DocCreator: A New Software for Creating Synthetic Ground-Truthed Document Images

Nicholas Journet, Muriel Visani, Boris Mansencal, Kieu Van-Cuong, Antoine Billy

Journal of *Imaging*

MDPI

Article

DocCreator: A New Software for Creating Synthetic Ground-Truthed Document Images

Nicholas Journet [1,*,†], Muriel Visani [2,†], Boris Mansencal [1,†], Kieu Van-Cuong [3] and Antoine Billy [1]

1 Laboratoire Bordelais de Recherche en Informatique UMR 5800, Université de Bordeaux, CNRS, Bordeaux INP, 33400 Talence, France; boris.mansencal@labri.fr (B.M.); antoine.billy@labri.fr (A.B.)
2 Laboratoire Informatique, Image et Interaction (L3i), Université de La Rochelle, 17000 La Rochelle, France; muriel.visani@univ-lr.fr
3 LIPADE Laboratory, Paris Descartes University, 45, rue des Saints-Pères, 75270 Paris, CEDEX 6, France; van-cuong.kieu@parisdescartes.fr
* Correspondence: journet@labri.fr
† These authors contributed equally to this work. Other authors: Kieu Van-Cuong worked on degradation models, Antoine Billy worked on synthetic document reconstruction.

Received: 30 October 2017; Accepted: 5 December 2017; Published: 11 December 2017

Abstract: Most digital libraries that provide user-friendly interfaces, enabling quick and intuitive access to their resources, are based on Document Image Analysis and Recognition (DIAR) methods. Such DIAR methods need ground-truthed document images to be evaluated/compared and, in some cases, trained. Especially with the advent of deep learning-based approaches, the required size of annotated document datasets seems to be ever-growing. Manually annotating real documents has many drawbacks, which often leads to small reliably annotated datasets. In order to circumvent those drawbacks and enable the generation of massive ground-truthed data with high variability, we present DocCreator, a multi-platform and open-source software able to create many synthetic image documents with controlled ground truth. DocCreator has been used in various experiments, showing the interest of using such synthetic images to enrich the training stage of DIAR tools.

Keywords: synthetic image generation; document degradation models; performance evaluation; data augmentation for retraining and fine-tuning; DIAR

1. Introduction

Almost every researcher in the field of Document Image Analysis and Recognition (DIAR) had to face the problem of obtaining a ground-truthed document image dataset. Indeed, many DIAR tools (image restoration, layout analysis, text-graphic separation, binarization, OCR, etc.) rely on a preliminary stage of supervised training. Moreover, ground-truthed document image datasets are needed to evaluate these DIAR tools. Digital curators are the first users of these tools, e.g., for announcing expected OCR recognition rates together with automatic transcriptions of books [1]. One common solution is to use ground-truthed training and benchmarking datasets publicly available on the internet. For document images, the following databases are the most commonly used. For printed documents: Washington UW3 [2], LRDE [3], RETAS-OCR [4], PaRADIIT [5], etc.; for handwritten documents IAM database [6], RIMES [7], GERMANA [8], etc.; for graphical documents: chemical symbol database [9], logo databases [10,11], architectural symbol database [12] or musical symbol database CVC-MUSICMA [13]; camera-based document image analysis [14,15]. The International Association for Pattern Recognition, for instance, gathered some interesting datasets [16] mostly used for different conference competitions over the last two decades. The main international conference in document image analysis, ICDAR, references on its websites many contest datasets. However, very few of them

are reliably annotated, copyright-free, up-to-date or easily available to download. An alternative for researchers and digital curators is to create their own ground truth by manually annotating document images. In order to assist them in the tedious task of ground truth creation, multiple software have been proposed during the last two decades.

As detailed in Table 1, some are fully manual stand-alone software (Pink Panther (1998) [17], trueViz (2003) [18]), while others provide semi-automatic annotation modules (GEDI (2010) [19], Aletheia (2011) [20,21]). Some of the most recent solutions are based on an online collaborative platform (Transcriptorium (2014) [22], DIVADIAWI [23] (2015), [24] (2016), Recital manuscript platform [25] (2017)). Among non open-source solutions, some have an academic licence: [20,26]. These software assist the user in creating the ground truth associated with real documents, intrinsically limited in number because of acquisition procedures and copyright issues. Moreover, despite the use of such software, manual annotation remains a costly task that cannot always be performed by a non-specialist.

Another solution is available for getting (quickly and with lower human cost) large ground-truthed document image datasets. This solution, investigated since the beginning of the nineties [27], is to generate synthetic images with controlled ground truth. The authors of [28,29] propose two similar systems. They consist of using a text editor (e.g., Word-office, Latex, etc.) to automatically create multiple documents with varied contents (in terms of font, background, layout). Alternative approaches consist of re-arranging, in a new way, elements extracted from real images so as to generate (manually, semi-automatically or automatically) multiple semi-synthetic document images [12,30]. Recently, in particular with the advent of deep learning techniques which require huge masses of training data, the need for synthetic data generation seems to be ever-growing. In [31], among the 60,000 character patches that were used to train a convolutional network for text recognition, only 3000 were real.

In this paper we present DocCreator, an open-source and multi-platform software that is able to create virtually unlimited amounts of different ground-truthed synthetic document images based on a small number of real images.

Table 1. Technical and functional characteristics of existing annotation software. Six features are presented: export format, source availability, desktop/online software, groundtruthing assistance (whether the software provides features that help the user to quickly create the groundtruth), collaborative/crowd-sourcing software, and year of distribution.

	Export	Open-Source	Desktop/Online	Groundtruthing Assistance	Collaborative	Year
Software for manual ground truth creation						
Pink Panther [17]	ASCII	n/a	desktop	no	no	1998
TrueViz [18]	XML	yes	desktop	no	no	2003
PerfectDoc [32]	XML	yes	desktop	?	no	2005
PixLabeler [33]	XML	no	desktop	no	no	2009
GEDI [19]	XML	yes	desktop	yes	no	2010
DAE [34]	no	yes	online	yes	yes	2011
Aletheia [20,26]	XML	no	online/desktop	yes	no	2011
Transcriptorium [22]	TEI-XML	no	online	yes	yes	2014
DIVADIAWI [23]	XML	n/a	online	yes	n/a	2015
Recital [25]	no	yes	online	yes	yes	2017
Algorithms for synthetic data augmentation						
Baird et al. [27]	no	n/a	n/a	n/a	no	1990
Zhao et al. [28]	no	n/a	n/a	n/a	no	2005
Delalandre et al. [12]	no	n/a	n/a	n/a	no	2010
Yin et al. [30]	no	n/a	n/a	n/a	no	2013
Mas et al. [24]	no	n/a	n/a	n/a	yes	2016
Seuret et al. [35]	no	yes	n/a	n/a	no	2015
Software for semi-automatic ground truth creation and data augmentation capabilities						
DocCreator	XML	yes	online/desktop	yes	no	2017

As illustrated in Figure 1, DocCreator can handle the creation of ground-truthed synthetic images from a limited set of real images. Various realistic degradation models can be applied on

original document images, the resulting images being called semi-synthetic images in the rest of the paper. If there is no ground truth associated to the real images, DocCreator can create, with a given text, synthetic images that look like the real ones and their associated ground truth. Depending on the needs and expertise of the user, DocCreator can be used in a fully automatic mode, or in a semi-automatic mode where the user can interact with the system and tune its parameters. Visual feedback of the results is returned by the system. Degradations available in DocCreator can be applied on any type of document images. The DocCreator ability to create synthetic documents that mimic real ones is effective for typewritten and handwritten characters (as long as the characters are apart from one another). Images created with DocCreator have already been used in many DIAR contexts: text/background/image pixel classification [36]; staff removal [13,37,38]; and handwritten character recognition [39]. In this article we present how DocCreator can be useful to enhance a binarization algorithm and for OCR performance prediction. DocCreator could also be used, for example, for camera-based document image analysis and word spotting.

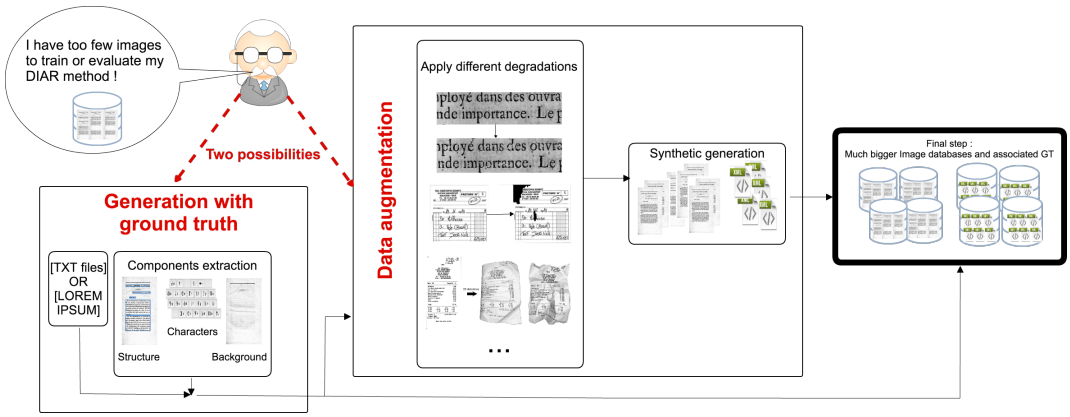


Figure 1. According to the needs of the DIAR researcher, it is possible to generate synthetic document images (and their ground truth) in different ways. First possibility: if a researcher has real document images but without any ground truth, DocCreator can generate synthetic images that look like the real ones, and of course, with the associated ground truth. Second possibility: a researcher has a ground-truthed database but it is too small or not heterogeneous enough. DocCreator provides several degradation algorithms to augment the dataset. By degrading text ink, paper shape or background colours it is possible to create a representative document image database where many defects are present. This complete database is finally useful for very precise performance evaluation or to provide multiple cases for retraining processes (in algorithms embedding a learning step).

DocCreator features compared to existing software are highlighted in Table 1. First of all, DocCreator is the only one that can create synthetic documents that mimic real ones. Besides, as it includes several degradation models, it provides an integrated solution to carry out data augmentation. DocCreator thus makes quickly available ground-truthed databases. It makes DocCreator a unique software that can be seen as a complementary tool to those mentioned in Table 1.

This paper is organized as follows. In Section 2, we present the methods used to extract document characteristics and to generate synthetic documents, while in Section 3 document degradation models are discussed. Section 4 highlights the advantages of DocCreator on various DIAR tasks, both for benchmarking and for retraining DIAR tools using data augmentation.

2. How to Create a Synthetic Document (with Ground Truth) That Looks Like a Real One?

The left part of Figure 1 illustrates the pipeline used in order to generate synthetic documents that look realistic. Given an original image, we extract the three main required components: (1) the font; (2) the background; and (3) the layout of the document. The system can then write any text with

the extracted font, onto the reconstructed background, with a layout similar to one of the original documents (see Figure 2).

The font is extracted using a semi-automatic method. First, an Optical Character Recognition system (OCR) automatically associates a label (Unicode value) to each character on the image. Tesseract OCR engine [40] is used. Then the user can modify the character properties (label, baseline, margins, etc.). Our software allows us to pair a given symbol with several character images in the extracted font. Thus, when the font is used to write a text, the software will randomly choose an image for the required symbol. This will make the final output look more realistic by reducing the strict uniformity between similar letters.

In order to correctly write text with this font, the baseline of each character has to be computed. The baseline is the imaginary straight line on which semi cursive or cursive text are aligned and upon which most letters "sit" and, below which, descenders extend. In order to extract each character baseline and deal with various documents, we propose a different approach from classical ones [40,41]. The main originality of our baseline extraction method is that the baseline is computed for each character individually instead of finding a baseline for the whole line. We evaluated this method on more than 5000 manually annotated baselines considered as the ground truth. The baseline extraction error rate is relatively close the one obtained using [40] (the same baseline extraction error rate). However, our method has the main advantage of being robust to skew orientation, the hand-written wavy pattern and unaligned columns in a same page. Besides, the inter words distance is automatically computed as the average of characters width. For the inter characters distance, none is specified by default. However, the user can interactively specify the left and right margin to better position the character relative to others. The GUI of DocCreator also allows the user to modify several parameters to improve the extracted characters. The user can change the baseline or the letter assigned to a character and smooth the border of a character. Via this semi-automatic font extraction method, the user is able to correct mistakes made by the OCR (frequent on old documents). From several testing sessions, we evaluate the time needed to correctly extract a font between 30 seconds (when the OCR works accurately) and 60 min (when the OCR fails and the user has to manually extract the characters).

Once the font is extracted, the background of the document can be computed. This background extraction is performed completely automatically. For that purpose, we apply an inpainting method to remove all the characters. We use the OpenCV implementation of [42].

To construct a realistic document, the layout of the document image is also extracted. Document image physical layout analysis algorithms can be categorized into three classes: top-down approaches [43], bottom-up approaches [44] and hybrid approaches [45,46]. As word segmentation is already available via Tesseract OCR (but not the complete document layout), we use a hybrid approach proposed by [45]. With only one parameter to adjust the number of extracted blocks, this method ensures a good layout segmentation of many different classes of typewritten documents. DocCreator, as an interactive software, leaves once more the possibility to adapt to the wished segmentation results. This method has the advantage of a very low computational cost, without any preprocessing training required.

At this point, the three characteristics used in the synthetic image generation process have been extracted (background, font and layout). The next step is to assemble these elements with a given text in order to build the final output, which is the created synthetic image and the associated XML ground truth. Figure 2 illustrates a synthetic image (right) created automatically from a given original document image (left). As this example illustrates, a complete automatic generation may still produce perfectible results. In particular, if the original image suffers from local deformations (as the original image in Figure 2), the characters extracted to build the font may have different forms or sizes, and, when assembled to compose the final document, may locally look too different and thus not realistic.

Obviously, one can combine fonts, background images, layout from different images and various texts, to generate many of synthetic document images.

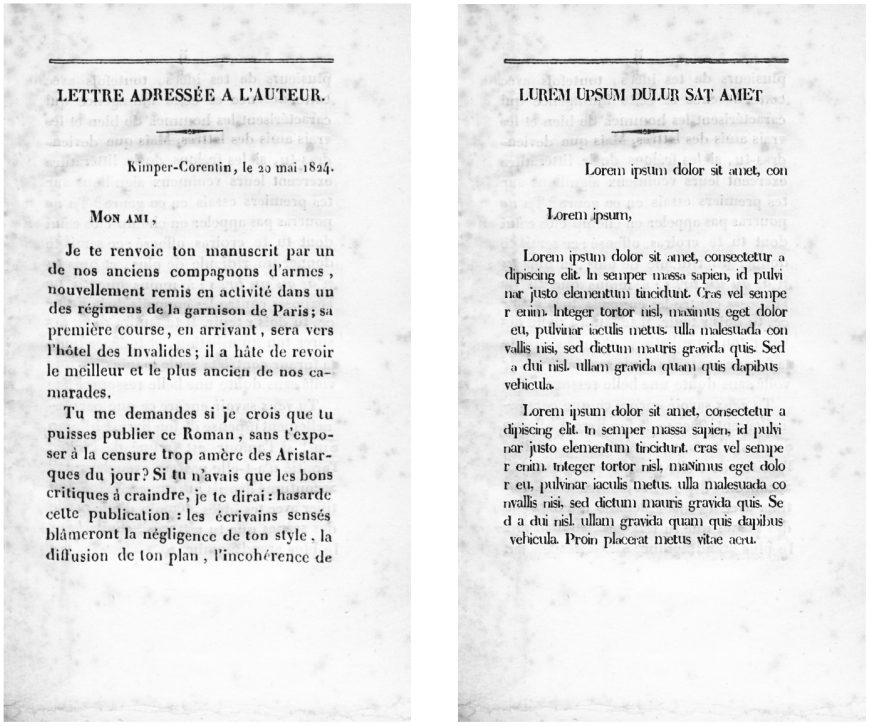


Figure 2. Synthetic document image generation. (**Left**) original document image. (**Right**) synthetic document image generated automatically with the random text "Lorem ipsum". The automatically generated image looks similar to the original one. The result is still perfectible. Here, as the original image suffers from local deformations, the characters extracted to build the font are quite different and may look too random when assembled on the synthetic document. A better font extraction or composition using the context to choose new characters may alleviate this problem.

3. Document Degradation Models

Physical degradation due to ageing, storage conditions or poor quality of printing materials may be present on documents.

DocCreator currently proposes seven degradation models.

As detailed in Figure 1 (right part), all these models can be applied on real images to extend any document image database. The user can interact with DocCreator in order to set the quantity of defects to generate.

In the following sections, we describe the main ideas of these seven degradation models. As DocCreator is an open source software, readers can consult the source code to get more details about the implementation of these models.

3.1. Ink Degradation

DocCreator provides a grayscale ink degradation model (detailed in [41]) able to simulate the most common character degradations due to the age of the document itself and printing/writing process, such as ink splotches, white specks or streaks. This model locally degrades the image in the neighbourhood of the characters boundaries. Noise is then generated to create some small ink spots near characters or to erase some characters ink area. Contrary to the well known Kanungo noise model [47] that works only on black and white images, this degradation method can process grayscale images. See Figure 3 for an ink degradation example.

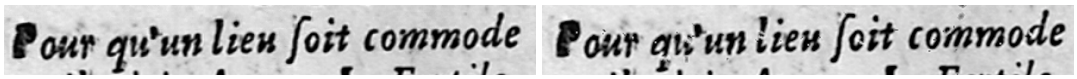


Figure 3. Ink degradation on an old document. (**Left**) original image. (**Right**) degraded image.

3.2. Phantom Character

The phantom character is a typical defect that appears on documents that have been manually printed (using a wooden or metal character). After many uses, a printing character can be eroded. It is thus possible that ink reaches the borders of the piece; borders are then printed on the sheet of paper. DocCreator provides an algorithm that reproduces such ink apparition around the characters. To be as realistic as possible, we have manually extracted more than 30 phantom defects from real images. These defects are then automatically put between characters following a patch-based algorithm.

The degradation algorithm works as follow: (1) the user provides an image and the percent of character to degrade; (2) characters are extracted using a connected component algorithm; (3) a list of characters is randomly set; (4) for each selected character; (4.1) a phantom defect is randomly selected from the manually extracted available defects; (4.2) the phantom defect is resized to fit with the character size; (4.3) to be realistic, the phantom defect is used only as a pattern; the pixels within the pattern are transformed using a patch algorithm inspired from [48] where a zone from another part of the document image is selected and copied within the patch.

See Figure 4 for an example.

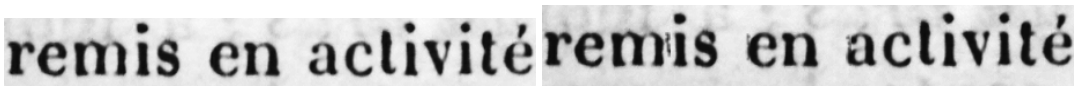


Figure 4. Phantom character apparition. (**Left**) original image. (**Right**) degraded image.

3.3. Paper Holes

Many old or recent document images contain holes. These holes have different shapes, sizes and locations. DocCreator provides an algorithm that creates different kinds of holes in a document image. This algorithm simply randomly applies holes extracted from real document images on a given document image. See Figure 5 for examples.

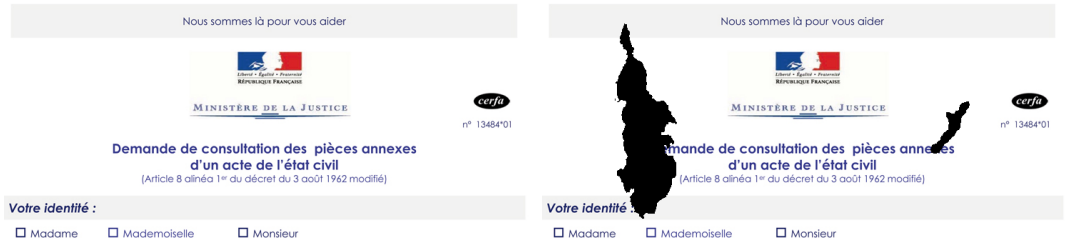


Figure 5. *Cont.*

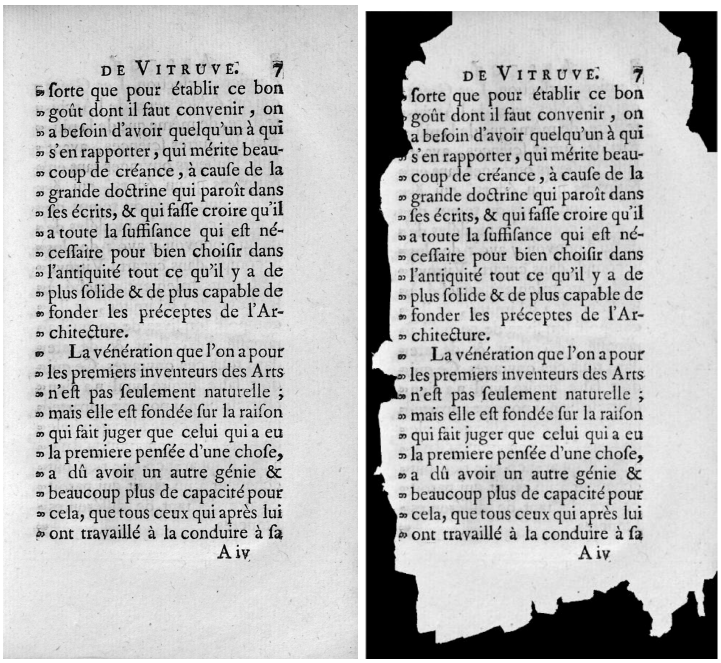


Figure 5. Hole degradation. (**Left**) original images. (**Right**) degraded images.

3.4. Bleed-Through

With DocCreator it is possible to add bleed-through defects. This algorithm is directly inspired from [49] that initially proposes an algorithm for erasing the bleed-through from a document image. By just giving an input recto image, an input verso image and the amount of wished degradation, a physical model is applied. This model mimics the verso ink that seeps through the recto side. This model simulates an anisotropic diffusion and each pixel at time $t+1$ is modified according to the pixels values at time t with the following equation:

$$I_{i,j}^{t+1} = I_{i,j}^{t} + \lambda * (c_{N_{i,j}}^{t} \cdot \bigtriangledown_N I_{i,j}^{t} + c_{S_{i,j}}^{t} \cdot \bigtriangledown_S I_{i,j}^{t} + c_{E_{i,j}}^{t} \cdot \bigtriangledown_E I_{i,j}^{t} + c_{W_{i,j}}^{t} \cdot \bigtriangledown_W I_{i,j}^{t})$$

where I is the recto image, V is the verso image, lambda a constant value in $[0; 0.25]$ and N, S, E, W are the mnemonic subscripts for North, South, East, West. $\bigtriangledown_N I_{i,j}^{t}$ and $c_{N_{i,j}}^{t}$ are defined as: $\bigtriangledown_N I_{i,j}^{t} = V_{i-1,j}^{t} - I_{i,j}^{t}$ and $c_{N_{i,j}}^{t} = \frac{1}{1+\frac{(V_{i-1,j}^{t} - I_{i,j}^{0})^2}{\sigma^2}}$. The user sets the number of iterations and thus the quantity of generated bleed-through.

See Figure 6 for a bleed-through example.

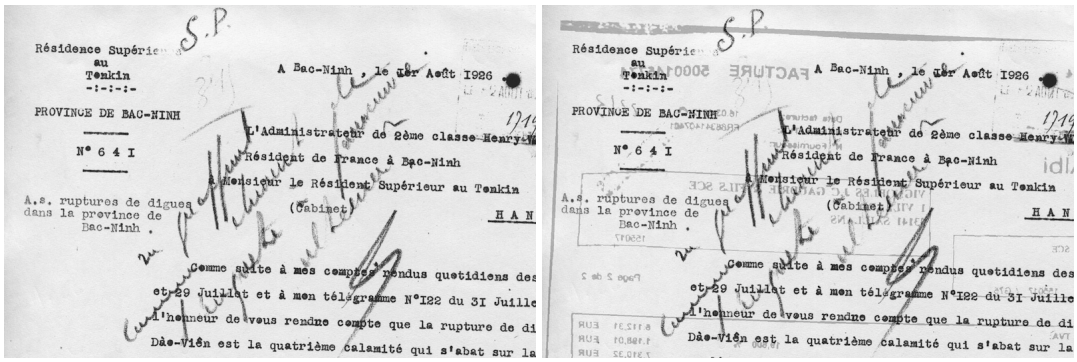


Figure 6. Bleed-through defect. (**Left**) original image. (**Right**) degraded image.

3.5. Adaptive Blur

The blur defect is a very common defect encountered during digitization campaigns. The difficulty here is to create a realistic blur defect that mimics the very slight blur that appears when the scanner is incorrectly set (a large blur is easily detected by scanners). To do so, we propose a method inspired by the blur detection from [50]. First, the user chooses a blur image to mimic among a real blur example available in DocCreator. Then, using a dichotomic algorithm, we compute the size of the kernel of a Gaussian blur that, once applied on the input image, produces a blur similar to the chosen real blur image. In this method, the Fourier Transform of the image is first computed. Then the module of the Fourier Transform is binarized according to its mean. The resulting binarized image produces a disc for images with only text. As the high frequencies decrease when the blur increases, the disc radius in the binarized image also decreases when the blur increases. This radius is used to characterize the images. The dichotomic algorithm is used to search the kernel size that produces a radius similar to the one found on the selected example image. See Figure 7 for an example.

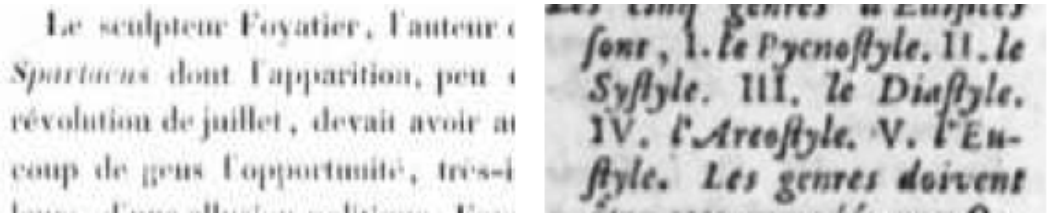


Figure 7. Adaptive blur defect. (**Left**) image with real blur. (**Right**) image with synthetic blur that mimics the real one

3.6. 3D Paper Deformation

The paper on which a book is printed may have several types of deformation (along curvature, rotation, fold, hole, etc.). We propose a 3D deformation model that generate realistic small or large paper deformations.

The full process is detailed in [51]. The main idea is: first, a 3D scanner is used to acquire a 3D mesh from a real document. This mesh preserves all representative distortions. Then, the mesh is unfolded into a 2D plan. Therefore, each vertex in the mesh has a corresponding 2D point. The coordinates of such a point are considered as texture coordinates. Finally, the mesh can be rendered with any 2D image mapped as a texture. For the rendering, we use the Phong reflection model as the illumination model. Changing light properties and position allows to accentuate or minimize distortion effects. DocCreator currently provides 17 parameterized meshes, enabling one to produce numerous distorted images. Figure 8 shows such a deformation.

This 3D paper deformation model can be used to simulate mobile document capture. The user can add a background plane with texture, on top of which the document stands. By changing viewpoint and light positions, the user can generate many images. These images can be used for camera-based document image analysis. Figure 9 shows examples of two points of view generated with the same document image.

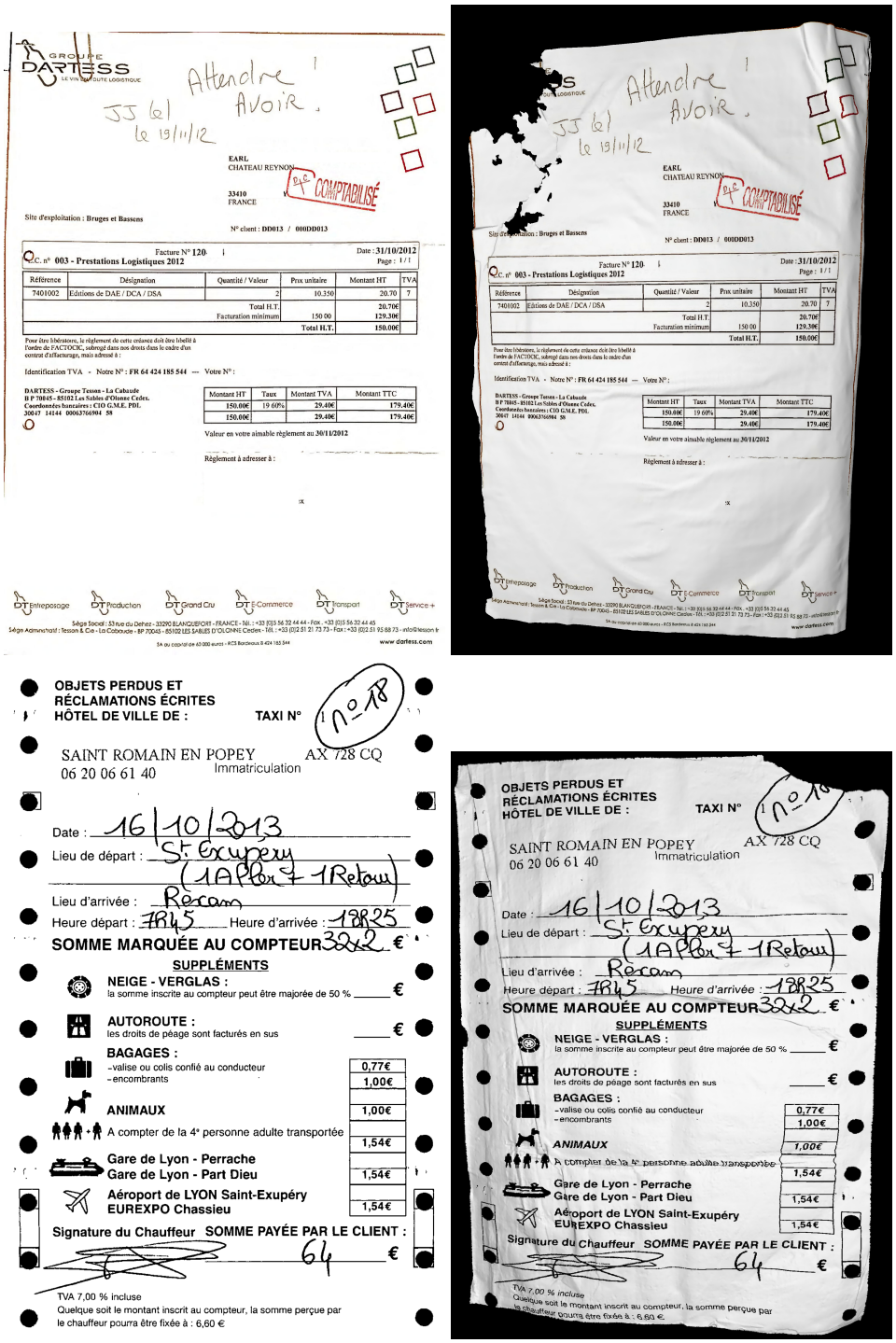


Figure 8. Examples of 3D deformations of a 2D receipt images. (**Left**) original images. (**Right**) degraded images.

Figure 9. Examples of two viewpoints of the same document image, that could be used in the context of camera-based document image analysis.

3.7. Nonlinear Illumination Model

DocCreator provides an implementation of the nonlinear illumination model proposed in [52]. When scanning thick documents, the page to be photocopied may not be flat on the document glass and thus the illumination is not constant on the whole document. This model consider that a border of the document is bend in one direction by a radius ρ. The illumination at a point P′ on the document pages is inversely proportional to the distance of point P′ from the light source L. The illumination at point P′ is computed with the following equation:

$$I_{P'} = I_0\left(\frac{l_0}{(l_0 + \rho(1 - cos\phi))}\right)^2$$

where I_0 is the original intensity, l_0 the distance between the document glass and the light source L, and ϕ the angle between the normal at P′ (where the page is curved) and the normal to the glass document. Figure 10 shows an example of this illumination defect. It is noteworthy that this illumination defect simulates just a particular case of what our 3D paper deformation model presented in the previous section can portray.

Except for the ink degradation model, the other degradation models work both on grayscale and colour document images.

DocCreator aims at providing other degradation models. In particular, we are currently working on the integration of a colour ink spot generation model described in [35].

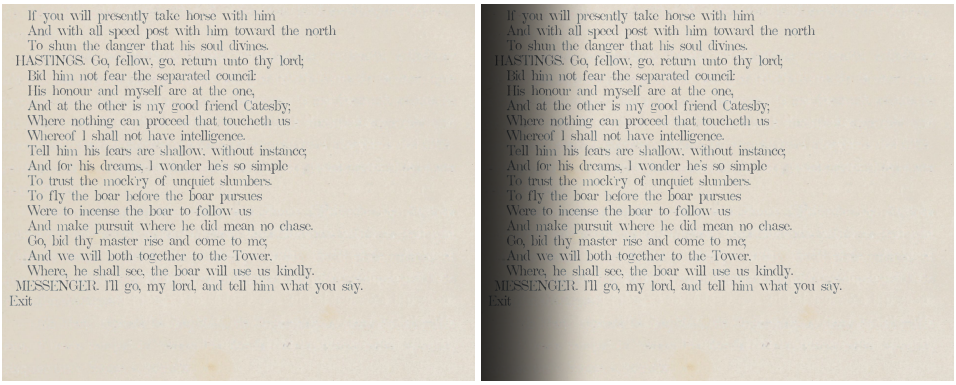


Figure 10. Examples of nonlinear illumination model defect. (**Left**) original image. (**Right**) degraded image with illumination defect applied on the left border.

4. Use of DocCreator for Performance Evaluation Tasks or Retraining

Here, we describe rapidly how DocCreator was used by other researchers and the conclusions they drew.

4.1. Published Results Using DocCreator

4.1.1. Document Image Generation for Performance Evaluation

The segmentation system proposed by [36] is based on a texture feature extraction without any a priori knowledge on the physical and logical document layout. To assess the noise robustness of their system, they used DocCreator and applied the character degradation model. From 25 simplified real document images, they generated a semi-synthetic database of 150 document images. This database is made up of several subsets where the degradation levels are different. The performance evaluations presented in [36] highlight that the texture descriptors are slightly perturbed by the degradations. When characters are highly disconnected (our algorithm has erased important character ink areas), a drop of the segmentation performances was observed.

DocCreator was also used during the ICDAR contest: staff-line removal from musical scores. The 3D distortion and the character degradation models were used in order to generate an extended database from the 1000 images of the MUSCIMA database [13]. As a result, the extended database contains 6000 semi-synthetic grayscale images and 6000 semi-synthetic binary images. This database has been used in the second edition of the music score competition ICDAR 2013 [37]. Five participants submitted eight methods. Participants were given a training set of 4000 semi-synthetic images and then 2000 semi-synthetic images to test their methods on. Regarding the results on the 3D distortion set, the submitted methods seem less robust to global distortion than to the presence of small curves and folds. For more details about the participants, the methods and the contest protocol, refer to [37]. This database has already become a benchmark database for musical document images analysis and recognition, as stated in [53]. So far, the database has indeed been used for benchmarking in multiple scientific publications about musical document processing and recognition [38,53–56] and even in the more general field of machine learning [57].

4.1.2. Document Image Generation for Retraining Task

The IAM-HistDB [58] database contains 127 handwritten historical manuscript images together with their ground truth. This database consists of three sets: the Saint Gall set containing 60 images (1.410 text lines) in Latin, the Parzival set containing 47 images (4.477 text lines) in Medieval German, and the Washington set containing 20 images in English. The authors of [39] used the character degradation model to create two extended databases of the IAM-HistDB. The first one is composed of 17.661 images degraded with the ink model. The 1.524 images from the second dataset have been

created using the 127 original images and transformed using our 3D distortion model. The tests presented in [39,59] confirm the conclusion of [60] about the impact of the degradation level on re-training, either for a task of character recognition or layout extraction.

4.2. New Results on Performance Prediction Using DocCreator

Here, we show whether DocCreator can be useful for performance prediction of existing methods.

4.2.1. Increase the Prediction Rate of Predictive Binarization Algorithm

In [61], we have presented an algorithm to predict error rates of 11 binarization methods on given document images so that the best binarization method is automatically chosen for any image depending on its quality. This method requires ground-truthed data as input of the training step. The DIBCO database [62] was used. However, the DIBCO database contains only 36 images.

We propose here to extend the original DIBCO database by using the ink degradation model. Since the DIBCO database contains 36 images, we extend it with the same number of semi-synthetic document images. This extended dataset is then used to train the prediction model of [61].

Our retraining tests show that the use of this extended dataset allows one to increase the performance of the prediction model of [61]. More precisely, the error rate of the prediction model decreases (until it levels off) when the number of semi-synthetic images in the training set increases. On average, the error rate drops of about 15% compared with using only real images in the training set. The error rate converges when the proportion of semi-synthetic images is around 50% of the training set.

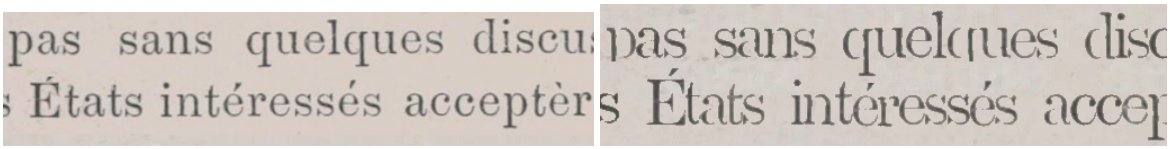
4.2.2. Predict OCR Recognition Rate Using Synthetic Images

Many on-line digital libraries propose a text search engine. To this end, the text within the document images has to be transcribed. Depending on the OCR recognition rate quality, three options are available: (1) directly use the OCR result when the recognition rate is close to 100%; (2) manually correct the OCR result when the automatic transcription gives "acceptable quality", or (3) do a complete manual transcription (often quite expansive). As a consequence, it is very important to be aware of the OCR recognition rate before deciding between one of these three solutions. The amount of recent publications on this subject ([63–66]) reflects the scientific interest in predicting OCRs recognition rate.

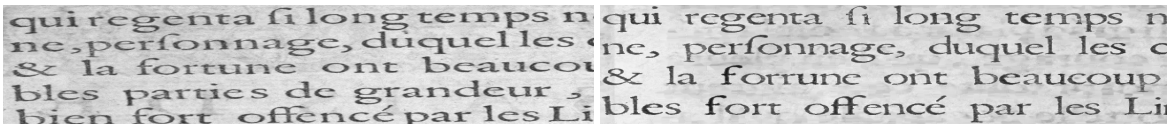
We propose here to use synthetic images to predict the OCR rate of a digitized book as follows: (1) font, background and layout are extracted from original images (with methods described in II). It is noteworthy to mention that the fonts were extracted thoroughly, in particular to include even characters not recognized by the OCR, or even to adjust margins of correctly labeled characters. (2) An adapted Lorem ipsum text is randomly generated and used to create synthetic images with the font and background previously extracted. This adapted Lorem ipsum is generated with accentuated characters (é, à, ù, etc.) and old characters (ff, fi, ſ, fl, ffi) if the original text contains such characters. Generating such characters is important to have a representative dataset for fair OCR testing. As a result, images like the one presented in Figure 2 are generated with the associated XML ground truth. (3) An OCR (Tesseract) is finally used to recognize the text on these synthesized images. This text is compared with the Lorem ipsum ground truth text, giving an OCR recognition rate. We consider that this recognition rate is a prediction of the OCR rate if the OCR software was applied on original images. Table 2 Column 1 provides the average OCR recognition rate obtained on the original images, Table 2 Column 3 refers to the average OCR rates computed on the synthetic "Lorem ipsum" images versions.

We also propose to evaluate the capacity of our method to correctly predict the OCR recognition rate by comparing original images with their synthetic version generated with exactly the same text (see Figure 11 to compare the original images and their synthetic versions). These images are generated following this protocol: (1) pages from three books (2 typewritten and 1 manuscript book) have been manually transcribed; (2) font, background and layout are automatically extracted from original

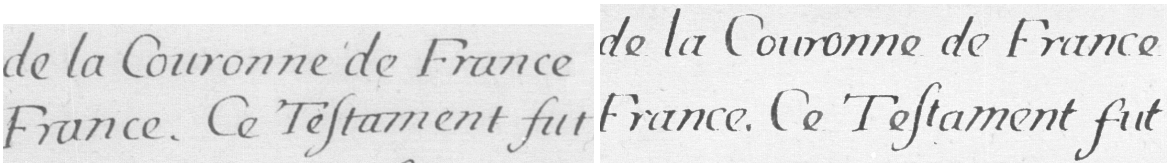
images; and (3) font, background and layout are finally used with the transcribed texts to automatically generate synthetic versions of the original ones. The whole database is composed of 93 document images containing 18.240 words and 115.622 characters.



(**a**) (DB1) Contemporary french typewritten document.



(**b**) (DB2) Old french typewritten document.



(**c**) (DB3) Old french Manuscript document.

Figure 11. Images extracted from the database used for testing our prediction algorithm. (**Left**) original images. (**Right**) synthetic generated images.

To evaluate the OCR text recognition rate, we use the Levenshtein distance (a metric measuring the difference between two strings) between the whole original transcribed text and the whole recognized text. We compute the mean of the Levenshtein distances for the N documents of each database. Using this Levenshtein distance, the difference between the OCR text recognition rate computed on real images and the one computed on "Lorem ipsum" version (Table 2 Column 1 and Column 2) is, on average, only overestimated by 0.04. The difference between the real OCR rate and the one computed on the synthetic versions (Table 2 Column 1 and Column 3) is, on average, only overestimated by 0.03. Most of the success of different existing OCR prediction methods ([63–66]) are related to the quality and quantity of the needed ground truth. Our prediction method presented here provides comparable results with the ones form the state of the art.

Table 2. Comparison between OCR recognition rates obtained on three different books original images and their synthetic versions. Column 1: OCR recognition rate on original images, Column 2: OCR recognition rate on synthetic images generated with both the text and the font from the original images, Column 3: OCR recognition rate on synthetic images generated with lorem ipsum random text and the font from the original images.

	Original Image	Font From	Same Text	Lorem Text
DB1	0.95	DB1	0.94	0.88
DB2	0.80	DB2	0.85	0.84
DB3	0.24	DB3	0.21	0.23

5. Conclusions

DocCreator gives to DIAR researchers a simple and rapid way to extend existing document image databases or to create new ones avoiding the tedious task of manual ground truth generation. DocCreator embeds many fonts, backgrounds, meshes and realistic degradation models which, when combined, result in an interesting combination of ground-truthed databases. The experiments

detailed in this paper show semi-synthetic and synthetic documents created with DocCreator are useful for performance evaluation, retraining tasks or performance prediction. In future work, we plan to improve the synthetic document creation to avoid to have too different characters in the composed document. For example, we should investigate if adding some constraints on the font extraction phase or taking into account the context when adding new characters to the synthetic document may lead to more realistic synthetic documents. We also consider to set up a cognitive experiment to evaluate the perceived realness of the degraded documents or even the created synthetic documents. We are also planning to investigate how the generation of highly diversified data can improve the results of tasks based on deep learning methods.

DocCreator (source, Linux, Mac, Windows packaged versions), all the databases used for the tests, a video and an extra database (31.000 synthetic images generated with William Shakespeare sonnet text files) are available at [http://doc-creator.labri.fr/].

Author Contributions: Nicholas Journet, Muriel Visani and Boris Mansencal contributed in equal proportion to the creation and tests of DocCreator (model degradation and synthetic document reconstruction) ; they also wrote this article. Kieu Van-Cuong contributed to the degradation models, Antoine Billy contributed to the synthetic document reconstruction.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. L'Affaire Alexis. Available online: http://gallica.bnf.fr/ark:/12148/bpt6k8630878m/f1.item.texteImage (accessed on 9 December 2017).
2. Shahab, A.; Shafait, F.; Kieninger, T.; Dengel, A. An Open Approach Towards the Benchmarking of Table Structure Recognition Systems. In Proceedings of the 9th IAPR International Workshop on Document Analysis Systems, Boston, MA, USA, 9–11 June 2010; ACM: New York, NY, USA, 2010; pp. 113–120.
3. Lazzara, G.; Levillain, R.; Géraud, T.; Jacquelet, Y.; Marquegnies, J.; Crépin-Leblond, A. The SCRIBO Module of the Olena Platform: a Free Software Framework for Document Image Analysis. In Proceedings of the 2011 International Conference on Document Analysis and Recognition (ICDAR), Beijing, China, 18–21 September 2011.
4. Yalniz, I.; Manmatha, R. A Fast Alignment Scheme for Automatic OCR Evaluation of Books. In Proceedings of the 2011 International Conference on Document Analysis and Recognition (ICDAR), Beijing, China, 18–21 September 2011; pp. 754–758.
5. Roy, P.; Ramel, J.; Ragot, N. Word Retrieval in Historical Document Using Character-Primitives. In Proceedings of the 2011 International Conference on Document Analysis and Recognition (ICDAR), Beijing, China, 18–21 September 2011; pp. 678–682.
6. IAM Handwriting Database. Available online: http://www.iam.unibe.ch/fki/databases/iam-handwriting-database (accessed on 9 December 2017).
7. Grosicki, E.; Carré, M.; Brodin, J.M.; Geoffrois, E. Results of the second RIMES evaluation campaign for handwritten mail processing. In Proceedings of the 2009 10th International Conference on Document Analysis and Recognition (ICDAR), Barcelona, Spain, 26–29 July 2009.
8. Perez, D.; Tarazon, L.; Serrano, N.; Castro, F.; Terrades, O.R.; Juan, A. The GERMANA Database. In Proceedings of the 2009 10th International Conference on Document Analysis and Recognition (ICDAR), Barcelona, Spain, 26–29 July 2009; IEEE Computer Society: Washington, DC, USA, 2009; pp. 301–305.
9. Nakagawa, K.; Fujiyoshi, A.; Suzuki, M. Ground-truthed Dataset of Chemical Structure Images in Japanese Published Patent Applications. In Proceedings of the 9th IAPR International Workshop on Document Analysis Systems, Boston, MA, USA, 9–11 June 2010; ACM: New York, NY, USA, 2010; pp. 455–462.
10. Eurecom. Available online: http://www.eurecom.fr/huet/work.html (accessed on 9 December 2017).
11. University of California, San Francisco. *The Legacy Tobacco Document Library (LTDL)*; University of California: San Francisco, CA, USA, 2007.
12. Delalandre, M.; Valveny, E.; Pridmore, T.; Karatzas, D. Generation of Synthetic Documents for Performance Evaluation of Symbol Recognition & Spotting Systems. *Int. J. Doc. Anal. Recognit.* **2010**, *13*, 187–207.

13. Fornés, A.; Dutta, A.; Gordo, A.; Lladós, J. CVC-MUSCIMA: A ground truth of handwritten music score images for writer identification and staff removal. *Int. J. Doc. Anal. Recognit.* **2012**, *15*, 243–251.
14. Burie, J.C.; Chazalon, J.; Coustaty, M.; Eskenazi, S.; Luqman, M.M.; Mehri, M.; Nayef, N.; Ogier, J.M.; Prum, S.; Rusiñol, M. ICDAR2015 competition on smartphone document capture and OCR (SmartDoc). In Proceedings of the 2015 13th International Conference on Document Analysis and Recognition (ICDAR), Nancy, France, 23–26 August 2015; pp. 1161–1165.
15. Nayef, N.; Luqman, M.M.; Prum, S.; Eskenazi, S.; Chazalon, J.; Ogier, J.M. SmartDoc-QA: A dataset for quality assessment of smartphone captured document images-single and multiple distortions. In Proceedings of the 2015 13th International Conference on Document Analysis and Recognition (ICDAR), Nancy, France, 23–26 August 2015; pp. 1231–1235.
16. TC11 Online Resources. Available online: http://tc11.cvc.uab.es/datasets/ (accessed on 9 December 2017).
17. Yanikoglu, B.; Vincent, L. Pink Panther: A complete environment for ground-truthing and benchmarking document page segmentation. *Pattern Recognit.* **1998**, *31*, 1191–1204.
18. Ha Lee, C.; Kanungo, T. The architecture of TRUEVIZ: A groundTRUth/metadata Editing and VIsualiZing toolkit. *Pattern Recognit.* **2003**, *36*, 811–825.
19. Doermann, D.; Zotkina, E.; Li, H. GEDI—A Groundtruthing Environment for Document Images. In Proceedings of the 9th IAPR International Workshop on Document Analysis Systems (DAS 2010), Boston, MA, USA, 9–11 June 2010.
20. Clausner, C.; Pletschacher, S.; Antonacopoulos, A. Efficient OCR Training Data Generation with Aletheia. In Proceedings of the International Association for Pattern Recognition (IAPR), Tours, France, 7–10 April 2014.
21. Garz, A.; Seuret, M.; Simistira, F.; Fischer, A.; Ingold, R. Creating ground truth for historical manuscripts with document graphs and scribbling interaction. In Proceedings of the 2016 12th IAPR Workshop on Document Analysis Systems (DAS), Santorini, Greece, 11–14 April 2016; pp. 126–131.
22. Gatos, B.; Louloudis, G.; Causer, T.; Grint, K.; Romero, V.; Sánchez, J.A.; Toselli, A.H.; Vidal, E. Ground-truth production in the tranScriptorium project. In Proceedings of the 2014 11th IAPR International Workshop on Document Analysis Systems Document Analysis Systems (DAS), Tours, France, 7–10 April 2014; pp. 237–241.
23. Wei, H.; Chen, K.; Seuret, M.; Würsch, M.; Liwicki, M.; Ingold, R. *DIVADIAWI— A Web-Based Interface for Semi-Automatic Labeling of Historical Document Images*; Digital Humanities: Sydney, Australia, 2015.
24. Mas, J.; Fornés, A.; Lladós, J. An Interactive Transcription System of Census Records using Word-Spotting based Information Transfer. In Proceedings of the 12th IAPR International Workshop on Document Analysis Systems (DAS 2016), Santorini, Greece, 11–14 April 2016.
25. Recital Manuscript Platform. Available online: http://recital.univ-nantes.fr/ (accessed on 9 December 2017).
26. Clausner, C.; Pletschacher, S.; Antonacopoulos, A. Aletheia—An Advanced Document Layout and Text Ground-Truthing System for Production Environments. In Proceedings of the International Conference on document Analysis and Recognition, Beijing, China, 18–21 September 2011; pp. 48–52.
27. Baird, H.S. Document Image Defect Models. In Proceedings of the IAPR workshop on Syntatic and Structural Pattern Recognition, Murray Hill, NJ, USA, 13–15 June 1990; pp. 13–15.
28. Jiuzhou, Z. *Creation of Synthetic Chart Image Database with Ground Truth*; Technical Report; National University of Singapore: Singapore, 2005.
29. Ishidera, E.; Nishiwaki, D. A Study on Top-down Word Image Generation for Handwritten Word Recognition. In Proceedings of the 2003 7th International Conference on Document Analysis and Recognition (ICDAR), Edinburgh, UK, 3–6 August 2003; IEEE Computer Society: Washington, DC, USA, 2003.
30. Yin, F.; Wang, Q.F.; Liu, C.L. Transcript Mapping for Handwritten Chinese Documents by Integrating Character Recognition Model and Geometric Context. *Pattern Recognit.* **2013**, *46*, 2807–2818.
31. Opitz, M.; Diem, M.; Fiel, S.; Kleber, F.; Sablatnig, R. End-to-End Text Recognition Using Local Ternary Patterns, MSER and Deep Convolutional Nets. In Proceedings of the 2014 11th IAPR International Workshop on Document Analysis Systems (DAS), Tours, France, 7–10 April 2014; pp. 186–190.
32. Yacoub, S.; Saxena, V.; Sami, S. Perfectdoc: A ground truthing environment for complex documents. In Proceedings of the Eighth International Conference on Document Analysis and Recognition, Seoul, Korea, 31 August–1 September 2005; pp. 452–456.
33. Saund, E.; Lin, J.; Sarkar, P. Pixlabeler: User interface for pixel-level labeling of elements in document images. In Proceedings of the International Conference on Document Analysis and Recognition, Barcelona, Spain, 26–29 July 2009; pp. 646–650.

34. Lamiroy, B.; Lopresti, D. An Open Architecture for End-to-End Document Analysis Benchmarking. In Proceedings of the 2011 International Conference on Document Analysis and Recognition (ICDAR), Beijing, China, 18–21 September 2011; pp. 42–47.
35. Seuret, M.; Chen, K.; Eichenbergery, N.; Liwicki, M.; Ingold, R. Gradient-domain degradations for improving historical documents images layout analysis. In Proceedings of the 2015 13th International Conference on Document Analysis and Recognition (ICDAR), Tunis, Tunisia, 23–26 August 2015; pp. 1006–1010.
36. Mehri, M.; Gomez-Krämer, P.; Héroux, P.; Mullot, R. Old Document Image Segmentation Using the Autocorrelation Function and Multiresolution Analysis. In Proceedings of the IS & T/SPIE Electronic Imaging, Burlingame, CA, USA, 3–7 February 2013.
37. Visani, M.; Kieu, V.; Fornés, A.; Journet, N. The ICDAR 2013 Music Scores Competition: Staff Removal. In Proceedings of the 2013 12th International Conference on Document Analysis and Recognition (ICDAR), Washington, DC, USA, 25–28 August 2013; pp. 1407–1411.
38. Montagner, I.d.S.; Hirata, R., Jr.; Hirata, N.S.T. A Machine Learning based method for Staff Removal. In Proceedings of the 2014 22nd International Conference on Pattern Recognition (ICPR), Stockholm, Sweden, 24–28 August 2014; pp. 3162–3167.
39. Fischer, A.; Visani, M.; Kieu, V.C.; Suen, C.Y. Generation of Learning Samples for Historical Handwriting Recognition Using Image Degradation. In Proceedings of the the 2nd International Workshop on Historical Document Imaging and Processing, Washington, DC, USA, 24 August 2013; pp. 73–79.
40. Smith, R. An Overview of the Tesseract OCR Engine. In Proceedings of the 2007 9th International Conference on Document Analysis and Recognition (ICDAR), Parana, Brazil, 23–26 September 2007; pp. 629–633.
41. Bahaghighat, M.K.; Mohammadi, J. Novel approach for baseline detection and Text line segmentation. *Int. J. Comput. Appl.* **2012**, *51*, doi:10.5120/8013-1039.
42. Telea, A. An image inpainting technique based on the fast marching method. *J. Graph. Tools* **2004**, *9*, 23–34.
43. Mehri, M.; Héroux, P.; Lerouge, J.; Gomez-Krämer, P.; Mullot, R. A structural signature based on texture for digitized historical book page categorization. In Proceedings of the 2015 13th International Conference on Document Analysis and Recognition (ICDAR), Tunis, Tunisia, 23–26 August 2015; pp. 116–120.
44. Breuel, T.M. Two geometric algorithms for layout analysis. In *International Workshop on Document Analysis Systems*; Springer: Berlin, Germany, 2002; pp. 188–199.
45. Ramel, J.Y.; Leriche, S.; Demonet, M.; Busson, S. User-driven page layout analysis of historical printed books. *Int. J. Doc. Anal. Recognit.* **2007**, *9*, 243–261.
46. Garz, A.; Seuret, M.; Fischer, A.; Ingold, R. A User-Centered Segmentation Method for Complex Historical Manuscripts Based on Document Graphs. *IEEE Trans. Hum.-Mach. Syst.* **2017**, *47*, 181–193.
47. Kanungo, T.; Haralick, R. Automatic generation of character groundtruth for scanned documents: A closed-loop approach. In Proceedings of the 13th International Conference on Pattern Recognition, Vienna, Austria, 25–29 August 1996; Volume 3, pp. 669–675.
48. Shakhnarovich, G. Learning Task-Specific Similarity. Ph.D. Thesis, Massachusetts Institute of Technology, Cambridge, MA, USA, 2005.
49. Moghaddam, R.F.; Cheriet, M. Low Quality Document Image Modeling and Enhancement. *Int. J. Doc. Anal. Recognit.* **2009**, *11*, 183–201.
50. Lelégard, L.; Bredif, M.; Vallet, B.; Boldo, D. Motion blur detection in aerial images shot with channel-dependent exposure time. In Proceedings of the ISPRS-Technical-Commission III Symposium on Photogrammetric Computer Vision and Image Analysis (PCV), Saint-Mandé, France, 1–3 September 2010; pp. 180–185.
51. Kieu, V.; Journet, N.; Visani, M.; Mullot, R.; Domenger, J. Semi-synthetic Document Image Generation Using Texture Mapping on Scanned 3D Document Shapes. In Proceedings of the 2013 12th International Conference on Document Analysis and Recognition (ICDAR), Washington, DC, USA, 25–28 August 2013; pp. 489–493.
52. Kanungo, T.; Haralick, R.M.; Phillips, I. Global and Local Document Degradation Models. In Proceedings of the 1993 2nd International Conference on Document Analysis and Recognition (ICDAR), Tsukuba City, Japan, 20–22 October 1993; pp. 730–734.
53. Calvo-Zaragoza, J.; Micó, L.; Oncina, J. Music staff removal with supervised pixel classification. *Int. J. Doc. Anal. Recognit.* **2016**, *19*, 1–9.
54. Bui, H.N.; Na, I.S.; Kim, S.H. Staff Line Removal Using Line Adjacency Graph and Staff Line Skeleton for Camera Based Printed Music Scores. In Proceedings of the 2014 22nd International Conference on Pattern Recognition (ICPR), Stockholm, Sweden, 24–28 August 2014.

55. Géraud, T. A morphological method for music score staff removal. In Proceedings of the 2014 IEEE International Conference on Image Processing (ICIP), Paris, France, 27–30 October 2014; pp. 2599–2603.
56. Zaragoza, J.C. Pattern Recognition for Music Notation. Ph.D. Thesis, Universidad de Alicante, Alicante, Spain, 2016.
57. Montagner, I.S.; Hirata, N.S.; Hirata, R., Jr.; Canu, S. Kernel approximations for W-operator learning. In Proceedings of the International Conference on Graphics, Patterns and Images (SIBGRAPI), Sao Paulo, Brazil, 4–7 October 2016.
58. IAM-HistDB Database. Available online: https://diuf.unifr.ch/main/hisdoc/iam-histdb (accessed on 9 December 2017).
59. Wei, H.; Baechler, M.; Slimane, F.; Ingold, R. Evaluation of SVM, MLP, and GMM Classifiers for Layout Analysis of Historical Documents. In Proceedings of the 2013 12th International Conference on Document Analysis and Recognition (ICDAR), Washington, DC, USA, 25–28 August 2013; pp. 1220–1224.
60. Varga, T.; Bunke, H. Effects of training set expansion in handwriting recognition using synthetic data. In Proceedings of the 11th Conference of the International Graphonomics Society, Scottsdale, AZ, USA, 2–5 November 2003; pp. 200–203.
61. Rabeux, V.; Journet, N.; Vialard, A.; Domenger, J.P. Quality Evaluation of Degraded Document Images for Binarization Result Prediction. *Int. J. Doc. Anal. Recognit.* **2013**, 1–13.
62. Pratikakis, I.; Gatos, B.; Ntirogiannis, K. ICDAR 2011 Document Image Binarization Contest (DIBCO 2011). In Proceedings of the 2011 11th International Conference on Document Analysis and Recognition (ICDAR), Washington, DC, USA, 25–28 August 2011; pp. 1506–1510.
63. Bhowmik, T.K.; Paquet, T.; Ragot, N. OCR performance prediction using a bag of allographs and support vector regression. In Proceedings of the 2014 11th IAPR International Workshop on Document Analysis Systems (DAS), Tours, France, 7–10 April 2014; pp. 202–206.
64. Peng, X.; Cao, H.; Natarajan, P. Document image OCR accuracy prediction via latent Dirichlet allocation. In Proceedings of the 2015 13th International Conference on Document Analysis and Recognition (ICDAR), Tunis, Tunisia, 23–26 August 2015; pp. 771–775.
65. Ye, P.; Doermann, D. Document image quality assessment: A brief survey. In Proceedings of the 2013 12th International Conference on Document Analysis and Recognition (ICDAR), Washington, DC, USA, 25–28 August 2013; pp. 723–727.
66. Clausner, C.; Pletschacher, S.; Antonacopoulos, A. Quality prediction system for large-scale digitisation workflows. In Proceedings of the 2016 12th IAPR Workshop on Document Analysis Systems (DAS), Santorini, Greece, 11–14 April 2016; pp. 138–143.