Open access • Proceedings Article • DOI:10.1109/ICDAR.2009.277

Clutter Noise Removal in Binary Document Images — Source link

Mudit Agrawal, David Doermann

Institutions: University of Maryland, College Park

Published on: 26 Jul 2009 - International Conference on Document Analysis and Recognition

Topics: Constant false alarm rate, Clutter, Background noise, Distance transform and Noise

Clutter Noise Removal in Binary Document Images

Mudit Agrawal and David Doermann
Institute of Advanced Computer Studies
University of Maryland
College Park, MD, USA
{mudit, doermann}@umd.edu

Abstract

The paper presents a clutter detection and removal algorithm for complex document images. The distance transform based approach is independent of clutter's position, size, shape and connectivity with text. Features are based on a residual image obtained by analysis of the distance transform and clutter elements, if present, are identified with an SVM classifier. Removal is restrictive, so text attached to the clutter is not deleted in the process. The method was tested on a collection of degraded and noisy, machine-printed and handwritten Arabic and English text documents. Results show pixel-level accuracies of 97.5% and 95% for clutter detection and removal respectively.

This approach was also extended with a noise detection and removal model for documents having a mix of clutter and salt-n-pepper noise. [1]

1. Introduction

Real world signals often deviate from the ideal signals that were produced by the source. These deviations, may manifest themselves during scanning, transmission, storage or conversion from one form to another and are referred to as noise. Irrelevant content can also be viewed as noise, making the problem of detection and removal very much application dependent.

Document analysis algorithms such as page segmentation and character recognition, for example, often work best on the assumption of a clean document and use principle of connected components as basic units. Unfortunately, noise often interferes with these assumptions.

Noise in binary document images can be viewed as dependent or independent of underlying document content. Ink blobs, salt-n-pepper [3], stray marks, marginal noise [4] are, in general, independent of location, size or other properties of text data in the document image. Recorded images having this type of noise, can be expressed as the sum of true image $I(i,j)$ and the noise $N(i,j)$ as $R(i,j) = I(i,j) + N(i,j)$ Blur, pixel-shift or bleed-through [12] on other hand, manifest themselves differently depending on the content. Such content-dependent noise is comparatively more difficult to model, mathematically non-linear and often multiplicative.

Noise can also be classified based on its consistency in properties like periodicity of occurrence in the document, its shape, position and gray-values. If noise shows a consistent behavior in terms of these properties, it is called regular noise. Unwanted punched holes and stray marks exhibit regularity in their shapes while ruled lines [13, 14] show periodicity in their positions as well. On the other hand, noise such has ink blobs, complex background binarized patterns, marginal noise [4] and salt-n-pepper [3, 1] often lack any consistent property. This 'irregular noise' has typically been classified with simple rule based features. Ozawa and Nakagawa [9], Wang and Tang [12], Negishi et. al. [8] use gray level to distinguish foreground from background. Fan et. al. [4] assumes length, position and neighborhood of noise to detect and remove the noise. Liang et al. [5] depend on periodicity and regularity of noise to get rid of it. However, there has not been much work reported on the removal of irregular noise from binary document images. In this paper, we focus on the removal of clutter from binary images.

Clutter is a general term we use to refer to unwanted foreground content which is typically larger than text in binary images. It can result from numerous sources. While some forms of clutter like punched holes (Figure 1a), ink seeps (Fig. 1b), ink blobs (Fig. 1c) and copier borders typically are present before the scanning process, other types of marginal noise may result from the scanning of bound or skewed documents (Fig. 1a,f) where the gap between the gutter and scanner or between edges of paper and scanner bed causes lighting variations. Other scanning and bina-

[1] The partial support of this research by DARPA through BBN/DARPA Award HR001108C0004 and the US Government through NSF Award 1150713501 is gratefully acknowledged

978-0-7695-3725-2/09 $25.00 © 2009 IEEE
DOI 10.1109/ICDAR.2009.277

rization artifacts may give rise to clutter as well (Fig. 1d,e). Clearly, clutter is predominantly independent and irregular.

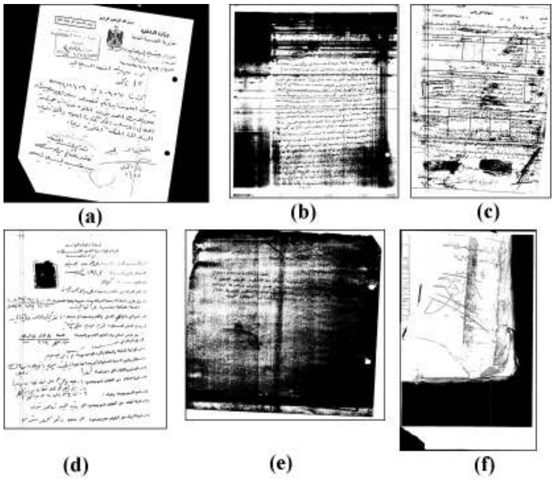

Figure 1: Examples of clutter

One of the major issues with clutter is its connectivity with text. Clutter often touches or overlaps some parts of the text. In case of ruled line documents with clutter, a single connected component connecting clutter, ruled lines and text may appear (Figure 2). Complete removal of the connected component in such cases may result in tremendous loss of content while morphology can degrade the text. As far as we know, there has been no collective work on the detection and removal of clutter, without removing or further degrading the attached text, in binary document images. Fan, Wang and Kay [4] detect and remove marginal noise regions based on three assumptions of shape, length and position. The technique does fairly well at removing only the marginal noise without the attached text. In contrast, our technique achieves the same on all forms of clutter, while being independent of clutter's position, size, shape and connectivity with text. It is also independent on the inclusion of any other type of noise.

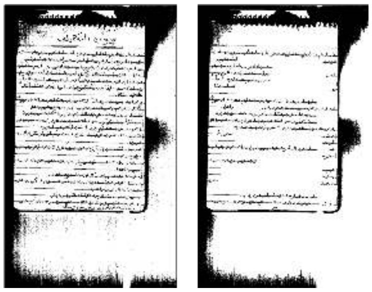

Figure 2: (a) Image (b) shows a single connected component with text attached to it

This paper is organized as follows. Section 2 describes the problem definition, clutter detection and removal. Section 3 extends this approach to propose a generic noise removal model in which several forms of noise can be removed iteratively, without interfering in the detection and removal of clutter. This is followed by experiments and evaluation in Section 4 and by conclusion and future work in Section 5.

2. Clutter Detection and Removal

Our approach to clutter removal is two-phased. In first phase, we detect the components which contain clutter. The second phase consists of removing only the clutter from the detected component(s) with minimal content deletion.

2.1. Problem Definition

Before framing the problem mathematically, we define our distance functions for digital images. Let p be a pixel in the document image I, located at (x, y) position, where $0 \leq x \leq imageheight$ and $0 \leq y \leq imagewidth$. Let $d(p_i, p_j)$ be a positive definite, symmetric and triangular measure of the distance from pixel p_i to p_j [11] such as the Euclidean distance. We are particularly interested in the integer approximations of Euclidean distance for every pixel on image. Rosenfeld and Pfaltz [11] proved that *octagonal distance* d_o is a better approximation to Euclidean distance than *city block, square, hexagonal* and *ceil of Euclidean* distance functions. Also, *nearest integer to Euclidean* and *floor of Euclidean* are not distance functions as they violate triangular property. Octagonal distance is defined as:

$$d_o = \max([2(|x_i - x_j| + |y_i - y_j| + 1)/3], \max(|x_i - x_j|, |y_i - y_j|)) \tag{1}$$

Using some pixel properties (e.g. grayvalue), an image can be divided into different sets of pixels. Distance Transform [2] associates distances to every pixel of a set P from other sets as follows:

$$D_P(p) = \min_{q \epsilon I}(d_o(p, q) + f(q)) \tag{2}$$

where initially,

$$f(q) = \begin{cases} \infty & \text{if q } \epsilon \text{ P} \\ 0 & \text{otherwise} \end{cases}$$

For binary images, there are only two sets of pixels, depending on foreground pixels (fg) background (bg):

$$I = \{P, P'\}, \quad P = \{p | I(p) = fg\}, \quad P' = \{p' | I(p') = bg\}$$

The distance transform is used both for detection and removal. We define D_I as the foreground distance transform of image I, where foreground pixels are labeled by their distance to the closest boundary and all background pixels are labeled 0. $D_{I'}$ is defined as the background distance transform of image I, where background pixels are labeled by their distance to the closest boundary and all foreground pixels are labeled 0. The distance transform can be computed efficiently with a two pass algorithm presented in [10].

2.2. Detection

Clutter, by definition, is larger than maximum text-stroke width present in the document, whereas thickness of ruled-lines, salt-n-pepper, stray-marks, bleed-through etc. can be of the order of text-stroke width. It is interesting to note that this property of clutter differentiates it from other types of noise and text. We assume clutter is bigger than twice the text's maximum stroke width present in the document. Hence, thinning the foreground pixels to half the maximum distance transform value, will erode all other text and other noise pixels from the document and will leave behind only a core of each clutter element. On the other hand, in the absence of any clutter, text strokes will be thinned to half their maximum width, maintaining a text-like pattern (albeit broken). After performing the distance transform, removing all pixels which are less than half of the maximum transform distance in the image, results in a residual image, we call the *half residual* (Figure 3(b)). It can be computed as follows:

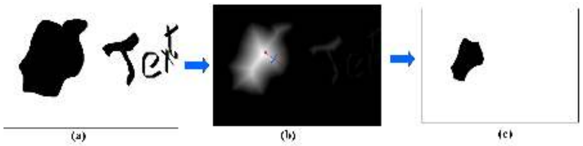

Figure 3: (a) An image with clutter and text. (b) the distance transform on the image with distances normalized to gray values [0-255] (c) the result of half-residual

1. Perform distance transform D_{Io} on the original binary image Io, as illustrated in Equation 2
2. Calculate the maximum value $dtMax = \max(D_{Io})$
3. Set all pixels p with $D_{Io}(p) < dtMax/2$ to background. The half-residual image Ih is obtained.

Next, we compute the features from this half-residual image Ih for clutter detection. Table 1 shows how the selected features, based on connected components, inherently distinguish Ihs of clean and clutter images. We train a 2-class SVM on these features to classify Ih as having clutter or not. Figure 4 shows clutter detection and removal model. Since in practise, several clutter components of varying sizes might occur in the same document, early iterations of half-residual image may erase the smaller clutter components. Hence, in such cases, detection and removal need to be performed iteratively till a clean document is detected, as shown in Figure 4.

2.3. Removal

Once the document image is classified as having clutter noise, the components from the half-residual image Ih, called *HR-cores*, are replaced with their corresponding (and larger) connected components in the original image I. Resulting image Ic has only these candidate clutter components from their original image I.

Table 1: Features selected for clutter detection

Features of CC	Clean Img	Cluttered Img
Number	high	very low
Avg size	low	high
Variance in size by Average Size	low	At extremes (zero or high)
Variance in positions of centroids of CC	very high	low
Average ratio of area by perimeter	low	high
Ratio of CC before and after half-residual	near to 1	very high

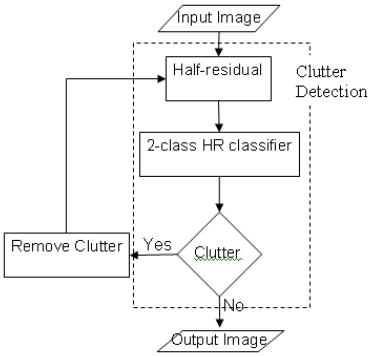

Figure 4: Clutter Detection and Removal

Our goal is to identify those pixels of these clutter components which belong to the clutter, and isolate the non-clutter (text) pixels. We observe that if we "regenerate" the clutter, from the half-residual core obtained in the previous step, by introducing the pixels from the original component for successive distances, then the clutter pixels will be encountered in roughly the same numbers in every step, as when we removed them.

As we approach the boundary of the clutter, this no longer holds true, because the original removal would have been eliminating both text and clutter pixels. Alternatively, we can consider, for an original removal step, how many regeneration steps (unique distances) would be required to regenerate it. We note that as we approach the clutter boundary, and attempt to regenerate it, the number of steps required for regeneration increases significantly. This is due to the fact that text-branches protrude out of the clutter's shape. The original removal step at which this number increases sharply is the minimum distance ρ from clutter's boundary at which all text is completely removed. This process can be shown as follows:

1. Compute D_{Ic}
2. Compute $D_{Ih'}$
3. $f(d) = |distinct(D_{Ih'}(p))|,\ \forall\ d\ \epsilon\ D_{Ic}(p)$ where $p\ \epsilon\ \{Ih' \cap Ic\}$,

As shown in Figure 5(b), moving outwards towards the boundary of clutter-component, there is a sharp rise in $f(d)$ at ρ. This function is a monotonically decreasing function.

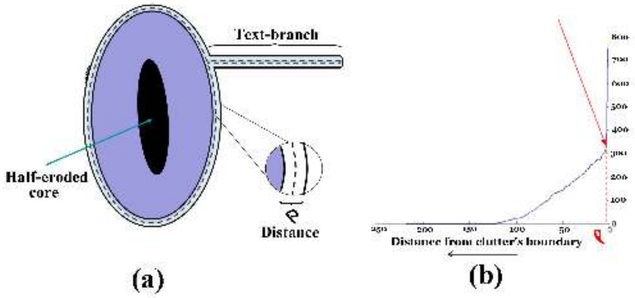

Figure 5: (a) Clutter component showing a text-branch attached to main clutter body (b) Frequency graph showing a sharp rise at ρ

$f'(d)$ is the rate of change of the function, which slows down at ρ. If $g(x) = f''(d)$, ρ is the index of *first* maxima of g(x).

$$\frac{d}{dx}(g(x)) = 0, \quad \frac{d^2}{dx^2}(g(x)) > 0 \tag{3}$$

It is not important that half-residual core should maintain the exact shape of the clutter. The point of *first* sudden drop in the function can predict the distance from the real boundary. The depth of the drop is proportional to the length of the text-branch. Once this distance ρ is obtained, shrinking and expanding the clutter-component by this distance, gets the clutter without its text-branches. If ρ is zero, these operations are not performed, as there is no text attached to the clutter and clutter component can directly be removed.

1. Obtain image Id is obtained by removing all pixels p from Ic such that $D_{Ic}(p) \leq \rho$
2. Compute D_{Id}
3. Removing all pixels in Id and pixels with $D_{Id} \leq \rho$, removes the clutter from Io without removing the text attached to it

2.4. Complete Removal of Clutter with Blurring Boundaries

Clutter does not always form contrasting boundaries with the background. Their edges sometimes blur into the background, creating spray-like appearances. These spray-like noise near a clutter's boundary, act as openings for distance transform (D_{Ic}), keeping parts of clutter intact around them (in the order of ρ), as shown in Figure 6(b). Closing these openings and applying clutter removal (as in Section 2.3) on the resultant clutter-component will remove this noise along with clutter. Though these openings resemble salt-noise, known methods for salt-noise removal like median filtering [3] can not be used, as these openings are extremely dense (unlike SnP noise) or big enough to escape removal in a prefixed size mask. Also, closing using a smaller structuring element may not close all these openings while using a bigger element may close the text-loops as in 'g','o' etc. Hence, we need a structuring element of the order of most frequent size of the openings. CC' contains the empty area outside clutter component and the various openings inside it. The radius of every opening is determined by performing a $3X3$ sized mask $M(p)$ on $D_{Ic'}$ which picks out the maximum distant pixel of every opening as follows:

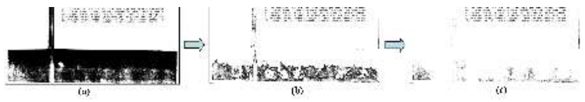

Figure 6: Complete Clutter Removal

$$fMax(p) = \begin{cases} D_{Ic'}(p) & \text{if } \max\limits_{l\epsilon M(p)}(D_{Ic'}(l)) = D_{Ic'}(p) \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

Histogram of these distances shows a sudden dip after the frequent openings' radius. This dip can be calculated using the same method described in Equation 3. The radius so obtained is used to close the openings and perform clutter removal thereafter. The improvement in results is shown in Figure 6(c).

3. Generic Noise Removal Model

An individual detection process for each kind of noise can be expensive, and at the same time, due to various shapes and forms of noise, it is difficult to *identify* noise as a whole, than *reject* it as non-content. Training a generic recognizer on any kind of noise is hence not possible. Larry and Malik [7] designed a single class classifier which is trained on positive samples only and rejects any sample not in the *trained* class as *other*. Using the same principle, it should be possible to train a single-class SVM on clean document images, which can then reject a document with any kind of noise (non-clean). The clutter removal approach, due to its independent nature with respect to other types of noise, can be combined with this generic noise removal framework. Since half-residual on a clutter document removes anything with similar or lesser width than text-stroke width, noises like salt-n-pepper, ruled lines, bleed-through and stray marks will be removed in the process, and will not interfere in clutter detection and removal.

4. Evaluation and Results

We evaluated the clutter detection and removal approach on datasets of printed and handwritten documents in English and Arabic scripts from five different sources. The dataset contains a representative set of 50 images with all forms of clutter and a set of 50 clean images. 30 images from each set are used for training. Clutter detection accuracy on the remaining 40 images is 97.5%. For clutter removal algorithm, we use an xml-based LAMP's GEDI tool [6] for pixel-based labeling and visualization. Each image is labeled into clutter and non-clutter (text, ruled-lines etc.) pixels. First evaluation criteria is pixel-based where reported accuracy of 95% is obtained as the percentage of clutter pixels removed. Figure 7 shows the clutter removal results of images in Figure 1. Second evaluation criteria is purposive, where we evaluate the improvement in successive stages due to clutter removal. Pixel based ruled line removal improves its processing time by 4 times due to much lesser foreground pixels in clutter-cleaned documents. Comparative evaluation of connected component based approaches like page segmentation and text-line extraction on clean and clutter documents will be a part of our future work.

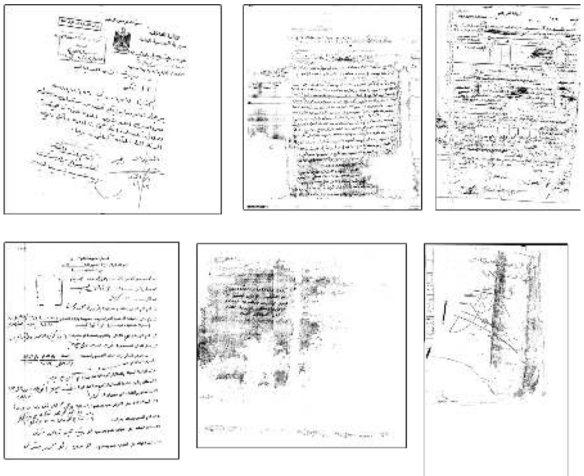

Figure 7: Clutter Removed

5. Conclusion and Future Work

We have presented a novel approach toward clutter detection and removal for complex binary documents. Our distance transform based approach aims at the removal of irregular and non-periodic clutter noise from binary document images and is independent of clutter's position, size, shape and connectivity with text. We use an SVM classifier to detect clutter. The novelty of this approach is in its restrictive nature to remove clutter, as text attached to the clutter is neither degraded nor deleted in the process. Clutter detection and removal accuracies were reported greater than 95% on machine-printed and handwritten documents of English and Arabic scripts. We would like to extend this approach to incorporate various other noise models to achieve our goal of a generic noise removal system. Better feature extraction and classification schemes for clutter detection will be another direction to enhance accuracy.

References

[1] M. Ali. Background noise detection and cleaning in document images. *Proc. 13th Int'l Conf. Pattern Recognition, (ICPR 1996)*, 3:758–762 vol.3, Aug 1996.

[2] G. Borgefors. Distance transformations in digital images. *Comput. Vision Graph. Image Process.*, 34(3):344–371, 1986.

[3] K. Chinnasarn, Y. Rangsanseri, and P. Thitimajshima. Removing salt-and-pepper noise in text/graphics images. *Asia-Pacific Conference on Circuits and Systems (IEEE APCCAS'98)*, pages 459–462, Nov 1998.

[4] K.-C. Fan, Y.-K. Wang, and T.-R. Lay. Marginal noise removal of document images. *Proc. Sixth Int'l Conf. Document Analysis and Recognition (ICDAR'01)*, pages 317–321, 2001.

[5] S. Liang, M. Ahmadi, and M. Shridhar. A morphological approach to text string extraction from regular periodic overlapping text/background images. *Proc. IEEE Int'l Conf. Image Processing (ICIP-94)*, 1:144–148 vol.1, Nov 1994.

[6] S. J. M. Roth and D. Doermann. Gedi: Ground truth. editor and document interface. In *Summit on Arabic and Chinese Handwriting Recognition*, 2006.

[7] L. M. Manevitz and M. Yousef. One-class svms for document classification. *J. Mach. Learn. Res.*, 2:139–154, 2002.

[8] H. Negishi, J. Kato, H. Hase, and T. Watanabe. Character extraction from noisy background for an automatic reference system. *Proc. Fifth Int'l Conf. Document Analysis and Recognition (ICDAR'99)*, pages 143–146, Sep 1999.

[9] H. Ozawa and T. Nakagawa. A character image enhancement method from characters with various background images. *Proc. Second Int'l Conf. Document Analysis and Recognition (ICDAR'93)*, pages 58–61, Oct 1993.

[10] A. Rosenfeld and J. L. Pfaltz. Sequential operations in digital picture processing. *Journal of the Assoc. for Comp. Mach.*, 13(4):471–494, 1966.

[11] A. Rosenfeld and J. L. Pfaltz. Distance functions on digital pictures. *Pattern Recognition*, 1(1):33–61, 1968.

[12] Q. Wang and C. L. Tan. Matching of double-sided document images to remove interference. *Proc. Comp. Vision and Patt. Recognition (CVPR'01)*, 1:I–1084–I–1089 vol.1, 2001.

[13] Y. Zheng, H. Li, and D. Doermann. A model-based line detection algorithm in documents. *Proc. Seventh Int'l Conf. Document Analysis and Recognition (ICDAR'03)*, pages 44–48 vol.1, Aug. 2003.

[14] Y. Zheng, C. Liu, X. Ding, and S. Pan. Form frame line detection with directional single-connected chain. *Proc. Sixth Int'l Conf. Document Analysis and Recognition (ICDAR'01)*, pages 699–703, 2001.