A Character Degradation Model for Color Document Images

Do Thi Luyen, Elodie Carel, Jean-Marc Ogier, Jean-Christophe Burie

A Character Degradation Model for Color Document Images

Do Thi Luyen, Elodie Carel, Jean-Marc Ogier and Jean-Christophe Burie
L3i, University of La Rochelle
La Rochelle, France
Email: luyendtit@gmail.com, {elodie.carel, jean-marc.ogier, jean-christophe.burie}@univ-lr.fr

***Abstract*—Degradation models are widely used to create large datasets with ground-truth for designing optimal denoising algorithms, and to assess the performance of different document analysis and recognition methods (OCR, segmentation, and so on). Since the processing of document images is usually performed in grayscale or in black and white, some degradation models have been proposed in order to simulate noise on these specific images. However, there is a growing interest in the study of color document images. In this context, there is a real need of tools able to simulate the effects of noise on color for an evaluation purpose. In this paper, we propose to extend a model from the state-of-the-art to color document images. Our model includes three main steps. First, seed-points are selected depending on color information. Then, they are associated to different kinds of noise in order to simulate the degradation effect occurring on different content. Last, the values of pixels located inside an elliptic region around the seed points are modified in order to obtain degraded regions.**

I. Introduction

Corruptions of document images usually result from various types of noise during document generation and copying processes. Even if a region looks homogeneous, colorimetric variations appear at a pixel-level. Noise alters colors. False colors and artifacts appear on the images and more specifically along the edges. This can affect the whole image processing flow performed on the images. For instance, for a segmentation task, noise can lead to an over-segmentation. Fig. 1 shows an example of a magnified portion of a noisy image. Small square blocks occur in the image. Discontinuities located at adjacent blocks are visible. The image also exhibits pixels with red color around the edges of the character.

To reduce the degradation effects, denoising approaches can be applied on the images. However, the choice of a method and of its parameters depend on the quality of the images to be processed. In this context, it is really important to be able to evaluate the performance of an image denoising algorithm, and there is a real need in noise models. Knowing the noise model can enable us to design algorithms for restoring degraded documents. Besides, by simulating the noise, we can automatically generate big datasets of images with ground-truth. Numerous image degradation models have been proposed previously. Kanungo et al. [1] proposed a statistical model applied on binary images in order to simulate degradations that occur when documents are printed, scanned, and digitized. The state of the art of some document image degradation models can be found in [2]. Recently, Kieu et al. [3], [4] have proposed a new degradation model. This model is able to create gray-level defects such as dark specks near characters or ink discontinuities. The images generated by this model satisfy the user's wishes and look realistic.

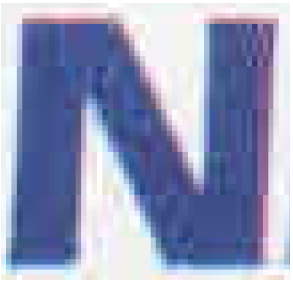


Fig. 1. An example of noise in a real color image

The models described in the above papers have been designed for binary and grayscale images. But they cannot be used for color images while almost all images today are in color. The main difficulty raised when dealing with color values rather than grayscale values. Color can be processed by considering each channel one by one independently but in this case, the correlation between the channels is lost. The second possibility is to process color as a vector, but some properties such as the total order relationship is not verified anymore. Some classical image processing approaches, used with gray level images, cannot be applied with color images. In this paper, we present a model that can be used to generate degradations on color document images by respecting their vectorial representation (at the opposite of marginal one). This work aims at creating a large number of color images with different degradation levels for the evaluation purpose.

This paper is organized as follows: Section 2 specifies our new degradation model for color document images. Experimental results are given in Section 3. Finally, conclusion are drawn in Section 4.

II. Color noise model

In this section, our noise model dedicated to color document images is described. We aim at modelling real noise that appear in the neighborhood of the characters by defining seed-points as the centers of the regions on which our model will be applied. The difficulty is to find relevant seed-points in regards to the reality. Hence we are going to present in the paper three categories of seed-points. The three main steps of this model are: 1) Seed-point selection; 2) Noise region classification; 3) Noise generation.

The first step concerns the seed-point position selection. We aim at studying the color information directly. Main color layers of the documents are extracted as a set of binary images in which all pixels belonging to the same original color are set to the black value and the rest of the pixels are considered as a white background [5]. An adaptive Kanungo noise model [1] is applied on each layer in order to generate seed-points. These points are used as input of the noise generation process. Second, if we process all the seed-points in the same manner, this may create the same kind of noise. So, the idea is to classify noise into different classes to reproduce a noise as realistic as possible. In this method, we consider three noise types that usually appear in the neighborhood of the characters: independent spot, overlapping spot, and disconnection spot (see section B for more details). A noise region has a center which is a seed-point belonging to a layer. The size and dimension of a noise region depend on the gradient vector at its center. As our approach relies on color information, we decided to use a color gradient which considers the color as a vector. We used the multidimensional gradient defined by Di Zenzo [6], which permits to respect the vectorial representation of the color. The assignation of a type of noise for each seed-point depends on the content of the surrounding area. Finally, the noise is generated. In the noise regions, the values of pixels are modified in order to obtain degraded regions. The Gaussian blurring is applied on the noise region in order to make the two adjacent pixel values more coherent.

A. *Seed-point selection*

In this step, we aim at selecting the seed-points which will be used to generate noise regions. First of all, main color layers of the documents are extracted as a set of binary images in which all pixels belonging to the same original color is set to black in a white background image [5] (see Fig. 2). Seed-points are more likely to be selected close to the characters. So the background layer is not considered. An adaptive Kanungo noise model is applied on each layer in order to select the centers of the degraded region in the neighborhood of the characters. The output of this process is a set of inverted pixels (from foreground to background and vice versa) which are called seed-points. Fig. 3 illustrates the position of seed-points in three layers of a zoomed part of image in Fig. 7. Seed-points are denoted by red color. Let d be the distance transform of a pixel P(x, y). The probability of flipping the pixel P(x, y) is:

$$p = \begin{cases} \alpha_0 \times e^{-\alpha d^2} + \eta_1 & \text{if P is foreground} \\ \beta_0 \times e^{-\beta d^2} + \eta_0 & \text{if P is background} \end{cases} \quad (1)$$

where η_0 is the constant probability of flipping for all background pixels, η_1 is the constant probability of flipping for all

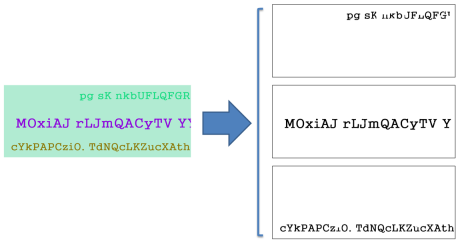


Fig. 2. An example of color layers

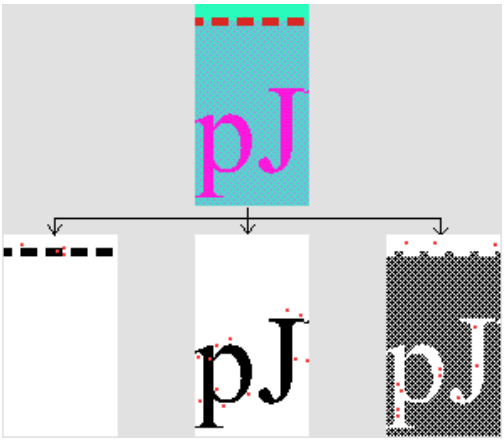


Fig. 3. An example of seed-points in three main color layers of the image (red points are seed-points)

foreground pixels, α_0 and α control the flipping probabilities of the background pixels, α_0 is the initial value for the exponential, α controls the decay speed of the exponential, β_0 and β control the flipping probabilities of the foreground pixels.

A random number r is generated by a number generator with a uniform distribution. A pixel P will be a seed-point if its probability p is greater than or equal to the random number r. As a result, this process will produce two sets of seed-points in each layer, P_{fb} and P_{bf}. P_{fb} is the center of each future background spot and P_{bf} is the center of each future foreground spot. These points are used as input of noise generation process.

B. *Noise region classification*

At this point, we have selected a set of points used as seeds to generate noise. Noise does not alter all the regions of the images in the same way depending on their features. For example, ink drop-out may generate characters with many disconnections; or ink spots can create noise between characters, words or lines. In this model, we consider three noise types: independent spot, overlapping spot, and disconnection spot. With a given number of independent spot, overlapping spot, and disconnection spot chosen by the user depending on the application, a heuristic rule is used to classify each seed-point to one of three types. This classification is applied on layers of images separately.

The three types of noise are defined as follows:

- The independent spots are regions where noise will appear inside or outside of characters.
- The overlapping spots overlap characters. They will introduce changes at the location of the edges of characters.
- The disconnection spots break the edges (the connectivity) of characters.

Fig. 4 shows examples of the three types of noise in real color document images.

(a)

(b) (c)

Fig. 4. Examples of the three types of noise in real color document images (a) independent spot, (b) overlapping spot, (c) disconnection spot

From the definition of three types of noise above, their classification will be implemented by considering the elliptic noise regions, the distance between seed-points and the edges of characters. The major axis of the elliptic noise region (see Fig. 5, 6) is defined by the gradient vector at its center and the parameter a_0, where a_0 is an input parameter controlling the size of the noise regions. The minor axis depends on the size of the major axis and of the flattening factor g of ellipse. g controls the flatness of the ellipse (the noise region). To benefit from the color information, Di Zenzo's multiscale gradient method [6] is used to define the size and dimension of the noise region.

Let a be the semi-major axis, and b be the semi-minor axis of an ellipse. The two values a and b are calculated as follows:

$$a = a_0 * (1 + \frac{v}{V}), \tag{2}$$

where v is the gradient value at C, and V is the maximum gradient value of all seed-points.

$$b = a * (1 - g), \ (0 \leq g \leq 1) \tag{3}$$

From (2), the size of the major axis of the spot, a_i, is calculated. Let a_{01i} be the distance between the center of the spot and the nearest edge of the stroke; a_{02i} be the distance between the center of the spot and the second edge of the stroke (see Fig. 5). Thus, the type of each seed-point is assigned depending on two thresholds a_{01i} and a_{02i}.

- If one spot is totally outside the character or one spot is totally inside the edges of stroke ($0 < a_i < a_{01i}$), it will be an independent spot.
- If $a_{01i} \leq a_i \leq a_{02i}$, it is classified as an overlapping spot.
- If $a_i > a_{02i}$, and the center of the spot belongs to the future background spot (P_{fb}), it will be a disconnection spot.

Another problem arising is that if too many large noise regions are generated, it might generate non-realistic images. So, a heuristic rule is used in order to assign a type of noise for each seed-point based on its two thresholds. Let N_{sp} be a set of seed-points. N_{is}, N_{os}, N_{ds} is percentage of independent spots, overlapping spots, disconnection spots (given by the user), respectively. Each seed-point has the two values a_{01}, a_{02}. The procedure to assign the type of seed-point is as follows:

- Choose N_{ds} seed-points where a_{02} is minimal.
- Choose N_{os} seed-points where a_{01} is minimal.
- The rest of the seed-points (N_{is}) will be independent spots.

C. Noise generation

In this step, the values of pixels inside the noise regions are modified in order to obtain degraded regions. The size of the ellipse depends on the type of the associated noise which has been determined as explained in the previous section. We aim at simulating the different effects that the noise can have on a region depending on its content.

Let a_i, a_j, a_k be the size of the major axis of an elliptic independent spot, overlapping spot, and disconnection spot region, respectively. We have:

$$\begin{cases} a_i = a_{01i} \times \mu_i, 0 < \mu_i < 1 \\ a_j = a_{01j} + \mu_j \times (a_{02j} - a_{01j}), 0 < \mu_j < 1 \\ a_k = a_{02k} + \mu_k \times \delta, 0 < \mu_k < 1 \end{cases} \tag{4}$$

where μ_i, μ_j, μ_k are randomly generated, $(0 < \mu_i, \mu_j, \mu_k < 1)$.

The minor axis of an elliptic independent spot, overlapping spot, and disconnection spot region is calculated using (3). The flattening factor is randomly chosen (g $\in [0, 1]$).

Since the sizes of elliptic noise regions are set, the values of pixels inside the noise regions are modified in order to obtain degraded regions. Let $\bar{c_i}(r, g, b)$ be the color value at the center C_i of an ellipse (Fig. 6). Considering the layer l containing C_i, $\bar{c_i}$ is set to the average color values of all background pixels of layer l if $C_i \in P_{fb}$ and to the average of all foreground pixels of j if $C_i \in P_{bf}$. Let B_j be a pixel at the edge of ellipse, $\bar{b_j}(r, g, b)$ be the color value at B_j. $\bar{b_j}$ is calculated as follows:

Considering the table I, let $(L_1^\star, a_1^\star, b_1^\star)$ be the value at I(x,y), $(L_2^\star, a_2^\star, b_2^\star)$ be the value at I(x+i,y+j) in $L^\star a^\star b^\star$ color space, (i, j = {-1, 0, 1}). The Euclidean distance between two pixels is:

$$d_{ij} = \sqrt{(L_1^\star - L_2^\star)^2 + (a_1^\star - a_2^\star)^2 + (b_1^\star - b_2^\star)^2} \tag{5}$$

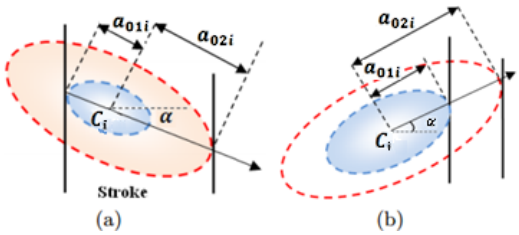


Fig. 5. Noise region in two cases: (a) spot inside a character. (b) spot outside and close to a character

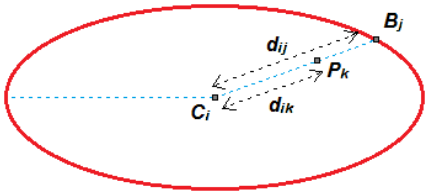


Fig. 6. Elliptic noise region

TABLE I. MASK FOR CALCULATING VALUES AT EDGE OF ELLIPSE

I(x-1, y-1)	I(x, y-1)	I(x+1, y-1)
I(x-1, y)	$B_j(x, y) \equiv$ I(x, y)	I(x+1, y)
I(x-1, y+1)	I(x, y+1)	I(x+1, y+1)

Then $\bar{b}_j$ is calculated as the value of one pixel in the 8-neighboring of B_j where, in $L^\star a^\star b^\star$ space, its distance to B_j corresponds to the median value of distances between B_j and each of 8-neighboring of B_j. $\bar{b}_j$ is used for calculating values of all pixels inside the line C_iB_j of the ellipse.

The new value p_k at pixel P_k in the line C_iB_j is calculated using Gaussian random distribution to make the final result more realistic.

$$p_k = N(\mu, \sigma^2) \tag{6}$$

where the standard deviation σ is the input parameter and the mean μ is calculated as:

$$\mu = \bar{c}_i + (\bar{b}_j - \bar{c}_i) * (\frac{d_{ik}}{d_{ij}}) \tag{7}$$

where d_{ik} is the distance between pixel P_k and the center of the ellipse, C_i; d_{ij} is the distance between C_i and B_j.

At the end, a Gaussian blurring is applied on the noise region in order to make the two adjacent pixel values more coherent, and to obtain a generated noise more realistic.

D. Noise level

The seed-points are likely selected in the neighborhood of the characters. So, the level of noise on the image depends on the number of characters appearing in the document image. This value is usually close to the number of connected components of the image. Let N_{sp} is the number of seed-points, N_{cc} is the number of connected components. The number of seed-points of the image is given by:

$$N_{sp} = l * N_{cc} \tag{8}$$

where l is used to control the level of noise on the image. If l is too large, it will generate too many seed-points, the result might look too much synthetic.

III. EXPERIMENTAL RESULTS

The experimentations have been done on synthetic images. They are high resolution and high quality color images of size 2480 x 3508 (width x height) pixels. These images will be used to generate noisy images. Fig. 7 shows an example of these images.

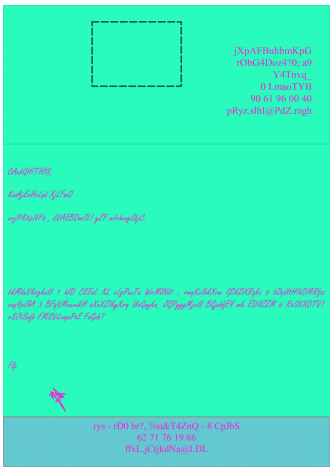
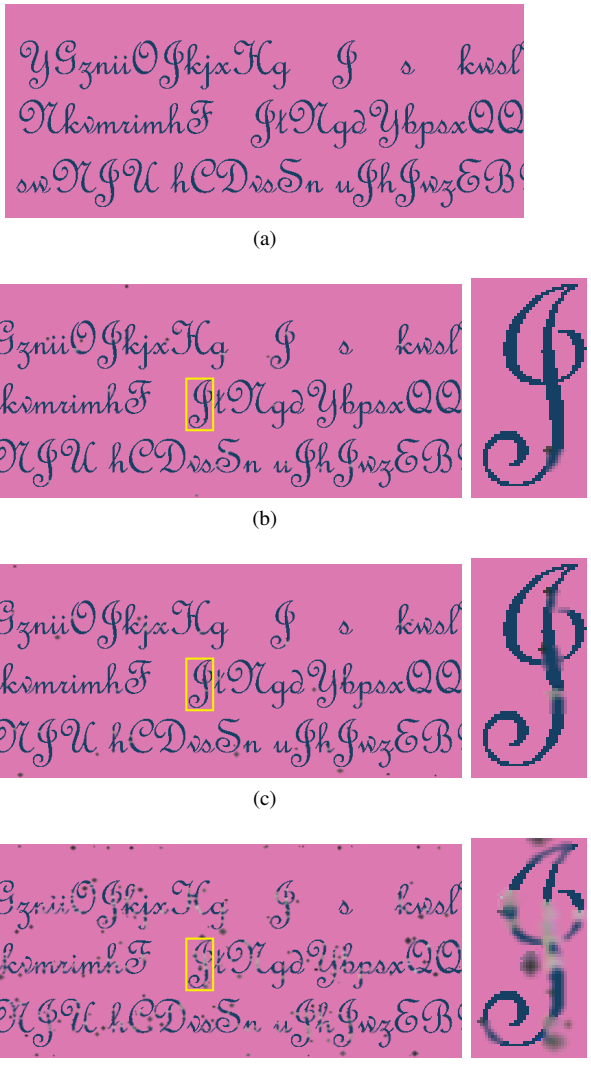


Fig. 7. An example of synthetic image

A. Test image with different number of seed-points

Let's take the parameters of the model as follows: the number of seed-points is successively equal to 112, 192, and 1616. The proportion of the three types of degradation is equal to 15% for independent spot, 60% for overlapping spot, and 25% for disconnection spot. The percentages have been chosen experimentally but with the purpose of being close to the reality. Fig. 8-b,c,d provide the degraded images of the original one in Fig. 8-a with the different number of seed-points (N_{sp}). In Fig. 8-b,c,d, the right images represent the zoomed parts of the left images.



(a)

(b)

(c)

(d)

Fig. 8. The original image and noisy images (a) Original image (b) $N_{sp} = 112$ (c) $N_{sp} = 192$ (d) $N_{sp} = 1616$

Parameters of the model are tuned depending on the application and on the amount of noise that is needed on

the images. Visually, the results look similar to what we can observe on real documents. As it can be seen from the Fig. 8, the increasing of the number of seed-points leads to the increasing of degradation level of the image. In Fig. 8-b,c, the images seem similar to the real document images. Nevertheless, Fig. 8-d visually looks too much synthetic due to the large number of seed-points. So, it is very important to find the proper parameters so as to get more realistic images.

B. Visualization of the effects of the different types of noise taken separately

In this test, the image is tested with three options: the image is degraded by generating only independent spots, overlapping spots, and disconnection spots, respectively.

Fig. 9-b shows the generation of noise regions near the characters or inside the characters of the image. In Fig. 9-c, overlapping spots occur in the image. These spots are connected with the edge of characters or locally modify the ink of the characters. Fig. 9-d corresponds to the disconnection spots. They break strokes of characters completely. In Fig. 9-b,c,d, the right images represent the zoomed parts of the left images.

In addition, in Fig. 9-b, the image still contains several overlapping spots and disconnection spots. This is also true with the Fig. 9-c,d. Image in Fig. 9-c is degraded by not only overlapping spots, but also some independent spots and disconnection spots. Fig. 9-d has the appearance of several independent spots and overlapping spots. It is because seed-points are chosen independently in main color layers of the original images, so some seed-points connected together may be classified as different kinds of spots according to the layers where they are processed. Thus, they will be treated in different ways.

IV. Conclusion

In this work, an efficient color image degradation model is proposed. This model is parametrized with four values: the number of seed-points (the degradation level), the percentage of each type of noise (independent spot, overlapping spot, and disconnection spot) in the final degraded image. The outputs obtained by this model look very realistic. For now, parameters can be tuned to generate more or less noise regions depending on the application since images are different in terms of colors, and content. A future work could be the automatic adaptation of the parameters to the images in order to make the resulting images more realistic. Besides, the validation of the model on real images should be taken into consideration.

Acknowledgment

This work was supported by Laboratory L3i, University of La Rochelle, France.

References

[1] T. Kanungo, R. M. Haralick, and I. Phillips, "Global and local document degradation models," *In Proceedings of the International Conference on Document Analysis and Recognition*, 1993.

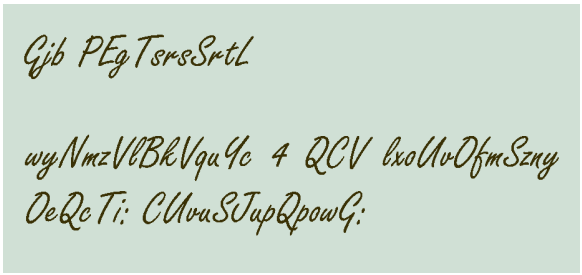
[2] H. S. Baird, "The state of the art of document image degradation modeling," *In Proc. of 4th IAPR International Workshop on Document Analysis Systems, Rio de Janeiro*, 2000.

[3] V. Kieu, M. Visani, N. Journet, J. Domenger, and R. Mullot, "A character degradation model for grayscale ancient document images," *In Proceedings of the International Conference on Pattern Recognition (ICPR)*, 2012.

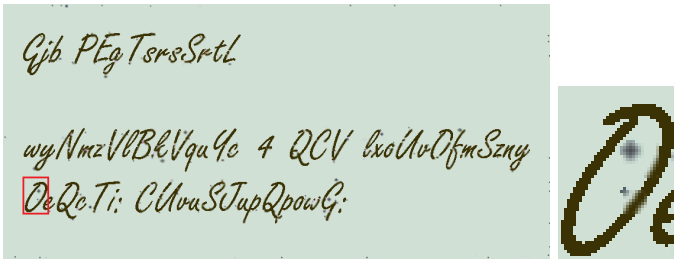
[4] V. Kieu, M. Visani, N. Journet, R. Mullot, and J. Domenger, "An efficient parametrization of character degradation model for semi-synthetic image generation," *In Workshop on Historical Document Imaging and Processing 2013 (HIP)*, 2013.

[5] E. Carel, V. Courboulay, J.-C. Burie, and J.-M. Ogier, "Dominant color segmentation of administrative document images by hierarchical clustering," in *ACM Symposium on Document Engineering 2013, DocEng '13*, Florence, Italy, 2013, pp. 115–118.

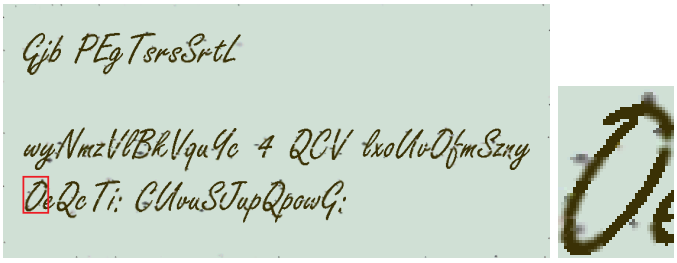
[6] S. D. Zenzo, "A note on the gradient of a multi-image," *Computer Vision, Graphics, and Image Processing*, 1986.



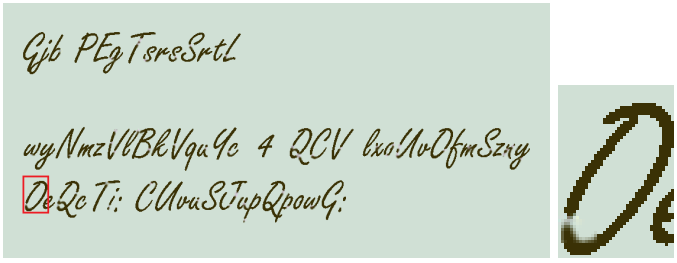
(a)



(b)



(c)



(d)

Fig. 9. The original image and noisy images with different percentages of types of noise (a) Original image (b) I = 100%, O = 0%, D = 0% (c) I = 0%, O = 100%, D = 0% (d) I = 0%, O = 0%, D = 100%