A New Deblurring Algorithm for Textual Document Images

Daniel M. Oliveira, Rafael Lins, Gabriel P. Silva
Departamento de Eletrônica e Sistemas - UFPE
Recife - Brazil
{daniel.moliveira, rdl, gabriel.psilva}@ufpe.br

Jian Fan[1], Marcelo Thielo[2]
Hewlett-Parckard Labs.
[1]Palo Alto - USA, [2]Porto Alegre - Brazil
{jian.fan, marcelo.resende.thielo}@hp.com

Abstract—**Documents digitalized by portable cameras or flatbed scanners may exhibit some blurred areas. Most deblurring algorithms are hard to implement and slow. Often they try to solve the problem for 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 deblurring of textual documents; there is no need to estimate the PSF and the filter can be directed applied to the image.**

Deblurring; blur; camera documents; scanner documents;

I. INTRODUCTION

The recent paper [9] presents a taxonomy for noises in document images and, besides providing an explanation of how such noise appeared in the final image, may provide pointers to the literature that show ways of avoiding or removing it. Noise is defined here as any phenomenon that degrades document information. In the classification proposed [9], there are four kinds of noise:

1. *The physical noise – whatever "damages" the physical integrity and readability of the original information of a document. The physical noise may be further split into the two sub-categories proposed in as internal and external.*
2. *The* digitalization *noise – the noise introduced by the digitalization process. Several problems may be clustered in this group such as:* inadequate digitalization 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). The introduction of colors not originally present in the document due to arithmetic manipulation or overflow is an example of such a noise.*
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.*

Blur is the effect of unsharpening images, which may fit 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 digitalization, due to several reasons: non-flat objects, digitalization errors, out of focus, motion etc. The presence of blur may be an indicator of low quality digitalization, 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 deblurring 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 times the computational complexity is prohibitively high and can yield undesirable artifacts such as ringing [7] as presented in Figure 1.

Figure 1. Ringing artifact [7]

The most successful approaches to blur removal point at focusing 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 digitalization noise.

In this paper, to increase the chances of better deblurring, the application domain is restricted to monochromatic scanned documents in which the blur is a digitalization noise originated from the unsuitable document placing on the

scanner flatbed due to a number of factors, one of which is book binding warping [10]. 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.

II. The New Method

A. Blur effect

The study performed here on the compensation of the blur effect was done for scanned images. Patterns were arranged in an elevated plane model [1] as illustrated in Figure 2 with $\psi = 0$. A HP PhotoSmart C4280 and HP ScanJet 5300c scanners were used. Some of the results are shown in Figure 3, where one may observe that in this case blur kernel can be approximated to a 1D horizontal function.

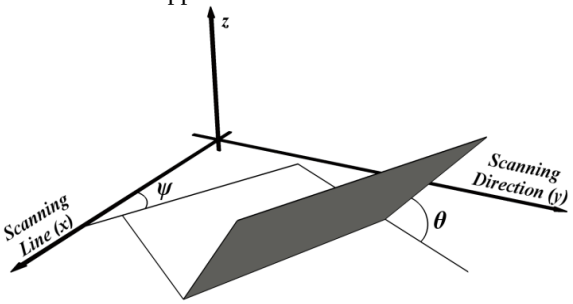
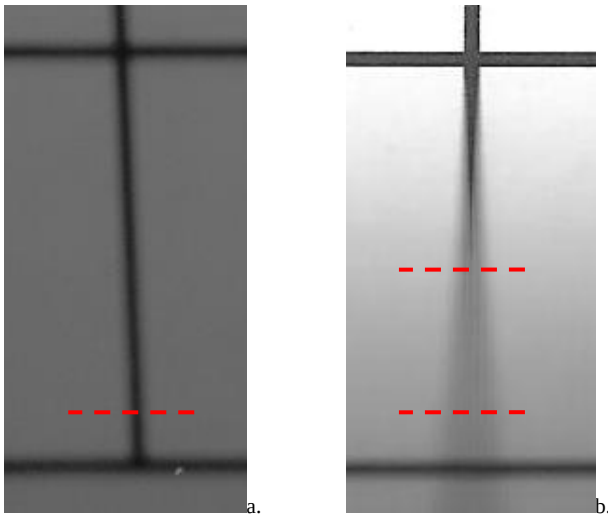


Figure 2. Elevated plane [1]



a. b.

Figure 3. Grid image: ScanJet, $\theta = 45°$ (a), PhotoSmart C4280, $\theta = 30°$ (b)

In both images shown in Figure 3, as the paper is farther away from the scanner flatbed the blur increases and illumination is weaker; this happen due to the scanning device being calibrated to digitalize documents at a pre-defined distance that is exactly the flatbed surface.

Figure 4 shows cross sections of the image of Figure 3 on an area with and without blur. One may observe that the signal in part 4.e the blur kernel is greater than the grid thickness, in this case is not possible to recover the unblurred signal. Besides that, the blur is stronger in the HP Photosmart than in the ScanJet as the first is a "All-in-one" device (printer, fax and scanner) and has less space to place the components whereas the latter is a single purpose scanner.

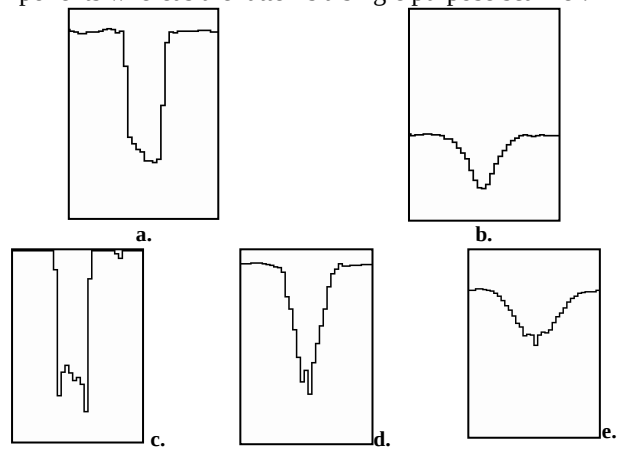


Figure 4. Cross section signals: Figure 3.a unblurred (a) and blurred (b); Figure 3.b unblurred (c), blur kenel close to limit (d) and totally blurred (e)

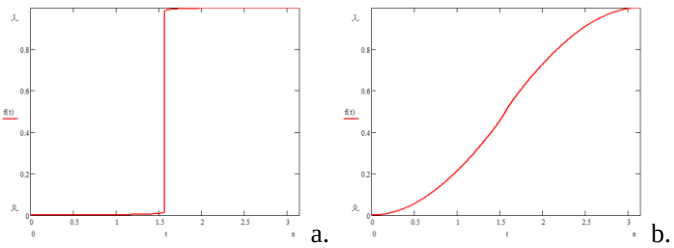
B. Reconstruction function

Figure 4 shows that if a signal has a blur kernel smaller than the grid thickness it is possible to improve or even recover the un-blurred signal. In the case of characters, the scenario is similar in a smaller scale.

For the values that belong to the paper background, the blurred intensity value is closer to the unblurred ones. In the same way, for blurred stokes values are closer to undistorted part. In this way a 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.

This work proposes function $S(t)$ 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 5 shows two plots for $p = 0.06$ and $p = 1.00$.

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



a. b.

Figure 5. S-function plots: $p = 0.06$ (a); $p = 1.0$ (b)

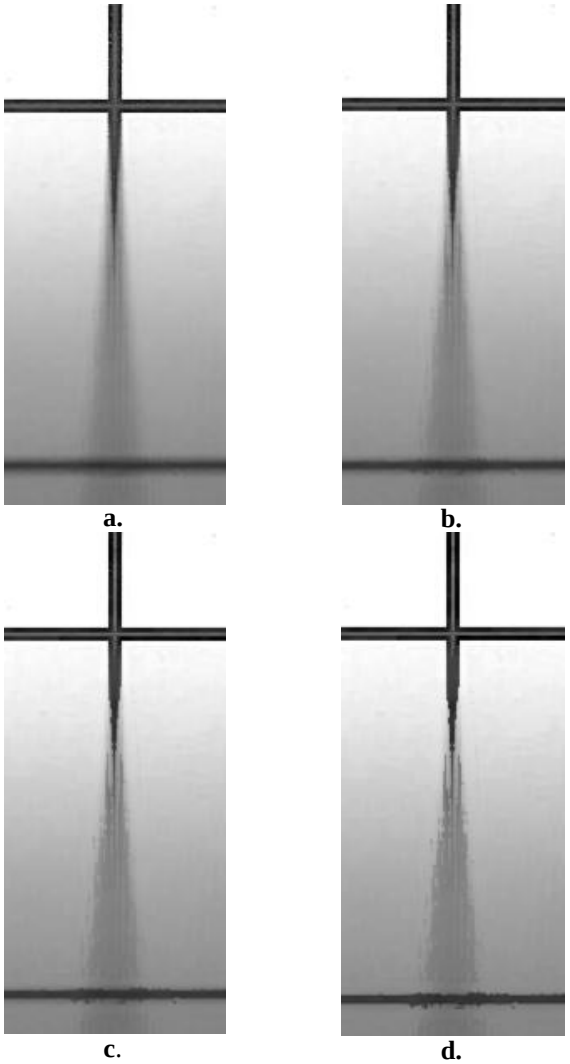
To apply the S-shape function, two reference values must be determined for the paper background and character stroke. This is done by looking out in a window for the pixel with largest and lowest intensity, then to obtain the unblurred

value eq. (2) is applied, 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.

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

III. RESULTS AND ANALYSIS

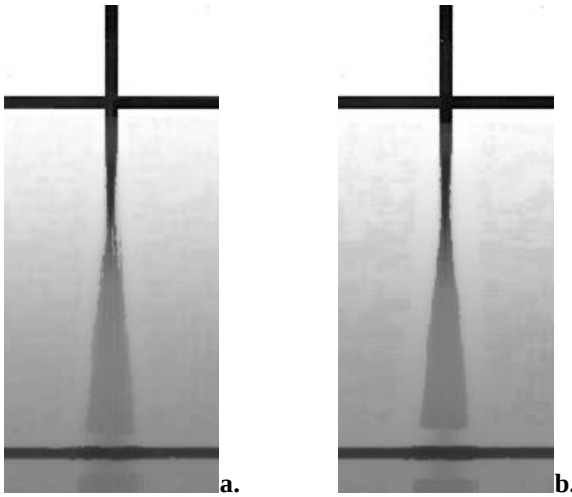
Figure 6 shows the results of applying the proposed algorithm to Figure 3.b using a 7x7 window. One may observe that vertical line grid was recovered until blur kernel got larger than a 7x7 window; although the blurred horizontal line on the bottom part could be partially recovered.



a. **b.** **c.** **d.**

Figure 6. Results for a 7x7 window: $p = 1.00$ (a); $p = 0.5$ (b); $p = 0.25$ (c); $p = 0.06$ (d)

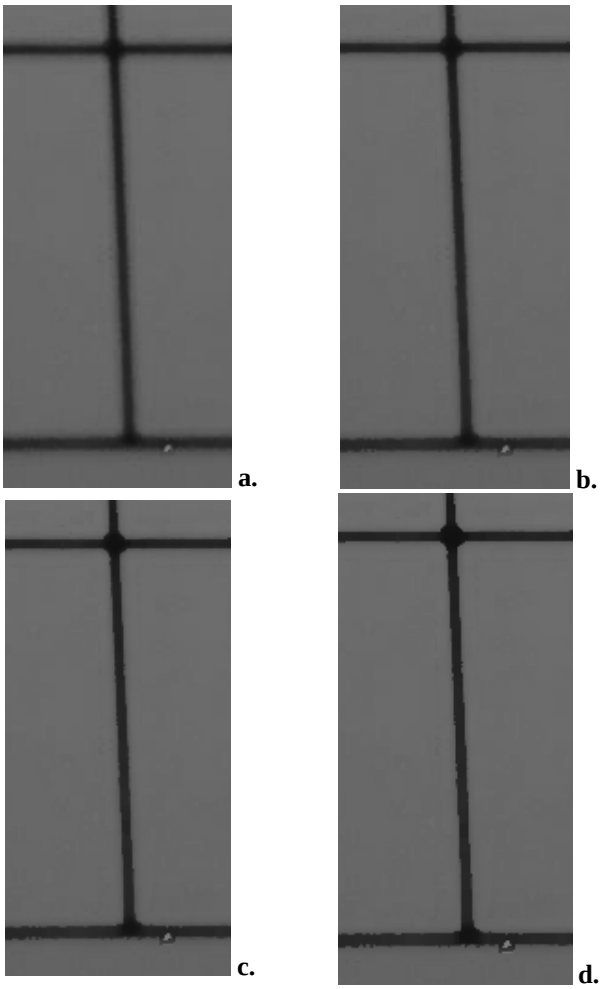
Increasing the window size is possible to recorver the area were the blur is larger, which is shown in Figure 7 with windows of sizes 11x11 and 19x19 and $p = 0.25$.



a. **b.**

Figure 7. Results with $p = 0.25$: 11x11 window (a); 19x19 window (b);

For figure 2.a, in which the blur is weaker than in 2.b, Figure 8 provides the results for same parameters. One may note that satisfactory results were obtained for $p = 0.5$ with a 7x7 window.



a. **b.** **c.** **d.**

Figure 8. Results for 7x7 window: $p = 1.00$ (a); $p = 0.5$ (b); $p = 0.25$ (c); $p = 0.06$ (d)

Finally, Figure 9 shows some examples of de-blurring applied to real document images.

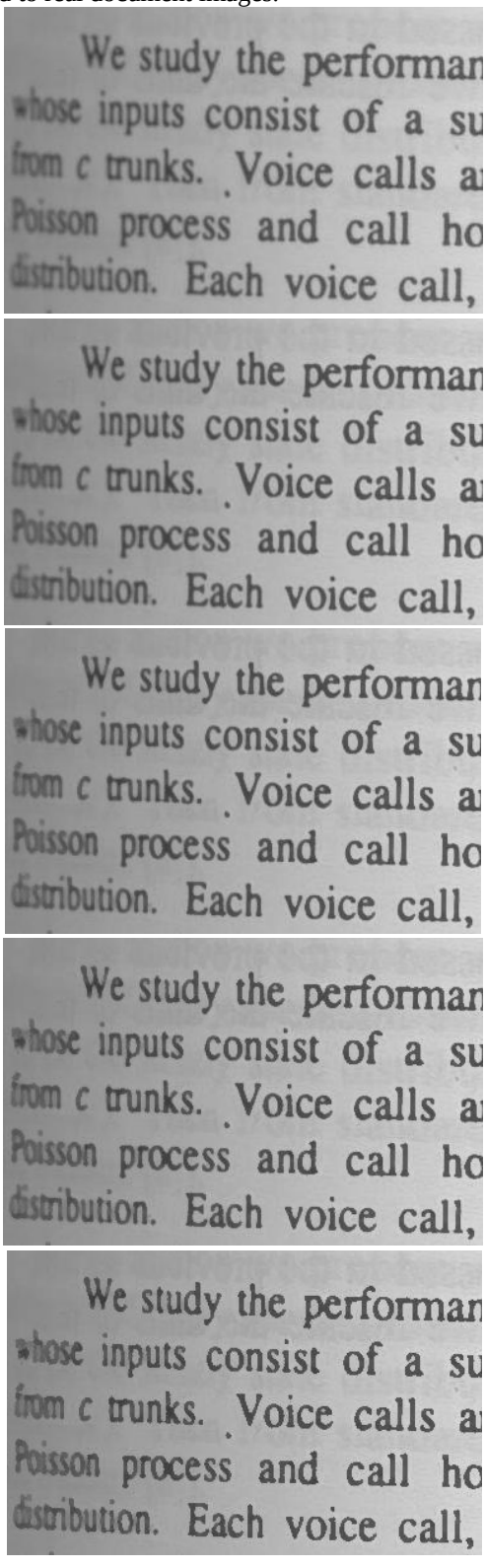


Figure 9. Document processed with 5x5 window: original image (a); $p = 1.0$ (b); $p = 0.50$ (c); $p = 0.25$ (d); $p = 0.06$ (e)

IV. CONCLUSIONS

The study performed here shows that focusing the scope of application of deblurring algorithms stand a better chance of better solving such a complex noise that may appear due to several sources: physical, digitalization, filtering and storage. This paper presented an algorithm to compensate the digitalization blur in scanned monochromatic documents with satisfactory results.

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.

REFERENCES

[1] Ukida, H. and Konishi, K. . 3D Shape Reconstruction Using Three Light Sources in Image Scanner. IEICE Trans. on Inf. & Syst., Vol.E84-D, No. 12, pp.1713-1721, Dec. 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., and Baraniuk, R. G.. Wavelet-based deconvolution for ill-conditioned systems. Proc. of IEEE ICASSP, Vol. 6, pp. 3241–3244, March 1999.

[4] Chambolle, A., and 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., and Fatemi, E.. Nonlinear total variation based noise removal algorithms. Physica D, 60, 259–268, 1992.

[6] Roth, S., and Black, M. J.. Fields of experts: a framework for learning image priors. CVPR, Vol. 2, pp. 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. and Thielo, M. Automatically Detecting and Classifying Noises in Document Images, ACM-SAC'2010, ACM Press, v.1. p.33 – 39, March 2010.

[9] Lins, R.D. A Taxonomy for Noise Detection in Images of Paper Documents - The Physical Noises. ICIAR 2009. LNCS v. 5627. p. 844-854, Springer Verlag, 2009.

[10] Lins, R.D., Oliveira, D. M., Torreão, G., Fan, J., and Thielo, M., Correcting Book Binding Distortion in Scanned Documents. ICIAR 2010. LNCS 6112, pp. 398-408, Springer Verlag, 2010.

[11] Chang, M. M., Tekalp, A. M., and Erdem, A. T., Blur identification using the bispectrum, *IEEE Trans. Signal Process.*, Vol. 39, No. 10, 1991, pp. 2323-2325.

[12] Mayntz, C., Aach T., and Kunz D., Blur identification using a spectral inertia tensor and spectral zeros, *Proc. of IEEE ICIP*, 1999.

[13] Cannon M., Blind deconvolution of spatially invariant image blurs with phase, *IEEE Trans. Acoust. Speech Signal Process.*, Vol. 24, No. 1, 1976, pp. 56-63.

[14] Biemond, J., Lagendijk, R. L., and Mersereau R. M., Iterative methods for image deblurring, *Proc. of the IEEE*, 1990, pp. 856-883.

[15] Rekleities, I. M., Optical flow recognition from the power spectrum of a single blurred image, *Proc. of IEEE ICIP*, 1996.

[16] Moghaddam, M.E. and Jamzad, M., Motion blur identification in noisy images using fuzzy sets, *Proc. IEEE ISSPIT*, Athens, 2005.

[17] Lokhande, R., Arya, K.V., Gupta, P. Identification of parameters and restoration of motion blurred images, ACM-SAC'2006, Dijon, 2006.

[18] Jain, A.K. , Fundamentals of digital image processing, Prentice-Hall, Inc., Upper Saddle River, NJ, 1989.