2009 10th International Conference on Document Analysis and Recognition

Pre-processing of degraded printed documents by Non-local Means and Total Variation

Laurence Likforman-Sulem
Telecom ParisTech
likforman@telecom-paristech.fr

Jérôme Darbon
UCLA
jerome@math.ucla.edu

Elisa H. Barney Smith
Boise State University
EBarneySmith@boisestate.edu

Abstract

We compare in this study two image restoration approaches for the pre-processing of printed documents: namely the Non-local Means filter and a total variation minimization approach. We apply these two approaches to printed document sets from various periods, and we evaluate their effectiveness through character recognition performance using an open source OCR. Our results show that for each document set, one or both pre-processing methods improve character recognition accuracy over recognition without preprocessing. Higher accuracies are obtained with Non-local Means when characters have a low level of degradation since they can be restored by similar neighboring parts of non-degraded characters. The Total Variation approach is more effective when characters are highly degraded and can only be restored through modeling instead of using neighboring data.

1 Introduction

Documents in libraries or archives present many defects due to aging, ink fading, holes, spots and bleed-through ink. Printing methods have existed from the mid-15th century and there are a number of printed documents from this era which include such defects. Pre-processing such documents is often necessary prior to document segmentation and recognition. High-pass filtering may remove stains and holes. Average filtering can reduce additive Gaussian noise. Adaptive binarization can also reduce noise while segmenting text from background. For reducing noise level in document images, Markov Random field (MRF) approaches have been successfully used [14][16]. The MRF approach is suitable because it includes in a single model both the data (the spatial local context of a pixel) and a degradation model. In [14], a MRF-based blind deconvolution is used to segment touching characters. In [16], both character repairation and binarization are performed through MRF modeling. Restoring character edges by PDE-based approaches has been proposed in [8]. This approach regularizes a document image using an anisotropic diffusion filtering.

We introduce two methods for the pre-processing of gray level document images. These methods called Non-local means and Total Variation have been successfully applied to image restoration [7, 5]. While a variational approach has been applied to the restoration of degraded character contours [4], as far as we know this is the first time that Non-local Means and Total Variation methods have been applied directly to document images.

The Non-local means and Total Variation methods differ from classical filtering and the MRF-based approaches cited above. The Non-local Means (NLmeans) method 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 [5]. Total Variation (TV) is a minimization-based method for image restoration which preserves edges and sharp boundaries. This method can also be expressed as a minimization-based problem invoking a MRF but it enhances the MRF formalism since it uses a set of binary MRFs [7].

Our study consists of evaluating the NLmeans and Total Variation methods as a preprocessing step to document recognition. We apply these methods to printed documents from various periods. We first study the impact of parameters according to character resolution. Then we evaluate quantitatively the effectiveness of the preprocessing methods by passing restored images through an open source OCR. Our evaluation is based on the recognition rate at the character level.

The paper is organized as follows. Section 2 introduces the NLmeans and TV restoration approaches. In Section 3, we describe our experiments and their results. Section 4 concludes the paper.

978-0-7695-3725-2/09 $25.00 © 2009 IEEE
DOI 10.1109/ICDAR.2009.210

2 Document restoration

We present here two approaches for restoring document images. They both aim at reducing background noise while preserving character shape. The reduction of background noise also increases the ability to extract textual data at different levels: zones, text lines and characters.

The NLmeans method is based on image data while the TV method mostly relies on image modeling and a variational point of view. Before presenting these two approaches we introduce common notation. It is assumed that an image is defined on a 2D regular grid S. The original image and its restored version are denoted by v and $\hat{u}$, respectively. The grid S is endowed with a neighborhood system $\mathcal{N}$ and we denote by $\mathcal{N}(s)$ the set of sites that are neighbors of s (relatively to $\mathcal{N}$).

2.1 Non-Local Means

NLmeans capitalizes on the redundancy present in most images [5]. Document images may contain even more redundancy than other forms of images. The NLmeans filter considers the similarity of a block of neighboring pixels to the block centered on the pixel under evaluation. We denote such a block by Δ. For the sake of clarity we assume it is a squared patch whose side is $(2P+1)$. The similarity measure $w(s,t)$ for the two sites s and t is defined as

$$w(s,t) = g_h\left(\sum_{\delta\in\Delta}\left(v(s+\delta)-v(t+\delta)\right)^2\right) \quad , \qquad (1)$$

where g_h is a Gaussian of standard deviation h, i.e., $g_h(x) = \exp -\frac{x^2}{h^2}$. The parameter h is used to control the amount of filtering. Note that in [5], common convolutional filters weight the neighboring pixel values by their distance from the pixel in question. Instead, we have considered the version without these weights because it leads to much faster algorithms [6].

Once the similarity measures are available they are convexly combined to produce the filtered image. The value of the filtered image at site s is

$$\hat{u}(s) = \frac{1}{Z(s)}\sum_{t\in\mathcal{N}(s)} w(s,t)v(t) \quad , \qquad (2)$$

where $Z(s)$ is a normalization constant defined as $Z(s) = \sum_{t\in\mathcal{N}(s)} w(s,t)$ for all sites. In the original NLmeans version [5], the similarity is assumed to be computed for every pair of pixels, and thus $N(s) = S \setminus \{s\}$ (all image sites not including s). This is extremely time consuming.

Instead, the similar sub images for pixel are searched for in a region around the considered pixel. Let us assume that this search window is a square whose side is $2K+1$. The size of this search window affects the pixels that could be found and the level of regularization on the image. A larger searching window possibly allows more similar patches to be found and thus yield a filter that better preserve the features. However, this tends to produce more regularized images since the restored pixel will be a weighted average of more values. This behavior gives a background that is more homogeneous. This also produces a small loss of contrast. Let us emphasize that this loss is much *less* strong than for Total Variation based filter (see Section 2.2). Nevertheless, there will be less contrast between background and characters due to the averaging with more data patches. As a consequence, the fading characters lose more pixels after binarization. Figure 1 shows the effect of the size-of-search-window parameter. A sample word has been preprocessed and binarized using the document threshold (see Section 3).

Figure 1. Influence of the size of search window parameter in NLmeans. Filtered and binarized sample word. Large windows regularize more but degrade fading character parts.

2.2 Total Variation

Variational and Markovian formulations of the image restoration problem consist of minimizing an energy that is generally a weighted combination of two terms, namely the data fidelity and the regularization

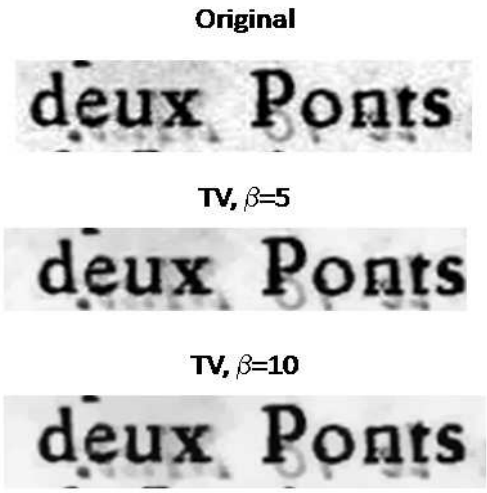

Figure 2. Total Variation: influence of regularization parameter β. Original and regularized words.

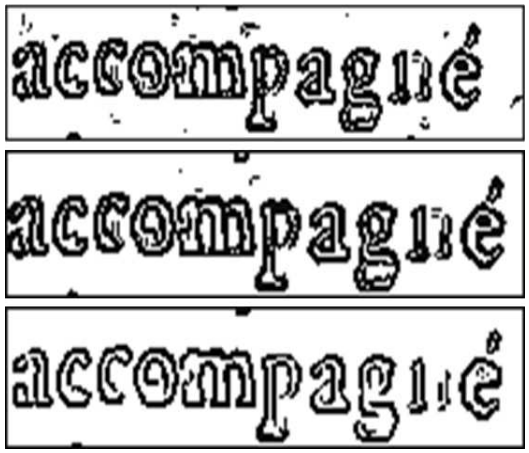

Figure 3. Edges from a highly degraded word from set XVIII-b. From top to bottom: without pre-processing, after restoration by NLmeans, after restoration by TV.

(also called prior). Since a discrete framework is considered in this paper, we consider a Markovian point of view as presented in [7]. The data fidelity D measures how far the current solution u is from the observed image v. It is defined from the nature of noise that corrupts the image. For instance, a separable quadratic data fidelity term corresponds to the assumption that the noise is additive Gaussian noise. We assume here such noise and thus the data fidelity is defined as

$$D(u|v) = \frac{1}{2}\sum_{s\in S}(u(s) - v(s))^2 \ . \tag{3}$$

The prior embeds the knowledge we have of the results. Among many regularization terms that have been proposed (see [15], for instance), the Total Variation has been a popular one [10]. The main characteristics of the TV prior is that the solution lives in the space of functions of Bounded Variation that allows for sharp edges and discontinuities. We follow [7] and define TV as the weighted l^1-norm of a discrete gradient. More formally we have

$$TV(u) = \sum_{s}\sum_{t\in N(s)} w_{st}|u(s) - u(t)| \ , \tag{4}$$

where w_{st} are some non-negative coefficients. Recall that $N(s)$ denotes the set of pixels that are neighboring the site s. In this paper, we consider the 4-nearest neighbors and set all the weights to 1.

The restored image $\hat{u}$ is the minimizer of the energy $E(u|v)$ that is a weighted combination of the two above terms

$$E(u|v) = D(u|v) + \beta TV(u) \ , \tag{5}$$

where the parameter β is non-negative. The latter coefficient is a parameter that governs the balance between the data fidelity and the regularization; a large value for β will produce an image with few details while a tiny one will yield an image that leaves v almost unchanged. It was shown in [9] that such an approach is prone to a loss of contrast. This prevents us from using a high regularization value since it would both remove small features that can be text, and reduce too much the contrast so that the binarization process would fail to produce the desired result. Figure 2 shows the effect of the regularization parameter β on sample words. The background has been highly regularized with both values of β. However for characters with lower resolution (around 13-14 pixel height), β should not be set to a high value in order to not destroy little character components. For highly degraded documents (see Fig. 3), TV can perform better than NLmeans since the noise in the background and on character contours is better reduced. However small details are better preserved with NLmeans.

3 Experimental results

We evaluate in this section the proposed approaches on sample printed documents. We use three sets of real degraded documents. The sets are built according to the period in which the documents were created. Each

set currently includes text images from two document collections, so that each set can be separated into set-a and set-b (see Fig. 4). Set XVII (century) includes 1,457 characters from the electronic collection of the British library [2]. Set XVII-a comes from a Hamlet theater piece, while set XVII-b is a festival book in French. Set XVIII includes 1,128 characters of French Gazettes, newspapers from the 18th century [3]. Set XVIII-a is less degraded, while set XVIII-b includes more degraded characters. Set XX includes 492,080 characters of twentieth century documents. Sample images from a French journal whose publishing period is around 1930 are used to form set XX-a [1] and the whole set News.3G provided by ISRI forms set XX-b [13]. The resolution of these documents is roughly the same, except for set XVII-a where characters are mid-sized (an x-character is 9x10 instead of 18x20).
To evaluate the performance of the proposed approaches, we pass the pre-processed images through the open source OCR Tesseract [11] hosted by Google. This OCR was originally developed by HP and obtained good results at the UNLV accuracy test in 1995 [13]. The set News.3G is one of the sets tested in the UNLV evaluation. We consider two tunings for this OCR. One tuning consists of reducing the influence of dictionaries in order to test the improvement brought by the proposed restoration approaches at the pattern recognition level, character by character. This is done by setting Tesseract configuration variables *ok_word*, *good_word*, *non_word*, and *garbage* to one. This setting is suggested by [12] and allows us to run the OCR without dictionary-based corrections. Another tuning consists of using the Tesseract dictionary of the document language (here English or French) since dictionaries are often used to compensate OCR errors. The improvement brought by both preprocessing and dictionary-corrections is then observed.

We provide to Tesseract the grey level document images, original or restored. The OCR uses the Otsu thresholding algorithm to binarize images. However for set XX-b (News.3G), this binarization method was not suitable as we obtained recognition accuracies below 3% for the original images. We thus binarized them prior to Tesseract recognition to a common global threshold (value=75). This threshold is the one chosen in [13] for this data set.

The results shown in Table 1 show that the proposed preprocessing methods improve document recognition at the character level. In these experiments the NLmeans parameters have been set to $P = 3, K = 4$, and TV parameter $\beta = 20$, except for set XVII-a for which $\beta = 5$ because of its lower resolution. In all cases when the documents were preprocessed by NLmeans an improvement in recognition accuracy was observed. In all but two of the datasets for OCR both with and without the dictionary corrections TV provided a recognition improvement. TV preprocessing is more effective on documents from earlier periods while NLmeans is more effective on modern documents. This can be explained by the fact that ancient documents include more fading characters; parts of these characters are not restored by neighboring ones or parts of the same character. Results also show that the improvement brought by NLmeans or TV is more important when dictionary-based corrections are turned off. While disabling of the dictionary is generally not desirable for practical applications, it shows that isolated recognition is enhanced with the proposed preprocessing methods. In particular for the most degraded set (XVIII-b) the recognition performance is the lowest when no preprocessing is applied. The improvement brought by the preprocessing is highlighted in Table 1. Even if the number of correctly recognized characters improves a little, the dictionary-based correction yields a much higher number of corrected characters.

no dictionary-based correction			
test set	no pre-processing	NLmeans	TV
XVII-a	43.5	49.9	40.2
XVII-b	66.2	66.6	68.6
XVIII-a	77.0	80.1	80.1
XVIII-b	31.7	38.0	44.7
XX-a	70.7	78.0	71.3
XX-b	92.5	92.7	88.7
with dictionary-based corrections			
test set	no pre-processing	NLmeans	TV
XVII-a	48.9	58.7	48.2
XVII-b	87.2	90.0	91.7
XVIII-a	92.8	93.3	94.2
XVIII-b	44.7	65.9	68.7
XX-a	92.4	95.7	95.4
XX-b	98.5	98.7	96.4

Table 1. Recognition rates (%) according to pre-processing approach. Top: Dictionary-based corrections turned off, bottom: turned on.

4 Conclusion

The NLmeans and Total Variation preprocessing methods have been evaluated for their ability to correct image defects found in document collections from

many periods. Both are capable of producing images that yield higher OCR recognition rates with and without dictionary post processing. It was shown that the TV preprocessing method was more effective on more degraded documents since it relies on an image model. NLmeans is more effective when small parts of characters are degraded since the method can restore degraded parts from neighboring image parts.
Future work will consist of comparing the proposed enhancement methods to classical ones such as PDE-based approaches [8]. The TV β parameter must be adapted to character resolution in order to preserve character details. Automatically determining this parameter will also be investigated in future work.

Figure 4. Sample documents. From top to bottom, left to right: XVII-a, XVII-b, XVIII-a, XVIII-b, XX-a, XX-b.

Acknowledgements

The authors wish to thank Marc Sigelle from Telecom ParisTech for fruitful discussions. This work has been supported by grant N000140710810.

References

[1] Archives Departementales des Yvelines. http://www.yvelines.fr/archives/home.html.

[2] British Library: Treasures in Full. http://www.bl.uk/treasures/treasuresinfull.html.

[3] Les Gazettes europeennes du 18eme siecle. http://gazettes18e.ish-lyon.cnrs.fr/.

[4] I. Bar-Yosef, A. Mokeichev, K. Kedem, U. Erlich, and I. Dinstein. Global and local shape prior for variational segmentation of degraded characters. In *ICFHR*, Montreal, 2008.

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

[6] J. Darbon, A. Cunha, T.F. Chan, S. Osher, and G.J. Jensen. Fast nonlocal filtering applied to electron cryomicroscopy. In *Proc. of ISBI*, pages 1331–1334, 2008.

[7] J. Darbon and M. Sigelle. Image restoration with discrete constrained total variation, Part I: fast and exact optimization. *JMIV*, 26:261–276, 2006.

[8] F. Drira, F. Lebourgeois, and H. Emptoz. Ocr accuracy improvement through a PDE-based approach. In *Proc. of ICDAR'07*, pages 1068–1072, Brasil, 2007.

[9] Y. Meyer. *Oscillating patterns in image processing and nonlinear evolution equations*, volume 22 of *University Lecture Series*. American Mathematical Society, 2001.

[10] L. Rudin, S. Osher, and E. Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D.*, 60:259–268, 1992.

[11] R. Smith. An overview of the Tesseract OCR engine. In *Proc. of ICDAR'07*, pages 629–633, (Brasil), 2007.

[12] M. Sturgill and S. Simske. An optical character recognition approach to quantifying thresholding algorithms. In *Document Engineering 08*, pages 263–266, 2008.

[13] K. Taghva, T. Nartker, J. Borsack, and A. Condit. UNLV-ISRI document collection for research in OCR and information retrieval. In *Document recognition and retrieval VII*, pages 157–164, San Jose CA, 2000.

[14] A. Tonazzini, S. Vezzosi, and L. Bedini. Analysis and recognition of highly degraded printed characters. *IJDAR*, 6:236–247, 2004.

[15] G. Winkler. *Image Analysis, Random Fields and Dynamic Monte Carlo Methods. A Mathematical Introduction*. Springer, 2006.

[16] C. Wolf and D. Doermann. Binarization of low quality text using a Markov random field. In *Proceedings of ICPR*, pages 160–163, 2002.