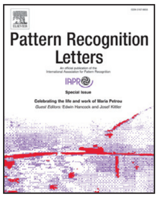
Pattern Recognition Letters 170 (2023) 9–16

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec



Processing a large collection of historical tabular images☆

Emilio Granell[a,*], Verónica Romero[b], José Ramón Prieto[a], José Andrés[a,c,d], Lorenzo Quirós[a], Joan Andreu Sánchez[a], Enrique Vidal[a,c]

[a] *Pattern Recognition and Human Language Technology Research Center, Universitat Politècnica de València, 46022 Valéncia, Spain*
[b] *Departament d'Informàtica, Universitat de València, València, Spain*
[c] *Valencian Graduate School and Research Network of Artificial Intelligence, Camíde Vera s/n, 46022 Valencia, Spain*
[d] *tranSkriptorium AI, Valencia, Spain*

ARTICLE INFO

Article history:
Received 10 August 2022
Revised 12 April 2023
Accepted 18 April 2023
Available online 20 April 2023

Edited by: Maria De Marsico

Keywords:
Table recognition
Historical tabular images
Text recognition
Probabilistic indexing

ABSTRACT

Processing automatically historical document images to allow the search of textual information requires the preparation of ground-truth data for training and evaluation. This process is an expensive and arduous task, especially when the historical document images contain specialized vocabulary and/or tabular information. In the latter case, relevant decisions have to be taken to annotate the tabular parts. This paper presents a complex collection of historical document images and the resulting database, which is called HisClima. In this database, half of the images are in tabular format and half as running text. Both types of images contain pre-printed and handwritten text. The textual information is plenty of abbreviations and specific vocabulary related to weather conditions and old ships. This database can be used to research technologies related to historical document image processing and analysis, both for tabular and running text recognition. Baseline results are presented for Document Layout Analysis, Text Recognition, and Probabilistic Indexing. Although these results are good, there is still room for improvement and some indications are provided in this direction.

1. Introduction

Table image recognition in historical documents is a prominent problem related to pattern recognition which interest has recently increased noticeably. Several reasons make this problem very relevant and attractive. From the point of view of archives, this problem is very relevant since archives contain thousands of collections composed of large amounts of table images related to borders, census, military information, visas, commerce, marriages, deaths, births, etc. Therefore, they are an invaluable information source if they are adequately processed. From the point of view of pattern recognition, table recognition in historical documents is a very challenging problem since it poses many difficult challenges not solved yet: complex layouts that change from collection to collection (or even in the same collection), lines written with different orientations, text recognition in absence of context, reading order, information extraction, dates written with different formats, abbreviations, hyphenated words in the same cell, stamps, etc., just to name a few.

Current solutions to tackle table image recognition are mainly based on machine learning techniques. To advance in the processing of historical tabular images with machine learning methods, it is required to prepare databases with their corresponding ground-truth (GT), and this is an expensive and tedious manual work. To this end, various databases for printed [1] and handwritten [2] text have been proposed in the past. Some of these databases have the corresponding GT, and preliminary experimental results are provided for some of the problems that have been mentioned above [3–5]. An entire pipeline for information extraction in handwritten tabular images was presented in [6]. They relied on the fact that the used corpus was very homogeneous and without variations in layout, with only one type of table. This made it possible to learn a different language model per column, as well as to detect an entire line per row and cut it regularly. However, the pipeline cannot deal with tables with different layouts. A relevant point related to the databases that have been mentioned above is the ground-truth preparation. Specific annotation tools (such as

☆ Editor: Maria De Marsico.
* Corresponding author.
E-mail addresses: emgraro@prhlt.upv.es (E. Granell), veronica.romero@uv.es (V. Romero), joprfon@prhlt.upv.es (J.R. Prieto), joanmo2@prhlt.upv.es (J. Andrés), lquiros@transkriptorium.com (L. Quirós), jandreu@prhlt.upv.es (J.A. Sánchez), evidal@prhlt.upv.es (E. Vidal).

https://doi.org/10.1016/j.patrec.2023.04.007

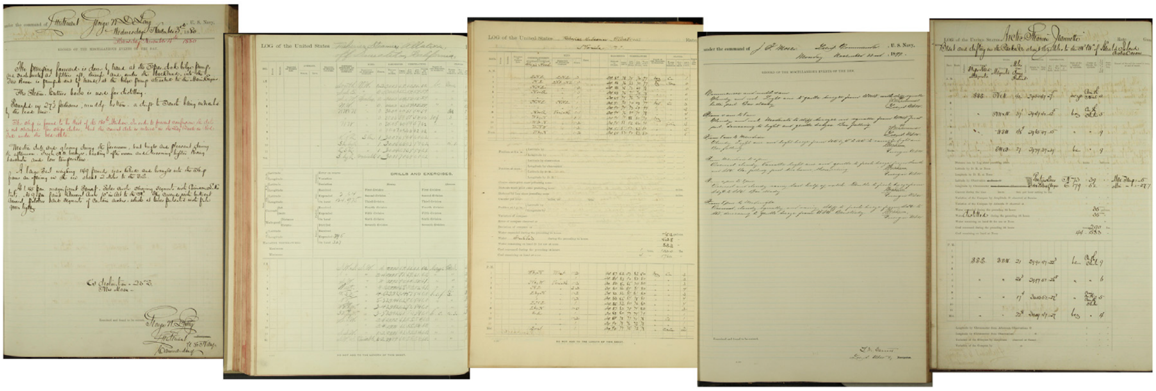


Fig. 1. Example of ship logs with annotation about the weather conditions.

Transkribus[1] [7] and t-truth[2]) were used in [1,2] where the annotation process was manually performed while others used a synthetic method as [3]. Crowdsourcing techniques [8] have been proposed in the past but this idea is very limited when the text is difficult to read, and in such case, experts paleographers have to carry out this process. Another alternative explored in [9] was to align text and image lines. None of these papers provide information about the decisions taken for annotating the GT except [2].

This paper introduces a database for table image recognition and preliminary recognition results for some of the problems that we listed previously. The tabular images are accompanied with images that have running text that complements and enriches the information that is written in the tables. It is worth noting that the handwritten information in both parts is filled with plenty of acronyms and jargon that makes the text recognition step more difficult.

The HisClima database is a set of table images related to logbooks that have a pre-printed template. They have handwritten daily weather conditions for every hour of the day: sea temperature, atmospheric pressure, wind direction, etc. It is filled with abbreviations, numeric information, quotes, and other artifacts distributed into cells. Each table page image has associated another page image that has a description of the day in plain text. Figure 1 shows an example of the images.

The whole collection has been used for collecting climate information on a crowdsourcing initiative of the OldWeather project.[3] Millions of entries were collected manually but only specific parts were collected given the effort required to extract all the information.

A preliminary version of HisClima was introduced in [10] together with some experiments on information extraction. Next, line detection experiments were reported in [11]. Finally, new information extraction experiments were reported in [12]. This preliminary research was important to decide the necessary ground-truth for the tabular documents contained in HisClima.

This paper extends [10–12] by considering a larger number of images from two ships, *Jeannette* and *Albatross*. In [10–12], only images from the *Jeannette* ship were considered. These images have the same template. This paper includes images from the *Albatross* ship, that has seven different templates although they are quite similar. Moreover, different writing styles are present in this set and the tables are more densely filled up than the tables of *Jeannette*.

The rest of the paper is structured as follows: the HisClima database and the process followed to create it are described in Section 2, the technology used to create the database and to define a baseline is presented in Section 3, the used evaluation metrics are introduced in Section 4, the obtained baseline results are analyzed in Section 5, conclusions are drawn in Section 6, and finally, some indications to ease the reproducibility of the baseline experiments can be found in Section 7.

2. The HisClima dataset: Corpus creation

HisClima is a freely available database[4] composed of printed and handwritten documents related to logbooks of ships with climate information from the OldWeather collection, and it was compiled during the *HisClima* project.[5]

The HisClima database is composed by two types of pages. Firstly, there are table images containing handwritten daily weather conditions for every hour of the day. This information is included in a pre-printed template: sea temperature, atmospheric pressure, wind direction, etc. It is plenty of abbreviations, numeric information, quotes, and other artifacts distributed into cells. Secondly, each table page image has associated another descriptive text page image, that describes the same day in plain text (see Fig. 1).

A preliminary version of this database was presented in [10]. That preliminary version was composed by 208 table page images of the logbook of a ship called *Jeannette*. This ship navigated in the Arctic ocean between July of 1880 and February of 1881. In the new version presented in this paper, the database has been extended with the corresponding 208 descriptive text page images of the *Jeannette* logbook. In addition, in order to introduce variability to the database, 6 different table structures, coming from different logbooks of a ship called *Albatross*, have been included. For each table structure, 12 pages have been selected from the full collection, making a total of 72 additional table pages. Finally, the corresponding 72 descriptive pages associated to these tables have been also included to the database.

The different pages of the database have been manually annotated with two types of information: first, the layout of every page has been marked, including information of blocks, columns, rows and lines; second, all the marked text lines have been transcribed by an expert in paleography.

The annotation process was manually carried out with the Transkribus platform [7], that includes tools to segment images and add the corresponding transcriptions. All the GT has been

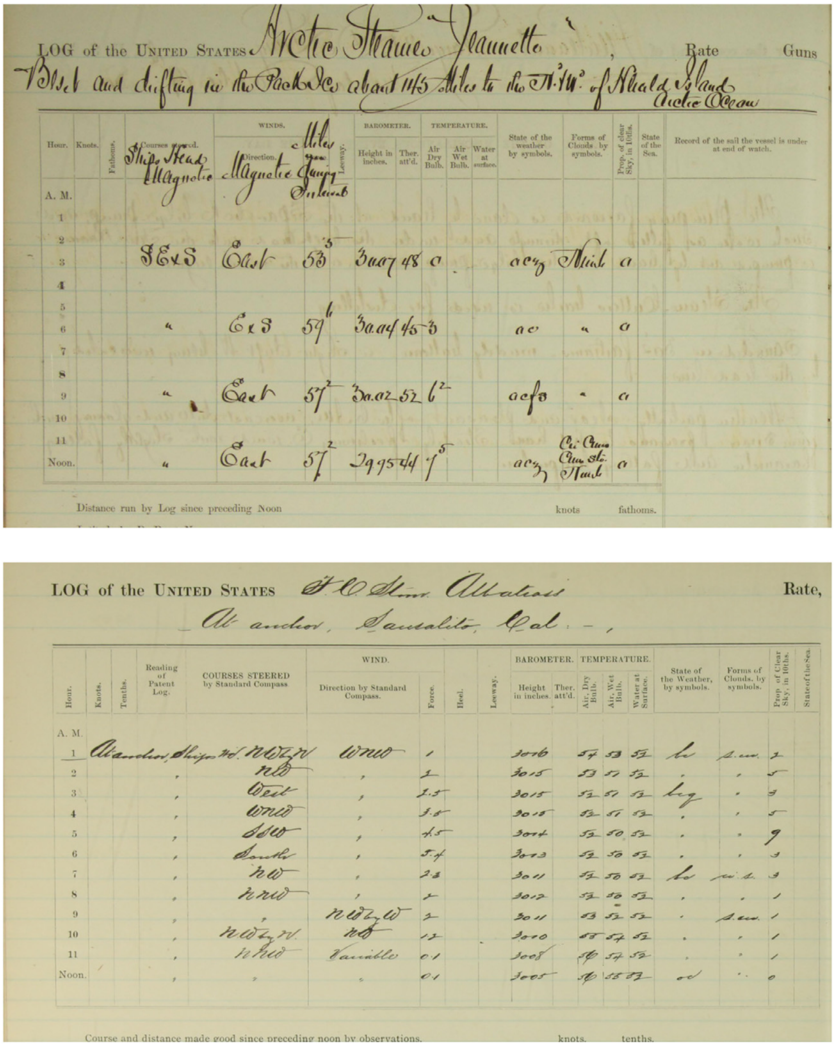
[1] https://readcoop.eu/transkribus.

[2] http://www.iapr-tc11.org/dataset/TableGT_UW3_UNLV/t-truth.tar.gz.

[3] https://www.oldweather.org/.

[4] https://zenodo.org/record/4106886.

[5] https://www.prhlt.upv.es/hisclima-dos-siglos-de-datos-climaticos/.



LOG of the UNITED STATES Arctic Steamer Jeannette , Rate Guns

Beset and drifting in the Pack Ice about 145 Miles to the N.W. of Herald Island Arctic Ocean

Hour.	Knots.	Fathoms.	~~Courses steered.~~ Ship's Head Magnetic	WINDS. Direction. Magnetic	~~Force~~ Miles Jumping Interval	Leeway.	BAROMETER. Height in inches.	Ther. att'd.	TEMPERATURE. Air Dry Bulb.	Air Wet Bulb.	Water at surface.	State of the weather by symbols.	Forms of Clouds by symbols.	Prop. of clear Sky, in 10ths.	State of the Sea.	Record of the sail the vessel is under at end of watch.
A. M.																
1																
2																
3			SE×S	East	53^5		30.07	48	0			acq	Nimb	a		
4																
5																
6			"	E×S	59^6		30.04	45	3			ac	"	a		
7																
8																
9			"	East	57^2		30.02	52	6^2			acfs	"	a		
10																
11																
Noon.			"	East	57^2		29.95	44	7^5			acq	Cir Cum Cum Str. Nimb	a		

Distance run by Log since preceding Noon knots fathoms.

LOG of the UNITED STATES F.C. Stmr. Albatross Rate,

At anchor, Sausalito, Cal. –,

Hour.	Knots.	Tenths.	Reading of Patent Log.	COURSES STEERED by Standard Compass	WIND. Direction by Standard Compass.	Force.	Heel.	Leeway.	BAROMETER. Height in inches.	Ther. att'd.	TEMPERATURE. Air, Dry Bulb.	Air, Wet Bulb.	Water at Surface.	State of the Weather, by symbols.	Forms of Clouds, by symbols.	Prop of Clear Sky, in 10ths.	State of the Sea.
A. M.																	
1				At anchor, Ship's Hd. NWbyN	WNW	1			3016		54	53	52	bc	s.cu.	2	
2			,	NW	,	2			3015		53	51	52	.	,	5	
3			,	West	,	1.5			3015		52	51	52	bcq	.	3	
4			,	WNW	,	3.5			3015		52	51	52	.	,	5	
5			,	SSW	,	4.5			3014		52	50	52	.	.	9	
6			,	South	,	5.4			3013		52	50	52	.	,	3	
7			,	NW	,	2.3			3011		52	50	52	bc	cu.s.	3	
8			,	NNW	,	2			3012		52	50	52	.	.	1	
9			,	,	NWbyW	2			3011		53	52	52	.	s.cu.	1	
10			,	NWbyN.	NW	1.2			3010		55	54	52	.	,	1	
11			,	NNW	Variable	0.1			3008		56	54	52	.	.	1	
Noon.			,	,	,	0.1			3005		56	55	52	oc	..	0	

Course and distance made good since preceding noon by observations. knots. tenths.

Fig. 2. Detailed view of two tables to be transcribed, one for the *Jeannette* ship (top), and another for the *Albatross* ship (bottom).

Table 1
Basic statistics of the HisClima database.

Number of:	Type of page		Total
	Table	Descriptive	
Pages	280	280	560
Lines	60 974	7 560	68 534
Running words	111 122	57 927	169 049
R.W. printed	63 754	7 231	70 985
R.W. handwritten	47 368	50 696	98 064
Lexicon	2 693	6 469	8 705
Lexicon printed	292	451	694
Lexicon handwritten	2 401	6 018	8 011

recorded in PAGE (Page Analysis and Ground-truth Elements) format [13]. It is a XML-base format that allows recording information about both layout structure and page content. Table 1 shows the basics statistics of the database. Note that we have distinguished between printed and handwritten text since this is important for evaluation purposes.

The next subsections describe the annotation process in more detail.

2.1. Table pages

Figure 1 shows how these pages are divided into two parts. The upper part of every table page contains a table for registering the information in the AM period of each day. The bottom part registers the information referred to the PM period of each day. Figure 2 shows a detailed view of a table page from the *Jeannette* ship (top) and a detailed view of a table page from the *Albatross* ship (bottom). Note that the table headers are only included on top of images and might contain vertical text.

These table images involve some challenges related to layout analysis, text recognition and information extraction.

One of the main handwritten text recognition difficulties of these pages corresponds to the fact that, in some cells, the information is replaced with quotation marks (''). This occurs when the data expected in a cell is the same data included in the previous filled row of the same column. These quotation marks are responsible for several difficulties. From the layout analysis point of view, these marks are very short and, therefore, difficult to detect. From the information extraction point of view, these marks refer to the relevant information in the previous row. Therefore, to extract the relevant information, it is necessary to identify both the quotation marks and their meaning.

Other difficulties of these types of pages are: text written between cells, cells with the information divided into several lines, crossed out column names, numbers written as superscripts, or the same information written in different ways. In addition, the lack of context between cells makes more difficult the recognition process. Note that every cell uses to contain a unique word or number and this information is not related with the cells of the same row.

Page layout GT

The first step in the ground-truth (GT) creation process corresponds with the annotation of the layout information. The position of the different regions, columns, rows and lines was marked for every page.

As explained in [10], given the specific characteristics of these types of pages, it was not clear if the lines at tables should be annotated and extracted at row or cell level. After some preliminary experiments [10], the best results were obtained labeling the lines at the cell level. Therefore, the layout GT included on every page was:

- The bounding box of the different regions: text and table regions. The number of text and table regions depends on the table structure page. For example, in the table pages of *Jeannette*, there are three text and two table regions.
- For every text region, the position of the different lines.
- For every table region: the bounding box of the rows and columns, the position of every line in the table with the corresponding cell information, and, lately, the reading order of the lines into every cell.

Transcription GT

The different text lines marked in the previous step were transcribed by an expert in paleography. This transcription contains both the diplomatic transcript and the relevant information. That is, for those lines whose content is a quotation mark (’’), its meaning is also included in the transcription between brackets. The database contains many numbers with decimals. Therefore, the decimal point was annotated in a different way to distinguish it from the usual point.

In addition, given that these pages include both printed and handwritten text, every word in the transcription was labeled with a tag indicating if the word is printed or handwritten. This information is very relevant to evaluate transcription results for both types of text independently. Basic statistics about the HisClima database are shown in the Table 1.

2.2. Descriptive text pages

These pages are composed of two parts. The upper part contains information about the commander and the date. It is important to remark that the date is not included in the tabular pages. The bottom part provides the description of the corresponding day. Fig. 3 shows examples of these types of pages.

As in the table pages, every descriptive page has been labeled with both GT, the layout and its corresponding transcription.

Page layout GT

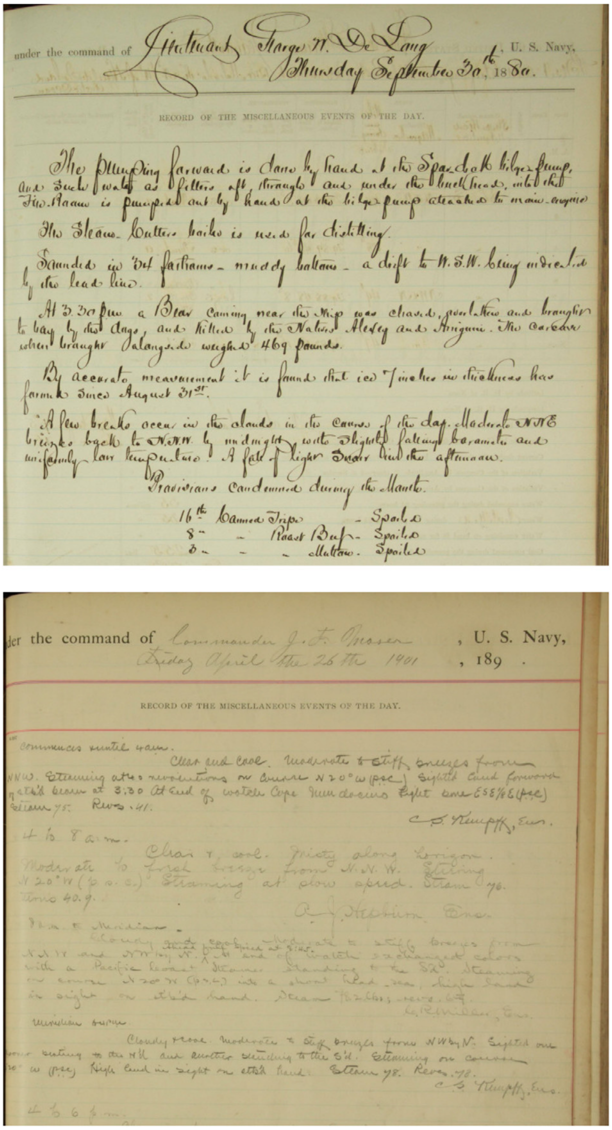
For all pages, the text lines coordinates and their corresponding reading order was manually annotated by a human expert.

Transcription GT

Every page was transcribed line by line by an expert paleographer. As in the table pages, given that these pages also include both printed and handwritten text, every word in the transcription was labeled with the corresponding text type tag. Basic statistics about this part of the database are shown in the Table 1.

2.3. Partitions

The full database is composed of 280 table pages and their associated 280 descriptive pages. In order to carry out experiments, the table pages were divided into three shuffled partitions: one for training the statistical models, another for validation, and the last one for testing the trained systems. The associated descriptive pages were also divided in the same partitions. In this way, there are 195 pages for training, 22 for validation, and 63 for testing. Tables 2 and 3 show the statistics of the partitions of the table and descriptive pages, respectively.



under the command of Lieutenant George W. De Long, U. S. Navy,
Thursday September 30th, 1880.

RECORD OF THE MISCELLANEOUS EVENTS OF THE DAY.

The pumping forward is done by hand at the Spar-deck bilge pump, and such water as filters aft, through and under the bulkhead, into the Fire Room is pumped out by hand at the bilge pump attached to main-engine

The Steam Cutter boiler is used for distilling.

Sounded in 34 fathoms - muddy bottom - a drift to N.S.W. being indicated by the lead line.

At 3.30 pm a Bear coming near the ship was chased, overtaken and brought to bay by the dogs, and killed by the Natives Alexey and Aniguin. The carcass when brought alongside weighed 469 pounds.

By accurate measurement it is found that ice 7 inches in thickness has formed since August 31st.

A few breaks occur in the clouds in the course of the day. Moderate N.N.E breezes back to N.N.W. by midnight with slightly falling barometer and uniformly low temperature. A fall of light snow in the afternoon.

Provisions condemned during the Month.

16th Canned Tripe - Spoiled
8 " - Roast Beef - Spoiled
3 " - - Mutton. Spoiled

der the command of Commander J. F. Moser , U. S. Navy,
Friday April the 26th 1901 , 189 .

RECORD OF THE MISCELLANEOUS EVENTS OF THE DAY.

commences until 4 am.

Clear and cool. Moderate to stiff breezes from NNW. Steaming at 80 revolutions on course N 20° W (psc) Sighted land forward of stb'd beam at 3:30 At end of watch Cape Mendocino light bore ESE 1/8 E (psc) Steam 75. Revs. 41.

C. S. Kempff, Ens.

4 to 8 a.m.

Clear & cool. Misty along horizon. Moderate to fresh breeze from N.N.W. Steering N 20° W (p.s.c.) Steaming at slow speed. Steam 76. Revs 40.9.

A. J. Hepburn Ens.

8 a.m. to Meridian.

Cloudy and cool. Moderate to stiff breezes from N.N.W and NW by N. Ahead full speed at 8:45. At end of watch exchanged colors with a Pacific Coast Steamer standing to the SE. Steaming on course N 20° W (p.s.c.) into a short head sea, high land in sight on stb'd hand. Steam [illegible], revs. [illegible].

G. R. Miller, Ens.

Meridian to 4 pm

Cloudy & cool. Moderate to stiff breezes from NW by N. Sighted one steamer standing to the N'd and another standing to the S'd. Steaming on course N 20° W (psc) High land in sight on stb'd hand. Steam 78. Revs. 78.

C. S. Kempff, Ens.

4 to 6 p.m.

Fig. 3. Detail of two descriptive text pages, one for the *Jeannette* ship (top), and another for the *Albatross* ship (bottom).

Table 2
Basic statistics of the partitions for the table pages.

Number of:	Train	Validation	Test
Pages	195	22	63
Lines	43 535	4 828	12 611
Running words	78 962	8 508	23 652
R.W. printed	44 842	4 873	14 039
R.W. handwritten	34 120	3 635	9 613
Lexicon	2 395	854	1 345
Lexicon printed	290	280	284
Lexicon handwritten	2 105	574	1 061

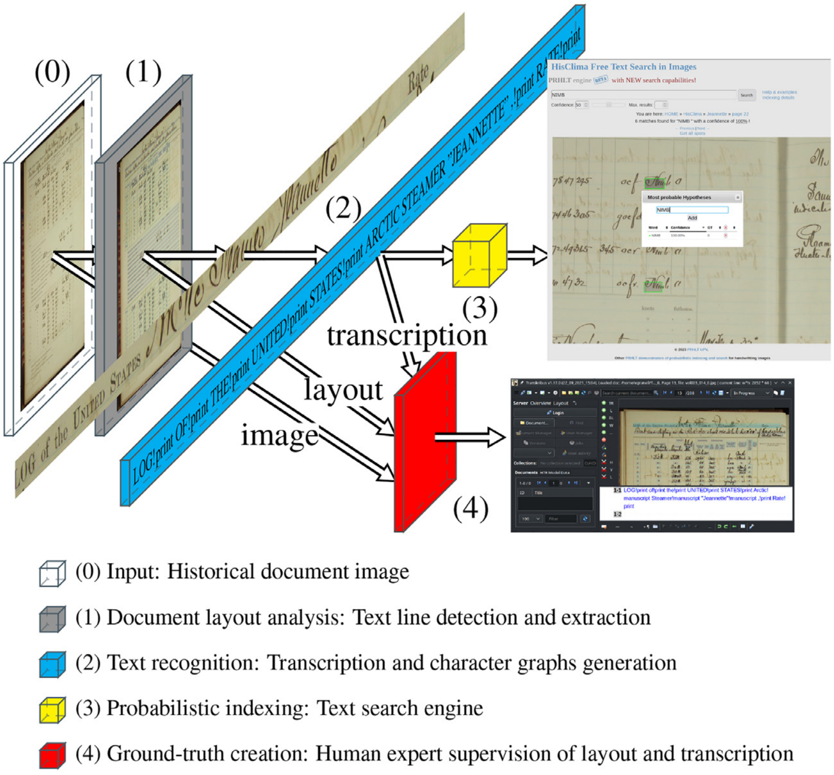

(0)
(1)
(2)
(3)
(4)
transcription
layout
image
LOG of the UNITED STATES Arctic Steamer Jeannette Rate
LOG!print OF!print THE!print UNITED!print STATES!print ARCTIC STEAMER "JEANNETTE" !print RATE!print
HisClima Free Text Search in Images
PRHLT engine with NEW search capabilities!
Most probable Hypotheses
(0) Input: Historical document image
(1) Document layout analysis: Text line detection and extraction
(2) Text recognition: Transcription and character graphs generation
(3) Probabilistic indexing: Text search engine
(4) Ground-truth creation: Human expert supervision of layout and transcription

Fig. 4. Diagram of the proposed method of ground-truth creation and information search in digitalised historical documents. The colored cubes and the associated numbers indicate the name of the techniques used in each block of the diagram.

Table 3
Basic statistics of the partitions for the descriptive pages.

Number of:	Train	Validation	Test
Pages	280	280	560
Lines	5 272	618	1 670
Running words	40 699	4 430	12 798
R.W. printed	5 007	597	1 627
R.W. handwritten	35 692	3 833	11 171
Lexicon	5 224	1 301	2 325
Lexicon printed	347	105	112
Lexicon handwritten	4 877	1 196	2 213

3. Technologies

This section describes the different technologies that have been employed for the creation of HisClima ground-truth and the baseline experiments. Figure 4 presents an overall diagram of the proposed method.

3.1. Document layout analysis: Line detection and extraction

Detecting and extracting the main document components (text areas and lines) of page images is a major stage in reading systems. Text line segmentation directly influences the accuracy of text recognition systems, whose basic inputs are usually full text line images.

As already mentioned in Section 2, the full HisClima collection comprises two types of pages: table and descriptive text pages. The table pages have a strict layout composed of zones of printed text lines and tables with very short lines (the lines must be detected at cell level). Therefore, the layout analysis method must be invariant to these differences and irregularities in script and writing style. The descriptive pages do not have such a regular structure, but many challenges appear for text line segmentation in this type of pages. For example, the variability in the skew angle between different text lines including the converse skew angles case in the same text line, the presence of overlapping words, and the adjacent text lines touching makes text line segmentation a difficult problem.

In this paper, we used the document layout analysis tool called P2PaLA [14] for text line detection and segmentation of the two types of page images that compose the database. This tool is able to detect the page lines by means of an artificial neural network and a basic baseline detector [15]. P2PaLA generates a PAGE [13] file with the detected lines information. It follows the popularized approach in recent years based on U-Net like neural networks, where all the main document components (text areas, records and text lines) are detected simultaneously [15–17]. In this approach, page segmentation is considered a pixel labeling problem, i.e., each pixel must be classified as one of the predefined classes. P2PaLA finally applies a connected component procedure in order to detect the text lines from the pixel map probabilities.

Once the text lines are detected, the extraction of them has been carried out using the method available in [18], that extracts lines from an image based on the indications of a PAGE [13] file, correcting the curvature of lines by straighting the lines based on the baseline segments.

3.2. Text recognition

Text recognition in historical documents is usually performed on the obtained text lines images without any prior segmentation into word neither characters [19,20].

The state-of-the-art in text recognition technology is based on neural network models. Generally, the systems are composed by a stack of several *convolutional* layers followed by several recurrent layers with Bidirectional Long Short Term Memory neurons units (BLSTM) [19,21–23]. Finally, an estimate of the probabilities of each character in the training alphabet is computed by a softmax output layer. All these architectures are generally called *Convolutional-Recurrent Neural Networks* (CRNN) [24].

In this paper, the text recognition *Laia Toolkit* [25] has been used for training and decoding with CRNN models. These models are used for the optical modeling of the characters with the following details: four convolutional layers with filters composed of different (16, 32, 48 and 64) maps of characteristics with kernels of 3×3 pixels and *LeakyReLU* as activation function, and four recurrent layers composed of 256 bidirectional long-short term memory units.

Most specifically, in order to minimize the so called *Connectionist Temporal Classification* (CTC) cost function [26], the RMSProp method [27] on mini-batches of size 10 is used to train the CRNN. In addition, in order to reduce the training over-fitting problem, a 0.2 dropout rate is used [28] on the convolutional layers, and 0.0003 as learning rate. Finally, it is interesting to note, that the validation set has been used to configure all the hyperparameters and to stop the training process.

The output of the trained CRNN for every text line image is a sequence of character posterior probability vectors. Although CRNN are able to capture lexical and linguistic context, classical language models (LM) can improve the results provided by the CRNNs. Character *N*-grams are the most traditional approach for this purpose.

The *N*-gram explicitly models contextual regularities and can be applied to the CRNN output in different ways [22]. In this paper we represent *N*-grams as a stochastic finite-state transducer, where the edge probabilities are computed by combining the estimated *N*-gram probabilities with the CRNN output [22]. Finally, the Viterbi decoding algorithm is used to obtain the optimal transcription hypothesis. LM were estimated as 10-gram of characters directly from the line transcripts included in the training and validation partitions using the SRILM toolkit [29]. Character graphs were computed with Kaldi [30]. These character graphs were later transformed into word graphs, as described in [25].

3.3. Probabilistic indexing

Basic Probabilistic Indexing (PrIx) ideas and developments have been presented in previous papers, such as [31–35], and more comprehensively in [36,37].

The idea behind this technology consists in detecting and storing any element in an image which is likely enough to be interpreted as a “word”, along with its *relevance probability* (RP) and location in the image. These text elements are referred to as *“pseudo-word spots”*. As explained in [37], the image-region word RP is denoted as $P(R=1 \mid X=x, V=v)$, or abbreviated as $P(R \mid x, v)$, where v is a (pseudo-)word, and x is an image region (e.g. a line region). Even though RPs are computed without taking into account where v may appear in x, the precise position(s) of v within x is (are) easily obtained as a by-product.

As discussed in [36] and [37], $P(R \mid x, v)$ can be computed using the same optical and language models as in a segmentation-free handwritten text recognition system. So the RP can be computed using state-of-the-art optical and language models and processing steps similar to those employed in text recognition (c.f. Sec 3.2), even though no actual text transcripts are explicitly produced.

4. Evaluation metrics

The quality of the obtained results was assessed by using the following measures: Precision, Recall and F1 for document layout analysis; character (CER) and word (WER) error rates, and Tag error rate (for the type of text, printed or handwritten) for text recognition; mean average precision (mAP), average precision (AP) and R-precision (RP) for probabilistic indexing. More information on these quality measures can be found in [15,25,34].

Table 4
Test-set results for document layout analysis experiments.

Measure	Table images	Descriptive images
Precision	0.89	0.90
Recall	0.87	0.95
F1	0.88	0.92

Table 5
Test set results (in %) for text recognition experiments. All 95% confidence intervals are smaller than ±1.0%, or within ±0.35% for the smaller (CER) values.

Text type	Overall			Printed		Handwritten	
	CER	WER	Tag ER	CER	WER	CER	WER
Tables	3.6	7.6	1.9	1.5	6.7	11.1	14.3
Descriptive	11.3	30.5	9.6	8.5	17.5	12.2	32.4

5. Experimental baseline

Baseline experiments were performed for the three main technologies described in Section 3. For these experiments, the text was processed to work only with capital letters, the punctuation marks were separated from the words and the nomenclatures of the winds and cloud formations were normalized to facilitate indexing. For instance, *NWxN* and *NWXN* were normalized to *NWBN*, and *CIRCUM, SIR.CUM, CIR,CUM, CIR.CIM* and *CIR.SUM* to *CIR.CUM*. Additional experimental details can be found in the dedicated repository [6].

5.1. Document layout analysis results

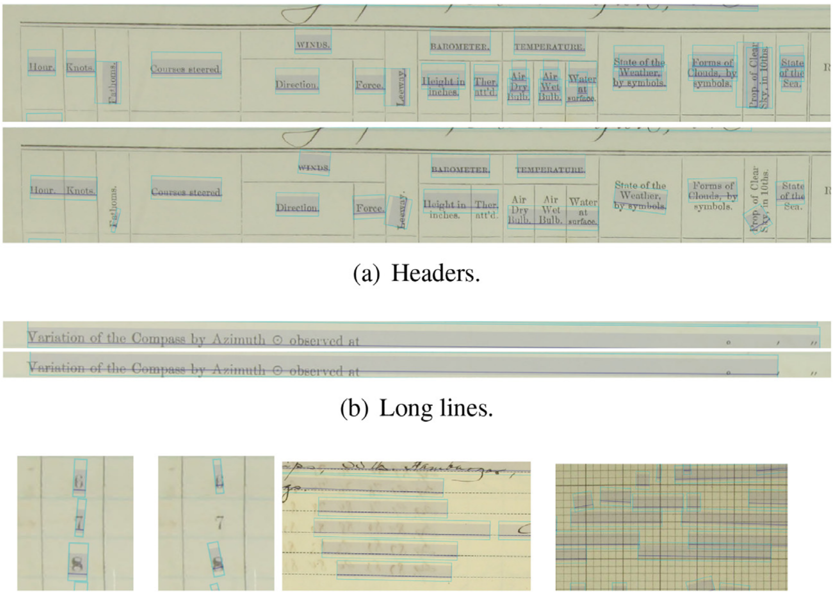
Table 4 presents the obtained baseline test results for document layout analysis. Evaluation of the baseline detection is measured with Precision, Recall and F1 measures, as explained in [38]. As can be observed, recognizing the layout on pages made up of tables is a much more complex problem than recognizing the layout on pages made up of descriptive text running on lines. Specifically, we obtained an F1 value of 0.88 and 0.92 for the test partition for the pages composed of tables and descriptive text, respectively.

The obtained results are quite good. However, if we look at the most common errors (see Fig. 5), we find that in the set of pages composed of tables, the most common errors are found in the headers, at the end of some long lines and in some row numbers. It is important to remark that, as explained in Section 2, the corpus includes seven different table structures, and every table type has its own header structure. In addition, for every different table structure from the *Albatross* logbooks only 12 pages are available. Therefore, we think that more training would be necessary to improve the results in the headers. It is specially relevant in the case of vertical lines, given that there are very few vertical lines in each image. On the other hand, the most common errors on pages made up of descriptive text are broken lines and false positives due to bleed-through and grids.

5.2. Text recognition results

Table 5 summarizes the overall test results for text recognition. It is important to remark that, for optical modeling, printed and handwritten characters were modeled without distinction. It

[6] https://github.com/PRHLT/hisclima_baseline.



(a) Headers.

(b) Long lines.

(c) Row numbers.

(d) False positives.

Fig. 5. Most common document layout analysis errors. Ground-truth on top or left and the obtained layout on bottom or right, except for (d).

Table 6
Sizes of the test query sets selected.

Measure	Table	Descriptive
Keywords	1 181	1 703
Total events	23 865	20 880
Relevant events	22 409	11 861

Table 7
Test-set results for probabilistic indexing experiments.

Measure	Table	Descriptive
mAP	0.55	0.53
AP	0.85	0.69
RP	0.90	0.73

can be seen that, in general, better results have been obtained on pages made up of tables compared to those made up of descriptive text. These differences may be due to the vocabulary size (see Table 1) and to the balance between the two types of text, printed and handwritten. This fact also explains the higher Tag ER on the descriptive pages.

Given that the HisClima database comprises printed and handwritten text, it is interesting to analyze the behavior of text recognition for each type of text. On the one hand, we have the pages made up of tables in which the two types of text are moderately balanced and where the lines are short. On the other hand, we have those composed of handwritten descriptive text in long lines and with little printed text.

In Table 5 it can be also observed that handwritten text recognition is a more difficult task than the recognition of printed text. Moreover, for handwritten text, although similar CER results were obtained for both types of pages, the WER is considerably worse for descriptive pages.

In addition, looking at the overall results (Table 5) again, the influence of printed text on the overall results can be noticed.

We believe that better results could be obtained in the handwritten text of the tables if more advanced LM were used, for example, based on conditional random fields, that are able to take into account not only horizontal relationships but vertical relationships. Note that the wind direction in a given hour uses not to be very different from the wind direction in previous hour, and this information is in the previous row.

5.3. Probabilistic indexing results

Table 6 presents the Probabilistic Indexing (PrIx) experiment setup. It shows the number of keywords, the total number of events, that is, the number of pairs composed of a page image and a keyword, and the number of relevant events (the keyword is actually written in the image).

The adopted criterion for selecting keywords for each type of page is to take all the words that appear in their corresponding test partitions. Given the text preprocessing, the number of keywords is reduced compared to the original corpus lexicon (see Tables 2 and 3).

The obtained test results for the PrIx baseline experiment are summarized in Table 7. Although mAP is practically the same for both types of pages, AP and RP are better in the pages composed of tables, partly because they have a smaller vocabulary and better results in text recognition.

6. Conclusions and future work

This paper introduces a new database called HisClima, which is composed of historical document images formed by tables and running text. The process followed to create it is explained with the aim of helping researchers who need to create new databases.

This database can be used to research and improve state-of-the-art in technologies related to historical documentation, such as, document layout analysis, text recognition and probabilistic indexing. For these technologies, some results that can be considered as a baseline were obtained. Although these results are good, there is still room for improvement.

In addition to the database, the scripts used during the baseline experimentation are freely available to facilitate replication by other researchers.

As future work, this database could also be used to research about other related technologies not covered in this paper, such as information extraction.

7. Reproducibility

To ease the reproduction of the baseline experiments carried out in this paper, the dataset is freely available on Zenodo [39], and the used scripts in the following repository: https://github.com/PRHLT/hisclima_baseline.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data is available in zenodo: https://zenodo.org/record/4106887.

Acknowledgments

Work partially supported by: the Universitat Politècnica de València under grant FPI-I/SP20190010 (Spain), the SimancasSearch project as Grant PID2020-116813RB-I00a funded by MCIN/AEI/ 10.13039/501100011033, grant DIN2021-011820 funded by MCIN/AEI/ 10.13039/501100011033, the valgrAI - Valencian Graduate School and Research Network of Artificial Intelligence and the Generalitat Valenciana, and co-funded by the European Union.

References

[1] J. Ziomek, S.E. Middleton, GloSAT historical measurement table dataset: enhanced table structure recognition annotation for downstream historical data rescue, in: 6th International Workshop on Historical Document Imaging and Processing (HIP), 2021, pp. 49–54.

[2] E. Lang, J. Puigcerver, A.H. Toselli, E. Vidal, Probabilistic indexing and search for information extraction on handwritten german parish records, in: ICFHR, 2018, pp. 44–49.

[3] S.R. Qasim, H. Mahmood, F. Shafait, Rethinking table recognition using graph neural networks, in: Proceedings of the International Conference on Document Analysis and Recognition (ICDAR), 2019, pp. 142–147.

[4] L. Gao, Y. Huang, H. Dejean, J.L. Meunier, Q. Yan, Y. Fang, F. Kleber, E. Lang, ICDAR 2019 competition on table detection and recognition (cTDar), in: Proceedings of the International Conference on Document Analysis and Recognition, ICDAR, 2019, pp. 1510–1515.

[5] A. Prasad, H. Dejean, J.L. Meunier, Versatile layout understanding via conjugate graph, 2019.

[6] T. Constum, N. Kempf, T. Paquet, P. Tranouez, C. Chatelain, S. Brée, F. Merveille, Recognition and information extraction in historical handwritten tables: toward understanding early 20th century paris census, in: Document Analysis Systems, Springer International Publishing, Cham, 2022, pp. 143–157.

[7] P. Kahle, S. Colutto, G. Hackl, G. Mhlberger, Transkribus - a service platform for transcription, recognition and retrieval of historical documents Vol. 04 (2017) 19–24, doi:10.1109/ICDAR.2017.307.

[8] J.A. Sánchez, E. Vidal, V. Bosch, Effective crowdsourcing in the EDT project with probabilistic indexes, in: Document Analysis Systems, 2022, pp. 291–305.

[9] Z. Ziran, X. Pic, S. Undri Innocenti, D. Mugnai, S. Marinai, Text alignment in early printed books combining deep learning and dynamic programming, Pattern Recognit. Lett. 133 (2020) 109–115.

[10] V. Romero, J.A. Sánchez, The HisClima database: historical weather logs for automatic transcription and information extraction, in: 2020 25th International Conference on Pattern Recognition (ICPR), 2021, pp. 10141–10148.

[11] E. Granell, L. Quirós, V. Romero, J.A. Sánchez, Reducing the human effort in text line segmentation for historical documents, in: J. Lladós, D. Lopresti, S. Uchida (Eds.), International Conference on Document Analysis and Recognition (ICDAR), 2021, pp. 523–537.

[12] J. Andrés, J.R. Prieto, E. Granell, V. Romero, J.A. Sánchez, E. Vidal, Information extraction from handwritten tables in historical documents, in: S. Uchida, E. Barney, V. Eglin (Eds.), Document Analysis Systems, 2022, pp. 184–198.

[13] S. Pletschacher, A. Antonacopoulos, The page (page analysis and ground-truth elements) format framework, in: 2010 20th International Conference on Pattern Recognition, 2010, pp. 257–260, doi:10.1109/ICPR.2010.72.

[14] L. Quirós, P2pala: page to page layout analysis tookit, 2017, https://github.com/lquirosd/P2PaLA, gitHub repository.

[15] L. Quirós, Multi-task handwritten document layout analysis, CoRR (2018). http://arxiv.org/abs/1806.08852

[16] S.A. Oliveira, B. Seguin, F. Kaplan, dhSegment: a generic deep-learning approach for document segmentation, ICFHR, 2018.

[17] K. Chen, M. Seuret, J. Hennebert, R. Ingold, Convolutional neural networks for page segmentation of historical document images, in: ICDAR, 2017, pp. 965–970.

[18] M. Villarreal, Curvature correction extractor, 2021. https://github.com/PRHLT/CurvatureCorrectionExtractor.

[19] A. Graves, M. Liwicki, S. Fernández, R. Bertolami, H. Bunke, J. Schmidhuber, A novel connectionist system for unconstrained handwriting recognition, IEEE Trans. PAMI 31 (5) (2009) 855–868.

[20] P. Voigtlaender, P. Doetsch, H. Ney, Handwriting recognition with large multidimensional long short-term memory recurrent neural networks, in: 2016 ICFHR, 2016, pp. 228–233, doi:10.1109/ICFHR.2016.0052.

[21] B. Moysset, T. Bluche, M. Knibbe, M.F. Benzeghiba, R. Messina, J. Louradour, C. Kermorvant, The A2iA multi-lingual text recognition system at the second Maurdor evaluation, in: ICFHR, 2014, pp. 297–302.

[22] T. Bluche, Deep neural networks for large vocabulary handwritten text recognition, Université Paris Sud - Paris XI, 2015 Theses. https://tel.archives-ouvertes.fr/tel-01249405

[23] T. Bluche, H. Ney, C. Kermorvant, The LIMSI/A2iA handwriting recognition systems for the HTRts contest, in: ICDAR, 2015, pp. 448–452.

[24] B. Shi, X. Bai, C. Yao, An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition, CoRR (2015). http://arxiv.org/abs/1507.05717

[25] J. Puigcerver, Are multidimensional recurrent layers really necessary for handwritten text recognition? in: ICDAR, 01, 2017, pp. 67–72.

[26] A. Graves, S. Fernández, F. Gomez, J. Schmidhuber, Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks, in: ICML, 2006, pp. 369–376.

[27] T. Tieleman, G. Hinton, Lecture 6.5-rmsprop: divide the gradient by a running average of its recent magnitude, COURSERA Neural Netw. Mach. Learn. 4 (2) (2012).

[28] V. Pham, C. Kermorvant, J. Louradour, Dropout improves recurrent neural networks for handwriting recognition, CoRR (2013). http://arxiv.org/abs/1312.4569

[29] A. Stolcke, SRILM-an extensible language modeling toolkit, in: Proceedings of the 3rd Annual Conference of the International Speech Communication Association (Interspeech), 2002, pp. 901–904.

[30] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, K. Vesely, The Kaldi speech recognition toolkit, ASRU, 2011.

[31] J. Puigcerver, E. Vidal, A.H. Toselli, Probabilistic interpretation and improvements to the HMM-filler for handwritten keyword spotting, in: 13th Int. Conf. on Document Analysis and Recognition (ICDAR), 2015, p. 731735.

[32] T. Bluche, S. Hamel, C. Kermorvant, J. Puigcerver, D. Stutzmann, A.H. Toselli, E. Vidal, Preparatory KWS experiments for large-scale indexing of a vast medieval manuscript collection in the HIMANIS project, in: 2017 14th IAPR Int. Conf. on Document Analysis and Recognition (ICDAR), Vol. 01, 2017, pp. 311–316, doi:10.1109/ICDAR.2017.59.

[33] E. Lang, J. Puigcerver, A.H. Toselli, E. Vidal, Probabilistic indexing and search for information extraction on handwritten german parish records, in: 2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR), 2018, pp. 44–49, doi:10.1109/ICFHR-2018.2018.00017.

[34] A.H. Toselli, E. Vidal, J. Puigcerver, E. Noya-García, Probabilistic multi-word spotting in handwritten text images, Pattern Anal. Appl. 22 (1) (2019) 23–32.

[35] A. Toselli, V. Romero, E. Vidal, J. Sánchez, Making two vast historical manuscript collections searchable and extracting meaningful textual features through large-scale probabilistic indexing, in: 2019 15th IAPR Int. Conf. on Document Analysis and Recognition (ICDAR), 2019.

[36] E. Vidal, A.H. Toselli, J. Puigcerver, A Probabilistic Framework for Lexicon-Based Keyword Spotting in Handwritten Text Images, Tech. rep., Univ. Polit. de València, 2017.

[37] J. Puigcerver, A probabilistic formulation of keyword spotting, Univ. Politcnica de Valncia, 2018 Ph.D. thesis.

[38] T. Grning, R. Labahn, M. Diem, F. Kleber, S. Fiel, Read-bad: a new dataset and evaluation scheme for baseline detection in archival documents, 2017, https://doi.org/10.48550/ARXIV.1705.03311.

[39] PRHLT, Hisclima dataset, 2022, https://doi.org/10.5281/zenodo.4106886.