



HAL
open science

Impact of Image Enhancement Methods on Automatic Transcription Trainings with eScriptorium

Pauline Jacsont, Elina Leblanc

► **To cite this version:**

Pauline Jacsont, Elina Leblanc. Impact of Image Enhancement Methods on Automatic Transcription Trainings with eScriptorium. 2023. hal-03831686v2

HAL Id: hal-03831686

<https://hal.science/hal-03831686v2>

Preprint submitted on 14 Feb 2023 (v2), last revised 4 Jul 2023 (v4)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Impact of Image Enhancement Methods on Automatic Transcription Trainings with eScriptorium

Pauline Jacsont & Elina Leblanc

University of Geneva

Abstract:

This study stems from the *Desenrollando el cordel (Untangling the cordel)* project, which focuses on 19th-century Spanish prints editing. It evaluates the impact of image enhancement methods on the automatic transcription of low-quality documents, both in terms of printing and digitisation. We compare different methods (binarisation, deblur) and present the results obtained during the training of models with the *Kraken* tool. We show that binarisation methods give better results than the other, and that the combination of several techniques did not significantly improve the transcription prediction. This study shows the significance of using image enhancement methods with *Kraken*. It paves the way for further experiments with larger and more varied corpora to help future projects design their automatic transcription workflow.

Keywords: image enhancement methods, binarisation, deblur, printed documents, Spanish literature

1. Introduction

The library of the University of Geneva holds a collection of almost 1'000 Spanish chapbooks, printed during the 19th century by several printers across Spain. Chapbooks, also known as *pliegos de cordel*, consist of a few pages (4 to 8) and are in quarto format. They recount real or fictitious events, share songs and poems, or prayers and other religious writings (Figure 1). The Geneva collection is the object of the *Untangling the cordel* project [1], which aims at studying and promoting these documents through a digital library.

As one of the main objectives of the project is to analyse the chapbooks' content, automatic transcription represents a significant step in our editorial workflow to publish diplomatic digital editions using the XML-TEI standard. After testing several tools, including *ABBYY FineReader* and *Transkribus* [2], we chose *Kraken* and its graphic interface *eScriptorium*¹ [3].

Several challenges arose during our initial experiments with this tool. First, the page segmentation phase is faced with the complex layout of the document. Chapbooks' pages can

¹ At the beginning of the project, we carried out our automatic transcription experiments with *ABBYY FineReader* and *Transkribus*. These tools allowed us to create our Ground Truth (GT) and transcribe a part of our collection. However, after one year, the University of Geneva developed an automatic transcription platform called FoNDUE, based on the *Kraken/eScriptorium* tools. We became beta-testers of this new platform, which explains the use of different tools during our project. All the experiments we describe in this paper were performed exclusively with *Kraken*.

display up to three columns. There can be variation in the layout of the pages even within the same chapbook.



Figure 1: Examples of chapbooks (From left to right: José María Moreno, Carmona, 1859; José María Moreno, Carmona, [s.d.]; Imp. El Abanico, Barcelone, [s.d.]; J. Jepús, Barcelone, 1884)

Then, during transcription, the variety of employed fonts used poses a challenge for *Kraken*, especially in the title sections where most of the errors produced by our model cluster. Regarding the core text, the quality of the print media (paper and ink) complicates character recognition. Indeed, Spanish chapbooks – also known as cheap prints – were printed in mass on poor-quality paper. This often results in bleed-through, which blends the text on the recto with that of the verso (Figure 2).

In our case, the quality of the digital facsimiles adds another layer of difficulty. Indeed, before the project began, several digitisation campaigns were carried out with various scanners, and by different staff members of the library and the Spanish Unit of the university. Therefore, the resulting digitised corpus is heterogeneous. While some chapbooks (33% of the corpus) are in TIFF with a resolution of 300 dpi, most of the corpus (67%) is in PDF format with a relatively low resolution (72 dpi). For dissemination, these PDF documents were converted into JPG, which resulted in a further deterioration in quality. The resulting images are indeed heavily pixelated, which adds noise to the recognition of the characters by our model (Figure 2).

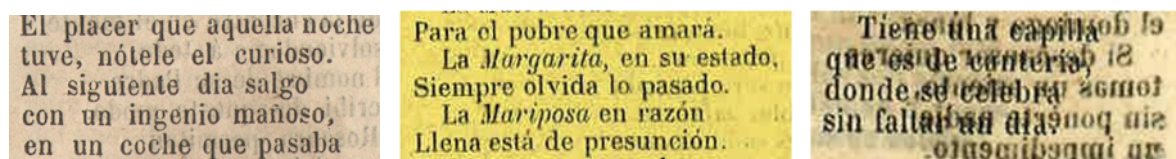


Figure 2: (From left to right) A TIFF image, a pixelated JPG image and a JPG image with bleed-through

In this paper, we focus on one of the problems mentioned above: the low quality of our data, both in terms of printing and digitisation. We propose a comparison of different image enhancement methods to address it, namely binarization and deblurring, and describe their impact on the accuracy of our transcription model.

After presenting related works, the third part explains our workflow and the image enhancement methods we included in our study. The fourth section presents the effectiveness of each method tested to improve our model, and the method we chose for our data. Finally, we discuss the replicability of this experiment with other types of documents. We also address

the possibilities offered by image enhancement methods to improve the results of models trained with *Kraken*.

2. Related works

With the advent of deep learning technologies, image enhancement is widely used to improve the effectiveness of many image processing tasks, such as segmentation, face detection or optical character recognition. In a recent survey about image enhancement methods [4], the authors distinguish six different tasks to improve the quality of an image, depending on the type of damage: binarization, deblurring, denoise, defading, watermark removal and shadow removal. For instance, binarization and denoising are preferred for document damage (wrinkles, stains, bleed-through); defading and deblurring are mostly used to improve the quality of a digitisation and its exposure.

From our analysis of the scientific literature, binarization appears as the main approach projects adopt to improve their damaged images of printed or handwritten historical texts prior to the automatic transcription. Early binarization approaches can be categorised into global and local thresholding methods. Global methods, amongst which the Otsu method [5] is the most well-known, apply the same threshold to the entire image and are effective for documents with a high contrast between the foreground and the background [6]. Local methods, such as Niblack [7] or Sauvola [8], determine a threshold based on local statistics within a specific window so that local variations in an image can be better taken into account [6]. A comparison of these different solutions for the improvement of OCR predictions with *ABBYY FineReader* can be found in Gupta, Jacobson, and Garcia [9]. They found that the Otsu algorithm gave the best results for automatic transcription of English newspapers, digitised at low-resolution and pixelated.

Recently, several studies have proposed to improve these methods, especially for handwritten documents. Boiangiu et al. [6] propose to modify the Niblack method by dynamically defining the window used to determine the threshold, instead of using a predefined value. They conclude that this method better harmonises the binarization of images with irregular brightness repartition. Ntirogiannis, Gatos and Pratikakis [10] propose a combination of global and local binarization. This method proves to be efficient in detecting faint characters and removing bleed-through from highly damaged handwritten documents. Adaptive binarization was also chosen by the developers of *Kraken*. Indeed, binarization is one of the available pre-processing options when a training is launched. This method dynamically calculates the difference between the highest and lowest thresholds for different regions of an image².

As for deblurring – which in our case seems to be a relevant approach to improve our pixelated digitisations –, most studies are related to object or face detection. Usually, for this type of task, methods rely on blind deconvolution methods. However, several works showed that convolutional neural networks (CNN) are better at deblurring text images [11, 12]. Other approaches suggest Generative Antagonistic Network (GAN) methods to deal with a heterogeneous corpus composed of faces pictures and text images to reconstruct high-resolution images from low-resolution ones [13].

² To our knowledge, no documentation or evaluation has been published about the method chosen by the developers of *Kraken*. These assumptions are based on the code available on the project's *GitHub*: <https://github.com/mittagessen/kraken/blob/master/kraken/binarization.py>

However, as pointed out by Anvari and Athitsos [4], while image enhancement methods prove their effectiveness in improving text images, few papers give details about their impact on the accuracy of automatic transcription models. We can mention some experiments with *ABBYY FineReader* and *Tesseract* in [9, 11, 12, 14]. In [6] and [10], the authors propose an evaluation with their own HTR systems. Furthermore, these works focus their attention on only one method to improve the images of their historical documents.

Therefore, in this paper, we study the benefit of different image enhancement techniques on the predictions of our model for Spanish printed documents. To achieve this, we use *Kraken*, an experiment that, to our knowledge, has never been carried out with this specific tool. It leads us to compare the native binarization method of *Kraken*, primarily thought for handwritten documents, with other solutions.

3. Method

To evaluate the impact of image enhancement on the effectiveness of our automatic transcription models, we conducted a series of tests comparing different binarization approaches, listed below. The Ground Truth (GT) consists of 198 pages transcribed with *ABBYY FineReader*, and then manually corrected. This corpus was divided into three subsets³. Each set is used in a different phase:

- The first set, 80% of GT, is used by *Kraken* to train the model;
- The second set, 10% of the corpus, is used by the tool to evaluate each iteration during training (validation set);
- The last set, 10% of the corpus, is used to evaluate the results of the model on documents it has never encountered (test set).

To compare the results, each training and test was carried out with the same sets. The division was done manually to ensure that all sets were equally difficult: each set has the same proportion of front pages (20%) and low-quality prints. Indeed, some pages have a more complex layout than others: front pages are more difficult to process with automatic transcription tools because of the variety of fonts.

Each time we pre-processed the data using a notebook⁴ for the entire corpus: we chose this method because it was easy to implement and allowed the use of open-source Python libraries - Open CV and Scikit-image - where a pre-processing collection is available. Our experiment is based on the work of Gupta, Jacobson and Garcia [9] on automatic transcription of English newspapers, which resemble our corpus in terms of layout and image quality.

To choose the best binarization method for our Spanish chapbooks, we reproduce a part of their work by comparing different solutions. We did not use exactly the same methods but adapted their experiment to our skills and time constraints. Thus, we chose methods that were easy and fast to implement, namely the thresholding binarization of OpenCV, Otsu, Niblack,

³ The details of the three sets are available on our *GitHub* repository : <https://github.com/DesenrollandoElCordel/FoNDUE-Spanish-chapbooks-Dataset/tree/main/Grountruth/Split>

⁴ The OpenCV, NumPy, SciPy and Scikit-image Python libraries are used in the notebook available here: <https://github.com/DesenrollandoElCordel/FoNDUE-Spanish-chapbooks-Dataset/blob/main/Varios-GroundTruth-TEST/ImagesTreatments.ipynb>.

Sauvola and the native binarization of *Kraken*. We also tested two types of Gamma correction: one to lighten the image (Gamma 1) and the other to darken it (Gamma 2) to experiment with contrasts.

We also tried a deblur technique, which follows the state of the art of Zahra Anvari and Vassilis Athitsos [4], that focuses, in part, on historical documents that, like those in our corpus, are degraded and damaged.

Each automatic transcription model⁵ trained in this test phase was run according to basic training⁶; the command used is:

```
Ketos train -f alto -t train.txt -e val.txt -d cuda data/*.xml
```

For each performed pre-processing, a training was followed by an evaluation test with the set prepared for this purpose. The same XML-ALTO files were used for each test; only the images changed depending on the type of pre-processing.

In a second phase, we conducted a series of tests to evaluate the effectiveness of different image enhancement methods when combined, using the pre-processes that had produced the best results in the first phase.

4. Results

The results of the models trained in the first test phase (i.e. the models created with the images having undergone only one pre-processing) are presented in Table 1. The results are given for the accuracy per character on the training tests (validation test) and the evaluation tests.

	<i>Validation test</i>	<i>Evaluation test</i>
Without pre-processing	94.20 %	91.73 %
Thresholding binarisation	95.99 %	94.42 %
Otsu binarisation	96.04 %	94.13 %
Niblack binarisation	97.80 %	94.36 %
Sauvola binarisation	96.29 %	94.33 %
Kraken binarisation	96.55 %	88.93 %
Deblur	95.94 %	93.47 %
Gamma 1	96.37 %	92.24 %
Gamma 2	96.54 %	93.43 %

Table 1: Models with single pre-processed images

⁵ The same tests were also carried out on the segmentation models, but this did not lead to improved results.

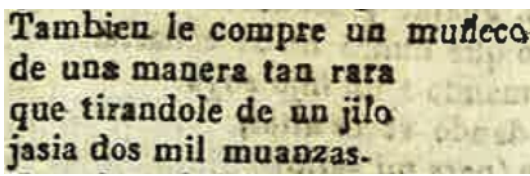
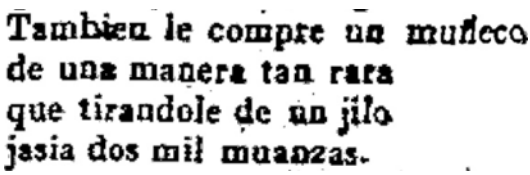
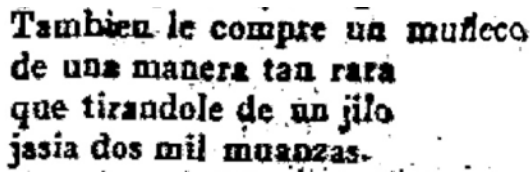
⁶ The submission script is available at the following address:
<https://github.com/DesenrollandoElCordel/FoNDUE-Spanish-chapbooks-Dataset/blob/main/Grountruth/submission-script.sh>

=== report ===				=== report ===				=== report ===			
27768	Characters			27768	Characters			27768	Characters		
2296	Errors			1549	Errors			1566	Errors		
91.73%	Accuracy			94.42%	Accuracy			94.36%	Accuracy		
983	Insertions			658	Insertions			630	Insertions		
75	Deletions			63	Deletions			81	Deletions		
1238	Substitutions			828	Substitutions			855	Substitutions		
Count	Missed	%Right		Count	Missed	%Right		Count	Missed	%Right	
22240	1490	93.30%	Latin	22240	1052	95.27%	Latin	22240	1029	95.37%	Latin
5528	731	86.78%	Common	5528	434	92.15%	Common	5528	456	91.75%	Common
Errors	Correct-Generated			Errors	Correct-Generated			Errors	Correct-Generated		
124	{ , } - { }			93	{ , } - { }			75	{ SPACE } - { }		
110	{ . } - { }			65	{ SPACE } - { }			66	{ i } - { i }		
97	{ . } - { , }			60	{ . } - { }			64	{ . } - { }		
94	{ SPACE } - { }			46	{ i } - { i }			62	{ , } - { }		
76	{ i } - { i }			45	{ a } - { }			46	{ a } - { }		
63	{ a } - { }			43	{ , } - { . }			36	{ , } - { . }		
49	{ l } - { }			32	{ r } - { }			34	{ } - { , }		
48	{ s } - { }			31	{ s } - { }			34	{ . } - { , }		
44	{ e } - { }			28	{ i } - { }			32	{ e } - { }		
39	{ r } - { }			27	{ l } - { }			31	{ s } - { }		

Figure 3: (From left to right) Test reports of the model without pre-processing; with the Thresholding binarization; with the Niblack model.

The detailed results (Figure 3) show that the models trained with the binarized images successfully recognise commas and dots. However, some errors remain, such as confusion between *i* and *í* and problems in recognising certain characters such as *a*.

To measure the efficiency of the models, we compared their ability to transcribe a few lines of text on the most complex documents (in this case, a chapbook with poor scan quality, block artefacts specific to the JPG format and smudges during printing).

Without enhancement		Tamfién le compre ua muøéro, de uns manera tan rará que tirandole de un øilo jasia dos mil muanzas.
Otsu		Tamfiea le compre nu munésó de una manera tan rará que tirandole de un jilo løsia dos mil muanzasø
Thresholding		Tambien le compre ua mulezo de uns manera tan rara que tirandole de un øilo jasia dos mil muanzas.

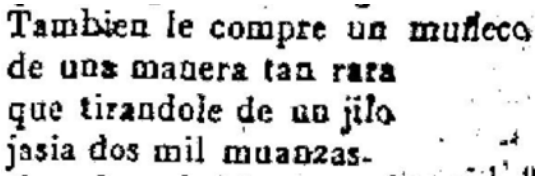
Niblack		Tambien le compre un muleco de uns manera tan rara que tirandole de un jilo jasia dos mil muanzas.
---------	---	---

Table 2: Transcription of complex lines. Ground truth: Tambien le compre un muñeco / de una manera tan rara / que tirandole de un jilo / jasia dos mil muanzas. Insertions are coloured in blue, substitutions in green and deletions in red.

The results shown in Table 2 confirm those obtained previously. The thresholding and Niblack models perform better but still have difficulty transcribing the letter *a*. Both make the mistake with *una* which is transcribed as *uns*. The model without pre-processing and the one with the Otsu method make more errors, especially on punctuation.

It is worth noting that none of the models manages to transcribe the tilde *n* (*muñeco*) correctly. Therefore, despite the similarity of our documents to those of Gupta, Jacobson and Garcia, which suggests that the Otsu method would also be most efficient for us, we do not get the same results. In our case the Niblack method seems to be the most appropriate.

A second series of tests was performed by combining binarization with another image enhancement method. We did not launch any automatic transcription model training if the image obtained was clearly unusable for the tool; this is the case, for example, when combining a deblurring method (sharpening kernel) with the Thresholding binarization. (Figure 4).



Figure 4: Example of unusable pre-processing (combination of techniques: Sharpening and Thresholding binarization)

Four other models were trained with double pre-processed images. The results obtained are shown in Table 3, with the accuracy per character in percent.

	Validation test	Evaluation test
Gamma 2 + Niblack	96.06 %	94.20 %
Gamma 2 + Sauvola	95.98 %	94.05 %
Gamma 2 + Otsu	95.53 %	93.44 %

Deblur + Niblack	96.02 %	94.05 %
------------------	---------	---------

Table 3: Models with multiple pre-processed images.

Contrary to our initial hypothesis, the multiple-image pre-processing method did not improve the results obtained in the first phase. However, the accuracy of the trained models remains higher than that of the model trained on images without pre-processing.

These different tests show the importance and influence of image pre-processing methods on automatic transcription predictions. Ultimately, the best method for improving our results, and the method applied to the rest of the images in our corpus, is the Niblack binarization. This model is interesting in that it performs better in the recognition of Latin characters than the thresholding binarization.

5. Conclusions and future works

In this paper, we have shown that the use of image enhancement methods can significantly improve the predictions of a model trained with *Kraken*. This empirical approach was carried out on a corpus of printed documents, with low resolution, bleed-through and JPG block artefacts.

By reproducing some of the experiments of Gupta, Jacobson, and Garcia, we found that the best pre-processing for our corpus was not Otsu binarization, but Niblack binarization. The choice and efficiency of image pre-processing in transcription depend largely on the specifics of the corpus and the quality of its digitisation. A similar experiment was carried out in another project, and the results obtained were very different: this experience made with a manuscript⁷, that was digitised at high resolution. The results showed that image enhancement methods did not improve the prediction of the models⁸.

These results pave the way for further experiments and research. Indeed, it would be useful to reproduce these experiments on a large scale to define recommendations on the best image enhancement methods to use with *Kraken* and other user-friendly automatic transcription tools, depending on the nature of texts and the quality of the images.

Finally, we used traditional and well-known methods for our binarization experiments. However, new approaches using Deep Learning architectures (CNN [17] or GAN [18]) are currently being tested. They are still in their infancy, but it would be interesting to include them in future work to analyse their impact on the accuracy of a model trained with *Kraken* compared to threshold and adaptive methods.

Authors' contributions

PJ has conceived and performed all the experiments with the different image enhancement methods. EL has elaborated the general workflow of the project and helped design this study.

⁷ This experiment has been done in the context of the digital edition project of a Latin manuscript [15] of the Bern Burgerbibliothek. This document from the 16th century was digitised in TIFF format (400 dpi).

⁸ The results of these tests are available in detail in [16].

Acknowledgements

We sincerely thank the Fondue project team, Simon Gabay, Pierre Künzli, Christophe Charpilloz and Jean-Luc Falcone, for their constant support and invaluable responsiveness, which made all these experiments possible.

Bibliography

1. Leblanc, Elina, and Constance Carta. 2021. Le projet « Démêler le cordel » : une bibliothèque numérique pour l'étude de la littérature éphémère espagnole du XIX e siècle. In *Humanistica 2021, Rennes, 10-12 mai 2021*, 100–101.
2. Kahle, Philip, Sebastian Colutto, Günter Hackl, and Günter Mühlberger. 2017. Transkribus - A Service Platform for Transcription, Recognition and Retrieval of Historical Documents. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, 04:19–24. <https://doi.org/10.1109/ICDAR.2017.307>.
3. Kiessling, Benjamin, Robin Tissot, Peter Stokes, and Daniel Stökl Ben Ezra. 2019. eScriptorium: An Open Source Platform for Historical Document Analysis. In *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, 2:19–19. <https://doi.org/10.1109/ICDARW.2019.10032>.
4. Anvari, Zahra, and Vassilis Athitsos. 2022. A Survey on Deep learning based Document Image Enhancement. arXiv. <https://doi.org/10.48550/arXiv.2112.02719>.
5. Otsu, Nobuyuki. 1979. A Threshold Selection Method from Gray-Level Histograms. *IEEE Transactions on Systems, Man, and Cybernetics* 9: 62–66. <https://doi.org/10.1109/TSMC.1979.4310076>.
6. Boiangiu, Costin-Anton, Alexandra Olteanu, Alexandru Stefanescu, Daniel Rosner, Nicolae Tapus, and Mugurel Andreica. 2011. Local Thresholding Algorithm Based on Variable Window Size Statistics. In *Proceedings of the 18th International Conference on Control Systems and Computer Science (CSCS)*, 2:647–652. Bucharest, Romania.
7. Niblack, Wayne. 1986. *An Introduction to Digital Image Processing*. Prentice-Hall International.
8. Sauvola, J., and M. Pietikäinen. 2000. Adaptive document image binarization. *Pattern Recognition* 33: 225–236. [https://doi.org/10.1016/S0031-3203\(99\)00055-2](https://doi.org/10.1016/S0031-3203(99)00055-2).
9. Gupta, Maya R., Nathaniel P. Jacobson, and Eric K. Garcia. 2007. OCR binarization and image pre-processing for searching historical documents. *Pattern Recognition* 40: 389–397. <https://doi.org/10.1016/j.patcog.2006.04.043>.
10. Ntirogiannis, K., B. Gatos, and I. Pratikakis. 2014. A combined approach for the binarization of handwritten document images. *Pattern Recognition Letters* 35. *Frontiers in Handwriting Processing*: 3–15. <https://doi.org/10.1016/j.patrec.2012.09.026>.
11. Hradiš, Michal, Jan Kotera, Pavel Zemčík, and Filip Sroubek. 2015. Convolutional Neural Networks for Direct Text Deblurring. In *Proceedings of BMVC*. Vol. 10. <https://doi.org/10.5244/C.29.6>.
12. Gangeh, Mehrdad J., Sunil R. Tiyyagura, Sridhar Dasaratha, Hamid Motahari, and Nigel P. Duffy. 2019. Document Enhancement System Using Auto-encoders. In *Proceedings of the 33rd Conference on Neural Information Processing Systems*. Vancouver, Canada.
13. Xu, Xiangyu, Deqing Sun, Jinshan Pan, Yujin Zhang, Hanspeter Pfister, and Ming-Hsuan Yang. 2017. Learning to Super-Resolve Blurry Face and Text Images. In *2017 IEEE International Conference on Computer Vision (ICCV)*, 251–260. <https://doi.org/10.1109/ICCV.2017.36>.
14. Souibgui, Mohamed Ali, Yousri Kessentini, and Alicia Fornés. 2021. A Conditional GAN Based Approach for Distorted Camera Captured Documents Recovery. In *Pattern Recognition and Artificial Intelligence*, ed. Chawki Djeddi, Yousri Kessentini, Imran Siddiqi, and Mohamed Jmaiel, 215–228. Communications in Computer and Information Science. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-71804-6_16.

15. Sardi, Gasparo. 1561. *Toponomasia*. Bern Burgerbibliothek.
16. Jacsont, Pauline. 2022. Mise en valeur du patrimoine textuel grâce aux éditions numériques « Le cas du codex 174 de la Bibliothèque de la Bourgeoisie de Berne ». Université de Genève.
17. Akbari, Younes, Somaya Al-Maadeed, and Kalthoum Adam. 2020. Binarization of Degraded Document Images Using Convolutional Neural Networks and Wavelet-Based Multichannel Images. *IEEE Access* 8: 153517–153534. <https://doi.org/10.1109/ACCESS.2020.3017783>.
18. Khamekhem Jemni, Sana, Mohamed Ali Souibgui, Yousri Kessentini, and Alicia Fornés. 2022. Enhance to read better: A Multi-Task Adversarial Network for Handwritten Document Image Enhancement. *Pattern Recognition* 123: 108–141. <https://doi.org/10.1016/j.patcog.2021.108370>.