Ancient documents bleed-through evaluation and its application for predicting OCR error rates

Vincent Rabeux, Nicholas Journet, Jean-Philippe Domenger

Ancient documents bleed-through evaluation and its application for predicting OCR error rates

Rabeux V.[a] , Journet N. [a] and Domenger J.P.[a]

[a]LaBRI, Université de Bordeaux, 351, cours de la Libération, Talence, France;

ABSTRACT

This article presents a way to evaluate the bleed-through defect on very old document images. We design measures to quantify and evaluate the verso ink bleeding through the paper onto the recto side. Measuring the bleed-through defect alows us to perform statistical analysis that are able to predict the feasibility of different post-scan tasks. In this article we choose to illustrate our measures by creating two OCR error rate predicting models based bleed-through evaluation. Two models are proposed, one for Abbyy FineReader * which is a very power-full commercial OCR and OCRopus † which is sponsored by Google. Both prediction models appears to be very accurate when calculating various statistic indicators.

Keywords: Document Image analysis, Quality, bleed through quantification, Ocr Prediction

1. INTRODUCTION

In scanning and indexing document chains, quality evaluation of the scanned document is a concerning issue. One could need to evaluate the image quality in order to : readjust the scanner parameters, predict the **O**ptical **C**haracter **R**ecognition (OCR) error rate, annotate a document image of quality meta-datas. Studies[1–4] were made on quality image evaluation in which metrics on characters degradations were introduced. These studies were made on modern documents suffering from holes in characters, broken and touching characters, and background speckle noise. Prior to this paper, image quality evaluation did not account for either the bleed-through defect nor ancients document images.

Bleed-through is a defect from which a lot of ancient documents suffer. The bleed-through defect can be explained by two combined reasons : the light reflected by the scanner backing,[5] and the ink of the verso side which is diffusing into the paper. Several bleed-through models[5,6] and restoration algorithms[7–10] exists. In Rsldi[8] document image classifiers are used to restore all background defects including bleed-through. This technic is very efficient but cannot be applied to our need since measuring the difference between the old image and the restored one will not only measure the bleed-through ; G. Shama[11] models the bleed-through with optical densities and applies a linear filter adapted (with the least mean squares) where recto pixels are ink on the verso side. We inspired ourselves with the method introduced by E.Dubois,[7] whom proposes an effective bleed-through restoration algorithm in which recto pixels that have an ink pixel on the verso side are replaced with an background estimation. Using this method implies studying the influence of the verso side ink diffused onto the recto side. Indeed, we evaluate the likely recto side bleed-through with different measures applied to the binarized recto and verso side of a document image page.

Further author information:
E-mails: {rabeux,domenger,journet}@labri.fr

*http://www.abbyy.com/

†http://code.google.com/p/ocropus/

The bleed-through effect is a major issue when *binarizing* a very old document image. Indeed, with a global binarization method some bleed-through pixels will be identified as ink pixels which will undoubtedly lead to **O**ptical **C**haracter **R**ecognition (OCR) errors. Worst, some adaptive binarization methods will identify bleed-through pixels as ink pixel in any zone of the page where only bleed-through is present. On the other hand, optical character recognition of very old documents is a major need, while the OCR error rate is a real economic issue. The quality of OCR softwares results are depending on the quality of the document image given as input. It is now stated that for good and moderns document, the average OCR recognition rate is close to 100%, unlike on poor quality documents where the recognition rate drops alongside quality. Since the binarization phase is critical for the OCR, the bleed-through defect is to be considered when predicting OCR results. Depending on the origin of the bleed-through , its cancellation can be hard : if it comes from the quantity of light used to digitize the document, it can be faded by tuning the scanner parameters. But, if it also comes from the ink diffusion into the paper, one can only cancel the bleed-through defect by applying a restoration process to the document. By measuring the bleed-through defect and predicting the OCR error rate, we could be able with statistic analysis, to tell whether or not a bleed-through restoration algorithm needs to be applied or if the scanner's parameters needs to be changed before trying to give the document to an OCR software. In our experiments we used two different OCR systems : OCRopus which is open source, and Abbyy FineReader. Even if it appears that Abbyy's OCR gives better results on very old document than OCRopus we decided to use both of them.

From observations made on a realistic corpus, we present in section 2 six measures able to efficiently evaluate the bleed-through defect. This vector of measures is then used in section 3, as an application, to create an OCR error rate predicting model. Finally, our research perspectives are discussed in section 4

2. BLEED-THROUGH EVALUATION

In order to measure the verso's ink pixel influence on the recto, we need to identify three classes of pixels : ink, bleed-through and background. We inspired ourselves with a bleed-through restoration method in which the registered[7] binarized verso and recto are needed. From the binarized verso P_v and recto P_r, we introduce I_r and I_v the set of ink pixels on respectively the recto side and the verso side, $\overline{I_v}$ is the projection of I_v onto the verso image. We also define T_r which is made out of all the pixels on the recto background that corresponds to an ink pixel on the verso side.

T_r can be defined by the folowing equation where p is a pixel of co-ordinated (x, y).

$$T_r = \{p \in \overline{I_v}, p \notin I_r\}$$

Finaly, B_r is introduced as the set of pixels corresponding to the background without any pixels likely to be bleed-through ones.

$$B_r = P_r \setminus [T_r \cup I_r]$$

Our document image quality measures are designed to quantify the document bleed-through defect which we observed on a corpus of ancient documents. Bleed-through has three principals characteristics : the intensity of the bleed through pixels, the amount of pixels and, at last, the location of the bleed through.

2.1 The bleed-through intensity

The intensity of bleed-through corresponds to the average bleed-through pixels grayscale level and varies from very light gray (close to the background level) to low dark gray (close to the ink level). The intensity of bleed-through is a major issue when trying to identify pixels corresponding to text in very old documents where that intensity is low, and we believe that it exists a value where the intensity of bleed-through is so low that the OCR will identify those pixels as ink pixels. This value must be in function of the average ink intensity and the average background intensity. Figure 1 shows the zones identified as text by the OCR on the same document images with different bleed-through intensity, and as we can see the OCR identifies some bleed through pixels as text pixels.

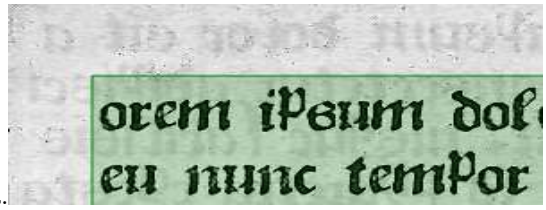
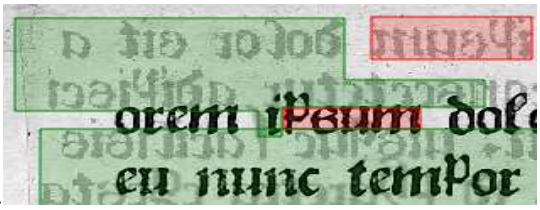
a.  b. 

Figure 1. OCR identified text zones (highlighted) : a. high value of intensity, the OCR correctly identifies the texts blocks, b. the same document with a lower intensity, the OCR fails to identify the text blocks. Green zones are recognized text-blocks, and red zones are unrecognized text-blocks.

To evaluate the bleed through intensity we consider two measures $\mathcal{MI}_i$ which corresponds to the distance to the average ink intensity and $\mathcal{MI}_b$ the distance to the background intensity. In the following equations let μ_{T_r}, μ_{I_r} and μ_{B_r} be the average intensity of respectively T_r, I_r and B_r.

$$\mathcal{MI}_i = \frac{\mu_{T_r} - \mu_{I_r}}{255} \quad \mathcal{MI}_b = \frac{\mu_{B_r} - \mu_{T_r}}{255}$$

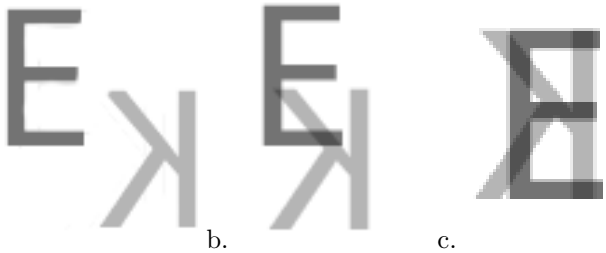
2.2 The bleed-through quantity

The bleed through intensity is not the only parameter that may produce OCR errors. Indeed, OCR systems will make more errors on a document with a lot of bleed-through that on one with just a few. Therefore, we also need to measure the quantity of bleed-through that may appear by introducing $\mathcal{MQ}$, the ratio of ink verso pixels that appear as bleed-through pixels on the recto side.

$$\mathcal{MQ} = \frac{\|T_r\|}{\|I_r\|}$$

2.3 The bleed-through location

We observed from our corpus of document images that the bleed-through letters location is also a caraceristic that needs to be considered and measured. Depending on the OCR used binarization methods and parameters (global or adaptative, widow-size, usage od dictionaries or not), the bleed-through location will create less or more OCR errors. Therefore, we propose measures to quantify three different bleed-through component locations.



a. b. c.

Figure 2. The different locations of a bleed-through letter on the page: a.the two letters do not overlap, b. the letters overlap, c.the letters are more overlapped than in b.

Let E and K two letters from respectively the recto and the verso image, we denote three different locations that gives different OCR outputs when the binariztion method fails :

- if E and K do not overlap (figure 2.a) : the component resulting from the wrong binarization of the letter K, will have the same characteristic (size, stroke-width) as the real text components. The result will be two ink connected components instead of one. The quantity of *non-touching* components is measured by $\mathcal{MA}$ (location measure 2.3.a).

- if E and K do overlap (figure 2.b) : the original letter will be altered. This leaves the OCR with one connected component that does not match any letter in the current page's font. The quantity of touching connected components is measured by $\mathcal{MS}$ (location measure 2.3.b).
- if E and K are completely overlapped (figure 2.c) : the original letter is completely modified. In this case, the OCR recognition phase is harder. The alteration's size is measured by $\mathcal{MSG}$ (location measure 2.3.c)

For the folowing measures, we define TC the set of connected components that are overlapping on the recto side where r and v are two connected components :

$$\begin{aligned} TC &= \{c = v \cup r \mid \exists v \subset \overline{I_v}, \exists r \subset I_r, r \cap v \neq \emptyset\} \\ \overline{TC} &= \{c = v \cup r \mid \exists v \subset \overline{I_v}, \exists r \subset I_r, r \cap v = \emptyset\} \end{aligned}$$

Let also be $\underline{I_r}$ and $\underline{I_v}$ the set of connected components of respectively I_r and I_v.

a) $\mathcal{MA} = \frac{\|\overline{TC}\|}{\|\underline{I_v}\|}$

$\mathcal{MA}$ is the ratio between the cardinal of the ink connected components on the verso side that are not overlapping any recto side ink components and the number of connected components on the verso side. In other words, $\mathcal{MA}$ measures the quantity of letters that may be added by the bleed-through defect onto the recto. This measure is not defined if the verso has no ink components. Otherwise, 0 means that no verso component will be added to the recto, and 1 means that all verso components will be added to the recto.

b) $\mathcal{MS} = \frac{\|TC\|}{\|\underline{I_r}\|}$

On the opposite $\mathcal{MS}$ is the ratio between the ink connected components that may be expended by any ink connected components from the verso and the total number of ink connected components on the recto side. if $\mathcal{MS} = 0$ no recto letters will be modified, if $\mathcal{MS} = 1$ all letters on the recto side will be modified.

c) $\mathcal{MSG} = \frac{\frac{1}{\|TC\|} \sum_{r \subset I_r, v \subset \overline{I_v}, r \cup v \subset TC} \|r \cup v\| - \|r \cap v\|}{\frac{1}{\|\underline{I_r}\|} \sum_{r \subset \underline{I_r}} \|r\|}$

As said, $\mathcal{MSG}$ will quantify the way the two letters overlaps by dividing the expended components mean area by the ink's components mean area. In other words, $\mathcal{MSG}$ measures the average expansion rate. The higher $\mathcal{MSG}$ is, the less the two connected components have pixels in common.

Table 1 shows the values of the three location measures on figure 2. $\mathcal{MA}$ is equal to 1 since 1 letter over 1 is added to the verso. The same reasoning can be done on $\mathcal{MS}$. $\mathcal{MSG}$ equals to 0 on figure 2.a since no black-component are overlapping. Latter lowers the more the two overlapping components have pixels in common.

	Figure 2.a	Figure 2.b	Figure 2.c
$\mathcal{MA}$	1	0	0
$\mathcal{MS}$	0	1	1
$\mathcal{MSG}$	0	2,04	1.62

Table 1. Example of the location measures on previous examples images. $\mathcal{MA}$ is none 0 in only the first case. $\mathcal{MSG}$ is higher the less the two connected components have pixels in common

2.4 Measures experiments on document images

Table 2 shows that our measures cover all the aspects of the bleed-through. In the first case (table 2.a), even though the locations and quantity measures are not good ($\mathcal{MQ}$ indicates that there are as many bleed-through as ink pixels on the recto side and $\mathcal{MS}$ that about 60% of the text will be modified) $\mathcal{MI}_i$ tells us that the bleed-through intensity is not close to the ink and as $\mathcal{MI}_b$ indicates that the bleed-through is close to the background, a global binarization method such as for example, *Otsu*,[12] can successfully identify ink pixels. The second (table 2.b) case as similar quantity and location measures, but $\mathcal{MI}_i$ and $\mathcal{MI}_b$ shows that the bleed-through pixel are

equal distance from the ink and the background color, the global binarization fails. The third case shows that even if the $\mathcal{MI}_i$ is similar to the second case, much more errors will appear : $\mathcal{MQ}$ shows that 300 times more bleed-through pixels will be added to the recto side and since $\mathcal{MA}$ equals 1, that none of the verso letters overlaps a recto letter. Two more cases are not shown on table 2 : the verso is a white page, the verso and the recto are white pages. In those two last cases no bleed-through from the verso can appear on the recto.

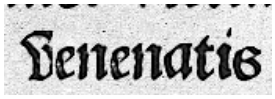
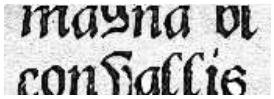
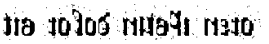
	Test Image	Verso	Measures	Value	Otsu Binarization
a.			$\mathcal{MI}_i$ $\mathcal{MI}_b$ $\mathcal{MQ}$ $\mathcal{MA}$ $\mathcal{MS}$ $\mathcal{MSG}$	0.716498 0.103556 1.19191 0.238095 0.606061 1.93156	
b.			$\mathcal{MI}_i$ $\mathcal{MI}_b$ $\mathcal{MQ}$ $\mathcal{MA}$ $\mathcal{MS}$ $\mathcal{MSG}$	0.327205 0.38078 0.948104 0.461538 0.215385 1.16205	
c			$\mathcal{MI}_i$ $\mathcal{MI}_b$ $\mathcal{MQ}$ $\mathcal{MA}$ $\mathcal{MS}$ $\mathcal{MSG}$	0.341211 0.266451 300.3 1 0 0	

Table 2. Experiments of measures document imagessorted from best (a) to worse (c) : a. A document where the bleed-through will not produce OCR errors, b. An other document with much more bleed-through that will produce OCR errors, and c. the worst case, all bleed-through pixels will be identified as ink pixels by a binarization method

3. USING THE MEASURES TO PREDICT THE OCR ERROR RATE

In this section our goal is to create an OCR error rate prediction model in order to prove the relevance of our measures. We used two OCR systems : the OCR Abbyy FineReader[‡] and OCRopus [§] giving us two different prediction model since both OCR do not have the same results. In order to correctly calculate the OCR error rate, we used an application of the Levenshtein string distance statistic presented in.[13] Creating a prediction model can be done in several steps :

- first, a dataset is created containing for each document image its ground truth (the text) and its measures. This dataset must be as heterogeneous as possible. Latter is then randomly separated in two sub-set: the prediction corpus and the validation corpus.
- secondly, a stepwise regression is done on the prediction corpus. This statistical technic will create an optimal prediction model based on our measures.
- finally, the prior prediction model is tested on the validation corpus in order to test its sharpness.

3.1 An heterogeneous ancient document image dataset and measures annalysis

The lack of OCR truth on the corpus of old document that is in our possession forced us to work on synthetic documents. We chose to use a software[?] able to create old documents with the truth associated. We extended this software to generate large amounts of random old documents :

[‡]http://france.abbyy.com/

[§]http://code.google.com/p/ocropus/

- for the text we used the web service *Lorem Ipsum*[¶] which is simply random text,
- we varied the size of the margins from 50 to 200 pixels,
- the line spacing from 30 to 50 pixels,
- the number of columns and rows from 1 to 3, in order to have multi-column and multi-row documents, the space between rows and columns was also random and varied from 30 pixels to 100 pixels.
- the percentage of text on the over all resulting page, randomly chosen from 0% (white page with only bleed-through) and 100%.
- the background was chosen randomly in a directory of realistic backgrounds (white pages from ancients documents),
- the font was chosen randomly between a set of three fonts.
- the quantity of bleed-through was created using a model[6] based on the diffusion of different layers : the recto ink, the background and the verso ink.

Where articles like *Prediction of OCR accuracy*[2] in which 21 pages are used to train their model, we created a set of 191 ancients documents. This set is divided in two subsets : the prediction corpus (90 %) and the validation corpus (10%) in which the documents were chosen randomly in the original set.

Measures	Mean	Standard-deviation	Minimum	Best Case	Maximum	Worst Case
$\mathcal{MI}_i$	0.57	0.089392	0.373813	1	0.822115	0
$\mathcal{MI}_b$	0.22	0.08041874	0.0536898	0	0.377166	1
$\mathcal{MQ}$	16,47	71.97754	0.698371	0	442.018	∞
$\mathcal{MA}$	0.49	0.1354490	0.290468	0	0.945854	1
$\mathcal{MS}$	0.37	0.1007081	0.0578035	0	0.587493	1
$\mathcal{MSG}$	2.23	0.5172265	0	0	2.60243	∞
OCR error rates[13]						
Abbyy Fine Reader	59,5	19,5	20,5	0	100	100
OCRopus	77,5	11.00037	48,5	0	100	100

Table 3. Measures distribution on the prediction corpus (90%) on the over all dataset (best and worst cases are theoritical). Concerning the measures, our dataset is heterogeneous. It is worth noting the mean error rate difference between Abby and OCRopus. OCRopus poorer performance suggest that a training could be done. $\mathcal{MQ}$ and $\mathcal{MSG}$ worst case values is due to the fact that both are normed but not between 0 and 1.

The study of table 3 on which the measures distribution is shown indicates that our dataset has heterogeneous bleed-through caracteristics. Indeed, measures have their values well distributed around their mean, minimums and maximums are close to best case and worst case theoretical values. *Worst Case* is the measure value that is likely to give a lot of OCR errors, and on the contrary *Best Case* is the measure value that is likely to give no OCR errors.

3.2 Prediction model of OCR error rate trained on 90% of our dataset

An univariate linear regression is done in order to analyze the correlation between each measure and the OCR error rate. This shows that all of the measures are linked to the OCR errors but none of them are able to individually predict the OCR error rates. Indeed, The proportion of variability Rsquare was often close to 0 instead of 1. This fact indicates that each measures has its relevance concerning the OCR errors rates. A stepwise regression[‖] includes regression models in which the choice of predictive variables is carried out by an automatic procedure.[14] This leads to a pure causality model without any confusion between our measure and the OCR error rate presented in table 4.

[¶] http://www.lipsum.com/

[‖] The stepwise regression method could be detailed and explained in the final manuscript.

Measures	OCR Abbyy prediction equation coefficient	p-value	OCRopus prediction equation coefficient	p-value
$\mathcal{MI}_i$	97,11	<0.0001	74,12	<0.0001
$\mathcal{MI}_b$	195,25	<0.0001	147,09	<0.0001
$\mathcal{MQ}$	-	-	0.16	0.003
$\mathcal{MA}$	19,8	<0.0001	-16,99	0.02
$\mathcal{MS}$	-	-	-74,5	<0.0001
$\mathcal{MSG}$	-22,19	0.0008	16,08	<0.0001

Table 4. Prediction model of the two OCRs (OCRopus and Abbyy Fine Reader). It is worth noting that for OCRopus, all of our measures are highly significant when trying to explain the OCR error rate. The lower the p-value, the less likely the result is if the null hypothesis is true, and consequently the more *significant* the result is.

Abbyy FineReader : The Measures $\mathcal{MI}_i$, $\mathcal{MA}$, $\mathcal{MSG}$ and $\mathcal{MI}_b$ are selected in the Abbyy FineReader prediction model. The fact the $\mathcal{MQ}$ and $\mathcal{MS}$ are removed does not indicates that those two are not significant. They are removed because they do not improve the model (to predict Abbyy FineReader error rate, $\mathcal{MA}$ and $\mathcal{MSG}$ are enough). Various indicators are calculated. Firstly, the R-square[15] which provides a measure of how well future outcomes are likely to be predicted by the model, equals to 0.96. Secondly, the Adjusted R-square which unlike the R-square, increases only if the new term improves the model more than would be expected by chance equals 0.96. Finally **R**oot **M**ean **S**quared **E**rror (RMSE) which measures the average of the square of the *error*, equals to 12.77. These values are very good, R-square best value is 1 and since the lower the RMSE value is the better the model is, we can say that our prediction model is very accurate.[15]

$$Prediction \quad = \quad 97,11 * \mathcal{MI}_i + 195,25 * \mathcal{MI}_b + 19 * \mathcal{MA} - 22.19 * \mathcal{MSG}$$

OCRopus : The stepwise regression on OCRopus error rates create a prediction model in which all measures are kept. Moreover, both model are a little different implies that both OCR error rates are not correlated. This is confirmed with a Pearson[16] correlation test and indicates that both OCRs use different methods (binarization, dictionaries). The same indicators have been calculated on this model : R-square and Adjusted R-square equal 0.99 and the RMSE equals to 7.5. These indicators show[15] that the prediction model is even more accurate the Abbyy FineReader's one.

$$Prediction \quad = \quad 74,12 * \mathcal{MI}_i + 147,09 * \mathcal{MI}_b + 0.16 * \mathcal{MQ} - 16,99 * \mathcal{MA} - 74 * \mathcal{MS} + 16,08 * \mathcal{MSG}$$

3.3 Predicting Model Validation on 10% of our dataset

In order to test if the prediction model can predict OCR error rates, it needs to be validated. The validation corpus corresponds to 10% of the original dataset. These documents were chosen randomly and are not used for the prediction model. From our prediction equations we create scores of OCR error rate for our validation corpus. This leads us with two set of values per OCR : the prediction, the OCR error rate truth.

	Slope Coefficient	R-square	Adjusted R-square	RMSE
OCR Abbyy FineReader	1.006	0.97	0.97	11,03
OCROpus	0.99	0.99	0.99	7

Table 5. Indicators show that our predictions using the two models are very close to the real OCR error rate.

A linear regression between the prediction scores and the truths shows that both prediction models are accurate enough to work on other documents (table 5). For Abbyy FineReader, the linear regression gives a slope coefficient of 1.006 (where a perfect correlation has a coefficient equal to 1), a R-square of 0.97, and a RMSE of 11.03. The prediction on the validation corpus is very accurate.[15] OCRopus validation is also very good : the slope coefficient is 0.99, R-Square equals 0.99, and RMSE equals 7.

4. CONCLUSION AND RESEARCH PERSPECTIVES

This paper introduce 6 measures able to measure the bleed-through defect on very old documents. Based on these measures, we are able to create prediction models of OCR error rates. Two models are given : one for Abbyy FineReader a commercial and very powerfull OCR and one for OCRopus, sponsored by google. The accuracy of the two models proves the relevance of our 6 measures concerning bleed-through evaluation. Therefore, we are able to predict OCR error rates from digitized very old documents.

We have many perspectives around this work, OCR error rate prediction is just one of many usages that can be done with these measures. It could also be interesting to study the effect of bleed-through on more simple tasks such as binarization methods. An other perspective is to run our measures on real ancient documents to create OCR prediction models that are even more accurate. To do so, and since we need the registered verso and recto, we will interest our-selves in future works with registration algorithms and the search of a corpus of very old documents with their OCR ground truth.

ACKNOWLEDGMENTS

We would like to thanks Arkhenum [**] for letting us observe, analyze and study the bleed-through defect on real digitized ancient documents. This work is done in the Polinum project context [††].

REFERENCES

[1] M. Cannon, J. Hochberg, and P. Kelly, "Quality assessment and restoration of typewritten document images," *International Journal on Document Analysis and Recognition* **2**(2), pp. 80–89, 1999.

[2] L. Blando, J. Kanai, T. Nartker, and J. Gonzalez, "Prediction of ocr accuracy," *Proceedings of the Third ...* , Jan 1995.

[3] S. Rice, J. Kanai, and T. Nartker, "An evaluation of ocr accuracy," *... Research Institute* , Jan 1993.

[4] M. Cannon, P. Kelly, and S. Iyengar, "An automated system for numerically rating document image quality," *Proceedings 1997 ...* , Jan 1997.

[5] K. Knox, "Show-through correction for two-sided documents," 1997.

[6] R. Moghaddam and M. Cheriet, "Low quality document image modeling and enhancement," *International Journal on Document Analysis and Recognition* **11**(4), pp. 183–201, 2009.

[7] E. Dubois and A. Pathak, "Reduction of bleed-through in scanned manuscript documents," *Conference* , pp. 177–180, 2001.

[8] R. F. Moghaddam and M. Cheriet, "Rsldi: Restoration of single-sided low-quality document images," *Pattern Recognition* , Jan 2009.

[9] A. Tonazzini, E. Salerno, and L. Bedini, "Fast correction of bleed-through distortion in grayscale documents by a blind source separation technique," *International Journal on Document Analysis and Recognition* **10**(1), pp. 17–25, 2007.

[10] H. Nishida and T. Suzuki, "Correcting show-through effects on document images by multiscale analysis," *Pattern Recognition* **3**, p. 30065, 2002.

[11] G. Sharma, "Show-through cancellation in scans of duplex printed documents," *IEEE Transactions on Image Processing* **10**(5), pp. 736–754, 2001.

[12] M. Gupta, N. Jacobson, and E. Garcia, "Ocr binarization and image pre-processing for searching historical documents," *Pattern Recognition* , Jan 2007.

[13] R. Soukoreff and I. MacKenzie, "Measuring errors in text entry tasks: an application of the levenshtein string distance statistic," *CHI'01 extended abstracts on Human factors in computing systems* , p. 320, 2001.

[14] M. Thompson, "Selection of variables in multiple regression: Part i. a review and evaluation," *... Statistical Review/Revue Internationale de Statistique* , Jan 1978.

[15] N. Draper and H. Smith, "Applied regression analysis," *explorer.csse.uwa.edu.au* , Jan 1981.

[16] G. Nahler, "Pearson correlation coefficient," *Dictionary of Pharmaceutical Medicine* , Jan 2009.

[**] http://www.arkhenum.fr/

[††] http://www.polinum.net/