

Document downloaded from:

<http://hdl.handle.net/10251/156936>

This paper must be cited as:

Castro-Bleda, MJ.; España Boquera, S.; Pastor Pellicer, J.; Zamora Martínez, FJ. (2020). The NoisyOffice Database: A Corpus To Train Supervised Machine Learning Filters For Image Processing. *The Computer Journal*. 63(11):1658-1667.
<https://doi.org/10.1093/comjnl/bxz098>



The final publication is available at

<https://doi.org/10.1093/comjnl/bxz098>

Copyright Oxford University Press

Additional Information

The NoisyOffice Database: A corpus to train supervised machine learning filters for image processing

M.J. CASTRO-BLEDA (1), S. ESPAÑA-BOQUERA (1),
J. PASTOR-PELLICER (2), F. ZAMORA-MARTÍNEZ (3)

(1) *VRAIN Valencian Research Institute for Artificial Intelligence,
Universitat Politècnica de València, Valencia, Spain*

(2) *Google, Zürich Area, Switzerland*

(3) *R&D Department, das-Nano S.L., Pol. Ind. Talluntxe II, Tajonar 31192, Spain*

Email: mcastro@dsic.upv.es, sespana@dsic.upv.es, joapaspe@gmail.com, pzamora@das-nano.es

This paper presents the “NoisyOffice” database. It consists of images of printed text documents with noise mainly caused by uncleanliness from a generic office, such as coffee stains and footprints on documents or folded and wrinkled sheets with degraded printed text. This corpus is intended to train and evaluate supervised learning methods for cleaning, binarization and enhancement of noisy images of grayscale text documents. As an example, several experiments of image enhancement and binarization are presented by using deep learning techniques. Also, double resolution images are also provided for testing super-resolution methods. The corpus is freely available at UCI Machine Learning Repository. Finally, a challenge organized by Kaggle Inc. to denoise images, using the database, is described in order to show its suitability for benchmarking of image processing systems.

*Keywords: Optical Character Recognition; Image Processing; Binarization; Denoising;
Super-resolution; Machine Learning; Neural Networks; Deep Learning*

Received 00 January 2009; revised 00 Month 2009

1. INTRODUCTION AND MOTIVATION

The field of offline handwriting recognition has been a topic of intensive research for many years (see some surveys in [1, 2, 3, 4, 5]). One of the first steps in the classical architecture of a text recognition system is preprocessing, where noise reduction and text normalization takes place. Preparing cleaner images for the recognition engines is often taken for granted. However, this step undoubtedly influences the overall performance of the system [6]. It can also improve very significantly the readability for human readers. Another important issue is to decide whether the recognition systems will process grayscale images or binary images. Let us take into account that a scanned image is usually stored in grayscale format, in which 0 means black and 1 white, and various shades of gray are represented between these two values. The process to convert the grayscale image to a binary image is called binarization, and is a common step in the overall recognition pipeline (although some architectures may retain the grayscale information through several stages). Besides, the enhancement of images should correct traces with low, non-uniform ink level produced by errors of the devices (low ink, low-resolution scanners, etc. . .), and remove background noise from the image.

Machine learning techniques (such as neural networks) can be used for denoising and enhancing image documents

(see [7, 8, 9] for a review on image processing with neural networks). To this end, noisy images and their corresponding clean or binarized groundtruth images are needed to train the models. The corpus presented in this work, the “NoisyOffice” dataset, consists of images of printed text documents with noise mainly caused by uncleanliness from a generic office (coffee stains and footprints on documents, folded and wrinkled sheets with degraded printed text, etc.), and their corresponding groundtruth images. The corpus is freely available at UCI Machine Learning Repository¹ [10].

Some significant related efforts have been done to create benchmarking datasets in order to compare current document image binarization practices. With that purpose, the Document Binarization Contest (DIBCO) [11, 12] and the Handwritten Document Binarization Contest (H-DIBCO) [13, 14] have been held in the context of the International Conference on Frontiers in Handwriting Conference (ICFHR) and International Conference on Document Analysis and Recognition conferences (ICDAR) since 2009. In each edition, new sets of supervised images have been provided for evaluation. These datasets are focused on historical documents and the kind of images is very heterogeneous.

¹Available at <http://archive.ics.uci.edu/ml/datasets/NoisyOffice>.

There exists other datasets for binarization purposes although many of them are not limited to this task and contain text line segmentation and Optical Character Recognition (OCR) output as well. The Research & Development Laboratory of EPITA (LRDE) Document Binarization Dataset [15] is a database composed of French documents extracted from a sole issue of “Le Nouvel Observateur” magazine with a degradation resulting from the scanning process. This database is in color and includes pictures. Several processed images are available: the images with pictures removed, the binarized document, the localization of lines and the OCR output. In [16], the authors describe the groundtruth creation for the IAM Historical Handwriting Database (IAM-HistDB). This database contains medieval manuscripts of the epic poem *Parzival* by Wolfram von Eschenbach and, besides the binarized images, text line segmentation and word segmentation are also provided.

There are also databases for non-Latin scripts such as the Persian heritage image binarization dataset² (PHIBD 2012) [17] or the AMADI_LontarSet [18]. The first one, written in Persian, comprises 15 old manuscript images which suffer several types of degradation, while the second one, written in Balinese script (an alphabet used in the island of Bali, Indonesia), has been written on palm leaves using a sharp pen and colored with natural dyes afterwards. As an example of a database acquired with a camera (using a smartphone) we can mention the LINX dataset³ [19]. As with the LRDE Document Binarization Dataset, this corpus is not limited to text documents.

The organization of the rest of the paper is as follows: next section gives a general overview of generation of supervised corpora for binarization and enhancement. Section 3 describes into detail the NoisyOffice database. Some experimentation performed with the corpus is presented in Section 4. Section 5 describes the “Denoising Dirty Documents” challenge by Kaggle Inc. and the obtained benchmarking. Finally, some conclusions and future work are drawn in Section 6.

2. IMAGE PROCESSING AND GROUNDTRUTH GENERATION

Generating synthetic data for the training and evaluation of document image processing systems has been widely addressed in recent years [6, 20, 21, 22, 23, 24]. In particular, image binarization evaluation is usually computed at pixel level, requiring an accurate groundtruth, with the inner complexity of data supervision at this detail level. To overcome this issue, there are several techniques to generate an accurate and useful groundtruth. There are two main strategies: *denoising* and *noising* (although it is possible to combine both). The former strategy starts with a dirty image and its groundtruth is obtained by cleaning it, while the latter takes a clean image and then noise

background and artifacts are added. On the one hand, the first approach has the advantage of using a real dirty image but, on the other hand, the supervised cleaning procedure is usually more costly and is subject to differences of criterion among human supervisors. The challenge for the second approach is, in turn, to find a realistic model to simulate the effects of problems which originates noise to an image.

2.1. Denoising supervision

Several approaches to remove noise from an image can be adopted in order to obtain its clean groundtruth:

- **Manual pixel image segmentation.** Basically it relies on removing noise by a human expert, usually assisted by a specific software. This method is discouraged and never used from scratch.
- **Combination of several methods and parameter tuning.** In this case the human expert could use some of the well known binarization/cleaning methods as starting point. The expert could even stack several of the most reliable approaches, in order to get better results and ease the final manual correction step.

For instance, image groundtruths of LRDE Document Binarization Dataset [15] have been obtained in a semi-automatic way by means of a global thresholding followed by a manual adjustment supervision. A similar technique has been applied to construct the groundtruth of the PHIBD 2012 dataset: first, a binarized image is obtained using an algorithm called PhaseGT and the resulting images are supervised afterwards [17]. In [16], several parameters of the binarization process have been manually adjusted for each manuscript of IAM-HistDB to find a good trade-off between reducing noise and maintaining the text detail.

- **Use layout information to extract the foreground.** One can take advantage of the layout and text line groundtruth data for text documents, since the foreground will be within the regions marked as text. This approach extracts the text regions and then applies the cleaning procedures only on these regions and marks the rest of the page as background.
- **Bootstrap denoising.** This approach relies on a semi-supervised method which improves in each iteration with new cleaned data. To start, one of the previous approaches could be used to get an initial small supervised subset of cleaned data, to be used to train the first version of the model. Then, an iterative training procedure of the model using the initial set, cleaning a larger set of images, manual supervision of the mistakes and retraining with the bigger set, is followed, until all the training data is used or a convergence criterion is reached.

All the above approaches require from human supervision, which is a very time consuming and error prone task, particularly the correction of frontier pixels between foreground and background. This problem becomes critical

²Available at <http://www.synchromedia.ca/PHIBD2012>.

³Available at <http://cmm.mines-paristech.fr/Projects/LINX>.

when denoising high resolution images, due to the presence of ambiguities at boundary pixels.

2.2. Noising methods

Using synthetic data or synthetically degraded data, has many advantages over human supervision, including rapid generation of datasets at lower cost, control of degradation level, and convenient testing of the same underlying document content with different corruption methods. Many different degradation effects can be used as defocusing, paper positioning variations, distortion of character strokes, non-uniform illumination, typesetting imperfections, perspective distortion, etc. [22, 20, 6, 25, 21, 23, 26]. These degradation models aim at generating synthetic noise that can be found in the real world and therefore to extend training sets to perform better on unseen scenarios.

3. THE NOISYOFFICE DATABASE

The NoisyOffice corpus is intended for cleaning, binarization and enhancement of noisy images of grayscale text documents using supervised learning methods. Double resolution images are also provided to test super-resolution techniques [27].

Since the aim of this database is not limited to binarization, the groundtruth for image enhancement purposes should contain gray levels at the edges for spatial anti-aliasing, as is usually applied when rasterizing computer fonts. This requirement makes the use of the denoising approach inappropriate. Indeed, this is the main reason why we have opted for the noising approach where real noisy backgrounds are combined with the images to generate the synthetic data, as detailed below.

The NoisyOffice corpus is divided into two datasets:

- “Simulated NoisyOffice” folder, which has been prepared for training, validation and test of supervised learning methods.
- “Real NoisyOffice” folder, which is composed of images printed, noised and scanned afterwards.

3.1. NoisyOffice: Simulated Noisy Image Dataset

A dataset of simulated noisy images was prepared by combining scanned images of noisy backgrounds with clean text images, following the scheme shown in Figure 1.

This process requires images of the noisy backgrounds, which were obtained by repeatedly folding and wrinkling white clean sheets of papers and making coffee and footprints stains on clean sheets, and scanning the noisy documents afterwards. Secondly, the noisy background images were combined with the clean text in order to obtain the simulated noisy image. For some types of noise, the foreground ink pixels are also noised as illustrated in the bottom of Figure 1. More sophisticated degradation mechanisms such as those cited in the previous section could also be applied. Examples of simulated noisy images are shown in Figure 2.

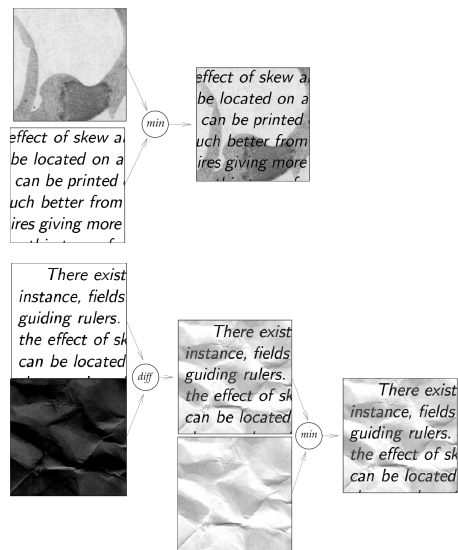


FIGURE 1: Simulated noisy process for “coffee-noise” (top) and “wrinkle-noise” (below).

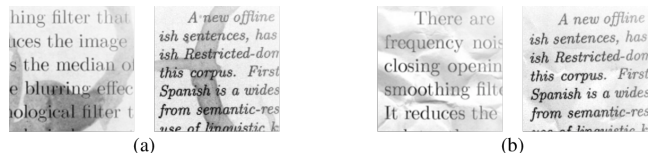


FIGURE 2: Simulated and real (a) “coffee-noise” and (b) “wrinkle-noise” images, respectively.

The “Simulated NoisyOffice” corpus was built by creating document images crossing different parameters of fonts and noise. The considered types of noise were: folded sheets, wrinkled sheets, coffee stains and footprints. About the font, the parameters were: font type (true-type, serif or roman), font size (footnote, normal or large) and yes/no emphasized font. Three different noisy documents of each type were created in order to get three sets for experimentation (training, validation and test sets) of 72 images each. Table 1 shows the parameters used for the corpus generation, along with some examples.

The final training, validation and test sets were composed of images of scanned simulated noisy images and their corresponding groundtruth images with the following variations:

- Binary groundtruth images at normal resolution (200 ppi), useful for binarization purposes.
- Grayscale groundtruth images with smoothing on the borders at normal resolution (200 ppi), suitable for image enhancement.
- Grayscale groundtruth images at double resolution (400 ppi) appropriate for testing super-resolution techniques.

Table 2 displays some statistic information about the simulated noisy images at normal resolution in grayscale.

TABLE 1: Parameters for the generation of the simulated noisy images and examples of some of the generated images.

		Noisy background			
		Folded sheet	Wrinkled sheet	Coffee stains	Footprints
Font type	True-type				
	San serif				
	Roman				
	Footnote				
Font size	Normal				
	Large				
Emphasized font	Yes				
	No				

3.2. NoisyOffice: Real Noisy Image Dataset

A portion of the corpus was built with real noisy images, provided for subjective evaluation, without the corresponding groundtruth⁴. This part of the corpus was obtained by printing clean text on white sheets and then adding real noise (coffee stains and footprints) to the paper documents. Moreover, the paper documents with printed text were also folded and wrinkled. The noisy paper documents were scanned afterwards. Examples of real noisy images are shown in Figure 2.

The real noisy documents were prepared by crossing the same parameters as when generating the simulated ones (see Table 1) resulting, again, in 72 noisy documents: 17 different files regarding the font type and size, crossed with the four types of noise. The documents were scanned at grayscale and are available at normal resolution (200 ppi) and double resolution (400 ppi).

4. EXPERIMENTATION WITH THE NOISYOFFICE CORPUS

Supervised machine learning techniques and, in particular, neural networks, have widely been used for image

restoration by learning appropriate filters from examples [28, 29, 7, 30, 31, 8, 32, 9, 33, 34]. The main property of neural networks, which is useful for preprocessing, is their ability to learn from examples complex nonlinear input-output relationships. They can be trained either as regression models, for noise filtering and image enhancement, or as classification models for binarization. A comprehensive review on image processing with neural networks can be found in [7, 8, 9].

Following, two different types of experiment which have used the NoisyOffice corpus are described: denoising and enhancing heterogeneous types of noise with a hierarchical system based on a cluster of Multilayer Perceptrons (MLPs) working at gray-level and, on the other side, some binarization experiments with neural networks. Besides, these last binarization techniques have been used in a mobile application aimed at capturing documents with the camera and enhancing the captured and dewarped obtained text images. All the neural networks have been trained by using the APRIL-ANN toolkit⁵ [35].

4.1. A “behaviour-based” clustering of MLPs for document enhancement

When document images are degraded by heterogeneous types of noise, a generic filter capable of cleaning up all types of noise is usually expected. However, it is possible to supply specific filters for each kind of noise provided that we can know which one has to be applied each time. This information can be given by using a noise classifier.

The idea of using a hierarchical clustering neural filters to restore images with diverse noise and degradation types was proposed in [36] and is based on the assumption that specific filters normally perform better and are easier to train than general ones.

In order to determine which is the specific filter to be applied at each image (or part of an image), a classifier of the kind of noise has to be trained as well. Neural networks were used in all cases: to train a general filter for comparison purposes, to train specific filters and, finally, to train the noise classifier, which achieved a classification rate of 68.05%. A remarkable aspect of this approach is that, instead of training a different classifier for each specific kind of noise, an agglomerative hierarchical clustering algorithm was employed to merge classifiers which behave very similarly [36].

The approach was objectively evaluated by using the simulated NoiseOffice dataset for training and the real noisy images for evaluation purposes by comparing the result of the proposed hierarchical system based on a clustering of neural filters and the result of applying a generic filter (trained with all types of noise). Since no groundtruth is available for the real noisy images, both systems have been compared with a reference: the result of cleaning each image with their most specific neural filter (trained only with their corresponding type of noise).

⁴The clean text cannot be considered as groundtruth since the scanned documents have not been registered to obtain an accurate alignment with the original clean image.

⁵<https://github.com/april-org/april-ann>

TABLE 2: Statistics of the simulated noisy images of the NoisyOffice dataset, at normal resolution (200 ppi) in grayscale. The Peak Signal to Noise Ratio (PSNR) quantifies the amount of noise presented on the images; and it is computed as $10 \cdot \log_{10}(1/\sqrt{(MSE)})$, where MSE is the mean squared error between the noisy input image and its groundtruth.

NoisyOffice	Train	Validation	Test
Images	72	72	72
Total Pixels	142.3Mp	142.3Mp	142.3Mp
Largest Image	540 × 420	540 × 420	540 × 420
Foreground pixels	1.51Mp (10.63%)	1.52 (10.62%)	1.44 (10.18%)
Peak Signal-to-Noise Ratio	13.55	13.78	12.34

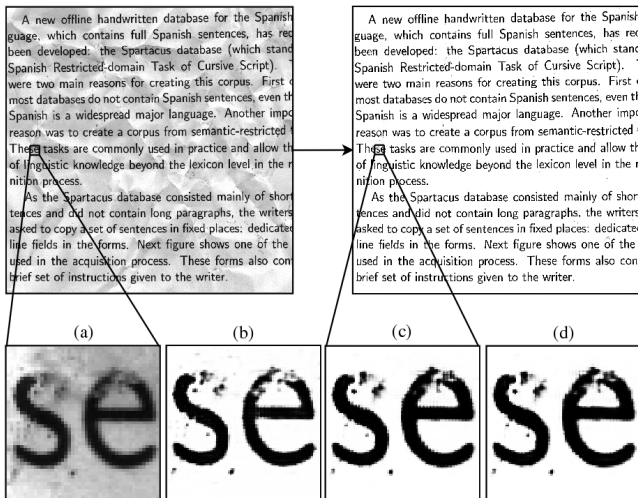


FIGURE 3: An example of the enhancement and cleaning process. (a) Original real noisy image. (b) Result of applying a neural filter trained with all types of noise. (c) Result of applying the proposed neural clustered filter. (d) Result of applying the neural filter trained with only that type of noise. (Figure from [36])

The average Euclidean distance of the images cleaned with the generic filter and the reference ones was 62.46, while the application of the proposed hierarchical neural system reduces the Euclidean distance to the reference down to 37.88, which is much lower.

Finally, in order to figure out the effect of the error caused by the filter classifier on the overall system, an upper bound was obtained by simulating the hierarchical enhancement system with an oracle error-free filter classifier. This simulation reduced the average distance from 37.88 to 28.92. An example of the performance of the proposed neural method is shown in Figure 3. As can be observed from the example, the result clearly improved the image quality.

4.2. Binarization with neural networks

Some image binarization neural network techniques have been investigated and tested with the NoisyOffice corpus. The most straightforward approach uses an MLP as a classifier in order to classify one pixel at a time. A sliding window centered on the pixel to be classified is usually fed to the MLP (see Figure 4). An extension of this approach

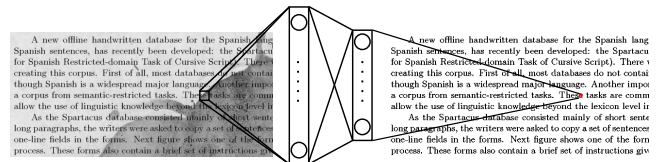


FIGURE 4: The input to the MLP is a window centered on the pixel to denoise. The output is one single value with the cleaned pixel.

has been investigated as well: the inclusion of more features as input. Besides MLPs, other connectionist models can also be envisaged: Convolutional Neural Networks (CNNs) [37] and Recurrent Neural Networks (RNNs). In the case of recurrent models, we have tried MultiDirectional Recurrent Neural Networks (MDRNNs) [38].

The first approach used an MLP which received a fixed size moving window centered at the pixel to binarize. In every experiment, the topology and parameters were estimated by a random search hyperparameter optimizer [39]. The final configuration used an input window of 9×9 pixels and two hidden layers of 32 and 16 ReLU neurons, respectively. The performance of the binarization process was measured by computing the F-Measure value, which is a combination between Precision and Recall.

Another experiment was performed by adding to the input window some additional features computed over a bigger region (see Figure 5). These features were:

- the value of the background estimation by a median filter with a radius of 10 pixels;
- four histogram values of the horizontal projection profile of one neighborhood column;
- four histogram values of the vertical projection profile of two neighborhood rows.

CNNs were also explored: The image was treated as one 2-dimensional input map (grayscale). A set of convolution-activation-pooling transformations were applied to the input maps to extract a new set of features. The CNN received a raw input window of the image in order to find a useful set of features in order to compute the predicted value of the current pixel (see Figure 6).

There are several advantages when using CNNs instead of MLPs: convolutional kernels usually operate on a smaller scale, and each one shares its weights at different

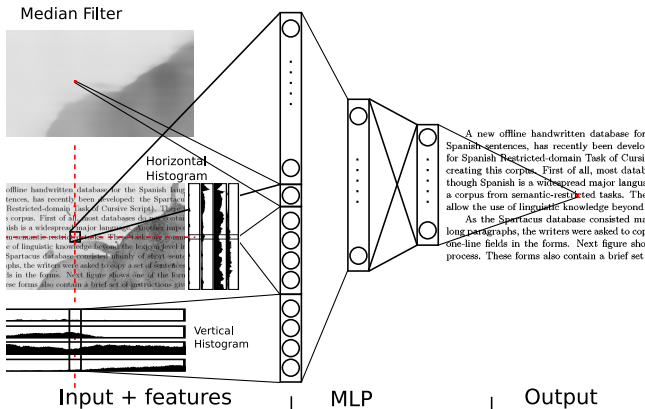


FIGURE 5: The input to the MLP is a window centered on the pixel to denoise plus some local features computed over a bigger region.

positions on the input window, which reduces the number of parameters decreasing the possibilities of overfitting and improving generalization. When using a sliding window, nearby pixels should have a significant number of features in common since they share the major part of the overlapped window. Thus, two consecutive input windows that look pretty similar have an entirely different representation of the input feature vector because of the window translation. This problem is handled better by CNNs because they maintain the 2-dimensional structure of the image and then the kernels can extract similar features from contiguous inputs. Also, max-pooling layers reduce the computational cost and provide translation invariance to the model. Therefore, with this approach, a combination of convolutional and pooling operators should be able to extract more significant features than the traditional MLPs.

The sliding window size was fixed to a 9 neighbors leading to a 19×19 window. The size of the kernel of the first convolution was set to 10 kernels of 6×6 . The size of the first sub-sampling layer was fixed to 2×2 . The sizes of the kernels of the second convolution were set to 20 kernels of 4×4 , before applying the last max-pooling sub-sampling layer of size 2×2 . The final pixel classifier in this approach is an MLP with two hidden layers of 32 and 16 ReLU neurons, respectively.

Our last experimentation makes use of recurrent neural networks which include feedback (or recurrent) connections in their hidden layers. This allows them to deal, for one-dimensional inputs, with sequences of arbitrary length. In order to work with images, we have opted for MDRNNs with Long Short Term Memories (LSTM) cells modified to keep 2-dimensional context introduced by [38]. The recurrence is done on the 4 possible orientations, as illustrated in Figure 7. The output value is extracted as a combination of these four orientations. Thus, when the value of a pixel sample is computed, each of the 4 recurrent neural networks has the information available from the 4 possible directions. The final configuration was a hidden layer with

TABLE 3: F-Measure scores for connectionist and baseline approaches.

F-Measure	Train	Validation	Test
Otsu	0.941	0.923	0.851
Sauvola	0.961	0.957	0.952
Wolf	0.928	0.919	0.929
MLP	0.984	0.981	0.976
MLP+Features	0.984	0.982	0.974
CNN	0.971	0.971	0.970
MDRNN	0.964	0.944	0.922
Ensemble	0.996	0.996	0.995

6 LSTM cells, which corresponds to 4×6 LSTMs (one for each direction).

Evaluation results of the binarization performed by the connectionist approaches are shown in Table 3, along with the performance achieved by other well-known binarization techniques: the global thresholding Otsu’s method [40], the adaptive Sauvola technique [41] and the Wolf binarization algorithm⁶ [42]. As can be observed, the neural filters performed much better than conventional methods. Finally, by combining the results from different methods, it is more likely to compensate their mistakes. We have made an ensemble of the best nets by using MERT [43]. As can be observed, the combination of approaches works very well. It is worth remarking that the smart ensemble showed an almost perfect performance. And finally, some examples together with the result of these binarization techniques are illustrated in Figure 8.

4.3. Mobile application to capture and enhance text images

The NoisyOffice corpus has been used in a mobile application developed for the Android platform and called esCam [44]. The goal of esCam is to preprocess the snapshots of text documents, in particular, perspective correction and image cleaning and enhancement. An MLP is applied to every pixel of the image, as described in previous section, and the NoisyOffice dataset was used to train it. Figure 9 shows an example of a cleaned and enhanced image by the esCam application. There are several applications designed for mobile platforms that are meant for the same purpose: Google Drive [45], CamScanner [46], or Scannable from Evernote [47] are some examples. Some of them allow to use an OCR engine or apply some enhancement filters.

5. THE “DENOISING DIRTY DOCUMENTS” CHALLENGE

We are pleased to report that, in the first two months of release at UCI, the corpus was adopted by Kaggle Inc. for the “Denoising Dirty Documents” challenge.⁷ The

⁶Using the code from <http://liris.cnrs.fr/christian.wolf/software/binarize/>.

⁷<https://www.kaggle.com/c/denoising-dirty-documents>

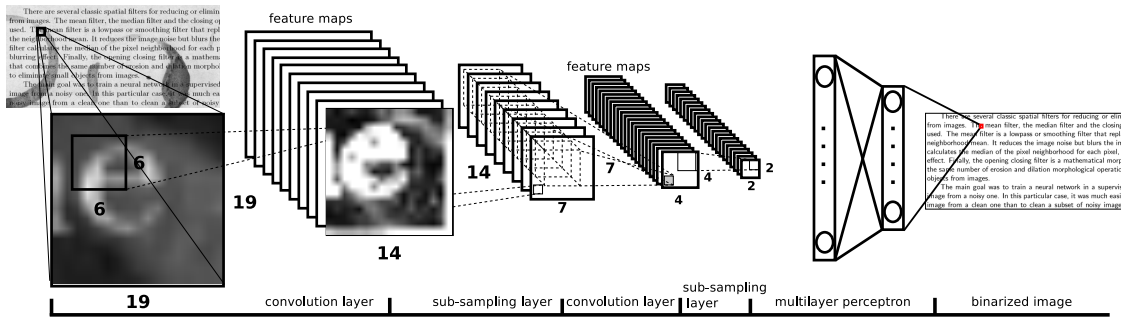


FIGURE 6: Enhancement with CNNs. The CNN was composed of two sets of convolution and sub-sampling layers, followed by an MLP with 2 hidden layers and single output neuron.

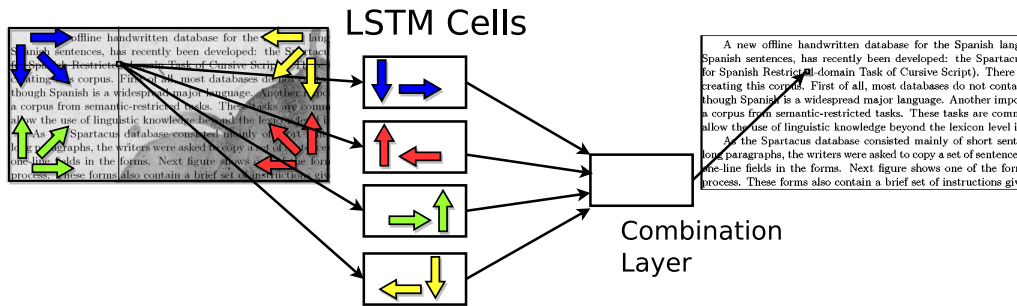


FIGURE 7: Enhancement with MDRNNs. Each of the 4 LSTM has a different context available.

- a)

A new offline handwritten database for the Spanish language sentences, has recently been developed: the Spartacus database Restricted-domain Task of Cursive Script). There were two main corpus. First of all, most databases do not contain Spanish sentences a widespread major language. Another important reason was to create restricted tasks. These tasks are commonly used in practice and knowledge beyond the lexicon level in the recognition process.

As the Spartacus database consisted mainly of short sentence paragraphs, the writers were asked to copy a set of sentences in fixed fields in the forms. Next figure shows one of the forms used in the forms also contain a brief set of instructions given to the writer.
- b)

A new offline handwritten database for the Spanish language sentences, has recently been developed: the Spartacus database Restricted-domain Task of Cursive Script). There were two main corpus. First of all, most databases do not contain Spanish sentences a widespread major language. Another important reason was to create restricted tasks. These tasks are commonly used in practice and knowledge beyond the lexicon level in the recognition process.

As the Spartacus database consisted mainly of short sentence paragraphs, the writers were asked to copy a set of sentences in fixed fields in the forms. Next figure shows one of the forms used in the forms also contain a brief set of instructions given to the writer.
- c)

A new offline handwritten database for the Spanish language sentences, has recently been developed: the Spartacus database Restricted-domain Task of Cursive Script). There were two main corpus. First of all, most databases do not contain Spanish sentences a widespread major language. Another important reason was to create restricted tasks. These tasks are commonly used in practice and knowledge beyond the lexicon level in the recognition process.

As the Spartacus database consisted mainly of short sentence paragraphs, the writers were asked to copy a set of sentences in fixed fields in the forms. Next figure shows one of the forms used in the forms also contain a brief set of instructions given to the writer.
- d)

A new offline handwritten database for the Spanish language sentences, has recently been developed: the Spartacus database Restricted-domain Task of Cursive Script). There were two main corpus. First of all, most databases do not contain Spanish sentences a widespread major language. Another important reason was to create restricted tasks. These tasks are commonly used in practice and knowledge beyond the lexicon level in the recognition process.

As the Spartacus database consisted mainly of short sentence paragraphs, the writers were asked to copy a set of sentences in fixed fields in the forms. Next figure shows one of the forms used in the forms also contain a brief set of instructions given to the writer.
- e)

A new offline handwritten database for the Spanish language sentences, has recently been developed: the Spartacus database Restricted-domain Task of Cursive Script). There were two main corpus. First of all, most databases do not contain Spanish sentences a widespread major language. Another important reason was to create restricted tasks. These tasks are commonly used in practice and knowledge beyond the lexicon level in the recognition process.

As the Spartacus database consisted mainly of short sentence paragraphs, the writers were asked to copy a set of sentences in fixed fields in the forms. Next figure shows one of the forms used in the forms also contain a brief set of instructions given to the writer.

FIGURE 8: Illustration of the binarization of the corpus with several methods. a) Original image, b) MLP, c) MLP+Features, d) CNNs, e) Ensemble.

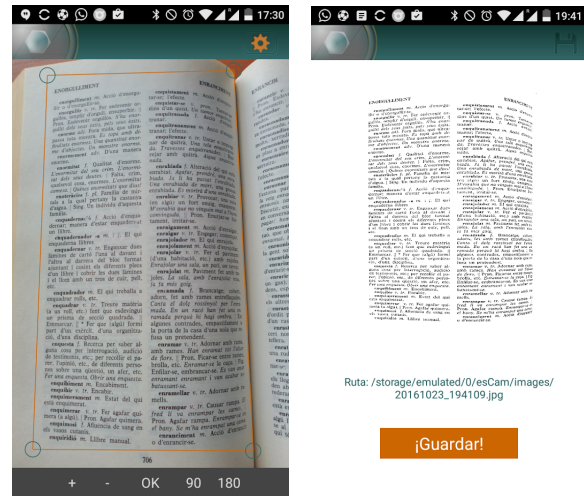


FIGURE 9: esCam application: a screenshot with an original image (left) and with the cleaned and enhanced image (right).

competition started 1 June 2015 and ended 5 October 2015, with a high participation (161 teams).

The algorithm to clean the images in the test set had to be submitted and they were evaluated on the root mean squared error between the cleaned pixel and the actual grayscale pixel intensities. Intensity values range from 0 (black) to 1 (white).

The top performances of the 161 participating teams are shown at Figure 10, where it can be observed that the best performance had an error of $4.16e^{-3}$.













#	Δ1w	Team Name	Score 	Entries	Last Submission UTC (next - Last submission)
1		 * <i>in the money</i>	0.00416	5	Sun, 04 Oct 2015 06:21:47
2		Colin	0.00512	29	Mon, 05 Oct 2015 14:14:57
3		Ironbar	0.01123	30	Mon, 05 Oct 2015 19:18:27 (-33.5h)
4		Altexsoft Team	0.01319	4	Wed, 23 Sep 2015 11:47:47
5		ShadowNet	0.01347	30	Mon, 05 Oct 2015 16:47:41 (-16.1h)
6		 toshi_k	0.01521	12	Sat, 03 Oct 2015 13:09:35 (-81d)
7		Cameron McKenzie	0.01624	18	Tue, 21 Jul 2015 23:19:50 (-2d)
8		nagadomi	0.01685	6	Sat, 11 Jul 2015 03:28:14
9	—	Richard Weiss	0.01739	12	Fri, 02 Oct 2015 23:39:01
10		zxytim	0.01794	27	Fri, 02 Oct 2015 12:46:52 (-4.1h)

FIGURE 10: Top performances for the “Denoising Dirty Documents” Kaggle challenge.

6. CONCLUSIONS

Acquisition of standard databases has become an important issue in the document analysis and recognition research community. In this paper, the NoisyOffice database has been presented in detail. The corpus consists of images of printed text documents with noise mainly caused by uncleanliness from a generic office such as coffee stains and footprints on documents and folded and wrinkled sheets with degraded printed text. This corpus is intended to train and evaluate supervised learning methods for cleaning, binarization and enhancement of noisy grayscale printed text image. To this end, the noisy images and their corresponding clean or binarized groundtruth images are provided. Double resolution groundtruth images are also provided in order to test super-resolution methods. The corpus is freely available⁸ at UCI Machine Learning Repository [10]. Experiments carried out with the corpus, regarding binarization, a clustering approach for enhancement, and its use in an App for preprocessing image snapshots, are also presented. More recently, the corpus was adopted by Kaggle Inc. for the “Denoising Dirty Documents” challenge with a very high participation and is becoming a standard for denoising tasks.

Finally, relating future lines of work, we believe that it is possible to overcome some of the shortcomings and limitations of the techniques applied to create this corpus, namely, the idea of limiting the need of manual supervision by using noising methods. This limitation came from the fact that both real noised and more complex artificially noised images require very sophisticated non-linear elastic/deformable image registration techniques. These techniques seem now more feasible thanks to recent developments relating the use of deep learning techniques (such as Convolutional Neural Networks and U-nets), for predicting non-linear registration mappings.

ACKNOWLEDGEMENTS

This research was undertaken as part of the project TIN2017-85854-C4-2-R, jointly funded by the Spanish MINECO and FEDER funds.

⁸<http://archive.ics.uci.edu/ml/datasets/NoisyOffice>

REFERENCES

- [1] Bozinovic, R. M. and Srihari, S. N. (1989) Off-Line Cursive Script Word Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, **11**, 68–83.
- [2] Plamondon, R. and Srihari, S. N. (2000) On-Line and Off-Line Handwriting Recognition: A Comprehensive Survey. *IEEE Trans. Pattern Anal. Mach. Intell.*, **22**, 63–84.
- [3] Vinciarelli, A. (2002) A survey on off-line cursive word recognition. *Pattern Recognition*, **35**, 1433–1446.
- [4] Bunke, H. (2003) Recognition of Cursive Roman Handwriting – Past, Present and Future. Proceedings of the 7th International Conference on Document Analysis and Recognition, Edinburgh, UK, 6-6 August, pp. 448–459, IEEE.
- [5] Impedovo, S. (2014) More than twenty years of advancements on frontiers in handwriting recognition. *Pattern Recognition*, **47**, 916–928.
- [6] Baird, H. S. (2007) The state of the art of document image degradation modelling. *Digital Document Processing*, pp. 261–279. Springer.
- [7] Egmont-Petersen, M., de Ridder, D., and Handels, H. (2002) Image processing with neural networks – A review. *Pattern Recognition*, **35**, 2279–2301.
- [8] Marinai, S., Gori, M., and Soda, G. (2005) Artificial neural networks for document analysis and recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, **27**, 23–35.
- [9] Rehman, A. and Saba, T. (2014) Neural networks for document image preprocessing: state of the art. *Artificial Intelligence Review*, **42**, 253–273.
- [10] Lichman, M. (2013). UCI machine learning repository.
- [11] Gatos, B., Ntirogiannis, K., and Pratikakis, I. (2009) ICDAR 2009 Document Image Binarization Contest (DIBCO 2009). Proceedings of the 10th International Conference on Document Analysis and Recognition, Barcelona, Spain, 26-29 July, pp. 1375–1382, IEEE.
- [12] Pratikakis, I., Gatos, B., and Ntirogiannis, K. (2013) ICDAR 2013 Document Image Binarization Contest (DIBCO 2013). Proceedings of the 12th International Conference on Document Analysis and Recognition, Washington, DC, USA, 25-28 August, pp. 1471–1476, IEEE.
- [13] Pratikakis, I., Gatos, B., and Ntirogiannis, K. (2010) H-DIBCO 2010-Handwritten Document Image Binarization Competition. Proceedings of the 12th International Conference on Frontiers in Handwriting Recognition, Kolkata, India, 6-18 November, pp. 727–732, IEEE.
- [14] Ntirogiannis, K., Gatos, B., and Pratikakis, I. (2014) ICFHR2014 Competition on Handwritten Document Image Binarization (H-DIBCO 2014). Proceedings of the 14th International Conference on Frontiers in Handwriting Recognition, Heraklion, Greece, 1-4 September, pp. 809–813, IEEE.
- [15] Lazzara, G. and Géraud, T. (2014) Efficient multiscale Sauvola’s binarization. *International Journal on Document Analysis and Recognition*, **17**, 105–123.
- [16] Fischer, A., Indermühle, E., Bunke, H., Viehhauser, G., and Stolz, M. (2010) Ground truth creation for handwriting recognition in historical documents. Proceedings of the 9th IAPR International Workshop on Document Analysis Systems, Boston, Massachusetts, USA, 9-11 June, pp. 3–10, ACM New York, NY, USA.
- [17] Nafchi, H. Z., Ayatollahi, S. M., Moghaddam, R. F., and Cheriet, M. (2013) An efficient ground truthing tool for binarization of historical manuscripts. Proceedings of the

- 12th International Conference on Document Analysis and Recognition, Washington, DC, USA, 25-28 August, pp. 807–811, IEEE.
- [18] Kesiman, M. W. A., Burie, J. C., Wibawantara, G. N. M. A., Sunarya, I. M. G., and Ogier, J. M. (2016) AMADI_LontarSet: The First Handwritten Balinese Palm Leaf Manuscripts Dataset. Proceedings of the 15th International Conference on Frontiers in Handwriting Recognition, Shenzhen, China, 23-26 October, pp. 168–173, IEEE.
- [19] Belhedi, A. and Marcotegui, B. (2016) Adaptive scene-text binarisation on images captured by smartphones. *IET Image Processing*, **10**, 515–523.
- [20] Zi, G. and Doermann, D. (2004) Document image ground truth generation from electronic text. Proceedings of the 17th International Conference on Pattern Recognition, Cambridge, UK, 26-26 August, pp. 663–666, IEEE.
- [21] Kieu, V. C., Visani, M., Journet, N., Mullet, R., and Domenger, J. P. (2013) An efficient parametrization of character degradation model for semi-synthetic image generation. Proceedings of the 2nd International Workshop on Historical Document Imaging and Processing, Washington, DC, USA, 24-24 August, pp. 29–35, ACM New York, NY, USA.
- [22] Varga, T. and Bunke, H. (2003) Generation of synthetic training data for an HMM-based handwriting recognition system. Proceedings of the 7th International Conference on Document Analysis and Recognition, Edinburgh, UK, 6-6 August, pp. 618–622, IEEE.
- [23] Fischer, A., Visani, M., Kieu, V. C., and Suen, C. Y. (2013) Generation of learning samples for historical handwriting recognition using image degradation. Proceedings of the 2nd International Workshop on Historical Document Imaging and Processing, Washington, DC, USA, 24-24 August, pp. 73–79, ACM New York, NY, USA.
- [24] Journet, N., Visani, M., Mansencal, B., Van-Cuong, K., and Billy, A. (2017) DocCreator: A New Software for Creating Synthetic Ground-Truthed Document Images. *Journal of Imaging*, **3**, 62.
- [25] Walker, D., Lund, W., and Ringger, E. (2012) A synthetic document image dataset for developing and evaluating historical document processing methods. Proceedings of SPIE 8297, Document Recognition and Retrieval XIX, 829710, Burlingame, California, USA, 23-23 January, SPIE.
- [26] Seuret, M., Chen, K., Eichenbergery, N., Liwicki, M., and Ingold, R. (2015) Gradient-domain degradations for improving historical documents images layout analysis. Proceedings of the 13th International Conference on Document Analysis and Recognition, Tunis, Tunisia, 23-26 August, pp. 1006–1010, IEEE.
- [27] Dong, C., Loy, C. C., He, K., and Tang, X. (2016) Image super-resolution using deep convolutional networks. *IEEE Trans. Pattern Anal. Mach. Intell.*, **38**, 295–307.
- [28] Stubberud, P., Kanai, J., and Kalluri, V. (1995) Adaptive Image Restoration of Text Images that Contain Touching or Broken Characters. Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal, Quebec, Canada, 14-16 August, pp. 778–781, IEEE.
- [29] Chi, Z. and Wong, K. (2001) A two-stage binarization approach for document images. Proceedings of the International Symposium on Intelligent Multimedia, Video and Speech Processing, Hong Kong, China, 4-4 May, pp. 275–278, IEEE. Hong Kong, China, China
- [30] Suzuki, K., Horiba, I., and Sugie, N. (2003) Neural Edge Enhancer for Supervised Edge Enhancement from Noisy Images. *IEEE Trans. Pattern Anal. Mach. Intell.*, **25**, 1582–1596.
- [31] Hidalgo, J. L., España, S., Castro-Bleda, M. J., and Pérez, J. A. (2005) Enhancement and cleaning of handwritten data by using neural networks. In Marques J. S., Pérez de la Blanca N., Pina P. (eds), Pattern Recognition and Image Analysis. IbPRIA 2005. Lecture Notes in Computer Science, vol 3522. Springer, Berlin, Heidelberg
- [32] Banerjee, J., Nambodiri, A. M., and Jawahar, C. (2009) Contextual restoration of severely degraded document images. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20-25 June, pp. 517–524, IEEE.
- [33] Gupta, S. and Mandal, R. (2015) A survey on image enhancement techniques. *International Journal of Electrical and Electronic Engineering & Telecommunications*, **4** (1), 47–54.
- [34] Pastor-Pellicer, J., España-Boquera, S., Zamora-Martínez, F., Zeshan Afzal, M., and Castro-Bleda, M. J. (2015) Insights on the use of convolutional neural networks for document image binarization. In Rojas I., Joya G., Catala A. (eds), Advances in Computational Intelligence. IWANN 2015. Lecture Notes in Computer Science, vol 9095. Springer, Cham.
- [35] España-Boquera, S., Zamora-Martínez, F., Castro-Bleda, M. J., and Gorbe-Moya, J. (2007) Efficient BP Algorithms for General Feedforward Neural Networks. In Mira J., Álvarez J.R. (eds), Bio-inspired Modeling of Cognitive Tasks. IWINAC 2007. Lecture Notes in Computer Science, vol 4527. Springer, Berlin, Heidelberg.
- [36] Zamora-Martínez, F., España-Boquera, S., and Castro-Bleda, M. J. (2007) Behaviour-based Clustering of Neural Networks applied to Document Enhancement. In Sandoval F., Prieto A., Cabestany J., Graña M. (eds), Computational and Ambient Intelligence. IWANN 2007. Lecture Notes in Computer Science, vol 4507. Springer, Berlin, Heidelberg.
- [37] LeCun, Y., Bengio, Y., et al. (1995) Convolutional Networks for Images, Speech, and Time Series. In Arbib, M. A. (ed.), The Handbook of Brain Theory and Neural Networks, MIT Press, Cambridge, MA, USA.
- [38] Graves, A., Fernández, S., and Schmidhuber, J. (2007) Multi-dimensional Recurrent Neural Networks. In de Sá J.M., Alexandre L.A., Duch W., Mandic D. (eds), Artificial Neural Networks. ICANN 2007. Lecture Notes in Computer Science, vol 4668. Springer, Berlin, Heidelberg.
- [39] Bergstra, J. and Bengio, Y. (2012) Random search for hyperparameter optimization. *The Journal of Machine Learning Research*, **13**, 281–305.
- [40] Otsu, N. (1975) A threshold selection method from gray-level histograms. *Automatica*, **11**, 23–27.
- [41] Sauvola, J. and Pietikäinen, M. (2000) Adaptive document image binarization. *Pattern Recognition*, **33**, 225–236.
- [42] Wolf, C., Jolion, J.-M., and Chassaing, F. (2002) Text Localization, Enhancement and Binarization in Multimedia Documents. Proceedings of the 16th International Conference on Pattern Recognition, Quebec City, Quebec, Canada, 11-15 August, pp. 1037–1040, IEEE.
- [43] Och, F. (2003) Minimum Error Rate Training in Statistical Machine Translation. Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics, Sapporo, 7-12 Japan, July, pp. 160–167, ACL, Stroudsburg, PA, USA.

- [44] Pastor-Pellicer, J., Castro-Bleda, M. J., and Adelantado-Torres, J. L. (2015) esCam: A Mobile Application to Capture and Enhance Text Images. In Rojas I., Joya G., Catala A. (eds), Advances in Computational Intelligence. IWANN 2015. Lecture Notes in Computer Science, vol 9095. Springer, Cham.
- [45] Google Inc (2016). Google drive. Retrieved from <http://play.google.com>. [Mobile application].
- [46] INTSIG Information Co., Ltd (2016). Camscanner-phone pdf creator. Retrieved from <http://play.google.com>. [Mobile application].
- [47] Evernote (2016). Evernote scannable. Retrieved from <http://itunes.apple.com>. [Mobile application].