONCLUSION

We have proposed a new enhancement method based on the combination of two powerful preprocessing methods, namely the Total Variation regularization and the NLmeans filtering. We have taken advantage of the background

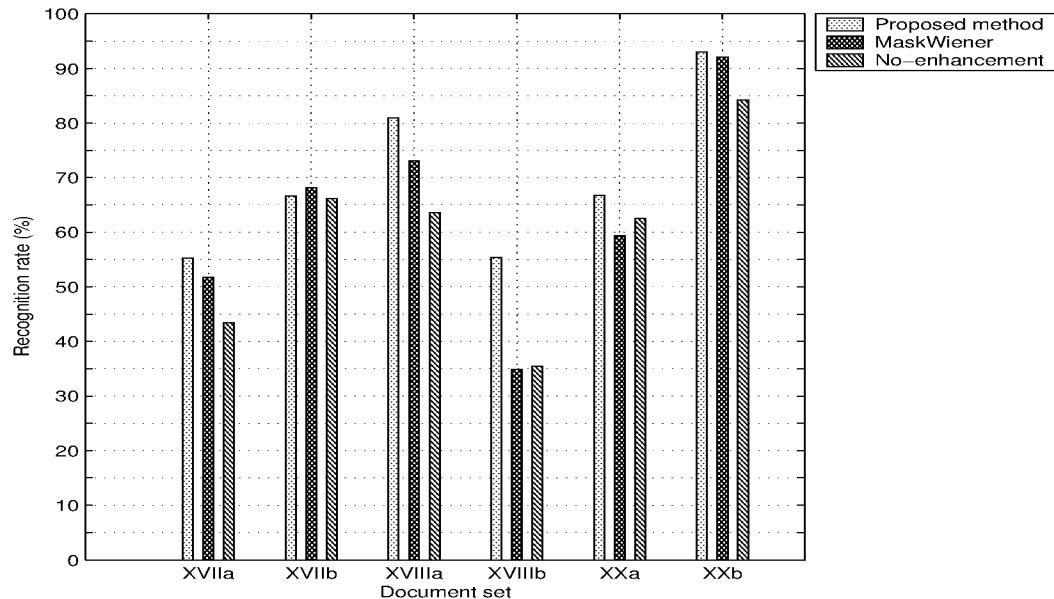


Figure 5. Performance of the proposed method and the same method where TV is replaced by the Wiener filter. Performance is provided through character recognition accuracy (in %).

noise reduction ability of TV and the character detail improvement of NLmeans. TV regularization is used for eliminating noise in the background through the construction of a mask. OCR improvement is observed with the proposed method for historical documents of various periods.

We have also conducted experiments, replacing TV by another regularization method such as the Wiener filter. We also observe an OCR improvement with the Wiener filter but a larger improvement with TV.

Future work will consist of testing our approach on other types of noise such as blurring.

Acknowledgements

The authors wish to thank Marc Sigelle from Telecom ParisTech for fruitful discussions. Research of Jérôme Darbon has been supported by the Office of Naval Research through grant N000140710810.

REFERENCES

- [1] “Googlebooks project.” <http://books.google.com/googlebooks/library.html>.
- [2] IMPACT, “Impact: Improving access to text, description of work.” www.impact-project.eu/.
- [3] Droettboom, M., “Correcting broken characters in the recognition of historical printed documents,” in [*Proc. of Joint Conference on Digital Libraries, JCDL’03*], 364–366 (2003).
- [4] Barney Smith, E., “Characterization of image degradation caused by scanning,” *Pattern Recognition Letters* **19**, 1191–1197 (1998).
- [5] Leung, C.-C., Chan, K.-S., Chan, H.-M., and Tsui, W.-K., “A new approach for image enhancement applied to low-contrast-low-illumination IC and document images,” *Pattern Recognition Letters* **26**(6), 769 – 778 (2005).
- [6] Sattar, F. and Tay, D., “Enhancement of document images using multiresolution and fuzzy logic techniques,” *Signal Processing Letters* **6**, 249–252 (1999).
- [7] Pan, X., Brady, M., Bowman, A. K., Crowther, C., and Tomlin, R. S. O., “Enhancement and feature extraction for images of incised and ink texts,” *Image Vision Comput.* **22**(6), 443–451 (2004).
- [8] Fan, K.-C., Lay, T.-R., and Wang, Y.-K., “Marginal noise removal of document images,” *Pattern Recognition* **35** (2002).

Un Courier extraordinaire de Madrid a apporté à la Cour de France à Paris la Nouvelle de ravitaillement de Gibraltar par le Flor de Anguilla, annoncée dans la Gazette de Madrid du 24. Avril: Mais en revanche ce même Courier a appris, que le Convoy de Marseille, au sujet duquel l'on étoit inquiet, est en sûreté dans le Port d'Alcante.

Un Courier extraordinaire de Madrid a apporté à la Cour de France à Paris la Nouvelle de ravitaillement de Gibraltar par le Flor de Anguilla, annoncée dans la Gazette de Madrid du 24. Avril: Mais en revanche ce même Courier a appris, que le Convoy de Marseille, au sujet duquel l'on étoit inquiet, est en sûreté dans le Port d'Alcante.

Il sera procédé le mardi vingt-sept janvier mil-neuf-cent-vingt à quatre heures, en l'étude et par le ministère de Mr DEBARRY notaire au Mesnil-Saint-Denis, permis à cet effet, à la vente aux enchères publiques, sans attribution de qualités, au plus offrant et dernier enchérisseur, du fonds de commerce, sous la désignation suit:

Il sera procédé le mardi vingt-sept janvier mil-neuf-cent-vingt à quatre heures, en l'étude et par le ministère de Mr DEBARRY, notaire au Mesnil-Saint-Denis, permis à cet effet, à la vente aux enchères publiques, sans attribution de qualités, au plus offrant et dernier enchérisseur, du fonds de commerce, sous la désignation suit:

The bird of heavening sings, as though he sang,
And that they fly, for France has beaten at sea,
The night we will inform, there we have France,
No France takes, nor which has power to change,
No protest, and its fallowed in that name.
Alas! So have I heard, and due to your business
But for the house no matter much to do,
What is the cause of pain for tomorrow's day,
Breaker would match up, and I'm sorry to say,
Let us hope that we shall see the day,
The time is coming, and I'm sorry to say,
The time is coming, and I'm sorry to say,
The time is coming, and I'm sorry to say,
The time is coming, and I'm sorry to say,
The time is coming, and I'm sorry to say,
The time is coming, and I'm sorry to say,
The time is coming, and I'm sorry to say,

The bird of heavening sings, as though he sang,
And that they fly, for France has beaten at sea,
The night we will inform, there we have France,
No France takes, nor which has power to change,
No protest, and its fallowed in that name.
Alas! So have I heard, and due to your business
But for the house no matter much to do,
What is the cause of pain for tomorrow's day,
Breaker would match up, and I'm sorry to say,
Let us hope that we shall see the day,
The time is coming, and I'm sorry to say,
The time is coming, and I'm sorry to say,
The time is coming, and I'm sorry to say,
The time is coming, and I'm sorry to say,
The time is coming, and I'm sorry to say,
The time is coming, and I'm sorry to say,
The time is coming, and I'm sorry to say,
The time is coming, and I'm sorry to say,

Tobacco chiefs still refuse


Steve Wesson, D-Conn. put that Republic's smoking ban on hold.



THE PLAN
Wesson's bill would allow states to opt out of the ban on flavored cigarettes.

Tobacco chiefs still refuse

Steve Wesson, D-Conn. put that Republic's smoking ban on hold.



THE PLAN
Wesson's bill would allow states to opt out of the ban on flavored cigarettes.

Figure 6. Enhancement results. Left: original image, right: enhanced image using a combination of TV and NLmeans.

[9] Nomura, S., Yamanaka, K., Shiose, T., Kawakami, H., and Katai, O., "Morphological preprocessing method to thresholding degraded word images," *Pattern Recognition Letters* **30**(8), 729 – 744 (2009).

[10] Kim, K., Jung, D. W., and Park, R. H., "Document image binarization based on topographic analysis using a water flow model," *Pattern Recognition* **35**, 265–277 (2002).

[11] Drira, F., Lebourgeois, F., and Emptoz, H., "OCR accuracy improvement through a PDE-based approach," in *[Proc. of ICDAR'07]*, 1068–1072 (2007).

[12] Moghaddam, R. F., Rivest-Henault, D., Bar-Yosef, I., and Cheriet, M., "A unified framework based on the level set approach for segmentation of unconstrained double-sided document images suffering from bleed-through," in *[Proc. of ICDAR'09]*, 441–445 (2009).

- [13] Tonazzini, A., Bianco, G., and Salerno, E., “Registration and enhancement of double-sided degraded manuscripts acquired in multispectral modality,” in [*Proc. of ICDAR’09*], 546–550 (2009).
- [14] Tonazzini, A., Bedini, L., and Salerno, E., “Independent component analysis for document restoration,” *IJDAR* **7**(1), 17–27 (2004).
- [15] Gatos, B., Pratikakis, I., and Perantonis, S. J., “Adaptive degraded document image binarization,” *Pattern Recognition* **39**, 317–327 (2006).
- [16] Tonazzini, A., Vezzosi, S., and Bedini, L., “Analysis and recognition of highly degraded printed characters,” *IJDAR* **6**, 236–247 (2004).
- [17] Wolf, C., “Improving recto document side restoration with an estimation of the verso side from a single scanned page,” in [*Proceedings of ICPR*], 1–4 (2008).
- [18] Wolf, C. and Doermann, D., “Binarization of low quality text using a Markov random field,” in [*Proceedings of ICPR*], 160–163 (2002).
- [19] Likforman-Sulem, L., Darbon, J., and Barney Smith, E., “Pre-processing of degraded printed documents by non-local means and total variation,” in [*Proc. of ICDAR’09*], 758–762 (2009).
- [20] Barney Smith, E. H., Likforman-Sulem, L., and Darbon, J., “Effect of pre-processing on binarization,” in [*DRR*], 1–10 (2010).
- [21] Buades, A., Coll, B., and Morel, J., “A review of image denoising algorithms, with a new one,” *SIAM-Multiscale Modeling and Simulation* **4**, 490–530 (2005).
- [22] “British Library: Treasures in Full.” <http://www.bl.uk/treasures/treasuresinfull.html>.
- [23] “Les Gazettes europeennes du 18eme siecle.” <http://gazettes18e.ish-lyon.cnrs.fr/>.
- [24] “Archives departementales des Yvelines.” <http://www.yvelines.fr/archives/home.html>.
- [25] Taghva, K., Nartker, T., Borsack, J., and Condit, A., “UNLV-ISRI document collection for research in OCR and information retrieval,” in [*Document recognition and retrieval VII*], 157–164 (2000).
- [26] Smith, R., “An overview of the Tesseract OCR engine,” in [*Proc. of ICDAR’07*], 629–633 (2007).
- [27] Sturgill, M. and Simske, S., “An optical character recognition approach to quantifying thresholding algorithms,” in [*Document Engineering 08*], 263–266 (2008).