A Historical Document Handwriting Transcription End-to-end System

Verónica Romero(✉), Vicente Bosch, Celio Hernández, Enrique Vidal, and Joan Andreu Sánchez

PRHLT Research Center,
Universitat Politècnica de València, Valencia, Spain
{vromero,vbosch,ceherto,evidal,jandreu}@prhlt.upv.es

Abstract. To provide access to the contents of the document collections that are being digitized, transcription is required. Unfortunately manual transcription is generally too expensive and, in most cases, current automatic techniques fail to provide the required level of accuracy. An alternative that can speed up and lower the cost of this process is the use of computer assisted, interactive techniques. These techniques work at line-level thus the transcription task assumes that the page images have been correctly decomposed into the relevant text line images. In this paper we present an end-to-end system that takes as input a page image and provides a fully correct transcript with the help of user interaction. The system automatically performs the text block and text line detection to be fed into the interactive computer assisted transcription. Experiments carried out show that the expected amount of user effort needed to produce perfect transcripts, can be reduced by using the proposed end-to-end system.

Keywords: Handwritten text recognition · Text line segmentation · Computer assisted transcription · Historical documents

1 Introduction

An increasing number of organizations are carrying out the digitization of large amounts of historical handwritten documents. However, for these raw digital images to be really useful, they need to be *transcribed* in order to provide new ways of indexing and querying the image collections. However, fully manual transcription requires highly qualified experts, making it a time-consuming and expensive process. Clearly, when the amount of text images to be processed is large, this is not a feasible solution. On the other hand, fully automatic transcription based on state-of-the-art Handwriting Text Recognition (HTR) methods is cheaper but often fails to provide the required level of transcription accuracy.

An alternative that can speed up and lower the cost of the process, while guaranteeing fully correct transcriptions, is the use of recently developed computer assisted, interactive HTR approaches such as CATTI (Computer Assisted Transcription of Text Images) [9]. For a given *text line image* to be transcribed an iterative interactive process is performed between a CATTI system and the user,

L.A. Alexandre et al. (Eds.): IbPRIA 2017, LNCS 10255, pp. 149–157, 2017.
DOI: 10.1007/978-3-319-58838-4_17

the system yields successively improved transcription hypotheses in response to the simple user corrective feedback.

For the successful use of CATTI in practice an accurate detection of the text lines of each page image is required. In the case of *historical handwritten* text images, line detection and extraction is in itself a difficult task. In these cases, advanced line detection techniques such as that proposed in [2] can be used.

The traditional HTR (and/or CATTI) workflow thoroughly decomposes the page image transcription task into two separated tasks: (a) image preprocessing and text block and line detection and extraction and (b) transcription of each extracted line image. This decomposition is very convenient for experimental purposes as each task can be tackled independently, but it is inappropriate for practical text image transcription tasks. In fact, real scenarios demand a system that accepts a full page image as input and provide full transcripts of all the text elements as output.

In this paper we present an end-to-end system that takes as input a page image and provides a fully correct transcript with the help of user interaction. This system have been assessed through experiments on a relatively small historical Spanish document, with encouraging results.

2 System Overview

As previously said, the system we are presenting takes as input a handwritten page image to be transcribed and returns its best transcription hypothesis. Then, the transcription errors can be interactively corrected in an assisted scenario. The system is composed of the following modules: (i) document image preprocessing [10]; (ii) layout analysis; (iii) line image recognition [10] (iv) and finally, a computer assisted transcription module [9].

2.1 Preprocessing

Each page image is preprocessed in order to reduce the noise, recover handwritten strokes damaged due to page degradation and correcting basic geometry distortions (see Fig. 1(b)). First, each image is converted to grey scale and the text is enhanced [11]. Then, a bi-dimensional median filter [5] is applied to the grey scale image to remove background and reduce the noise. At this step the global text image skew angle is also determined and corrected [3,8].

2.2 Layout Analysis

The layout analysis is performed in a two step top-down process.

Text Block Detection. To detect the different text blocks of each page an automatic localization of text areas is performed by means of horizontal and vertical line detection methods based on the use of enhanced profiles. The block information obtained is used in order to further preprocess the pages and eliminate all issues outside the text blocks (Fig. 1(c)). Finally, the text lines are detected on the cleaned images.

Text Line Detection and Segmentation. The text line analysis and detection (TLAD) approach is based on HMMs and finite state or N-gram vertical layout models [1,2]. It follows the same successful statistical framework which is firmly established for automatic speech and handwritten text recognition. In this context, each page must be represented as a feature vector sequence which conveys information about the vertical page layout; namely, information about where text-lines may appear along the vertical page dimension and of which kind these lines (or other non-textual objects) are. These features consist of horizontal projection profiles computed in several vertical slabs of the page [1,2].

We formulate the TLAD as the problem of finding the most likely line label sequence hypothesis, $\hat{\mathbf{h}}$, for a given handwritten page image, represented as a sequence of feature vectors $\mathbf{o}$. In addition to adequately labelling each horizontal region, we are also interested in actually determining their corresponding *vertical position* inside the page. Let $\mathbf{b}$ be the sequence of *boundary marks*, that define the different lines found in the page (see Fig. 1(d)). Following the same discussion presented in [1], from the decoding process we can obtain both the best label sequence, $\hat{\mathbf{h}}$, and the best segmentation, $\hat{\mathbf{b}}$:

$$(\hat{\mathbf{b}}, \hat{\mathbf{h}}) \approx \arg\max_{\mathbf{b},\mathbf{h}} P(\mathbf{h})\, P(\boldsymbol{o}_{b_0}^{b_1} \mid h_1) \ldots P(\boldsymbol{o}_{b_{n-1}}^{b_n} \mid h_n) \tag{1}$$

where $P(\mathbf{o} \mid \mathbf{h})$ is a *vertical line shape model* and $P(\mathbf{h})$ is a *vertical layout model* (VLM). $P(\mathbf{o} \mid \mathbf{h})$ is approximated by HMMs, while $P(\mathbf{h})$ is modelled by a finite-state model representing a-priory restrictions of how the different types of horizontal regions (called "line labels") are concatenated to form a text page. In our approach only three horizontal region are considered; namely BlankSpace (BS), Normal (base)Line (NL) and InterLine (IL). Accordingly, a very simple finite-state VLM is used which allows for any concatenation of pairs NL-IL, surrounded by BS regions.

Finally, for each detected baseline, an extraction polygon can be easily calculated (see Fig. 1(e)) by taking some pixels above and below the base-line as per the documents script.

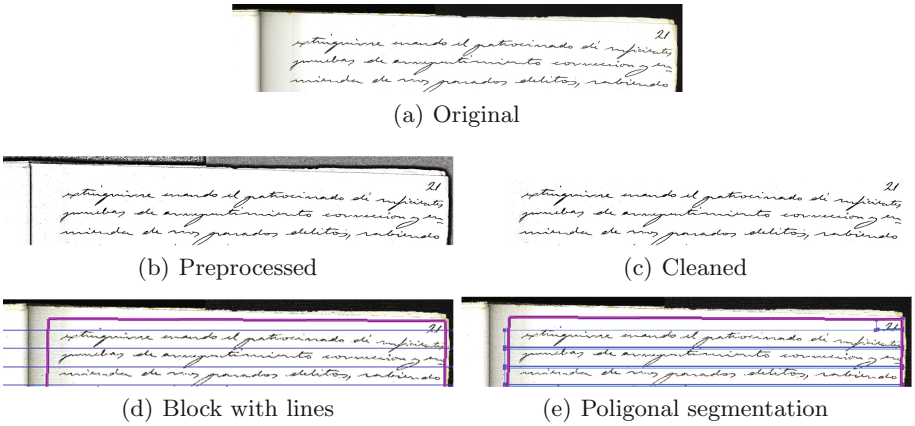
2.3 Handwritten Text Recognition

Recognition can be formulated as the problem of finding the most likely word sequence, $\widehat{\mathbf{w}} = (\hat{w}_1\, \hat{w}_2 \ldots \hat{w}_l)$, for a given handwritten sentence image represented by a feature vector sequence $\mathbf{x} = (\boldsymbol{x}_1\, \boldsymbol{x}_2 \ldots \boldsymbol{x}_n)$ [7]:

$$\widehat{\mathbf{w}} = \arg\max_{\mathbf{w}} P(\mathbf{w} \mid \mathbf{x}) \approx \arg\max_{\mathbf{w}} P(\mathbf{x} \mid \mathbf{w}) P(\mathbf{w}) \tag{2}$$

where $\mathbf{w}$ ranges over all possible sequences of words. $P(\mathbf{w} \mid \mathbf{x})$ is typically approximated by concatenated character HMMs [4]. On the other hand, $P(\mathbf{w})$ represents probabilistic syntactic knowledge and is approximated by an n-gram language model [4].

The search of Eq. (2) is carried out by using the Viterbi optimization algorithm [4]. In addition to the optimal solution, $\widehat{\mathbf{w}}$, a huge set of best solutions



(a) Original

(b) Preprocessed

(c) Cleaned

(d) Block with lines

(e) Poligonal segmentation

Fig. 1. Figure shows the process that a sample page section undergoes through out the document layout process.

can be obtained as a by-product in the form of a word graph (WG). WGs will be used in the interactive HTR.

2.4 Computer Assisted Transcription of Text Images

The last step of our system consists in assisting the human in the obtention of the perfect transcription as per the CATTI scenario [9].

In the CATTI approach, the process starts when the HTR system proposes a full transcription $\hat{s}$ of a feature vector sequence $\mathbf{x}$, extracted from a handwritten text line image. Then, the human transcriptor validates the longest prefix of the transcription which is error-free and introduces some amendments to correct the erroneous text that follows the validated prefix, producing a new prefix (p). Next, the HTR system takes into account the new prefix to suggest a suitable continuation (i.e., a new $\hat{s}$), thereby starting a new cycle. This process is repeated until a correct, full transcription t of $\mathbf{x}$ is accepted by the user.

In the CATTI framework, in addition to the given feature sequence, $\mathbf{x}$, a prefix p of the transcription is also available and the HTR module is asked to complete this prefix by searching for a most likely suffix $\hat{s}$ as:

$$\hat{s} = \arg\max_{s} P(s \mid \mathbf{x}, p) = \arg\max_{s} P(\mathbf{x} \mid p, s) \cdot Pr(s \mid p) \tag{3}$$

$P(s \mid \mathbf{x}, p)$ is modelled by HMM morphological words models [4] and $Pr(s \mid p)$ is modelled by an n-gram language model conditioned by p [9]. Using word-graph to implement these techniques, very efficient linear cost search is achieved.

3 The RSEAPV Database

The "Real Sociedad Económica de Amigos del País de Valencia" (RSEAPV) is a partnership that was established in 1776. It was a reference center for discussion and treatment of the most important and cutting-edge issues of that moment.

The RSEAPV posseses an archive composed of more than 8,000 documents that has been digitalized and made available to the public.[1] In this paper we have chosen a document of this collection to test our end-to-end system on it. The selected document was written by a single writer in Spanish in 1905 and it is composed of 170 pages.

To carry out layout, HTR and CATTI experiments we used a small set of the first 42 pages of the document. This set was annotated with two different types of annotations. First, a layout analysis of each page was manually done to indicate text blocks and lines, resulting in a dataset of 651 lines. Second, the dataset was transcribed line by line by an expert paleographer. The column "Total" of the Table 1 summarizes the basic statistics of the dataset text transcriptions.

Table 1. Basic statistics of RSEAPV dataset.

Number of:	Total	Train	Test	Cross-Val
Pages	42	22	20	5.25
Lines	651	303	348	81.4
Running words	4,573	2,150	2,439	572
Lexicon size	1,497	838	936	299.6
Out-of-vocabulary words	–	–	813	143

Two different partitions were defined in this RSEAPV dataset to carry out the experiments. The *train-test* partition was composed of two consecutive blocks. The first one, composed of the first 22 pages, was used to train the statistical models. The second one, composed by the remaining 20 pages, was used to test the system. The columns "Train" and "Test" of Table 1 summarizes the basic statistics of the two blocks.

Given the difficulty of this partition and in order to assess how increasing the number of training data affects to the quality of the automatic transcription, a second partition was defined. In this partition the 42 pages were divided into 8 blocks to carry out *cross-validation* experiments.

4 Experimental Framework

4.1 System Setup

Experiments were carried out to test the different steps of the end-to-end system.

[1] https://riunet.upv.es/handle/10251/18484.

With respect to the text line segmentation, the same feature extraction and HMM meta parameters determined in previous works [2] were adopted here. The VLM used was fixed on the base of prior knowledge of the general structure of the pages. In the text line segmentation experiments the *Train-Test* partition was used. The train block was used to train the vertical line shape HMMs using the Baum-Welch training algorithm [4].

For the HTR system, experiments with the two defined partitions were carried out. The training line images were used to train corresponding character HMMs for the HTR system, using also the standard embedded Baum-Welch training algorithm. Standard values, that have been proven to work well in previous experiments, were chosen [9].

In addition, two experimental condition were considered in each experiment: *Open* and *Closed Vocabulary* (OV and CV). In the OV setting, only the words seen in the training transcriptions were included in the recognition lexicon. In CV, all the words which appear in the test set but were not seen in the training transcriptions were added to the lexicon. In both cases a 2-gram with Kneser-Ney back-off smoothing [6] was estimated only from the training transcriptions. Finally, for each recognized test line image a WG was obtained [9].

4.2 Assessment Measures

Line Segmentation Evaluation Measures. In order to assess the quality of the line segmentation approach, two kinds of measures were adopted: *line error rate* (LER) and *Alignment Accuracy Rate* (AAR). LER is calculated as the number of incorrectly detected lines divided by the total number of actual lines. On the other hand, the AAR is a more quantitative measure which evaluates the geometrical accuracy of the detected horizontal baseline coordinates with respect to the corresponding (correct) reference marks. The AAR is computed in two steps: first, for each page, we find the best alignment between the system-proposed horizontal baseline positions and the corresponding references. In a second step we compute the final AAR as the ratio (in %) between the average error (computed in the first step) and the average text line height (also in pixels) for the whole corpus.

Handwritten Text Recognition Measures. The quality of the transcriptions given by the system with any kind of user interaction is assessed by the well know *Word Error Rate* (WER) and *Character Error Rate* (CER). They are defined as the minimum number of words/characters that need to be substituted, deleted or inserted to convert the text produced by the system into the reference transcripts, divided by the total number of words/characters in these transcripts.

Computer Assisted Transcription Measures. To asses CATTI effectiveness we use the *word stroke ratio* (WSR). It is defined as the number of required word level user interaction steps necessary to achieve the reference transcripts, divided

by the total number of reference words. The WSR gives an estimate of the (simulated) human effort needed to produce correct transcripts using CATTI.

The definitions of WSR and WER make these measures directly comparable. The relative difference between them provides a good estimate of the reduction in human effort that can be achieved by using CATTI with respect to using conventional HTR system followed by human post-editing (EFR).

5 Empirical Results

Line Segmentation Results: Table 2 shows the results obtained in the text line segmentation process. We obtained a LER of 2.6%, which means that, for every 100 lines less than 3 caused issues to the line detection system. Furthermore we provide the AAR measure that indicates the geometrical accuracy of the detected horizontal baseline coordinates with respect to the corresponding ground-truth baselines, in average the detected lines are close to the actual baseline reference.

Table 2. LER and AAR results of the text line segmentation process.

LER (%)	AAR (%)	
	Average	Standard deviation
2.6	11	30

An important aspect to remark in these experiments is the really few number of pages required to obtain good results. Only 22 pages were necessary to obtain a detection accuracy above the 2.6% mark without having to perform specific parameter tuning for the corpus.

HTR Results: Table 3 shows the results obtained in the HTR step. The first row (Aut. TLD) shows the recognition result of the lines automatically segmented in the previous step. The high WER obtained is mainly due to the few samples (only 22 pages) used to train both the HMM and the LM. Note that more than 33% of the words in the test set are OOV words, and these OOV words are sure errors in the recognition step. Note that the CV experiment is a very optimistic evaluation, but allows us to study the influence of the availability of a lexicon for the given task: it gives a lower bound for the error rate that could be obtained by the availability of a better lexicon. In the Automatic TLD experiment the WER could be reduced by more than 15 points.

In order to test how much errors are due to the automatic line segmentation, we have carried out the same experiment (train with only 22 pages and test with the remaining 20) using the perfect lines marked in the GT (row GT TLD). Only 5 points are lost using the automatic detected lines. Considering that verifying by a human expert and correcting the automatic detected line errors is a very tedious and slow task, this 5% of WER is an affordable cost.

Finally, we carried out some experiments increasing the number of pages used for training the models. Given that only 42 pages are available with annotated GT, we performed a cross-validation experiment with the 8 folds previously described. We carried out eight rounds, with each partition used once as test (5 pages) and the remaining 35 pages belonging to the other 7 folds used as the training data. The average results obtained can be seen in the last row (Cross-Val) of Table 3. We can see that increasing the number of training pages only in 15, the obtained WER is reduced by more than 20%. Finally, using a better lexicon, a WER around 36% can be obtained.

Table 3. WER and CER for the two scenarios considered (OV and CV) using both partitions Train-Test (Tr-Ts rows) and Cross-Validation (Cross-Val row).

		CV		OV	
		WER	CER	WER	CER
Tr-Ts	Aut. TLD	55.6	30.3	70.7	43.0
	GT TLD	45.3	25.2	65.8	39.1
Cross-Val	GT TLD	35.9	16.4	55.7	29.6

CATTI Results: With respect to the experiments carried out to assess the performance of the CATTI system, they were performed using the WGs generated only with the OV cross-validation experiment. The estimated human effort (WSR) obtained was 45.8%, and the estimated effort reduction (EFR) computed as the relative difference between WER and WSR is 18.1%.

According to these results, to produce 100 words of a correct transcription, a CATTI user should only have to type 46 words; the remaining 54 would be automatically predicted by the system. On the other hand, if interactive transcription is compared with post-edition approach: for every 100 word errors corrected in post-edition approach the CATTI user would interactively correct only 72. The remaining 18 words would be automatically corrected by CATTI.

6 Conclusions and Future Work

In this paper, we have presented an end-to-end system that takes as input an handwritten text line image and returns the corresponding transcription. The system is based on state-of-the-art preprocessing, layout analysis and handwritten text recognition techniques. We have studied the capability of this system when applied to historical handwritten documents. The obtained results are quite encouraging. In addition, the use of assisted technologies show that the expected amount of user effort can be reduced.

Acknowledgment. This work has been partially supported through the European Union's H2020 grant READ (Recognition and Enrichment of Archival Documents) (Ref: 674943), the MINECO/FEDER-UE project TIN2015-70924-C2-1-R, and the HIMANIS EU project, JPICH programme, (Spanish grant Ref. PCIN-2015-068).

References

1. Bosch, V., Toselli, A.H., Vidal, E.: Statistical text line analysis in handwritten documents. In: Proceedings ICFHR, pp. 201–206 (2012)
2. Bosch, V., Toselli, A.H., Vidal, E.: Semiautomatic text baseline detection in large historical handwritten documents. In: ICFHR, pp. 690–695, September 2014
3. Pastor, M., Toselli, A., Vidal, E.: Projection profile based algorithm for slant removal. In: Campilho, A., Kamel, M. (eds.) ICIAR 2004. LNCS, vol. 3212, pp. 183–190. Springer, Heidelberg (2004). doi:10.1007/978-3-540-30126-4_23
4. Jelinek, F.: Statistical Methods for Speech Recognition. MIT Press, Cambridge (1998)
5. Kavallieratou, E., Stamatatos, E.: Improving the quality of degraded document images. In: DIAL 2006, pp. 340–349, April 2006
6. Kneser, R., Ney, H.: Improved backing-off for N-gram language modeling. In: ICASSP 1995, Los Alamitos, CA, USA, vol. 1, pp. 181–184 (1995)
7. Kozielski, M., Forster, J., Ney, H.: Moment-based image normalization for handwritten text recognition. In: Proceedings of the ICFHR, pp. 256–261 (2012)
8. Rezaei, S.B., Sarrafzadeh, A., Shanbehzadeh, J.: Skew detection of scanned document images. In: IMECS, Hong Kong, vol. 1, March 2013
9. Romero, V., Toselli, A.H., Vidal, E.: Multimodal Interactive Handwritten Text Transcription. MPAI. World Scientific Publishing, River Edge (2012)
10. Toselli, A.H., et al.: Integrated handwriting recognition and interpretation using finite-state models. IJPRAI **18**(4), 519–539 (2004)
11. Villegas, M., Toselli, A.H.: Bleed-through removal by learning a discriminative color channel. In: ICFHR, pp. 47–52, September 2014