A character degradation model for grayscale ancient document images

V.C.Kieu[1], M.Visani[2], N.Journet[1], J.P.Domenger[1], R.Mullot[2]
[1]*Laboratoire Bordelais de Recherche en Informatique - LaBRI, University of Bordeaux I, France*
[2]*Laboratoire Informatique, Image et Interaction - L3i, University of La Rochelle, France*
{vkieu, journet, domenger}@labri.fr, {muriel.visani, remy.mullot}@univ-lr.fr

Abstract

Kanungo noise model is widely used to test the robustness of different binary document image analysis methods towards noise. This model only works with binary images while most document images are in grayscale. Because binarizing a document image might degrade its contents and lead to a loss of information, more and more researchers are currently focusing on segmentation-free methods (Angelika et al [2]). Thus, we propose a local noise model for grayscale images. Its main principle is to locally degrade the image in the neighbourhoods of "seed-points" selected close to the character boundary. These points define the center of "noise regions". The pixel values inside the noise region are modified by a Gaussian random distribution to make the final result more realistic. While Kanungo noise models scanning artifacts, our model simulates degradations due to the age of the document itself and printing/writing process such as ink splotches, white specks or streaks. It is very easy for users to parameterize and create a set of benchmark databases with an increasing level of noise. These databases will further be used to test the robustness of different grayscale document image analysis methods (i.e. text line segmentation, OCR, handwriting recognition).

1. Introduction

In this paper, we propose a local noise model to generate synthetically most common defects observed in real old document images (see Figure 1). The main interest of this model is to allow a creation of a large number of grayscale images with different degradation levels, and use them for benchmarking and comparing the robustness of different grayscale document analysis methods (i.e. text line segmentation, OCR, handwriting recognition…). Our model allows to control easily the intensity of the degradation.

Many degradation models have previously been proposed. Loce *et al* [6] proposed a perturbation model of the print reflectance modulation resulting from scanner mechanical disturbances. In [1], based on the physics of the image acquisition process, a ten parameter model for character degradation is discussed. In [3], a document degradation model which simulates 4 types of noise is proposed by Zhai *et al.* This model is used for testing the robustness of line detection algorithms. A model, based on an adaptation of a bleedthrough restoration method detailed in [7], is also presented in [8]. This model is used to compare the robustness of two OCR algorithms when the bleedthrough intensity is increasing.

Our new degradation model is able to simulate the most common degradations due to the age of the document itself and printing/writing process, such as ink splotches, white specks or streaks. The physical process of printing consisted of soaking wood-character stamps in ink and pressing it on the base (sheet of paper, parchment…). If stamps or base are not in a good condition or the quality of ink is not good, ink-specks might appear (see the first letter "a", "b", and "s" in Figure 1) or a letter looks "broken" (see the second letter "a" in Figure 1). Sometimes, both defects appear together (see the letter "u" in Figure 1). This kind of noise mostly appears in the neighbourhood of the characters.

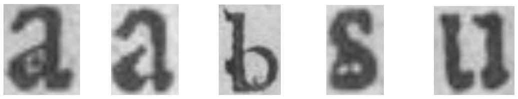

Figure 1. *Examples of defects in real ancient document*

To select the centers of the degraded region in the neighbourhood of the characters, we use the non-linear local selection presented by Kanungo *et al.* in [5], as a part of a more general local degradation model for binary images which was further validated in [4]. Dedicated to binary images, this degradation consists in adding "salt and pepper" noise, i.e. flipping pixels from background to foreground and vice-versa, in the neighbourhood of the characters. A flipping probability is firstly computed as a function of the pixel distance to the foreground boundary. The random perturbation process then flips some pixel's values (foreground pixel(1) changes to background(0) and vice versa). Finally, a morphological closing operation is applied for correlating the flipped pixels. Kanungo noise model works well with binary images and is widely used to simulate the presence of noise for assessing the performances of different document analysis methods, such as the feature extraction [9] or symbol recognition contests[1] [2].

This paper is organized as follows: in Section 2, our model is specified. Experimental results for visual validation and OCR robustness tests are given in Section 3 while conclusions and future extensions of this work are provided in Section 4.

2. Grayscale noise model

Our model, described in Figure 2, is composed of a seed-point selection (which relies on binarized images because seed points are more likely to be selected close to the characters), a noise region definition, and a noise generation process (which is applied directly to the grayscale images).

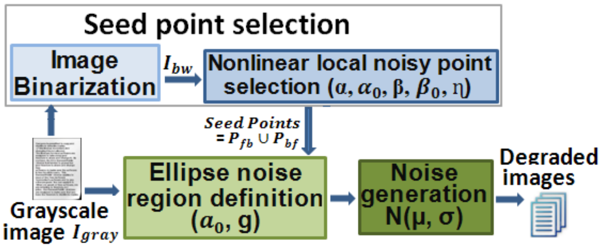
Seed point selection
Image Binarization
I_{bw}
Nonlinear local noisy point selection (α, α_0, β, β_0, η)
Seed Points = $P_{fb} \cup P_{bf}$
Ellipse noise region definition (a_0, g)
Noise generation N(μ, σ)
Degraded images
Grayscale image I_{gray}

Figure 2. *Grayscale noise model*

2.1 Seed-point selection

As the result of random perturbation process in Kanungo noise model, the inverted pixels (from foreground to background and vice versa) are called *seed-points* (they are the centers of the noise regions that we add to original grayscale images). To compute these points, we use a binarization method such as Otsu to get a binarized version I_{bw} of I_{gray}. I_{bw} is the input of nonlinear local noisy point selection (see Figure 2). Consequently, the computation provides two sets of seed-point that we use as input of our noise generation process. The first set (P_{fb}) represents the center of each future degraded area where ink will disappear. The second one (P_{bf}) represents the center of each future degraded area where ink will appear. The five parameters in the Kanungo noise model, $\theta = (\alpha_0, \alpha, \beta_0, \beta, \eta)$ (see [4] for more details) are used for controlling the amount of generated seed-points.

2.2 Noise region definition

Each seed-point is a center of a noise region. For a center C_i of noise region in I_{gray} and $C_i \in P_{fb} \cup P_{bf}$, the local gradient vector is calculated from the grayscale levels, shown as in Figure 3-b. The size and dimension of the noise region are related to the local gradient vector at the center. The shape of a noise region is considered as a parameter of our model. For example, in Figure 3-c, the elliptic noise region is given with 6 points ($|P_{fb}| = 4, |P_{bf}| = 2$). Let v be the gradient value at C_i and V be the maximum gradient value of all seed-points. Let a be the semi-major axis and b be the semi-minor axis of an ellipse. We define $a = a_0 * (1 + \frac{v}{V})$ where a_0 is an input parameter controlling the size of the noise regions, and $b = a * (1 - g)$ where g is the flattening factor ($0 \leq g \leq 1$), an input parameter controlling the "flatness" of the noise regions.

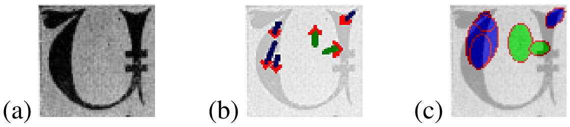

Figure 3. *(a) Original image, (b) Local gradient vector, (c) Ellipse noise regions (a_0 = 5, g=0.3) of 6 points ($|P_{fb}|, |P_{bf}|$) = (4, 2).*

2.3 Noise generation process

In a noise region, the grayscale value of pixels is changed in order to obtain a degraded region similar to one observed in a real old document. For example, we have the ellipse noise region given in Figure 4. Let $\overline{c_i}$ be the gray value at the center C_i. $\overline{c_i}$ is set to the average grayscale value of all background pixels if $C_i \in P_{fb}$ and to the average of all foreground pixels if $C_i \in P_{bf}$. For each pixel B_j at the edge of ellipse, the average value $\overline{b_j}$ of its 8-neighbours (in the initial grayscale image) is used for calculating values of all pixels in the line C_iB_j. Let us consider any point P_k belonging to the segment $[C_i, B_j]$. Let d_{ij} be distance C_iB_j and d_{ik} be

[1] www.cvc.uab.es/grec2003/SymRecContest/index.htm.
[2] www.cs.cityu.edu.hk/grec2005.

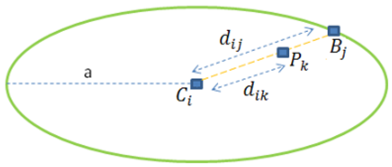

Figure 4. *Ellipse noise region*

distance C_iP_k. To produce a noise as realistic as possible, we introduce a random number generation function satisfying the normal distribution $N(\mu, \sigma^2)$ for generating the new grayscale value of pixel P_k in segment $[C_i, B_j]$. The standard deviation σ is an input parameter of our model whereas the mean μ of this function is calculated as in (1):

$$\mu = \overline{c_i} + (\overline{b_j} - \overline{c_i}) * (\frac{d_{ik}}{d_{ij}}) \qquad (1)$$

3. Experimental results

3.1 Visual validation of the model

For generating degraded images, it is possible to tune kanungo parameters $(\alpha_0, \alpha, \beta_0, \beta, \eta)$ in order to change the number of seed-points $(P_{fb} \cup P_{bf})$. On the other hand, users can only define the size of the elliptic noise region (a_0, g). In Figure 5, the ellipse size $(a_0 = 5, g = 0.3)$ and initial parameters $(\alpha_0 = \beta_0 = 1, \eta = 0, \sigma = 0.6)$ are fixed while α and β are progressively decreased. Obviously, when these two parameters decrease, light specks (noise regions with $C_i \in P_{fb}$) appear and break the connectivity of characters. Also, more dark specks (noise regions with $C_i \in P_{bf}$) appear and make fat or connected characters. Figure 5-d illustrates the importance of not generating too many seed-points. The visual result might look "too much" synthetic.

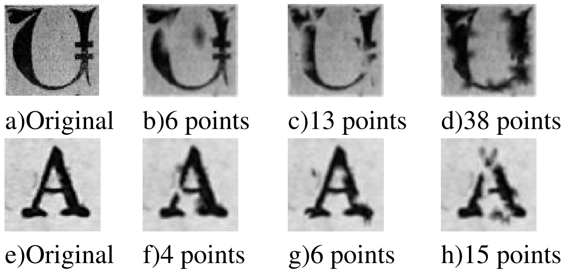
a)Original b)6 points c)13 points d)38 points
e)Original f)4 points g)6 points h)15 points

Figure 5. *Degraded images with the size of elliptic region fixed ($a_0 = 5, g = 0.3$) and $\sigma = 0.6$.*

Figure 6 shows several generated images with a fixed number of seed-points and an increasing size of ellipse (a_0). In Figure 6-b, c, f, and g, the stroke becomes thinner but is not broken with $a_0 \leq 4$. However, with $a_0 = 5$ the stroke of character is completely broken in Figure 6-d and h. In conclusion, the more a number of seed-points increases or the larger degrading region becomes, the more degraded an image will be. Except the case in which too many seed-points are generated, image looks really similar to the real degraded one (more results on[3]).

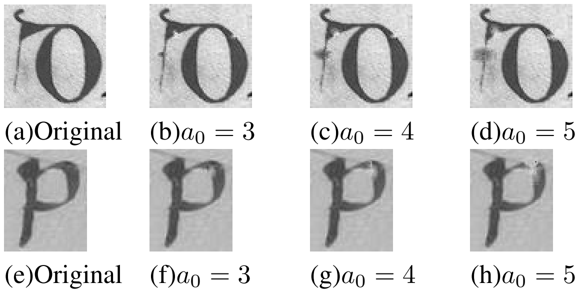
(a)Original (b)$a_0 = 3$ (c)$a_0 = 4$ (d)$a_0 = 5$
(e)Original (f)$a_0 = 3$ (g)$a_0 = 4$ (h)$a_0 = 5$

Figure 6. *Degraded images with the seed-points fixed ($|P_{fb}|, |P_{bf}|$) = (2, 1) for "d", ($|P_{fb}|, |P_{bf}|$) = (1, 0) for "p"), and $g = 0.3, \sigma = 0.6$.*

3.2 OCR robustness regarding our generated character defects

The performance of OCR algorithms decreases when the quantity of character defects increases. It is the reason why we expect to verify if our noise model has a coherent influence on OCR error rate. We perform two tests with OCR algorithms (which are implemented in the OCRopus software). For each generated image, the text is extracted by OCRopus. The Levenshtein distance is then used to calculate the difference between the extracted text in degraded image and the original one. The further the distance is, the higher the error rate is. In the first test, the database is divided into 8 categories of defect intensity. In one category, 10 images are created with the same parameter values. From category 1 to category 8, the number of seed-points increases (by decreasing the α and β parameters) while the size of noise region is fixed. As shows in Figure 7, the Levenshtein distance increases exponentially when the number of seed-points increases. In each defect category, the number of seed-points of each image is different due to a pseudo-random function in the seed-points selection step. Box-plots show that this function has no influence on the increment of the error rate.

In the second test, we generate a database of 7 degraded images in which the size (a_0) of a noise region

[3] http://www.labri.fr/perso/vkieu/content/demo.html

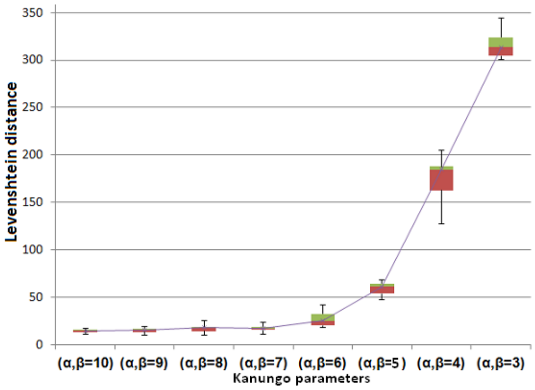

Figure 7. *OCRopus error rate evolution for an increasing number of seed-points*

increases progressively whereas the others are fixed. Figure 8 shows that for a fixed number of seed-points, if the size of an ellipse increases the error rate increases. These tests show that OCR algorithm error rate gradually increases when characters becomes more degraded as in the reality.

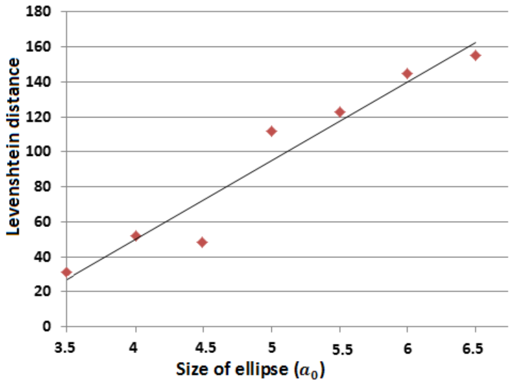

Figure 8. *OCRopus error rate evolution for an increasing size of noise region*

4. Conclusion

In this paper, we present a local noise model adapted to grayscale ancient document images. Instead of flipping a value of points between foreground and background on binary images (like Kanungo's model), our model generates grayscale noise regions defined as ellipses, where the level of degradation of each pixel is computed by a Gaussian random function. This model is not difficult to parametrize, so it might be widely applied to generate benchmark databases of different levels of degradation for assessing or comparing the robustness of different image analysis methods towards noise. Obtained images are very realistic and some of the most common defects observed in real ancient document images are mimicked. These defects might break the connectivity of strokes or, inversely, increase the connectivity. The evolution of the recognition rates of a state-of-the-art OCR is coherent with different levels of degradation that we generate. More work is needed to formally validate the model and estimate the parameters. The model will soon be integrated in a software dedicated to a semi-synthetic old document image generation.

Acknowledgements

This research is done within the DIGIDOC project financed by the ANR (Agence Nationale de la Recherche).

References

[1] H. S. Baird. *Structured Document Image Analyse*, chapter Document image defect models, pages 546–556. Springer-Verlag, New York, USA, 1992.

[2] A. Garz, A. Fischer, R. Sablatnig, and H. Bunke. Binarization-free text line segmentation for historical documents based on interest point clustering. In *IAPR Int. Workshop on Doc. Anal. Systems*, pages 95–99, Gold Coast, Australia, 2012.

[3] D. D. Jian Zhai, Liu Wenyin and Q. Li. A line drawings degradation model for performance characterization. In *Proc. of Seventh ICDAR*, pages 1020–1024, Edinburgh, Scotland, August 2003. IEEE Computer Society.

[4] T. Kanungo, R. Haralick, H. Baird, W. Stuezle, and D. Madigan. A statistical, nonparametric methodology for document degradation model validation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 22(11):1209 – 1223, 2000.

[5] T. Kanungo, R. M. Haralick, and I. Phillips. Global and local document degradation models. In *Proc. of the ICDAR*, pages 730–734, Tsukuba Science City, Japan, Oct. 1993.

[6] R. Loce and W. Lama. Halftone banding due to vibrations in a xerographic image bar printer. *Journal of Imaging Technology*, 16(1):6–11, 1990.

[7] C. M. Moghaddam R.F. Low quality document image modeling and enhancement. In *Int. J.Doc. Anal. Recognit*, volume 11, pages 183–201, Berlin, Heidelberg, March 2009. Springer.

[8] V. Rabeux, N. Journet, and P. Domenger. Document recto-verso registration using a dynamic time warping algorithm. In *ICDAR*, pages 1230–1234, Beijing, China, November, 2011.

[9] M. Visani, O. R. Terrades, and S. Tabbone. A protocol to characterize the descriptive power and the complementarity of shape descriptors. *Int. J. Doc. Anal. Recognit.*, 14(1):87–100, Mar. 2011.