# Automatic Anonymization of Printed-Text Document Images

Ángel Sánchez(✉), José F. Vélez, Javier Sánchez, and A. Belén Moreno

Rey Juan Carlos University, 28933 Móstoles, Madrid, Spain
angel.sanchez@urjc.es

**Abstract.** Nowadays, the storage and transmission of some types of documents requires the removal of personal information from involved users. Automatic text anonymization or de-identification is a solution for hiding all sensible information contained in the documents. Although the problem has been mainly studied for plain printed-text documents, there are not works where the de-identification task also produces anonymized document images with the same text fonts as those in the original documents. This data augmentation process could be applied to train a system for document image classification. In this paper, we describe an implementation of an automated anonymization modular system for printed-text image documents written in Spanish. System evaluation performed on a dataset of invoice images shows the viability of our proposal.

**Keywords:** Document image analysis · Printed-text anonymization
Regular expression · Font classification · Convolutional neural network

## 1 Introduction

The open use of personal data in processes such as Open Data, Big Data or Public Sector Transparency, requires a tradeoff between the right to information usage and the right to protection of personal data. Actually, Data Analytics results provide large benefits to some companies, which make usage of these personal information. The process of data anonymization consists in masking or removing the sensitive data in a document (or database), while preserving its original format [1]. This task is needed for sharing documents without exposing to third parties any type of sensitive information (i.e. personal names, addresses, ID document or bank account numbers, among others) contained in the original documents. Many national and international laws regulate the personal data protection. For example, the Spanish Royal Decree 1720/2007 on the protection of personal data or the European General Data Protection Regulation 2016/679 with the requirements for anonymizing personal data.

Nowadays, many public and private companies require the capture and digitization of diverse paper documentation they have. There are different Optical Character Recognition (OCR) systems which capture the text in those documents, but these systems are not able to distinguish the useful from the useless


A. Mansouri et al. (Eds.): ICISP 2018, LNCS 10884, pp. 145–152, 2018.
https://doi.org/10.1007/978-3-319-94211-7_17

information. That is why primary companies usually tackle (with help of specialized consultants) some ad-hoc development projects for digitizing the information contained in their paper documents. In this context, anonymization arises as a tool to mitigate the risks of massively capturing and processing personal data. Most work on automatic data anonymization has been carried out on unstructured or structured electronic documents both available as text files. Unstructured documents (e.g. email messages) contain plain text written in a natural language while structured documents (e.g. forms) also include some method of embedded mark-up coding, such as XML or HTML, to identify the contained text structures.

Data anonymization is specially important in the health domain for keeping patient confidentiality. Sharing medical information with third parties without causing a privacy problems is currently an great concern. For example, the US Health Insurance Portability and Accountability Act aims to protect patient data confidentiality. Moreover, the usage of these personal data for research purposes requires an informed consent of the patient, the approval of a internal board and the de-identification of sensible data [8]. Automatic text anonymization (or de-identification) is mostly based on pattern matching and machine learning techniques. These include analysis of regular expressions, grammar-based rules and dictionaries [8]. A recent survey by Garfinkel [3] describe some current data anonymization techniques and models to preserve the privacy of sensible personal information in documents. The re-identification risk, which can be estimated as the probability that any anonymized personal data can be re-identified, must also be taken into account. Anonymized documents are stored in databases of organizations which could give access to such information but ensuring confidentiality of handled personal information. Software architectures supporting data anonymization [11] should be adaptable to different domains and flexible enough to add new tools into the de-identification process.

Some common methods applied to sensible text anonymization [1,8] include: text suppression, tagging, random substitution, and text generalization. Text suppression is quite simple and consists in exchanging any sensible text component, referred as named entity (NE), by a neutral string which substitutes the original word (e.g. using "XXXX" or blank characters instead the original text element). Tagging aims to replace any NE by a unique identifier label that can mean its class followed by an identification number (e.g. replace "John" by "name1"). Random substitution changes any sensible NE by the name of another random entity of the same class stored in a dictionary (e.g. the name "John" is replaced by the name "Charles"). Finally, text generalization consists of replacing any sensible NE in the document by another NE that represents an object of a more general class that the original one (e.g. "Rey Juan Carlos University" can by replaced by "Public University" or even by "Institution").

In this paper, we describe an anonymization system for text image documents written in Spanish. We applied the above mentioned anonymization methods on invoice image documents. Our system produces for any given document image: (a) the anonymization of sensible text contained in the document and (b) a

transformed document image with its sensible data anonymized using the same fonts for these data that in the original document (in particular, letter type, size and style). Our goal is that this transformed document image presents as most as possible the appearance of the original document. For this last task, an automatic font analysis and classification stage [6] based on the use of deep learning models (in particular, convolutional neural networks) was included in the presented system.

Our approach can be used to create an augmented database of anonymized documents for research purposes (e.g. creating additional document samples for training different deep learning models for classification of de-identified document images). An additional challenging application of our system, in the field of Forensics Biometrics, could consist in developing automatic methods and tools to detect which were the modified documents [9] in the augmented database of document images, since the produced dataset will contain both genuine and changed documents.

## 2 Proposed Anonymization Approach

This section summarizes the overall architecture of our anonymization system, in special the method applied for text de-identification and how the fonts of sensible NE are detected in order to produce the resulting anonymized document images.

### 2.1 Overall Anonymization Process

Figure 1 shows a UML Activity Diagram sketching the main involved successive processes (white boxes in the central column), intermediate results and outputs in the proposed anonymization method (gray boxes). First, original paper document is properly scanned at a suitable spatial and radiometric resolution. The result is the corresponding *Image Document*. Next, we apply an OCR processing to this last image, and produce as its output a structured XML document with the contained machine-encoded printed text. This task is performed using the open-source OCR engine *Tesseract* [10], and the resulting *OCR-XML Output* contains the detected text elements in the document and some additional information (e.g. coordinate positions of text element in the document image). The fields of text to be anonymized are given as some "a priori knowledge" using some dictionaries (e.g. for names and surnames of persons, street names, and so on) or as regular expressions (e.g. identity documents numbers, dates, and so on). The following stage in our method is text anonymization which is detailed in Subsection 2.2. Next, the resulting *Anonymized OCR-XML Output* is a result of our method and it consists in a XML-structured text file containing the textual information of the original document image with the sensible information properly substituted according to the selected anonymization algorithm (i.e. suppression, tagging, substitution or generalization, respectively). The last stage of our method is producing as additional output a new similar document

image where the sensible text fields of the original scanned document are automatically substituted (using the same fonts as in the original text image) by the anonymized texts contained in the *Anonymized OCR-XML* file. This way, the resulting *Anonymized Image Document* will have a very similar aspect to the original scanned image. This last stage is detailed in Subsection 2.3.

### 2.2 Textual Anonymization Process

Although different secure hash ciphering methods, like Secure Hash Algorithms (SHA), could be used to cipher the text strings detected after the OCR process (e.g. SHA-256, SHA-512 or SHA-3), since the lengths of detected sensible text entities (NE) in the documents are relatively short, there is a risk of deciphering the original text strings [5]. In consequence, we opted to implement some of the anonymization methods described in [1]. In particular, suppression, tagging, substitution and text generalization, respectively. The considered sensible NE to be anonymized in printed documents were the following ones: personal first names and surnames, addresses, personal ID numbers, tax ID numbers, amounts, dates, phone numbers and bank account numbers. To anonymize each of these categories of sensible information, first it is necessary to recognize them and their type. For the purpose of recognizing the NE, given as regular expressions, we used the different grammar rules created by the lexical analyzer generator tool JFlex [4,7]. This software takes as input a specification with a set of regular expressions and the corresponding actions, and then generates a program (a lexer) that reads the inputs, matches them against the regular expressions in the specification file, and runs the corresponding actions when the regular expressions are matched.

### 2.3 Substitution of Anonymized Text in Document Image

Figure 2 presents the architecture of Convolutional Neural Network (CNN) defined for font type detection and classification. The input to this network consisted of a $32 \times 32$ window that slides over the word image to be classified with a slide of 8 pixels. The network outputs are three font types (in particular, Arial, Times and Courier) and two style variants (italic and bold, respectively). Note that the normal plain text style is embedded in a combination of values of output neurons. In addition, this has the following hidden layers: 3 convolutional layers of $3 \times 3$, interleaved with 3 pooling layers of size $2 \times 2$, and a final flat dense layer of 512 units as shown by Fig. 2.

In order to train this network, multiple printed samples have been created with characters of different letter fonts, styles, sizes and combinations of these features. The digitized text images have been split into words, and each word has been normalized to a height of 32 pixels by keeping its aspect ratio. Nest, each normalized word has been split into $32 \times 32$ windows with a slide of 8 to build the training and test samples for this CNN classifier.

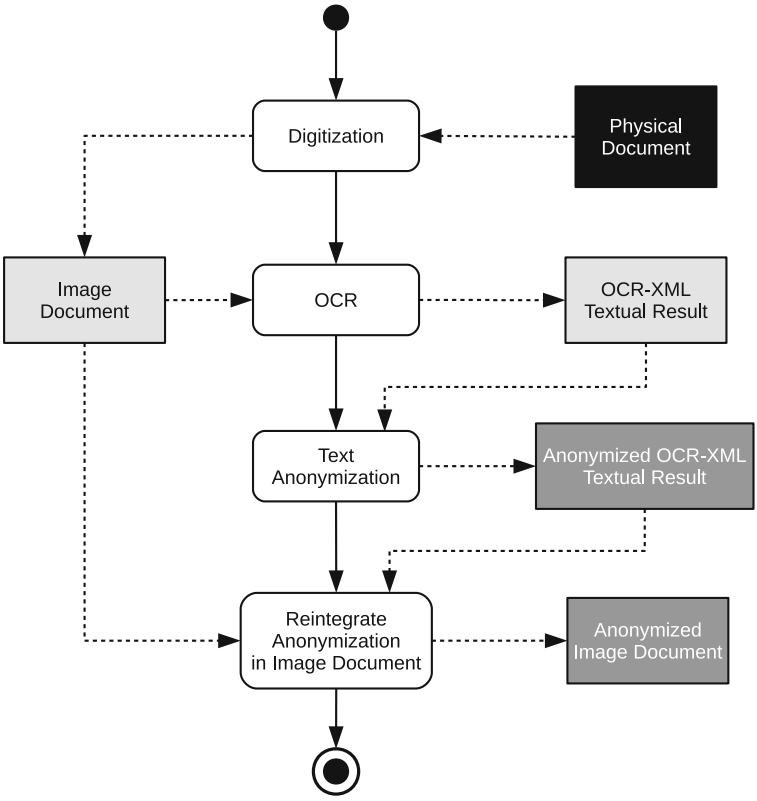


**Fig. 1.** UML Activity Diagram of Overall Anonymization Process.

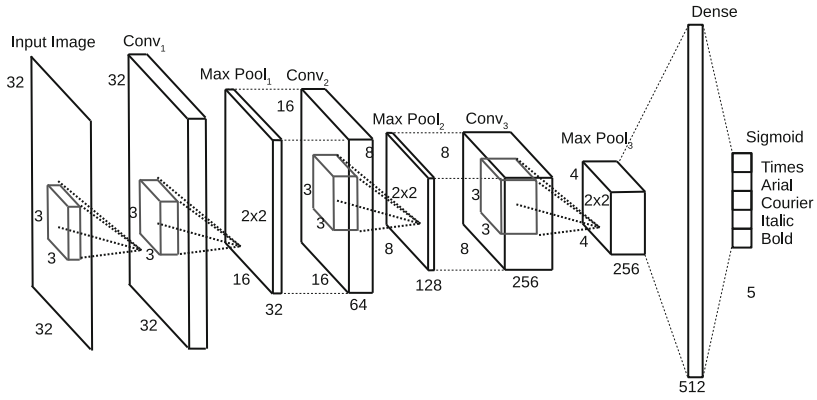


**Fig. 2.** CNN architecture used for font classification.

## 3 System Evaluation

In this section, we describe the evaluation results corresponding to the text anonymization in the document images. In our experiments, we used a number of 25 different Spanish invoice documents corresponding to different companies. Although we also experimented other additional types of documents with different fonts and our system was able to generalize, the here presented results correspond only to the set of invoices. The quality in the initial scanning of documents and the capabilities of OCR program used (particularly, the Tesseract software) influenced the recognition results. To analyze the anonymization results, we considered the *precision* and *recall* measures [2], defined as:

$$precision = \frac{TP}{TP + FP} \qquad recall = \frac{TP}{TP + FN} \tag{1}$$

where: $TP$ corresponds to sensible strings or NE which are correctly recognized as such (using the ground truth images), $FP$ are non-sensible strings recognized as sensible ones; and $FN$ corresponds to non-anonymized sensible NE strings since they were recognized as non-sensible entities. The F1-score [2] is the harmonic mean between precision and recall, and it is defined as:

$$F1 - score = 2 \cdot \frac{precision \cdot recall}{precision + recall} \tag{2}$$

By considering all of the test invoice images, we achieved an average precision and recall results of 0.886 and 0.806, respectively, in the detection of named entities (NE). For these results, the produced OCR errors when scanning the documents were also considered. Using these precision and recall values, the achieved F1-score was 0.8595. This last value represents a good tradeoff between the number of correctly classified named entities (precision) and the robustness when detecting a significant number of these sensible text entities in the analyzed documents (recall).

Figure 3 illustrates an example of anonymization applied to a scanned invoice. The left column shows the original document, and in the right column we show: a zoom of a region of it (top), the output produced by Tesseract OCR on that region (middle), and the result of substitution anonymization applied to the same invoice region (bottom).

Finally, a training set composed of 24,247 characters corresponding to times, arial and courier fonts sources, including the italic and bold variants, was built. Approximately, an 18% of these patterns (a number of 4,433 ones) was dedicated to test our CNN architecture, and an 82% (19,814 patterns) were used for training purposes. As network loss function the $binary_crossentropy$ was used, and the *Adadelta* function has been used as optimizer. A value of 0.25 dropout regularization ratio was applied to the first three convolutional layers, and a dropout ratio of 0.5 was applied to the dense layer. Using the Tensorflow open-source software library, our network was trained during 3 epochs, in which all patterns have passed through the learning algorithm using batches of size 128. The achieved accuracy for the test patterns, after the training process, was the

maximum value of 1.0 (which means a correct prediction for all the fonts of the test patterns).

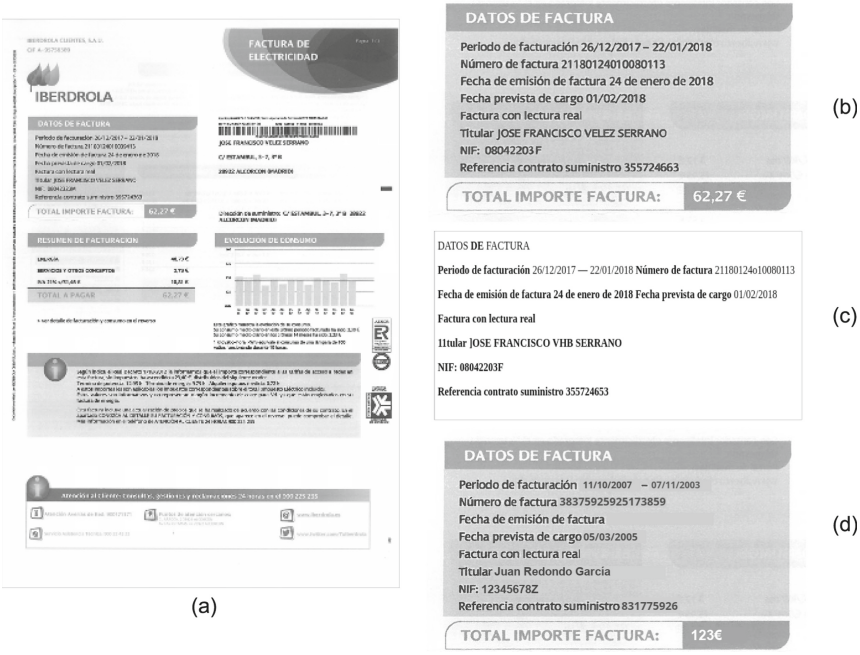


**Fig. 3.** Sample anonymization of sensible information contained in an invoice: (a) original invoice, (b) zoom of a region, (c) OCR output of this region and (d) substitution anonymization of the same region.

## 4 Conclusion

This paper described an automated anonymization system for printed-text image documents. Our approach analyzed four different anonymization methods (suppression, tagging, random substitution and text generalization, respectively), and produced an anonymized version of the documents which keeps the font styles of sensible elements in the original document. This was achieved through the training of a convolutional neural network for font type detection. Experiments were performed on image documents corresponding to Spanish invoices, and produced an average F1-score result of more than 0.85. In the future, we plan to extend our study to more different types of text-printed image documents and also investigate the problem of automatically detecting the altered (or forgered) documents in the augmented dataset of image documents produced by our system.

**Acknowledgements.** This work has been funded by the Spanish Ministry of Economy and Competitiveness under project number TIN2014-57458-R.

## References

1. Dias, F., Mamede, N., Baptista, J.: Automated anonymization of text documents. In: Proceedings of the IEEE Congress on Evolutionary Computation, pp. 1287–1294 (2016)
2. Fawcett, T.: An introduction to ROC analysis. Pattern Recogn. Lett. **27**, 861–874 (2006)
3. Garfinkel, S.L.: De-Identification of personal information. National Institute of Standards and Technology (NIST). Internal Report 8053 (2015)
4. Klein, G., Rowe, S., Décamps, R.: JFlex User's Manual. Version 1.6.1. (2015). URL: http://jflex.de/manual.html. Accessed 29 Feb 2018
5. Khovratovich, D., Rechberger, C., Savelieva, A.: Bicliques for preimages: attacks on Skein-512 and the SHA-2 family. In: Canteaut, A. (ed.) FSE 2012. LNCS, vol. 7549, pp. 244–263. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-34047-5_15
6. Lee, C.W., Jung, K.: NMF-based approach to font classification of printed English alphabets for document image understanding. In: Torra, V., Narukawa, Y., Miyamoto, S. (eds.) MDAI 2005. LNCS (LNAI), vol. 3558, pp. 354–364. Springer, Heidelberg (2005). https://doi.org/10.1007/11526018_35
7. Levine, J.: Flex & Bison. O'Reilly Media, Sebastopol (2009)
8. Meystre, S.M., Friedlin, F.J., South, B.R., Shen, S., Samore, M.H.: Automatic de-identification of textual documents in the electronic health record: a review of recent research. BMC Med. Res. Methodol. **10**, 70 (2010)
9. Saini, K., Kaur, S.: Forensic examination of computer-manipulated documents using image processing techniques. Egypt. J. Forensic Sci. **6**, 317–322 (2016)
10. Tesseract OCR: Tesseract Open Source OCR Engine (main repository). URL: https://github.com/tesseract-ocr. Accessed 10 Feb 2018
11. Vico, H., Calegari, D.: Software architecture for document anonymization. Electron. Notes Theoret. Comput. Sci. **314**, 83–100 (2015)