Comparing human and machine performances in transcribing 18th century handwritten Venetian script

Sofia Ares Oliveira and Frederic Kaplan

Digital Humanities Laboratory, Ecole Polytechnique Fédérale de Lausanne

{sofia.oliveiraares, frederic.kaplan}@epfl.ch

1 Introduction

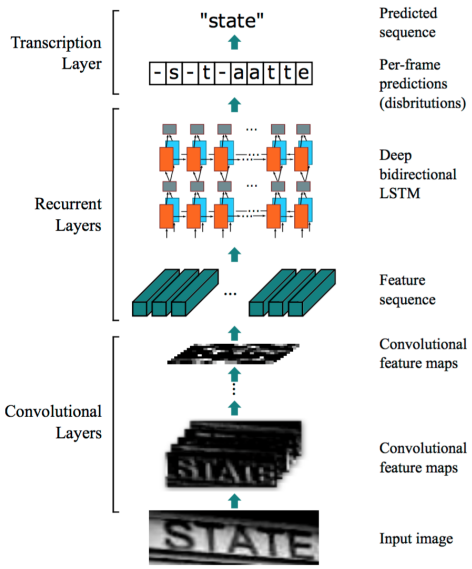
Automatic transcription of handwritten texts has made important progress in the recent years [1] [2] [3]. This increase in performance, essentially due to new architectures combining convolutional neural networks with recurrent neutral networks, opens new avenues for searching in large databases of archival and library records. This paper reports on our recent progress in making million digitized Venetian documents searchable, focusing on a first subset of 18th century fiscal documents from the Venetian State Archives (Condizione di Decima, Quaderni dei Trasporti, Catastici). For this study, about 23'000 image segments containing 55'000 Venetian names of persons and places were manually transcribed by archivists, trained to read such kind of handwritten script, during an annotation phase that lasted 2 years. This annotated dataset was used to train and test a deep learning architecture, with the objective of making the entire set of more than 2 million pages searchable. As described in the following paragraphs, performance levels (about 10% character error rate) are satisfactory for search use cases, which demonstrates that such kinds of approaches are viable at least for this typology of handwritten scripts. This paper compares this level of reading performance with the reading capabilities of Italian-speaking transcribers, preselected with a test based on 100 transcriptions. More than 8500 new human transcriptions were produced, confirming that the amateur transcribers were not as good as the expert. However, on average, the machine outperforms the amateur transcribers in this transcription tasks.

2 Machine performance

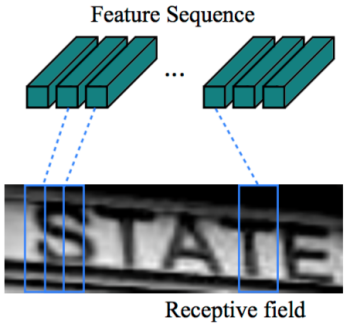
We developed a transcription system based on the combination of convolutional and recurrent neural networks as described in [4] for handwritten text (Fig.1a)[1]. On the one hand, convolutional neural networks (CNN) capture hierarchical spatial information, with the first layers capturing low level features and later ones capturing high level ones. On the other hand, recurrent neural networks (RNN) capture temporal data, with the ability to grab contextual information within a sequence of arbitrary length. Convolutional recurrent neural networks (CRNN) combine the best of both worlds to handle multi-dimensional data as sequences.
From an input image, the convolutional layers extract a sequence of compact representations which corresponds to the columns of the feature map. They are processed from the left to the right of

[1]The code is implemented in python and is available at https://github.com/solivr/tf-crnn

the image to form a sequence of local image descriptors (Fig.1b).



(a) Network architecture. The architecture consists of three parts: 1) convolutional layers, which extract a feature sequence from the input image; 2) recurrent layers, which predict a label distribution for each frame; 3) transcription layer, which translates the per-frame predictions into the final label sequence. [4]



(b) Feature Sequence [4]

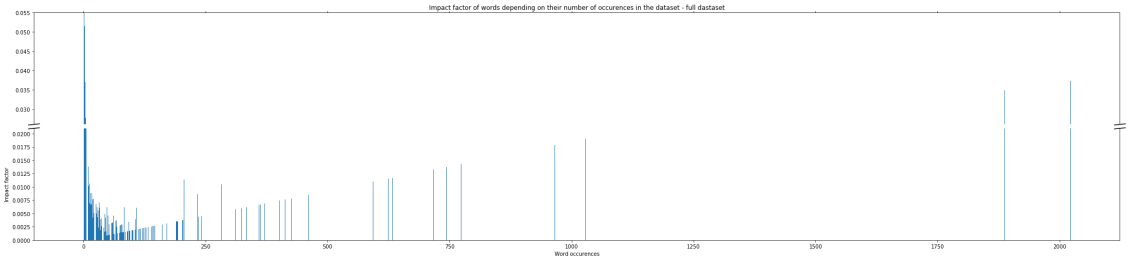
The sequence is then input to the recurrent layers which consist of stacked bidirectional long short-term memory (LSTM) cells [5]. LSTM cells have the ability to capture long-range dependencies but are directional, and thus only use past contexts. Since in image-based sequences context from both directions are useful and complementary, one forward and one backward LSTM cells are combined to form bidirectional LSTMs which are then stacked to have several recurrent layers. The recurrent network outputs per-frame predictions (probabilities) that need to be converted into a label sequence.
In the transcription layer, the connectionist temporal classification (CTC) [6] is used in order to obtain the "label sequence with the highest probability conditioned on the per-frame predictions". The sequence label is found by taking the most probable label at each time step and mapping the separated labels to the correct sequence label (see [6] to have the detailed explanation on how the repeated and 'blanks' labels are dealt with).
The CRNN was trained on data coming from various types of Venetian handwritten documents. The dataset is composed of image segments of mainly names and places that have been transcribed by archivists in Venice. Image segments are used in order to reflect only the performance of the transcriber system, without introducing possible errors from the segmentation process. Thus, the segmentation step is not part of the proposed experiment. The set was randomly split into training and testing set and the content of the image segments ranges from one to several words (Tab.1).

We show in Fig.2a and 3a how words are distributed in the dataset. We define the vocabulary to be the set of different words. The impact factor IF is a measure of the words' distribution in the

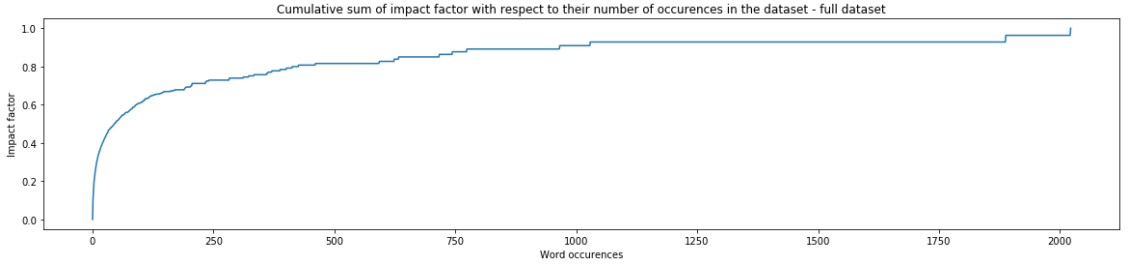
Set	# images segments	# total words	size of vocabulary
Training set	20712	48628	8848
Testing set	2317	5559	2157
Full set	23029	54187	9429

Table 1: Datasets used

dataset and is defined as $IF(i) = \frac{c(i)}{n} hist(i, c)$, with c the vector of counts of each vocabulary word, n the total number of words, $hist$ the histogram operation and $hist(i, c)$ the number of vocabulary words that occur i times. The left part of these plots shows that most of the words do not appear commonly but a few are very present in the dataset as it can be seen on the right of the figures (those are mainly prepositions such as 'di', 'de', 'in', etc). The cumulative sums (Fig.2b and Fig.3b) show that common words have limited impact, but also that the system does not suffer from over-fitting to the vocabulary since most of the words used for training are 'rare' in the dataset.



(a)

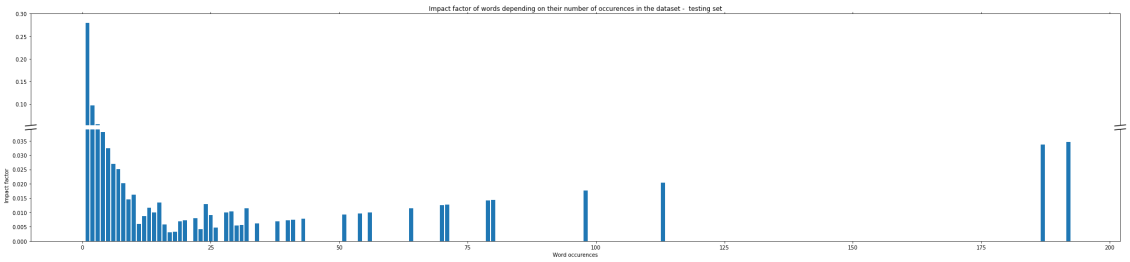


(b)

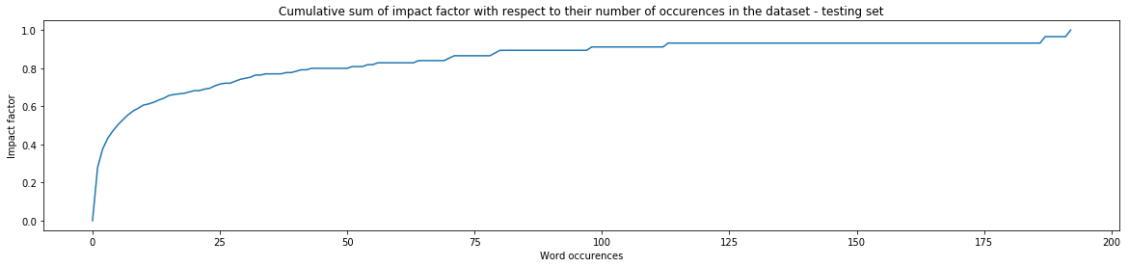
Figure 2: Word distribution and impact factor in full dataset. We observe that 70% of the dataset is represented by words appearing less than 250 times (out of 54187 words)

To evaluate the performance of the system we use the Character Error Rate (CER) measure on the test set defined as $CER = (i + s + d)/n$ with i, s, d, n the number of character insertions, substitutions, deletions and total characters respectively. The numerical results are shown in Tab. 2. Several experiments were performed using different sets of characters (called 'Alphabet' hereafter) and resulted in one model per Alphabet. A few randomly selected examples can be seen in Appendix A.

On this dataset, our transcription system is below 10% CER, which is sufficiently good to be able to search for entities in documents using regular expressions and fuzzy matching. Moreover, we believe this performance is better than the human average and in order to verify our hypothesis, we conducted an experiment described in the following section.



(a)



(b)

Figure 3: Word distribution and impact factor in the testing dataset

Alphabet	Set of characters	# image segments	CER
Capital-lowercase-symbols	A-Za-z'.,: -=	24035	0.089
Capitals-lowercase-digits-symbols	A-Za-z0-9'.,:; -_ =()[]/	96198	0.045
Digits	0-9	72326	0.013

Table 2: The Character Error Rate (CER) for each Alphabet

3 Human performance

In order to quantify the human average error rates on our dataset, we conducted an experiment on Crowdflower's platform, where Italian speaking persons were paid to transcribe image segments of the testing set (see examples in App. A). The contributors had to decipher a few units before being able to start the survey and during the experiment some of their transcriptions were evaluated. There were 103 evaluation questions that allowed to separate low accuracy contributors' answers from reliable ones. Each image segment was transcribed at least three times, and in total 11'727 units were transcribed. Only the answers of contributors maintaining at least 60% accuracy throughout the experiment and who transcribed at least 50 units were taken into account for the analysis. This resulted in a total of 8'674 valid transcriptions to analyse. The number of transcriptions (judgments) per contributor and its location can be seen in Fig.6.

We compare the performance of the system and the amateur transcribers in Tab.3 and Fig.4,5 (onesample t-test, $p < 0.005$). It is clear from the graphs that the CRNN system has a better CER and WER than the human average on this dataset, and only a few contributors have lower or comparable performance to the system, but is not yet as good as the expert. It is interesting to notice that the performance of the best amateur transcriber almost doubles when capital letters and punctuation are not considered (case 3) whereas the CRNN makes little improvement. Indeed, although the system has inferred some sort of weak language model, we have seen it producing unlikely transcriptions whereas the best contributor uses its knowledge of Italian proper nouns to deduce the correct transcription when some characters are difficult to read. Thus, the system's

CER and WER could be reduced by using a lexicon-based transcription, where the output of the neural network would be compared to a dictionary and the closest element would be chosen.

Case	CER		WER	
	CRNN	contributors	CRNN	contributors
0 : No modifications (Fig.4a)	0.0804	0.1328	-	-
1 : Capital letters replaced by lowercase (Fig.4b)	0.0768	0.1137	-	-
2 : All punctuation removed (Fig.4c, 5a)	0.0766	0.1241	0.2709	0.4318
3 : Combination of Case 1 and Case 2 (Fig.4d, 5b)	0.0718	0.1047	0.2551	0.3507

Table 3: Comparison of Character Error Rates (CER) and Word Error Rates (WER) considering different formatting cases of the transcriptions for our system and the mean of the contributors (ground-truth and predictions are formatted in the same way)

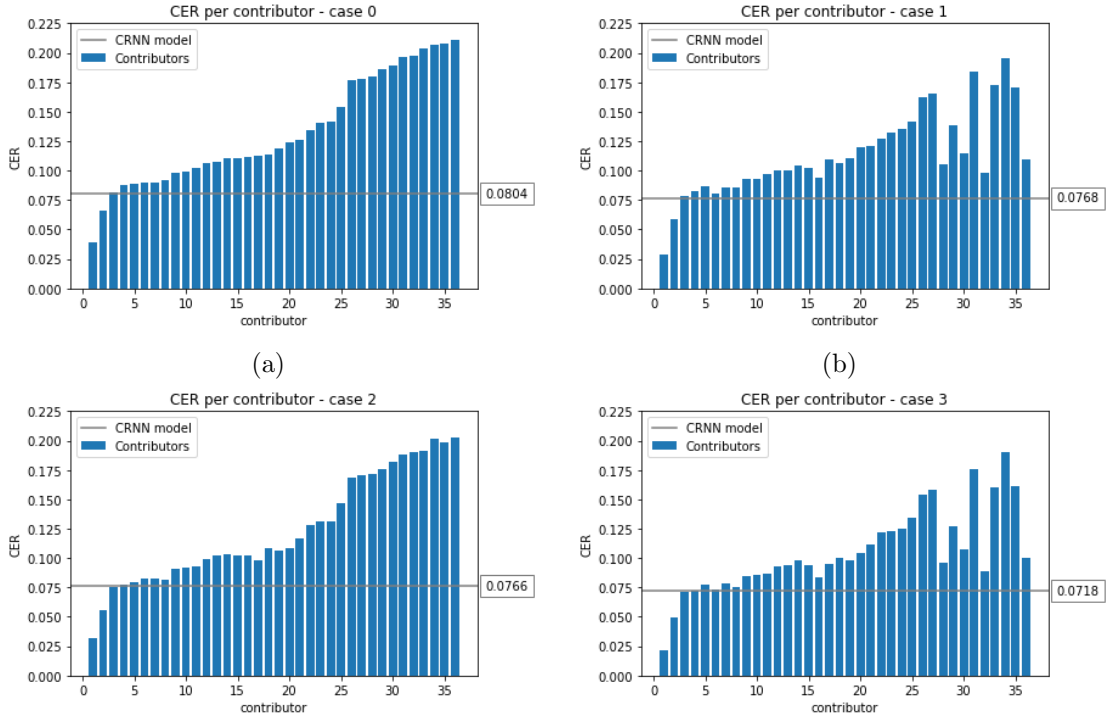
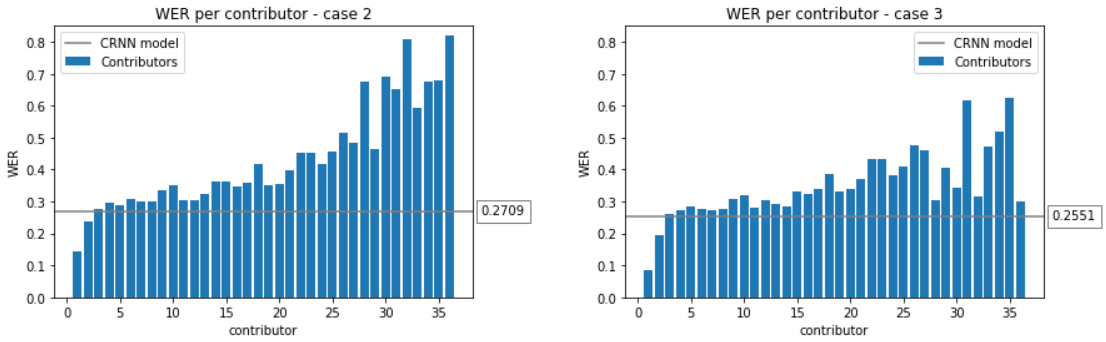


Figure 4: Character Error Rate per contributor for different cases (refer to Tab.3).

4 Perspectives

The developed system shows promising results to make possible the textual search on digitized handwritten documents. These results open up new prospects for massive indexing, analyse and study of historical documents. We showed that the system had lower Character and Word Error Rates than the human average, thus being sufficiently reliable to use for searching purposes. Further work will focus on improving the architecture of the model, especially the CNN. We will also explore



(a) (b)

Figure 5: Word Error Rate per contributor for different cases (refer to Tab.3).

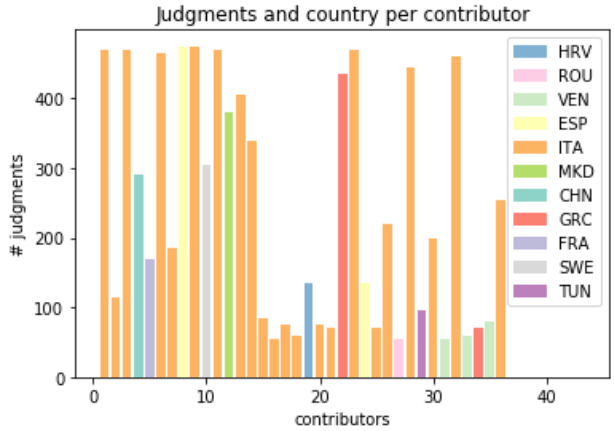


Figure 6: Number of judgements made (image segments transcriptions) by each contributor and its location. The contributors' ordering is the same as Fig.4a (by increasing CER)

the possibility of lexicon- or rule-based transcription to decrease error rates.
More generally, it seems that the automatic transcription is currently passing a threshold in terms of performance, now giving better results than good amateur transcribers. Future research will show how far this level of performance depends on the expert initial training set or whether, after some exposition with dozens of different scripts, the automatic transcriber may be able to generalize by itself without further specific training.

References

[1] J. A. Sánchez, V. Romero, A. H. Toselli, and E. Vidal, "Icfhr2014 competition on handwritten text recognition on transcriptorium datasets (htrts)," in *Frontiers in Handwriting Recognition (ICFHR), 2014 14th International Conference on*, pp. 785–790, IEEE, 2014.

[2] J. A. Sanchez, A. H. Toselli, V. Romero, and E. Vidal, "Icdar 2015 competition htrts: Handwritten text recognition on the transcriptorium dataset," in *Document Analysis and Recognition (ICDAR), 2015 13th International Conference on*, pp. 1166–1170, IEEE, 2015.

[3] J. A. Sánchez, V. Romero, A. H. Toselli, and E. Vidal, "Icfhr2016 competition on handwritten text recognition on the read dataset," in *Frontiers in Handwriting Recognition (ICFHR), 2016 15th International Conference on*, pp. 630–635, IEEE, 2016.

[4] B. Shi, X. Bai, and C. Yao, "An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 11, pp. 2298–2304, 2017.

[5] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[6] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *Proceedings of the 23rd international conference on Machine learning*, pp. 369–376, ACM, 2006.

A Transcription examples

Figure 7: 'P' the prediction of the system and 'GT' the ground-truth transcribed by Venetian archivists