Towards a Handwritten Text Interpretation Framework for Ancient Spanish Manuscripts

Eduardo Xamena[a,b⋆], Carlos Ismael Orozco[a], and Gastón Carrasco Cabrera[a]

[a]Departamento de Informática (DI) - Facultad de Ciencias Exactas - UNSa
[b]Instituto de Investigaciones en Ciencias Sociales y Humanidades (ICSOH) - CONICET - UNSa
Universidad Nacional de Salta (UNSa)
Av. Bolivia 5150, Salta, Argentina
examena@di.unsa.edu.ar

Abstract. Handwritten Text Recognition is an extensively studied research topic. We implement a widely known binarization method in order to preprocess handwritten text images efficiently and accurately, acquiring adequate binary black-white images for later recognition processes. Afterwards, the characters present in the documents are used to train and evaluate deep-learning mechanisms for the recognition task. Our framework provides good source images for the recognition phase in terms of noise removal and processing of low contrast images. Besides, the process of character recognition is also improved by means of deep-learning techniques.

Keywords: Handwritten text recognition, Convolutional Neural Networks, Binarization, Document images

1 Introduction

Historical documents tell stories about our ancestors and unsuspected facts are discovered by means of them. However, automatically processing text coming from low-quality images of printed or handwritten scanned documents constitutes a challenging task. Useful information can be extracted from documentary sources present in e.g. national archives in an automated manner. The Transcriptorium project [10] comprises a study-case of digital text extraction from a big data set, by means of widely proven methods for Off-line Handwritten Text Recognition (HTR).

There exist different approaches for the task of recognizing text from written or printed documents. On the one hand, documents are segmented in lines, lines in words and words in characters [12]. On the other hand, every recognized line of text is processed by means of a set of finite-state models or Long-Short Term Memory Networks (LSTM) [3].

⋆ Corresponding author: examena@di.unsa.edu.ar

Usually, the images of scanned documents (especially old documents) are noisy and present different illumination issues. In different works, authors proposed high performance algorithms for preprocessing document images, for both printed and handwritten sources [2]. An appropriate process of binarization over a document image is a very good starting point for character recognition tasks.

HTR can be performed both in on-line or off-line manners. The former corresponds to a context where a reader needs to understand a handwritten message in real time [1], e.g. recognizing written numbers in bank cheques or options in manually filled forms. The latter is related to the case of processing large volumes of handwritten texts [10], e.g. for information extraction tasks. The present work is oriented to an Off-line HTR task, given the nature of the collected documents. As a comprehensive guide of the state-of-the-art tools for HTR, the work of Sanchez et al (2019) [11] should be read.

This work is organized as follows: Background section explains the basic concepts about methods and techniques employed in this proposal. Next, the proposed framework is detailed and the current experiments and results are presented. Finally, conclusions and future work prospects are expressed.

2 Background and Related Work

The acquisition process of digital text from images of handwritten or printed documents can be carried on by different methods. Traditionally, applications with online requirements such as humanoid robots or autonomous vehicles employ Character Recognition methods, with previous segmentation stages [12]. On the other hand, for instance, in the task of offline transcription of handwritten texts, combinations of Stochastic Finite-State (SFS) mechanisms and LSTM networks perform better. Besides, a previous step of image binarization can lend better results. The next subsections explain methods for a preliminary phase of image denoising and the latter step of text recognition, for the task of Off-line HTR.

2.1 Binarization

The process of binarization involves the change of representation of an image, turning every pixel into 1 if it belongs to the foreground or 0 if it is part of the background. Apart from HTR or basic Character Recognition tasks, this way of representation has been used e.g. for face recognition tasks [7]. Some binarization methods compute a global threshold of pixel intensity [6, 2], and turn 0 a pixel value if it is above the threshold or 1 otherwise. Others calculate local measures over the neighbourhood of every pixel for determining the relative intensity, running a sliding window [13]. As stated in [6], the former group of methods, particularly those derived from Otsu's method, show better performance values for HTR tasks.

2.2 Handwritten text recognition

The traditional off-line text recognition mechanism consists of a machine learning technology for the recognition task itself, and a prior step of segmentation in various levels: Line, word and character segmentation. In the work of [12] this architecture is implemented for an on-line HTR task, with the use of histograms of gradients as features and a Support Vector Machine for the classification. A simpler architecture is described in [4], with a neural network with back-propagation training, for off-line character recognition. All these works make use of prior binarization techniques with the purpose of decreasing the noise or errors in the images.

There are manifold successful implementations of the mentioned approaches for both on-line and off-line text recognition in terms of segmentation phases and character classification. However, the state-of-the-art methods in the last years apply a different general approach, with techniques and tools taken from the field of Automatic Speech Recognition (ASR). Basically, the working methodology consists of a mixture of Hidden Markov Models (HMM), Stochastic Finite States Transducers (SFTS) and a variety of LSTM architectures, in combination with language models. In [14] the method proposed applies a set of HMM for the task, while in [5] and [9] a mixture of Artificial Neural Networks with different architectures and prior HMMs for character recognition are used.

3 Proposed framework

The first step in the process of recognition is the binarization of the document image. For this task, Otsu's algorithm [8] was selected, as it exhibits the best trade-off between efficiency and results among the consulted works. Otsu's method is a global binarization algorithm for separating background and foreground in an image. As it is a global method, an intensity threshold is calculated for the separation process, and each pixel is contrasted with this threshold.

After the binarization step, the characters of the image are recognized by means of a Convolutional Neural Network (CNN). This mechanism accepts normalized character images as input, and returns the corresponding character class as output. The character images need to be normalized to a predefined size. For this work, the tested sizes in pixels were 20x20, 30x30, 40x40 and 50x50. Finally, the size that allowed best results was 30x30 pixels. Figure 1 shows the architecture of the CNN developed for this task.

The proposed CNN consists of the following layers, depicted in Figure 1:

- A first convolution layer that applies different masks on the images, generating 200 features of 28x28.
- A Max pooling stage that reduces the features dimensions to 14x14.
- Another convolution layer that produces 100 artificial features of 12x12.
- A dense neuron network of three layers, with 64, 50 and 23 neurons over each layer respectively.

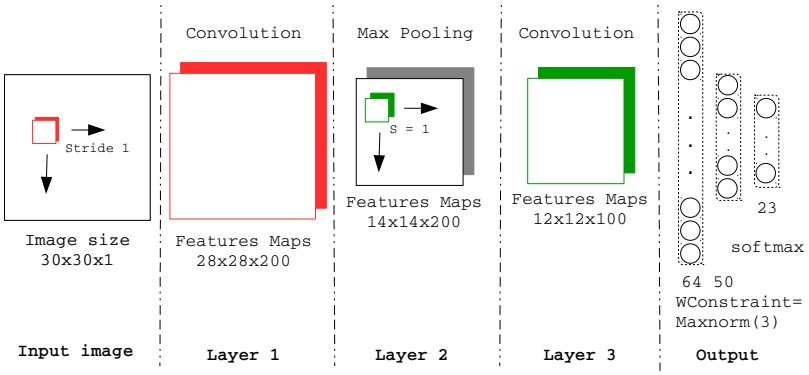

Fig. 1. CNN architecture for the task of character recognition.

The final layer of the CNN produces 23 outputs, each representing a character of the objective alphabet. Some characters of the spanish alphabet were not included as they do not appear in the texts, or there are very few instances of those to be included, as is the case of letters k and w. Then, the highest value of activation lends the current character detected by the entire network for an image. The total number of parameters to train in the CNN is 1,108,187.

The measure of success for the task of recognition is the precision obtained by the CNN. This is calculated as the number of well-classified characters over the total number of classification instances. Hence, a confusion matrix is built for this purpose.

4 Experiments and results

The software routines for the implementation of the proposed framework were built by means of Python open source libraries. For the CNN model, Keras and TensorFlow were employed. The programs were run on an Intel core i7-6700HQ processor with 16 gb of RAM.

The proposed platform has been evaluated over a set of pages of a handwritten letter of Pedro Cortés y Larraz, archbishop of the diocese of Guatemala, to the viceroy, in the year 1,771. This dataset has been collected from the PARES portal[1]. PARES is a digital repository of different Spanish Files, providing free access to diverse ancient documents related to several ages and places.

As can be observed in Figure 2, the binarization achieved by Otsu's method is very accurate. For most of the images, the background was removed with very high precision. Regarding character images, Figure 3 shows examples of extracted characters from the binarized document images.

[1] http://pares.mcu.es, Archivo General de Indias, GUATEMALA, 948, N.3

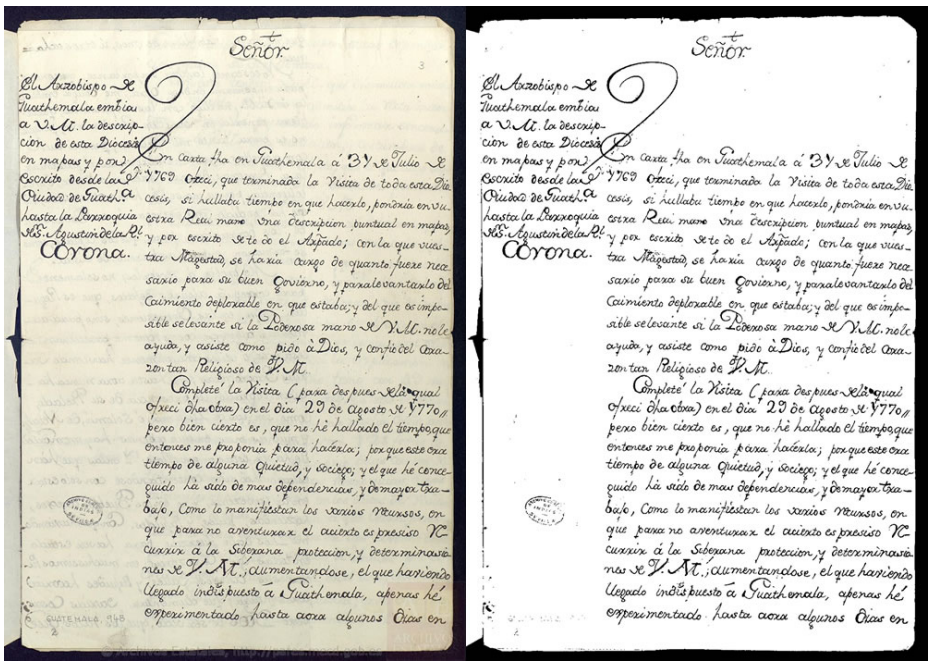

Fig. 2. Binarization applied to a handwritten document image. At the left side, the original image. At the right side, the binarized image.

Fig. 3. Examples of characters extracted from the handwritten document images.

The resulting dataset consists of 12 handwritten texts from the original letter and 710 extracted character images. The total number of character images for the CNN training phase was 664, and 43 images were left for the test phase. The training and test phases required few minutes for each parameter configuration of this small dataset. Regarding accuracy, the selected method method was cross-validation, with $k = 3$, previously shuffling the instances. The precission achieved was 90% for the training data and 71% for the test data, averaging the 3 results of cross-validation.

5 Conclusions and future work

A new method for Character Recognition has been presented. This method makes use of a global algorithm of binarization, the Otsu approach, that has high-performance values in other works. Besides, the recognition phase is carried out by the implementation of a Convolutional Neural Network over each character image. The results achieved in terms of binarization and recognition are encouraging. The stages of line and character segmentation have been executed by hand. As a matter of future work, different segmentation algorithms as well as line skew and slant correction routines will be tested. Besides, new approaches on the binarization task will be taken, and new datasets will be covered.

The purpose of the present project is building a complete framework for Offline Handwritten Text Recognition (HTR) and Transcription. In a wider scope, the texts collected will be of interest for other related projects including Text Mining and Visualization of the information that lies on those documents. The strong difficulties of the HTR task are widely known, such as the existence of writing styles as writers in a community, or the low quality of the calligraphy over early ages like the 15th or 16th centuries.

Acknowledgments

This work has been supported by Universidad Nacional de Salta (Proyecto CIUNSa C 2659).

References

1. Ahlawat, S., Rishi, R.: Off-line handwritten numeral recognition using hybrid feature set–a comparative analysis. Procedia computer science 122, 1092–1099 (2017)
2. Almeida, M., Lins, R., Bernardino, R., Jesus, D., Lima, B.: A new binarization algorithm for historical documents. Journal of Imaging 4(2), 27 (2018)
3. Castro, D., Bezerra, B.L., Valença, M.: Boosting the deep multidimensional long-short-term memory network for handwritten recognition systems. In: 2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR). pp. 127–132. IEEE (2018)

4. Choudhary, A., Rishi, R., Ahlawat, S.: Off-line handwritten character recognition using features extracted from binarization technique. AASRI Procedia 4, 306 – 312 (2013), 2013 AASRI Conference on Intelligent Systems and Control
5. Espana-Boquera, S., Castro-Bleda, M.J., Gorbe-Moya, J., Zamora-Martinez, F.: Improving offline handwritten text recognition with hybrid hmm/ann models. IEEE Transactions on Pattern Analysis and Machine Intelligence 33(4), 767–779 (April 2011)
6. Gupta, M.R., Jacobson, N.P., Garcia, E.K.: Ocr binarization and image preprocessing for searching historical documents. Pattern Recognition 40(2), 389–397 (2007)
7. Jee, H.K., Jung, S.U., Yoo, J.H.: Liveness detection for embedded face recognition system. International Journal of Biological and Medical Sciences 1(4), 235–238 (2006)
8. Otsu, N.: A threshold selection method from gray-level histograms. IEEE transactions on systems, man, and cybernetics 9(1), 62–66 (1979)
9. Romero, V., FornéS, A., Serrano, N., SáNchez, J.A., Toselli, A.H., Frinken, V., Vidal, E., LladóS, J.: The esposalles database: An ancient marriage license corpus for off-line handwriting recognition. Pattern Recognition 46(6), 1658–1669 (2013)
10. Sánchez, J.A., Mühlberger, G., Gatos, B., Schofield, P., Depuydt, K., Davis, R.M., Vidal, E., de Does, J.: transcriptorium: a european project on handwritten text recognition. In: Proceedings of the 2013 ACM symposium on Document engineering. pp. 227–228. Citeseer (2013)
11. Sánchez, J.A., Romero, V., Toselli, A.H., Villegas, M., Vidal, E.: A set of benchmarks for handwritten text recognition on historical documents. Pattern Recognition 94, 122–134 (2019)
12. Sarathy, S., Manikandan, J.: Design and evaluation of a real-time character recognition system. In: 2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI). pp. 519–525. IEEE (2018)
13. Singh, T.R., Roy, S., Singh, O.I., Sinam, T., Singh, K., et al.: A new local adaptive thresholding technique in binarization. arXiv preprint arXiv:1201.5227 (2012)
14. Toselli, A.H., Juan, A., Vidal, E.: Spontaneous handwriting recognition and classification. In: ICPR (1). pp. 433–436. Citeseer (2004)