SEURET — Gradient

<1 figure>

Gradient-Domain Noise for Realistic Degradations in Historical Document Images

Seuret, M., Eichenberger, N., Liwicki, M. and Ingold, R.

We developed a novel method for adding realistic degradations to color document images in order to generate large sets of training data for computational processing. We extract and process patches of real noise from degraded documents and paste them into the target document image using the gradient domain in order to achieve independence from the hue of the document. We will describe the novel algorithm and present an expert’s opinion on the data generated.

# \* \* \*

Historical documents suffer from different kinds of damages during their conservation, caused by either constant use (e.g., stains, fingermarks, ink strokes or spots, scratches, degradation of ink by abrasion), by inappropriate storage conditions (such as mildew caused by humidity), by historical catastrophes like wars, or by the material itself (such as ink corrosion). These damages can affect the readability of the manuscripts and also pose a severe challenge for digital processing of such documents.

Developing robust digital processing algorithms requires machine learning, which relies on huge amounts of realistic training data, i.e., large sets of document images together with their corresponding ground truth (which refers to the expected labeling of the document). The high cost of ground-truthing limits the number of available training and testing documents. This motivates the development of methods that artificially add noise to existing ground-truthed documents in order to increase the number of available training documents.

Fischer et al. (2013) presented a method to generate training samples by deforming binarized text lines. They showed that the accuracy of text recognition could be increased by up to 3.23% by their method.

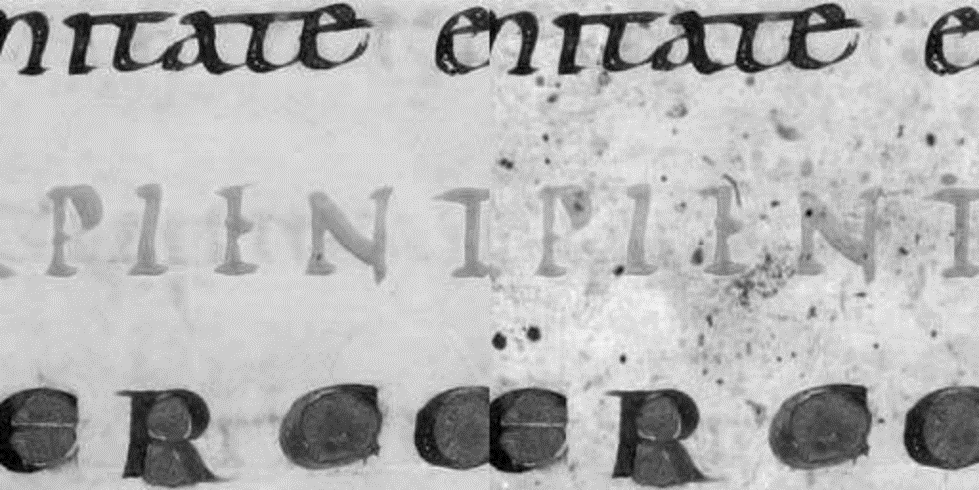


Figure 1. An document image sample (left) and its degraded version (right).

Kieu et al. (2012) have presented a local noise generation method for historical document gray-scale images. They evaluated their method with optical characters recognition (OCR) software and found a linear relation between the quantity of the noise and the recognition error rate. Several other methods have been developed, but they all work either on the red, green, and blue values of images, or on binary (black-and-white) images.

# Methodology

The main novelty of our method is, contrary to previous work, to work in the gradient domain of color images, i.e., using the difference of values between pixels, and to use patches of real degradations from existing document images.

To insert degradation patches in the gradient domain, we first compute the gradients of the document image and of the degradation patches, then paste the gradients of the degradations on the gradients of the document image, and finally reconstruct the document image from its modified gradients. This method is illustrated in Figure 1. Working in the RGB color space would lead to inconsistencies if the document from which we extracted the degradations does not have a very similar background color as our target document image. Working in the gradient domain to avoid this is the main novelty of our method.

Our method is inspired by the Poisson image editing method presented by Pérez et al. (2003). First, we compute the horizontal and vertical gradients of the image. Then we apply the degradations on these gradients. Finally, we construct the result image from the modified gradients.

# Scholar’s Point of View

From a manuscript scholar’s point of view, the method for adding synthetic noise presents advantages and disadvantages regarding its verisimilitude. One advantage is that the inserted patches are based on real degradations appearing in historical documents; thus, they represent realistic historical degradations. This method provides a rather realistic impression for small-sized details of a manuscript page. However, the overall impression of an entire page looks less realistic, because the degradation patches are spread repeatedly, following a uniform random distribution over the page, and can easily be recognized as artificially added. Thus, the gradient method is appropriate for simulating degradations such as stains, scratches, and ink spots, which mostly overlap the script or the background, but doesn’t have a direct impact on the script (as, for example, abrasion and humidity do). It works best on small-scale details.

# Discussion

Our method can be used to generate document images to train machine learning algorithms. The settings of the software that we developed are easy to manipulate by non–computer scientists. However, precautions have to be taken in order to clarify that the resulting degraded documents do not represent the actual state of the real historical document. As future work, we will compute statistics about the position of the degradations while extracting patches in order to generate more realistic distribution models.

# References

**Fischer, A., Visani, M., Kieu, C. V. and Suen, C. Y.** (2013). Generation of Learning Samples for Historical Handwriting Recognition Using Image Degradation. In *Proceedings of the 2nd International Workshop on Historical Document Imaging and Processing*, pp. 73–79.

**Kieu, C. V., Visani, M., Journet, N., Domenger J.-P. and Mullot, R.** (2012). A Character Degradation Model for Grayscale Ancient Document Images. In *21st International Conference on Pattern Recognition (ICPR),* pp. 685–88.

**Pérez, P., Gangnet, M. and Blake, A.** (2003). Poisson Image Editing. In *ACM Transactions on Graphics (TOG),* pp. 313–18.