

De-blurring Textual Document Images

Daniel M. Oliveira¹, Rafael Dueire Lins¹, Gabriel P. Silva¹,
Jian Fan², and Marcelo Thielo³

¹ Universidade Federal de Pernambuco
Recife - Brazil

{daniel.moliveira,rdl,gabriel.psilva}@ufpe.br

² Hewlett-Parckard Labs.
Palo Alto - USA

³ Hewlett-Parckard Labs.
Porto Alegre - Brazil

{jian.fan,marcelo.resende.thielo}@hp.com

Abstract. Document images may exhibit some blurred areas due to a wide number of reasons ranging from digitalization, filtering or even storage problems. Most de-blurring algorithms are hard to implement, slow, and often try to be general, attempting to remove the blur in any kind of image. In the case of text document images, the transition between characters and the paper background has a high contrast. With that in mind, a new algorithm is proposed for de-blurring of textual documents; there is no need to estimate the PSF and the filter proposed can be directed applied to the image. The presented algorithm reached an improvement rate of 17.08% in the SSIM metric.

Keywords: De-blurring, blur, camera documents, scanner documents.

1 Introduction

Noise is any phenomenon that degrades information. A taxonomy for noises in document images is proposed in reference [9] which besides providing an explanation of how different noises appeared in the final image, it gives pointers to the literature that show ways of avoiding or removing them. In the classification proposed [9], there are four kinds of noise:

1. *The physical noises – whatever “damages” the physical integrity and readability of the original information of a document. It may be further split into the two sub-categories proposed in as internal and external.*
2. *The digitization noises – introduced by the digitization process. Several problems may be clustered in this group such as: inadequate digitization resolution, unsuitable palette, framing noises, skew and orientation, lens distortion, geometrical warping, out-of-focus digitized images, motion noises.*
3. *The filtering noise – unsuitable manipulation of the digital file may degrade the information that exists in the digital version of the document (instead of increasing it).*
4. *The storage/transmission noise – the noise that appears either from storage algorithms with losses or from network transmission. JPEG artifact is a typical example of this kind of undesirable interference.*

The blur noise has the effect of unsharpening images. Depending on how it arises it may be included in any of the four categories above. The physical blur may be the result of document “washing”, for instance, in which a document, printed with water soluble ink, gets wet. Blur may also be the result of unsuitable digitization, due to several reasons: non-flat objects, digitization errors, out of focus, motion etc. The presence of blur may be an indicator of low quality digitization, but can also be associated with other problems such as the scanning of hard-bound volumes. Blur may be the result of unsuitable filtering, such as a Gaussian or low-pass filter. And finally, blur may appear as the result of storing images in a file format with losses that perceptually degrades the image.

The technical literature points at several approaches proposed for de-blurring images in general. To list a few of them: Demoment [2] uses statistics, Neelamani, Choi, and Baraniuk [3] use Fourier and wavelet transforms, references [4] and [5] apply variational analysis, and Roth and Black [6] use total variation and Field of experts. Most of times, the computational complexity of those algorithms is prohibitively high and can yield undesirable artifacts such as ringing [7] as presented in Figure 1.

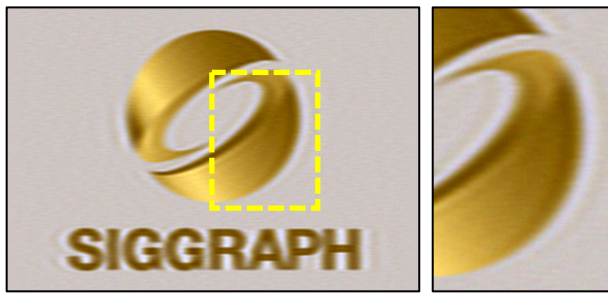


Fig. 1. Ringing artifact [7]

The most successful approaches to blur removal focus at one specific kind of blur. For instance, the literature presents several algorithms [11, 12, 13, 14, 15, 16, 17] that address the problem of motion blur, an specific kind of digitization noise.

In this paper, to increase the chances of better de-blurring, the application domain is restricted to monochromatic scanned documents with book binding warping [10]. The resulting image has uneven blur and illumination. The document images treated here are basically constituted by text and plain paper background. The transition between them in the original physical document is sharp. Using this fact a new algorithm is proposed by using nearby pixels to increase the difference between them. No Point Spread Function (PSF) [18] estimation is done and blur is minimized into a direct application of the image.

2 The New Method

The study performed here focus on the compensation of the blur noise which appears in scanning hardbound documents. Patterns were arranged in an elevated plane model

[1] as shown in Figure 2 to simulate the hardbound warp. Two HP scanners (PhotoSmart C4280 and a HP ScanJet 5300c) were used to digitize a pattern of lines in a grid. Two examples of blurred cross sections can be seen on Figures 3 and 4, corresponding to the two different scanning devices, respectively with two elevation and skew angles.

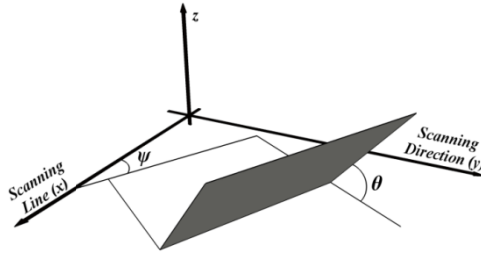


Fig. 2. Elevated plane [1]

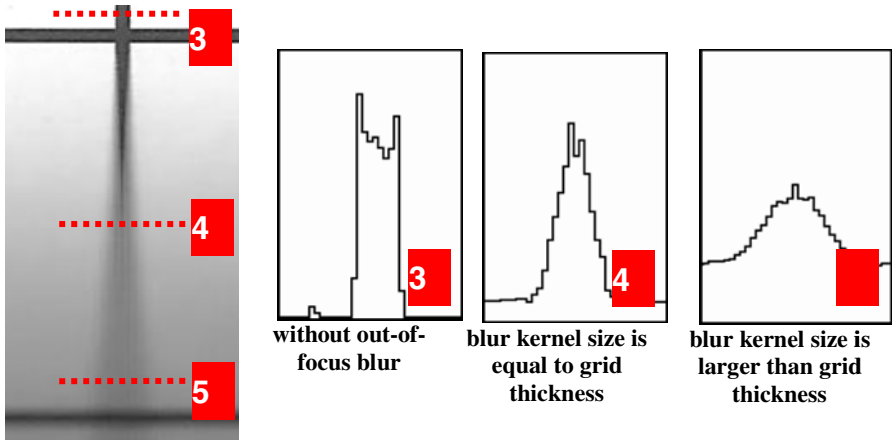


Fig. 3. Line grid scanned with HP PhotoSmart C4280 with $\psi = 0^\circ$ and $\theta = 30^\circ$

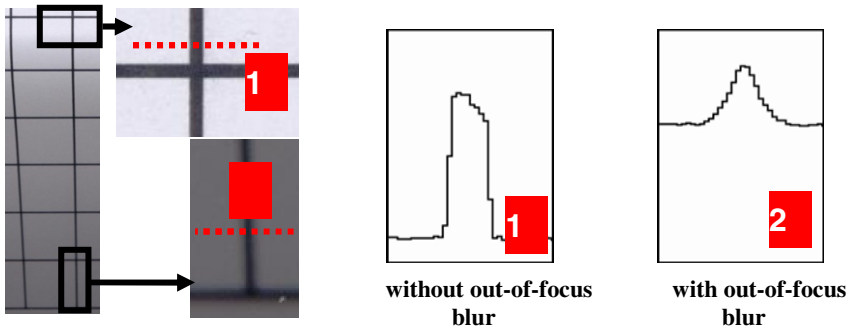


Fig. 4. Line grid scanned with HP ScanJet 5300c with $\psi = 0^\circ$ and $\theta = 45^\circ$

In Figures 3 and 4, as the paper is further away from the scanner flatbed, the blur increases and illumination is fades out; as the scanning device is calibrated to digitize documents at a pre-defined distance, which is exactly the flatbed surface.

These figures also present several cross sections at different parts of the calibration grid images. They show regions without blur (cross sections “1” and “3”) and regions with blur kernel size larger than the grid thickness. The line labeled with number “4” is the limit when is not possible to remove the blur totally.

In the case of characters, corners of the strokes are vulnerable regions to the blur. The kernel area in this region is dominated by the information not related to the given point. Figure 5 shows two kinds of corners in the letter “M” that can be irrecoverable.

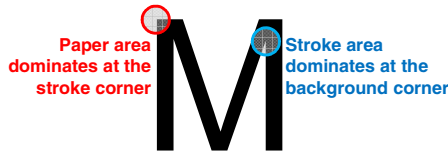


Fig. 5. Corners of upper case “M” which are vulnerable to the blur noise

2.1 Reconstruction Function

Thoulin and Chang [21] identify document background and foreground locally for the resolution expansion of document images. Proposed method obtains these colors by searching the maximum and minimum on the pixel neighborhood and uses it into the reconstruction function.

Most pixels that belong to the paper background have their intensities values closer to the background intensity. Similarly, for blurred stokes values are closer to the foreground intensity. In this way an S-function can be built, with input and output varying from 0 to 1, whereas the output is below the line of the identity function between 0 and 0.5, and above it between 0.5 and 1.0.

In this work the function $S(t)$ is defined by equation 1 with the fixed parameter p that varies between 0 and 1.0, which controls how strong the correction will be. For p values closer to 0, the function shape looks similar to a step function with higher transitions; for values closer to 1.0, the shape gets closer to a *sin* function scaled by π . Figure 6 shows two plots for $p = 0.06$ and $p = 1.00$.

$$S(t) = 0.5 - 0.5 \times \text{sign}(\cos t\pi) \times |\cos t\pi|^p \quad (1)$$

To apply the *S-shape* function, two reference values must be determined for the paper background and character stroke. This is done by looking out in a window for the pixel with largest and lowest intensity. The un-blurred value is obtained by eq. (2), where I_b is the blurred intensity value (i.e. original image); *min* and *max* are the lowest and the highest intensity values in the given window, respectively.

$$I_u = \left[S\left(\frac{I_b - \text{min}}{\text{max} - \text{min}}\right) \times (\text{max} - \text{min}) \right] + \text{min} \quad (2)$$

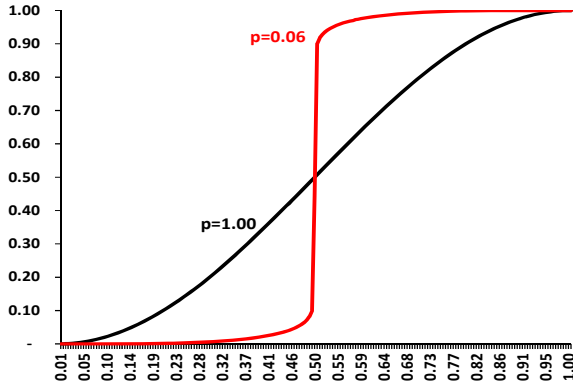


Fig. 6. S-function plots for p equal to 1.00 and 0.06

One may notice that using large windows and low values for p is more intrusive than the other way round. The choice of this parameter will depend on the blur level.

3 Results and Analysis

The evaluation of the proposed algorithm is done in three parts. The first part uses computer generated images to obtain blurred and un-blurred images; the latter is used as the ground-truth. At the second part, a study is done in scanned images using the elevated plane model [1]. At last, processing results of real documents are presented.

Unfortunately, comparing the method proposed here with other algorithms described in the literature was not possible as they do not offer enough details for granting their implementation, their executable code is not available, and the evaluation datasets used by them do not include document images.

3.1 Computer Generated Images

In order to provide objective quality measures of the proposed algorithm processing, letters and chess shape were generated. Several levels of blur were applied in these images using the GaussianBlur filter available in ImageJ [20]; two examples are presented in Figure 7. The Gaussian radius represents the region with 61% of the whole Gaussian, thus the kernel size is not restricted to this radius [20].

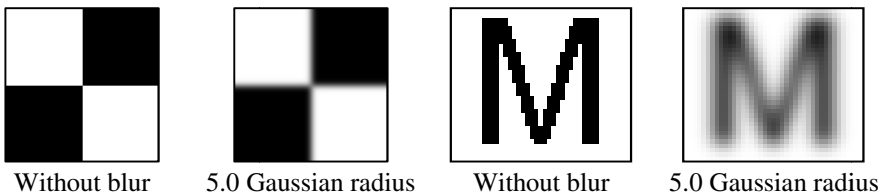


Fig. 7. Blur applied to chess pattern image and upper case “M”

The de-blurring method proposed herein was applied to the blurred images. Images corrected and without blur were compared using SSIM [19] index. It measures how images are perceptually different with values between 0 (different) and 1 (the same).

Tables 2 and 3 provide at each cell average SSIM between the ground truth and the de-blurred images of the chess and letters pattern, respectively. Cells in the same row represent the de-blurring of the same input but with different window sizes. The values in parenthesis are the amount of information “gain” computed by eq. (3). Table 1 shows the values of the average SSIM between de-blurred and without blur images. Value 0.1 was set to p in all tests. One may notice that the proposed algorithm could improve the SSIM metric by 17.08% at most.

$$\text{SSIM gain} = \frac{\text{Average of SSIM between **deblurred** and without blur images}}{\text{Average of SSIM between **blurred** and without blur images}} - 1 \quad (3)$$

Table 1. Average of SSIM between blurred and without blur images

Blur kernel size	1	2	3	4	5
Average of Chess pattern SSIM	0.999	0.965	0.926	0.888	0.852
Average of Letters SSIM	0.999	0.976	0.954	0.932	0.910

Table 2. Average of SSIM between de-blurred and without blur chess pattern images

5	0.923 (+8.41%)	0.970 (+13.94%)	0.989 (+16.16%)	0.996 (+16.96%)	0.997 (+17.08%)
4	0.961 (+8.25%)	0.991 (+11.68%)	0.997 (+12.36%)	0.998 (+12.47%)	0.999 (+12.50%)
3	0.988 (+6.75%)	0.998 (+7.86%)	0.999 (+7.96%)	0.999 (+7.96%)	0.999 (+7.96%)
2	0.999 (+3.55%)	1.000 (+3.64%)	1.000 (+3.64%)	1.000 (+3.64%)	1.000 (+3.64%)
1	1.000 (+0.08%)	1.000 (+0.08%)	1.000 (+0.08%)	1.000 (+0.08%)	1.000 (+0.08%)
Blur Win	1	2	3	4	5

Table 3. Average of SSIM between de-blurred and without blur letters pattern images

5	0.942 (+3.56%)	0.947 (+4.06%)	0.940 (+3.30%)	0.941 (+3.38%)	0.942 (+3.53%)
4	0.968 (+3.85%)	0.973 (+4.37%)	0.970 (+4.09%)	0.971 (+4.17%)	0.972 (+4.27%)
3	0.987 (+3.44%)	0.991 (+3.85%)	0.991 (+3.82%)	0.990 (+3.78%)	0.991 (+3.80%)
2	0.999 (+2.30%)	0.998 (+2.25%)	0.998 (+2.23%)	0.998 (+2.23%)	0.998 (+2.22%)

Table 3. (Continued)

1	1.000 (+0.07%)	1.000 (+0.07%)	1.000 (+0.07%)	1.000 (+0.07%)	1.000 (+0.07%)
Blur	1	2	3	4	5
Win					

Table 4. Letters with various blur sizes applied with different window sizes

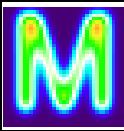
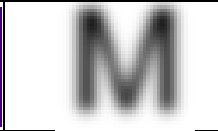
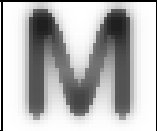


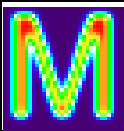
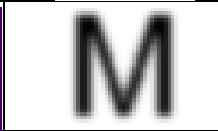
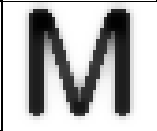
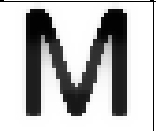
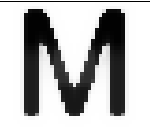
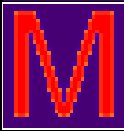
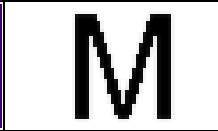
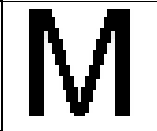
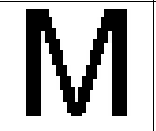
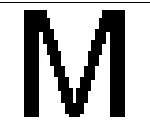
					
					
					
Thermal map	Blur	Win.	1	3	5

Table 4 shows some results of processing the examples used for Table 3. The first column shows the blurred images thermal map for each blur level, with red representing the stroke color (black) and blue the background (white). The second column show the blurred images used as inputs.

Images with the Gaussian radius set to 1, is possible to recover most of the stroke. The thermal map of blurred “M” with 3.0 radius shows that only the upper part of the stroke is red, thus only it can be totally recoverable. Although, the corrected images appear to be restored properly as the blurred pixels are too close to black.

The same is not observed to the correct images of radius set to 5. The thermal map shows that most part of the stroke are green, showing that the blur affected the character structure. The quality improvement is observable in all images on Table 4.

3.2 Blurred Images Generated from Scanned Images

The dataset presented on previous section has noise-free images. In order to evaluate the proposed algorithm with more realistic scenarios, 3 documents were digitized by a scanner flatbed yielding into images with low noise level, which is considered as the ground truth. Analogically to previous section, the blurred images and the effectiveness of the correction were obtained. Table 5 shows the SSIM performance results.

Table 5. Average of SSIM between de-blurred and scanned images

5	0.898 (+3.02%)	0.899 (+3.05%)	0.895 (+2.62%)	0.893 (+2.36%)	0.890 (+2.12%)
4	0.924 (+2.52%)	0.920 (+2.05%)	0.916 (+1.61%)	0.914 (+1.37%)	0.912 (+1.16%)
3	0.944 (+1.32%)	0.938 (+0.58%)	0.934 (+0.18%)	0.931 (-0.10%)	0.929 (-0.37%)
2	0.960 (-0.25%)	0.952 (-1.11%)	0.946 (-1.67%)	0.942 (-2.15%)	0.938 (-2.58%)
1	0.969 (-2.94%)	0.946 (-5.27%)	0.928 (-7.04%)	0.915 (-8.40%)	0.905 (-9.42%)
Blur Win	1	2	3	4	5

The scanned images present a low degree of blur; thus the similarities between the scanned and weakly blurred images are too high. Therefore, the gains of the SSIM quality are negative for the images with 1-3 radiuses on Table 5.

The de-blurred image provides image with sharper edges than the ground truth and input images, thus it has better visual quality than both of these images. Figure 8.a shows the scanned image; a zoomed part can be seen on (b); (c) shows the blur level 1 applied to (b) with the de-blurred version on (d). It is noticeable that Figure 8.d is more pleasant to see than Figures 8.b and 8.c.

For the strong blur (4 and 5 radius), the SSIM gain is positive; this shows the proposed algorithm improves the quality of the image. One can be seen on the de-blurred version of Figure 8.e (8.b blur level 5) on Figure 8.f.

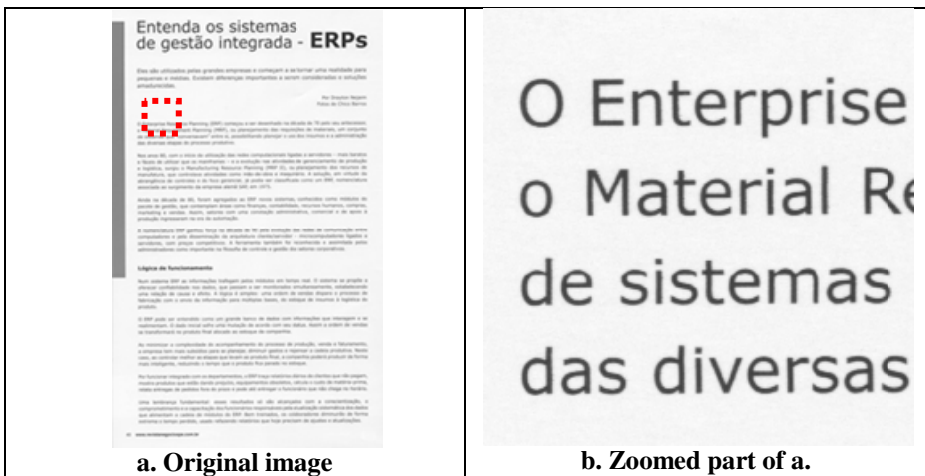


Fig. 8. Example of processing the scanned images

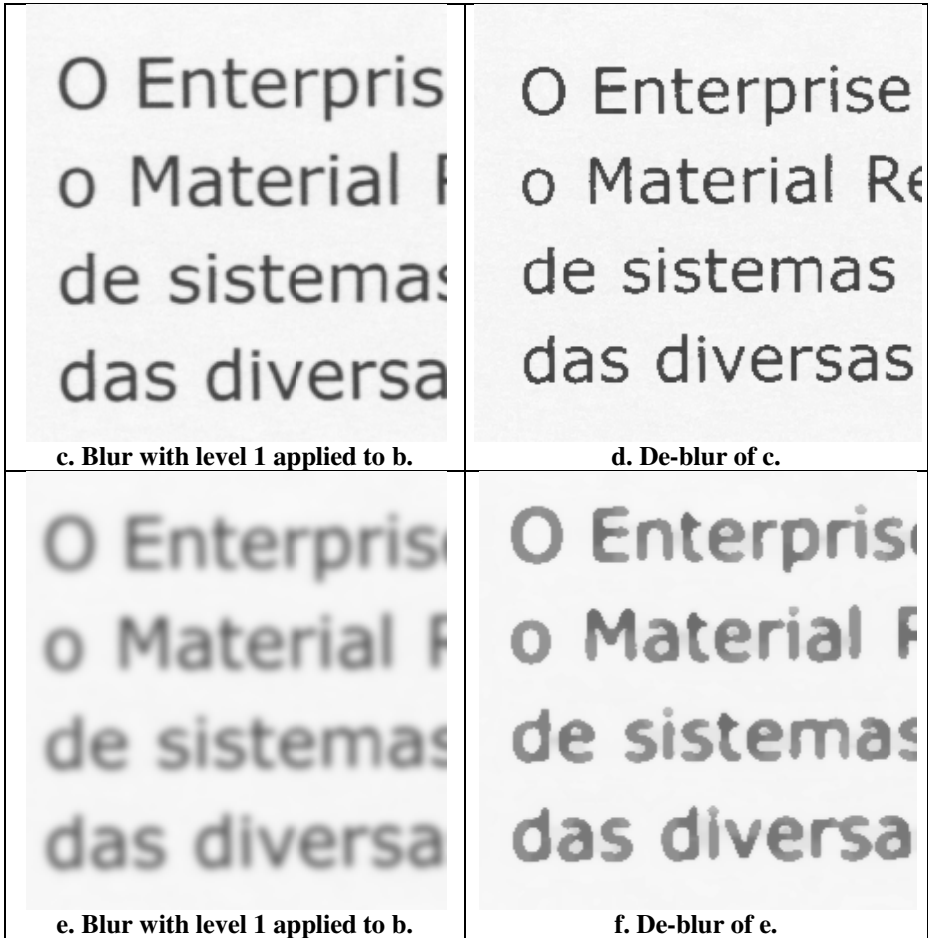


Fig. 8. (Continued)

3.3 Elevated Plane Image Analysis

The images obtained with the elevated plane model [1] shows the blurred stroke combined with other noises. Figure 9 shows the results of applying the proposed algorithm to Figure 3 using a 7×7 window for p equal to 1.00 (a) and 0.25 (b). One may observe that vertical line grid was recovered until blur kernel got larger than a 7×7 window (dashed rectangle of 8.b); although the blurred horizontal line on the bottom part could be partially recovered. Increasing the window size is possible to recover the area were the blur is larger, which is shown in Figure 9 with windows of sizes 11×11 (c) and 19×19 (d) for $p = 0.25$.

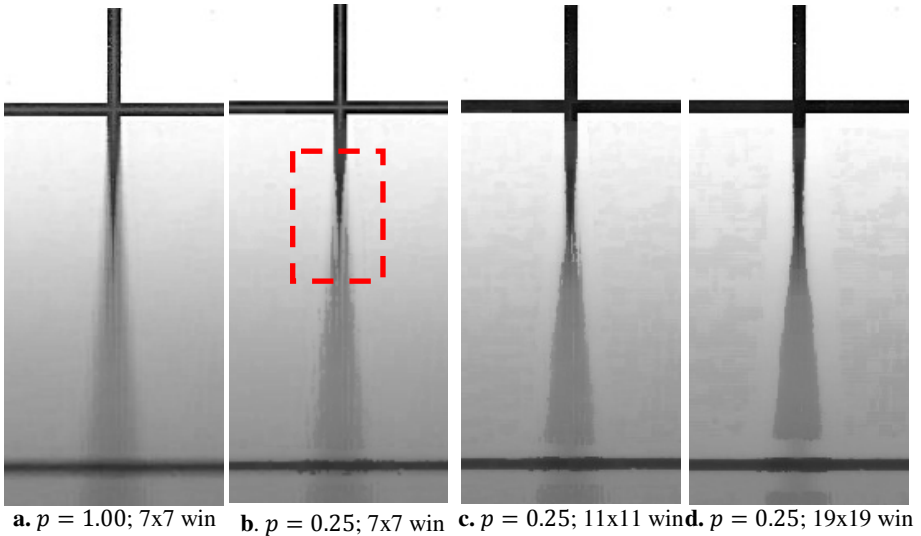


Fig. 9. Output with 7×7 window

3.4 Document Images

Finally, some examples of de-blurring applied to real document images are presented. Figures 10 and 11 show the output at different window sizes and values of p . The resulting image was improved by the proposed algorithm.

4 Conclusions

The study performed here shows that focusing the scope of the application of a de-blurring algorithm stands a better chance of more adequately and efficiently removing such complex noise that may appear due to several different sources: physical, digitization, filtering and storage. This paper presents an algorithm to compensate the digitization non-constant blur that appear in scanning bound grayscale documents, for instance. The algorithm performance was confirmed by high values of SSIM metric in computer generated images.

The automatic inference of the parameters of the algorithm through the use of blur intensity classifier such as the one described in reference [8] is under implementation and shows some promising results already.

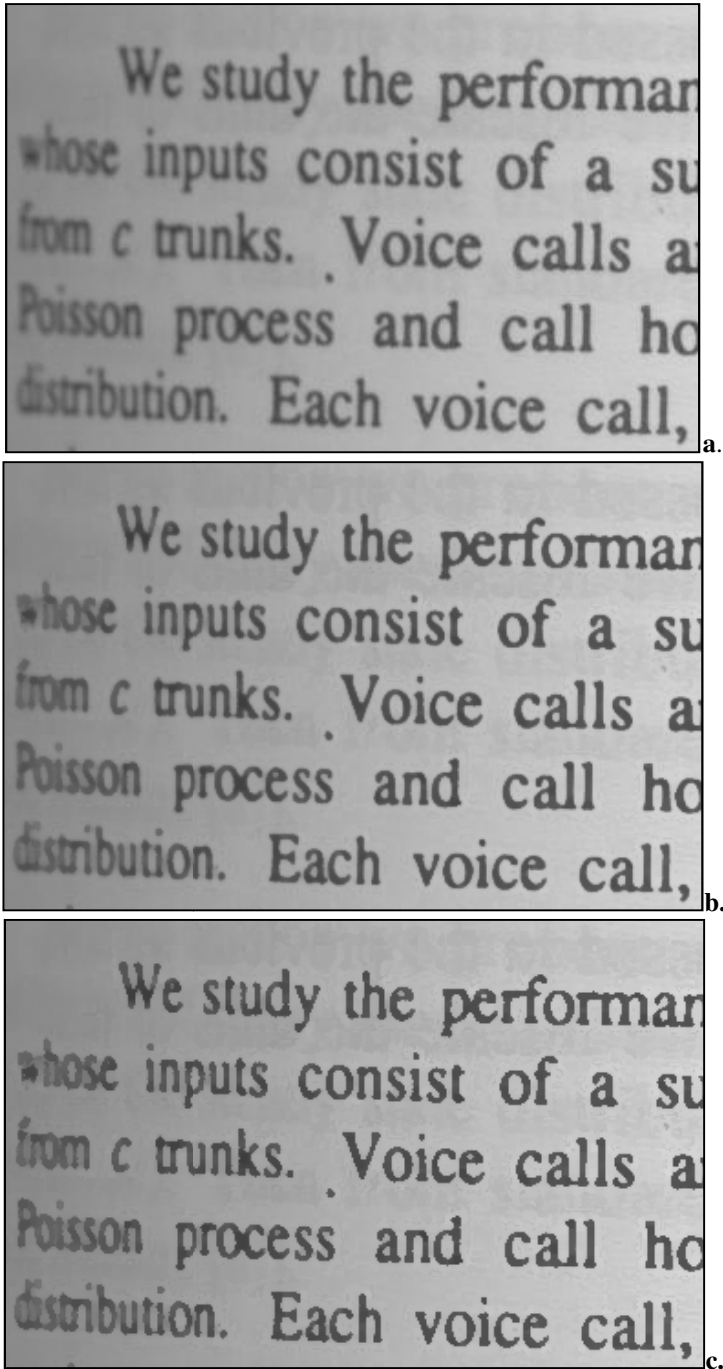


Fig. 10. Result with 5x5 window: original image (a); $p = 0.50$ (b); $p = 0.06$ (c)

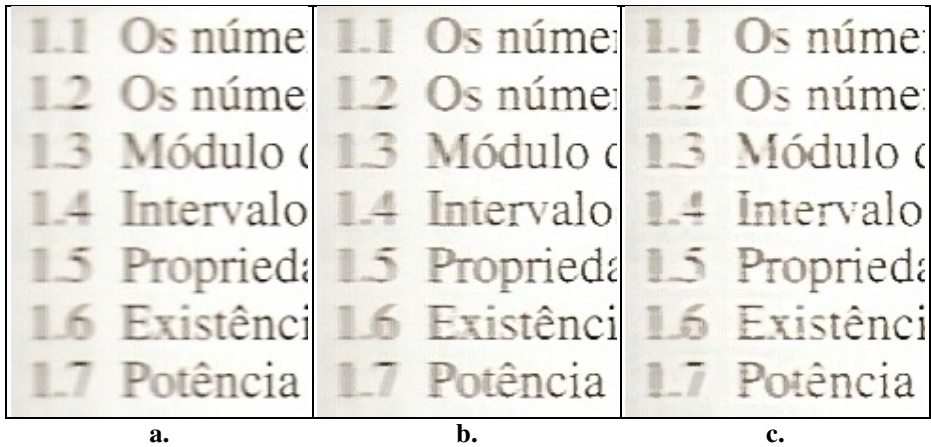


Fig. 11. Result with $p=0.5$: original image (a); 3x3 window (b); 7x7 window (c)

Acknowledgements. The research reported herein was partly sponsored by CNPq and MCT-Lei de Informática grants both from the Brazilian Government.

References

- [1] Ukida, H., Konishi, K.: 3D Shape Reconstruction Using Three Light Sources in Image Scanner. *IEICE Trans. on Inf. & Syst.* E84-D(12), 1713–1721 (2001)
- [2] Demoment, G.: Image reconstruction and restoration: Overview of common estimation structures and problems. *IEEE Transactions on Acoustics, Speech, & Signal Processing* 37(12), 2024–2036 (1989)
- [3] Neelamani, R., Choi, H., Baraniuk, R.G.: Wavelet-based deconvolution for ill-conditioned systems. In: *Proc. of IEEE ICASSP*, vol. 6, pp. 3241–3244 (1999)
- [4] Chambolle, A., Lions, P.L.: Image recovery via total variation minimization and related problems. *Numerische Mathematik* 76(2), 167–188 (1997)
- [5] Rudin, L.I., Osher, S., Fatemi, E.: Nonlinear total variation based noise removal algorithms. *Physica D* 60, 259–268 (1992)
- [6] Roth, S., Black, M.J.: Fields of experts: a framework for learning image priors. *CVPR* 2, 860–867 (2005)
- [7] Joshi, N.S.: Enhancing photographs using content-specific image priors. Phd thesis, University of California, San Diego (2008)
- [8] Lins, R.D., Silva, G.F.P., Banergee, S., Kuchibhotla, A., Thielo, M.: Automatically Detecting and Classifying Noises in Document Images. In: *ACM-SAC 2010*, vol. 1, pp. 33–39. ACM Press (March 2010)
- [9] Lins, R.D.: A Taxonomy for Noise in Images of Paper Documents - The Physical Noises. In: Kamel, M., Campilho, A. (eds.) *ICIAR 2009*. LNCS, vol. 5627, pp. 844–854. Springer, Heidelberg (2009)
- [10] Lins, R.D., Oliveira, D.M., Torreão, G., Fan, J., Thielo, M.: Correcting Book Binding Distortion in Scanned Documents. In: Campilho, A., Kamel, M. (eds.) *ICIAR 2010*, Part II. LNCS, vol. 6112, pp. 355–365. Springer, Heidelberg (2010)

- [11] Chang, M.M., Tekalp, A.M., Erdem, A.T.: Blur identification using the bispectrum. *IEEE Trans. Signal Process.* 39(10), 2323–2325 (1991)
- [12] Mayntz, C., Aach, T., Kunz, D.: Blur identification using a spectral inertia tensor and spectral zeros. In: *Proc. of IEEE ICIP* (1999)
- [13] Cannon, M.: Blind deconvolution of spatially invariant image blurs with phase. *IEEE Trans. Acoust. Speech Signal Process.* 24(1), 56–63 (1976)
- [14] Biemond, J., Lagendijk, R.L., Mersereau, R.M.: Iterative methods for image de-blurring. *Proc. of the IEEE*, 856–883 (1990)
- [15] Rekleities, I.M.: Optical flow recognition from the power spectrum of a single blurred image. In: *Proc. of IEEE ICIP* (1996)
- [16] Moghaddam, M.E., Jamzad, M.: Motion blur identification in noisy images using fuzzy sets. In: *Proc. IEEE ISSPIT, Athens* (2005)
- [17] Lokhande, R., Arya, K.V., Gupta, P.: Identification of parameters and restoration of motion blurred images. In: *ACM-SAC 2006, Dijon* (2006)
- [18] Jain, A.K.: *Fundamentals of digital image processing*. Prentice-Hall, Inc., Upper Saddle River (1989)
- [19] Wang, Z., Bovik, A.C., Sheikh, H.R.: Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing* 13(4), 600–612 (2004)
- [20] ImageJ. GaussianBlur (ImageJ API), <http://rsbweb.nih.gov/ij/developer/api/ij/plugin/filter/GaussianBlur.html>
- [21] Thouin, P.D., Chang, C.I.: A method for restoration of low-resolution document images. In: *IJDAR* (2000)