A novel preprocessing method for hectography prints based on independent component analysis

Thomas Kurbiel Iuliu Konya Stefan Eickeler

Fraunhofer Institute for Intelligent Analysis and Information Systems (IAIS)
Schloss Birlinghoven, 53754 Sankt Augustin, Germany
E-mail: {first-name.last-name}@iais.fraunhofer.de

***Abstract*—Archives and cultural facilities consists of vast spectra of different document classes, many of which are not encountered today anymore. The digitization therefore calls for image enhancement and preprocessing solutions so far not required and hence unsolved. A prominent document class in this context is hectography, which was a inexpensive printing and duplication method, widely used throughout the 20th century. The major challenge with hectography is poor contrast on the one side and multiple degradation effects on the other side.**

In this paper a novel preprocessing method for hectography duplicates is proposed which leads to better Optical Character Recognition (OCR) results compared to traditional methods that operate on grayscale images. The proposed method is based on the model conform with the independent component analysis. The problem of unwanted Gaussian noise components is considered as well.

***Index Terms*—independent component analysis; hectography; image enhancement; extraction; OCR**

I. Introduction

Millions of books, documents and papers are stored all around the world in libraries, museums and archives - an immense store of knowledge. The questions of how this cultural heritage can be made available to the greatest possible number of people and how it can be preserved for future generations are of utmost importance. The question of preservation is readily solved by mere digitizing of the original sources which has been steadily pursued in the last two decades. The more challenging question however is the call for availability which is closely interlinked with efficient searchability and therefore with meta-data generation. Given the fact that a cost-effective method requires an automatic extraction of meta-data, the requirement for excellent OCR recognition rates poses a conditio sine qua non. A major problem that arises in this context is the low-quality of many of the documents stored in archives, where the low quality can be traced back to two main reasons:

- the outdated printing or duplication process itself produced only documents of low quality,
- in general time-dependent degradation effects decreased the quality of the regarding documents.

In general it is not possible to resolve the described quality problems using some traditional generic image enhancement methods. The use of thresholding techniques to remove the background is often not effective since the intensities of the unwanted background can be very close to those of the foreground text [1]. In these conditions, thresholding either does not remove the background or also eliminates part of the information in the text of interest [1]. For these reasons novel preprocessing methods have to be devised. This is due to the unique nature of the documents and their unique state of degradation which requires tailord solutions.

In this paper we are dealing in particular with hectography documents (an outdated printing and duplication method) often encountered in libraries and archives thematically connected to the 19th and 20th century. In the past some attempts have been done to cope with a similar problem of the so-called bleed-through or show-through effects [1], where the focus was on palimpsests, ancient manuscripts that have been erased and then rewritten again. In [2] a Bayesian formulation for a joint blind-source separation and restoration of noisy mixtures of degraded images is proposed. Here edge-preserving Markov random field (MRF) image models are considered to describe local spatial auto-correlation. This method involves a linear data model, where multimodal observations of an object are seen as mixtures of all the patterns to be extracted.

In this contribution we propose a novel algorithm which allows for separation of the time-dependent degradation effects from the crucial text-component. In the proposed algorithm the extraction problem is treated as a classical blind-source separation problem. Therefore the independent component analysis (ICA) widely used in this context is applied. Furthermore, straightforward suggestions for an efficient implementation of this method are presented throughout.

This contribution is organized as follows: In section 2, we give a brief description over the hectography procedure as well as over the arising problems. Next, in section 3, we discuss the basic idea of the proposed approach followed by a detailed description of all required steps. Subsequently, in section 4, an illustrative example is shown combined with the evaluation of our method and the comparison with different standard methods. In section 5 concluding remarks are drawn.

II. Statement of the Problem

Hectography is an obsolete copying/duplication and printing method for text documents. It was widely used throughout the 20th century. Today hectography documents are primarily

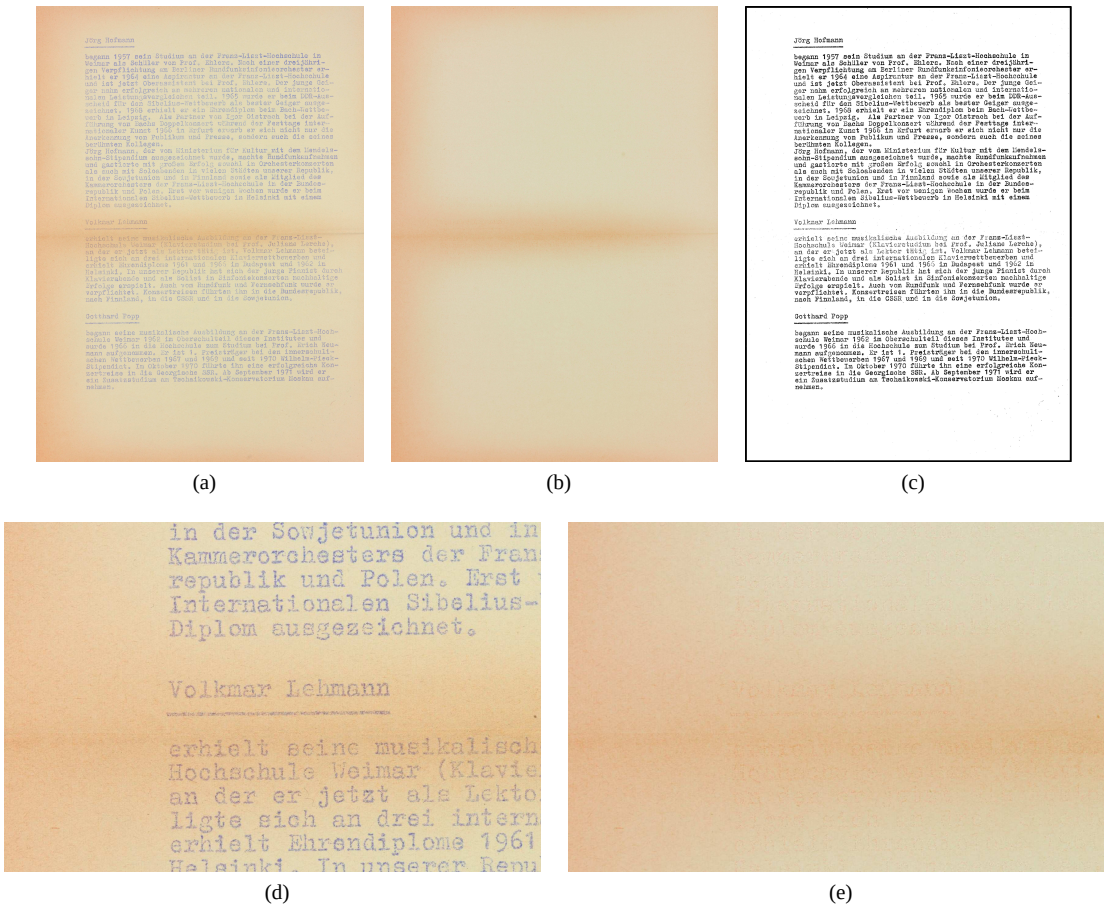


Figure 1: **(a)** A typical hectography duplicate with strong degradation effects and eminent folding **(b)** Background of a hectography document extracted with the proposed method **(c)** Binarized text component **(d)** Magnified section of *subimage (a)* **(e)** Magnified section of *subimage (b)* (source: Deutsche Nationalbibliothek DNB).

found in archives. Hectography uses a mastercopy consisting of paper coated with a special mixture of gelatin. The mastercopy is written or typed with a special aniline ink. Gelatin is used because its moisture keeps the ink from drying. Duplicates of the mastercopy are made by using a roller which presses the blank papers onto the gelatin. With each copy some ink is removed from the gelatin and consequently successive copies are progressively lighter. In practice, between 20 and 80 duplicates can be made from one mastercopy, depending upon the skill of the user and the quality of the original. In the past at least eight different colors of hectographic ink existed, but blue/violet was the most popular because of its density and contrast, as to be seen in. Fig. 1.

However the duplicates obtained by hectography are usually of low-quality only. Furthermore, the paper used for hectography becomes yellow over time, causing a poor contrast with regard to both colors and brightness. Strong degradation effects and foldings in the used paper are common as well. Therefore without any preprocessing, performing OCR on hectography duplicates leads to very low recognition rates. Combining all above factors an eligible preprocessing method has to incorporate the following minimum goals: separate the background from the text-component and enhance the contrast of the latter.

Concluding, we infer the main cause why PCA alone does not allow for a proper extraction of the text component. For document papers the size of the background area exceeds in general the text area by far, with the result that the variance of the background noise exceeds the variance of the text-component analogously. Therefore the first principal component inevitably is determined by the background noise, as are all other principal components due to the orthogonal nature of the PCA [3], [4].

III. Proposed Approach

The underpinning of our approach is the assumption that hectography duplicates can be modeled as weighted superposition of several independent components e.g. text-content, degradation effects, paper-texture and noise. In case this simple model holds, the independent component analysis (ICA) [5], [6] is an adequate tool for extraction. Since the RGB-scan files contain three observed signals, the number of independent components to be extracted from them is limited to three as

well [5]. This restriction however does not pose a major problem, since both paper-texture and degradation effects belong to related classes of phenomena, and their distinct extraction is not needed usually, i.e. it is sufficient to extract the sum of both components. In addition, the proposed approach does not consider the spatial information of the image and hence, the RGB components are treated as mere random variables. The RGB-components of each pixel in the image can therefore be seen as single observations.

Using the common ICA notation [5] the problem can be stated as follows:

$$\begin{aligned} R &= \alpha_{11} \cdot Z_{\text{text}} + \alpha_{12} \cdot Z_{\text{degrad}} + \alpha_{13} \cdot Z_{\text{noise}}, \\ G &= \alpha_{21} \cdot Z_{\text{text}} + \alpha_{22} \cdot Z_{\text{degrad}} + \alpha_{23} \cdot Z_{\text{noise}}, \\ B &= \alpha_{31} \cdot Z_{\text{text}} + \alpha_{32} \cdot Z_{\text{degrad}} + \alpha_{33} \cdot Z_{\text{noise}}, \end{aligned} \quad (1)$$

where R, G, B are the observed random variables and $Z_{\text{text}}, Z_{\text{degrad}}, Z_{\text{noise}}$ are the unknown generative components. The unknown mixing matrix is denoted by α_{ij}, $i, j \in \{1, 2, 3\}$. In consistence with the restrictions of the ICA we can assume that all components are non-Gaussian except for the noise-component Z_{noise}. Particularly the independent component associated with the text-content Z_{text} is non-Gaussian due to a high structure of the corresponding distribution functions. It has to be considered thoroughly how to deal with a possible noise problem. Preliminary results show that the variance of the noise-component Z_{noise} is negligible such that this component can be omitted after whitening. This additional dimension reduction has a significant impact on the time performance of the ICA [5], [7].

In the following all processing steps are described in detail:

A. Inverse Gamma Companding

In the first processing step an inverse gamma companding is performed in all three color channels:

$$v = \begin{cases} V/12.92 & V \leq 0.04045 \\ \left((V + 0.055)/1.055\right)^{2.4} & V > 0.04045 \end{cases} \quad (2)$$

where $V \in \{R, G, B\}$. This step is obligatory since otherwise our model in (1) would not hold due to the performed gamma correction. However the impact of the inverse gamma companding on the final binary image is only marginal.

B. Whitening and Dimension Reduction

The next processing step comprises both the whitening of the random variables R, G, B and the subsequent dimension reduction. Since we intend to reduce the dimension as well, we achieve this best by the means of the classical principal component analysis (PCA) [8], [9]. First to simplify matters it is common to remove the means of all three random variables [5], [8]. Since it is obvious from the context, we will use the same names for the original and the mean-free variables.

As it is well known from the theory of PCA [8], [9] in case the variance of one of the random variables R, G, B exceeds the variances of the others by far, then the first principal component is almost completely determined by this random variable. In such case determining the variance declared by each principal component is distorted and misleading. In order to avoid this phenomenon we have to standardize the variance before applying PCA:

$$\begin{aligned} R &= R/\sqrt{\mathcal{E}\{R^2\}}, \\ G &= G/\sqrt{\mathcal{E}\{G^2\}}, \\ B &= B/\sqrt{\mathcal{E}\{B^2\}}, \end{aligned} \quad (3)$$

such that all random variables R, G, B have the same impact on the principal components.

Subsequently the principal components P_1, P_2, P_3 of the random variables R, G, B have to be obtained. To this end any known method can be applied. We haven chosen the simple and straightforward method of eigenvalue decomposition of the corresponding covariance matrix of R, G, B. In hectography the first two principal components P_1, P_2 declare almost always more then 90% of the original total variance. For this reason the last component P_3 which most likely contains only Gaussian noise is removed.

To conclude the whitening process, we have to standardize the variances of the remaining components:

$$P_i = P_i/\sqrt{\lambda_i},\ i \in \{1, 2, 3\} \quad (4)$$

where $\lambda_1, \lambda_2, \lambda_3$ denote the eigenvalues of the covariance matrix of the scaled R, G, B.

C. FastICA

In the last computation step the independent components:

$$\begin{aligned} Z_{\text{text}} &= \beta_{11} \cdot P_1 + \beta_{12} \cdot P_2 + \beta_{13} \cdot P_3, \\ Z_{\text{degrad}} &= \beta_{21} \cdot P_1 + \beta_{22} \cdot P_2 + \beta_{23} \cdot P_3, \\ Z_{\text{noise}} &= \beta_{31} \cdot P_1 + \beta_{32} \cdot P_2 + \beta_{33} \cdot P_3, \end{aligned} \quad (5)$$

are obtained using an iterative procedure described in [5], [6]. As proposed in [5] we apply the symmetric approach and use the $\tanh$ non-linearity. Due to the dimension reduction described above P_3 and Z_{noise} are not considered usually.

Subsequently all preceding scalings and transformations are combined to one transformation matrix as described in [5], [6] and applied pixel-wise on the original image. Please note that the obtained independent components $Z_{\text{text}}, Z_{\text{degrad}}, Z_{\text{noise}}$ are still floating point numbers, which in the last step are rescaled and quantized according to:

$$y = \left\lceil \frac{255}{Y_{\max} - Y_{\min}} \cdot (Y - Y_{\min}) \right\rceil, \quad (6)$$

where $Y \in \{P_1, P_2, P_3\}$.

As explained in [5], [6] two ambiguities or interdeterminacies hold with respect to ICA: the ambiguity of the sign and the ambiguity of the order of the independent components.

D. Background Color

The first ambiguity, the ambiguity of the sign, has the effect, that in some cases some of the images representing the independent components have inverted grayscales values with respect to the original image, i.e. white background in the original image appears black in the component image.

Since in text documents the amount of pixels belonging to the background usually exceeds the amount of pixels belonging to the text, this effect can be compensated rather easily by first determining the background color of the original image (after the grayscale transform) and second comparing it with the background color of each component image. The previous scaling in (6) allows for using a simple static threshold of $Y_{\mathrm{thres}} = 128$ for this purpose. We then count the total amount of pixels in each grayscale-image satisfying $Y < Y_{\mathrm{thres}}$ and $Y \geq Y_{\mathrm{thres}}$.

E. Finding Text-Component

The main difficulty of ICA, as mentioned above, is the fact that the order of the independent components is arbitrary. Therefore it is not guaranteed that the first component always will be the text-component, which brings out the question for an automatical identification of this component. There are many possible ways of identifying the text component. The most obvious one would be to count the number of connected components [10], [11] in all component images. Since the text component ideally contains only characters and in contrast to the background and noise components almost no specks, the number of connected components should be much lower here. To render this identification method more stable more sophisticated methods, like determining the variance of the size of the connected components, could be incorporated.

Our approach however is based on two fundamentals result from information theory, *1.)* the more "random", i.e. unpredictable and unstructured a random variable is, the larger its (differential) entropy [5]:

$$H(X) = -\int p_X(\xi) \cdot \log p_X(\xi)\, \mathrm{d}\xi, \tag{7}$$

and *2.)* the fact that Gaussian variables have the larges entropy among all random variables with a given covariance matrix. Both fundamental results allow to define a measure that is zero for a Gaussian variable and always non negative [5]. This measure is called negentropy [5]:

$$J(X) = H(X_{\mathrm{gauss}}) - H(X). \tag{8}$$

Since the brightness of letters naturally strongly differs from the brightness of the background, text components will usually have a very structured probability density function and hence the negentropy (8) assumes a large value. On the other side, extracted background components will have less obvious inherent structure and therefore stronger resemble Gaussian variables. Consequently the negentropy will have a smaller value.

The main problem with negentropy is that it is computationally very difficult. In [5] several different approximations have therefore been proposed. We are employing the following one:

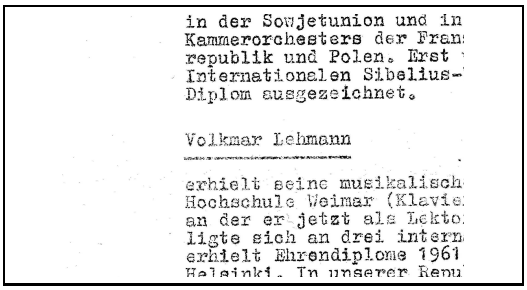
$$\begin{aligned} J(X) &= k_1 \cdot \left(\mathcal{E}\left\{X \cdot \exp\left(-X^2/2\right)\right\}\right)^2 \\ &+ k_2 \cdot \left(\mathcal{E}\left\{|X|\right\} - \sqrt{2/\pi}\right)^2, \end{aligned} \tag{9}$$

with $k_1 = 36/\left(8\sqrt{3} - 9\right)$ and $k_2 = 24/\left(16\sqrt{3} - 27\right)$. In practice the expectation is substituted by the sample average.

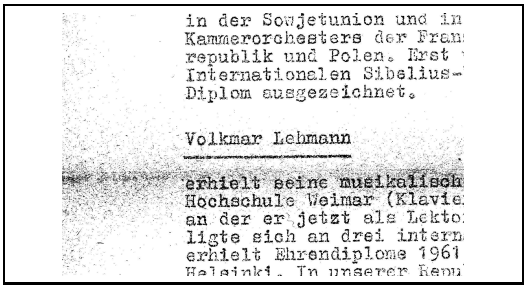
IV. Evaluation

Subsequently, we apply the above method to the hectography image in Fig. 1 (a). The original document was scanned with 600dpi and the pixel resolution of this image is 3815×5319. The corresponding demixing matrix is obtained by performing the FastICA on a downsampled version of this image using a downsampling factor of $M = 2$. This is done for reasons of computation time. Please note, that the subsequent transformation into the independent components is performed on the original image size again. The first two principal components, obtained with additional scaling of the variance according to the explanations in sec. 3, declare more then 98% of the original variance. The last principal component contains noise only and therefore is left out.

Fig. 1 (b) and (e) show the extracted background component. For clarity reasons the original colors of the document are retrieved, compare the original image in Fig. 1 (a). To this end the concerning independent component has been transformed back to the RGB-space. As can be seen the background component in Fig. 1 (e) contains both the paper texture and the degradation effects. This phenomenon is treated in sec. III.



(a) proposed method



(b) grayscale transformation & Otsu-binarization

Figure 2: Text component after binarization

The result of the subsequent binarization of the extracted text component is depicted in Fig. 2 (a). The binarization algorithm used is that one proposed in [12]. To demonstrate the effectiveness of our approach, we have chosen a especially distorted and noisy region of the original image Fig. 1 (a) whose magnification is depicted in Fig. 1 (c). However the only weak point to be discovered in the binary image are some too faint appearing characters in the upper border. These

can be corrected subsequently by using simple morphological operations. On the whole the text has been extracted properly and has sufficiently high quality for a subsequent OCR. This is especially demonstrative, inasmuch the prominent folding in Fig. 1 (b) is not visible at all.

Finally a thorough evaluation of the proposed method is presented which employs the OCR recognition rate. For this purpose we are using two different OCR engines: 1.) Abbyy Finereader, the best commercial engine and 2.) Google Tesseract, the best free engine available. Next, ground truth templates for 22 different hectography documents are generated manually, containing alltogether 42825 characters and 5833 words. In order to guarantee a fair comparison between the different preprocessing methods, the exact positions of all text lines are marked manually as well. Then the proposed method is compared with two classical approaches which involve a grayscale transformation, followed by a subsequent binarization. The two "tradional" binarization methods we use are 1.) Otsu's method, which is a global binarization method [12] and 2.) Sauvola's algorithm, which is an adaptive binarization method [13]. Since the hectography images are rather bright, we are adjusting the threshold level in Sauvola's algorithm [13] for each pixel by the factor of 1.2. The local window radius used here is 35x35.

	new method	Otsu's meth. [12]	Sauvola's alg. [13]
Finereader	2182 5%	15193 35%	13754 32%
Tesseract	17184 40%	35207 82%	38360 89%

Table I: Levenshtein distance and error rate

The results of the evaluation are summarized in Tab. I. The metric we use for measuring the amount of difference between the different OCR results and the ground truth is the Levenshtein distance [14]. The much higher error rate of Tesseract can be traced back to the fact that it uses connected components labeling for extraction. Therefore a pixelbased comparison approach would have been much more adequate here.

Concluding, in Fig. 2 (b) we contrast the result of our method with the classical Otsu's method. Apart from the apparent remaining of the paper folding it is to be noted how most of the characters are not clearly demarcated and how their margins/edges are not clearly defined.

ACKNOWLEDGMENT

This work was supported by the German Federal Ministry of Economics and Technology (BMWi) funded program Theseus in the project Contentus. In addition the authors would like to thank the Deutsche Nationalbibliothek for providing a vast amount of testing documents.

V. CONCLUSION

In this contribution a preprocessing method for low-quality hectography documents has been introduced. The approach is based on an additive generative model specifically suitable for hectography documents. As a consequence our approach does not require a sophisticated model for the noise component nor the knowledge of the spatial correlation of close-by pixels. The separation of the additive components is performed by the means of the independent component analysis often used for similar blind-source separation problems [7]. The problem of noise distortion is dealt with by a well-thought-out dimension reduction subsequent to the first processing step involving whitening of the original R, G, B signals. The results show that the presented method is adequate for increasing of the optical legibility of hectography images and therefore for increasing of the OCR performance. The main reason why the proposed method works especially well on hectography documents is the fact, that the printing or copying process used in hectography generates characters/objects which are not entirely opaque. Furthermore bleed-through or show-through effects almost do not occur, due to the single page copying technique. For these reasons hectography documents abide especially strong to the proposed additive generative model.

Our current research concentrates on the question how well the proposed algorithm is able to cope with old typewriter documents with strong noise components. Since the characters here are much more opaque then in hectography documents, a modification of the additive generative model is required.

REFERENCES

[1] A. Tonazzini, L. Bedini, and E. Salerno, "Independent component analysis for document restoration," in *Proc. IJDAR 2004*, vol. 7, pp. 17–27.

[2] A. Tonazzini, I. Gerace, and F. Martinelli, "Multichannel blind separation and deconvolution of images for document analysis," in *IEEE Transactions on Image Processing*, vol. 19, pp. 912–925.

[3] Y. Ohta, T. Kanade, and T. Sakai, "Color information for region segmentation," in *Computer Graphics and Image Processing*, vol. 13, pp. 222–241.

[4] M. S. Drew and S. Bergner:, "Spatio-chromatic decorrelation for color image compression." in *Sig. Proc.: Image Comm.*, vol. 8, pp. 599–609.

[5] A. Hyvaerinen, J. Karhunen, and E. Oja, *Independent Component Analysis*. John Wiley and Sons, 2001.

[6] A. Hyvaerinen, "Fast and robust fixed-point algorithms for independent component analysis," in *IEEE Transactions on Neural Networks*, vol. 3, pp. 626–634.

[7] A. Cichocki and S. I. Amari, *Adaptive Blind Signal and Image Processing*. John Wiley and Sons, 2002.

[8] I. T. Jolliffe, *Principal Component Analysis, 2nd ed.* Springer-Verlag New York, 2002.

[9] G. H. Dunteman, *Principal Components Analysis (Quantitative Applications in the Social Sciences) 1st ed.* SAGE Publications, 1989.

[10] M. B. Dillencourt, H. Samet, and M. Tamminen, "A general for arbitrary approach image to connected-component representations labeling," in *Journal of the ACM*, vol. 3, pp. 253–280.

[11] Chang and C.-J. Chen, "A component-labeling algorithm using contour tracing technique," in *Proc. ICDAR 2003*, Edinburgh.

[12] N. Otsu, "A threshold selection method from gray-level histograms," in *IEEE Trans. on Systems, Man, and Cybernetics*, vol. 9, pp. 62–66.

[13] J. Sauvola and M.Pietikainen, "Adaptive document image binarization," in *Pattern Recognition*, vol. 2, pp. 225–236.

[14] V. I. Levenshtein, "Binary codes capable of correcting deletions, insertions, and reversals." in *Doklady Akademii Nauk SSSR.*, vol. 163, pp. 845–848.