INTERNATIONAL CONFERENCE ON PATTERN RECOGNITION 2020

The HisClima database: historical weather logs for automatic transcription and information extraction

Verónica Romero[1] and Joan Andreu Sánchez[2]

[1] *Universitat de València, Spain*

[2] *PRHLT Research Center, Universitat Politècnica de València, Spain*

January-2021

Index

1 Introduction ▷ 0

2 The HisClima Dataset ▷ 1

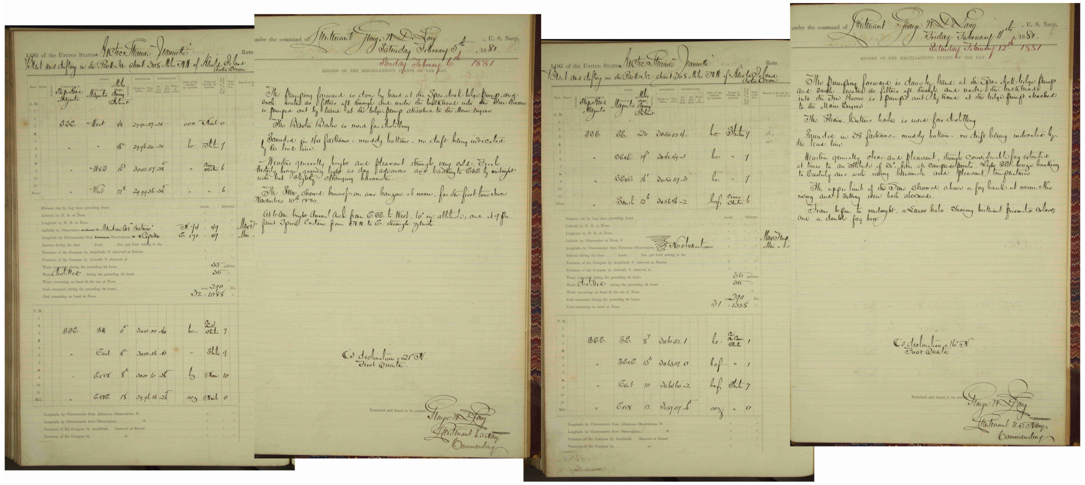
3 Technologies ▷ 3

4 Experimental Framework ▷ 6

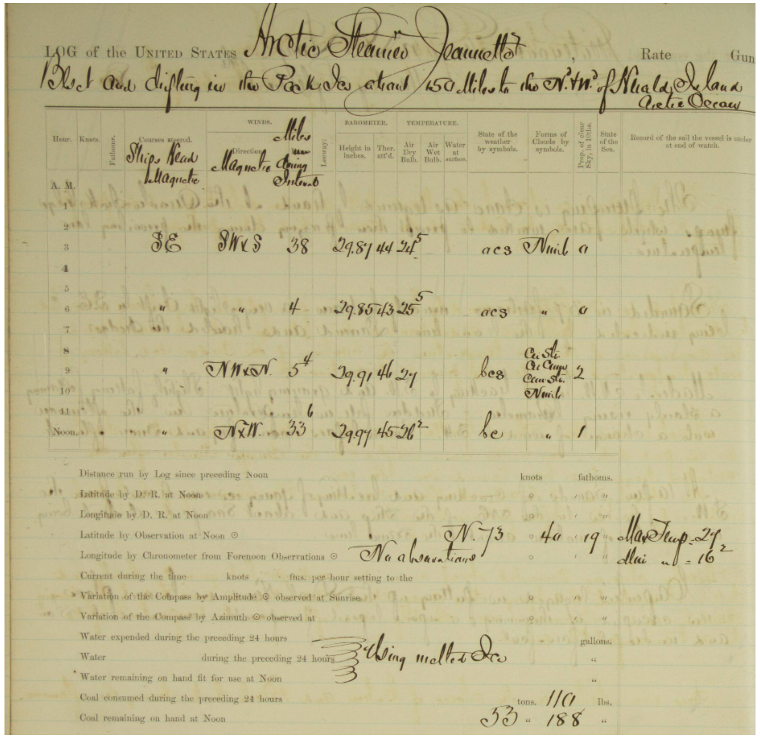
5 Conclusions ▷ 7

Introduction

- The state of the art in document digialization has increased the interest in preserving and providing access to handwriting historical documents.
- A particularly interesting and important type of historical documents are the ship log records, that were written daily when ships were sailing.
- Objective: extract the relevant semantic information contained in these documents about the climate of several centuries ago.
- This paper presents a new database of this type of documents and baseline results for state-of-the-art line segmentation, recognition and information extraction approaches.



The HisClima Dataset



LOG of the United States Arctic Steamer Jeannette, Rate Gun

Beset and drifting in the Pack Ice about 150 Miles to the N.&W. of Herald Island Arctic Ocean

Hour.	Knots.	Fathoms.	Courses steered. Ships Head Magnetic	WINDS. Direction Magnetic	WINDS. Miles during Interval	Leeway.	BAROMETER. Height in inches.	BAROMETER. Ther. att'd.	TEMPERATURE. Air Dry Bulb.	TEMPERATURE. Air Wet Bulb.	TEMPERATURE. 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	SW×S	38		29.87	44	24^5			acs	Nimb	0		
4																
5																
6			"	"	4		29.85	43	25^5			acs	"	0		
7																
8																
9			"	NW×N	5^4		29.91	46	27			bcs	Cu St. Cu Cum Cum St. Nimb	2		
10																
11																
Noon.			"	N×W	33^6		29.97	45	26^2			bc	"	1		

Distance run by Log since preceding Noon knots fathoms.

Latitude by D. R. at Noon ° ′ ″

Longitude by D. R. at Noon ° ′ ″

Latitude by Observation at Noon ⊙ N. 73 ° 40 ′ 19 ″ Max Temp = 27

Longitude by Chronometer from Forenoon Observations ⊙ No observations ° ′ ″ Min " = 16^2

Current during the time knots fms. per hour setting to the

Variation of the Compass by Amplitude ⊙ observed at Sunrise ° ′ ″

Variation of the Compass by Azimuth ⊙ observed at ° ′ ″

Water expended during the preceding 24 hours gallons.

Water during the preceding 24 hours Using melted Ice "

Water remaining on hand fit for use at Noon "

Coal consumed during the preceding 24 hours tons. 1 10 lbs.

Coal remaining on hand at Noon 33 " 188 "

- It is a freely available handwritten text database compiled from the log book of a XIV century ship.
- It is composed of 208 table pages.
- Te upper part of the page registers the information in the AM period of each day and the bottom registers the PM.
- Different challenges related with layout analysis, handwritten recognition and information extraction.

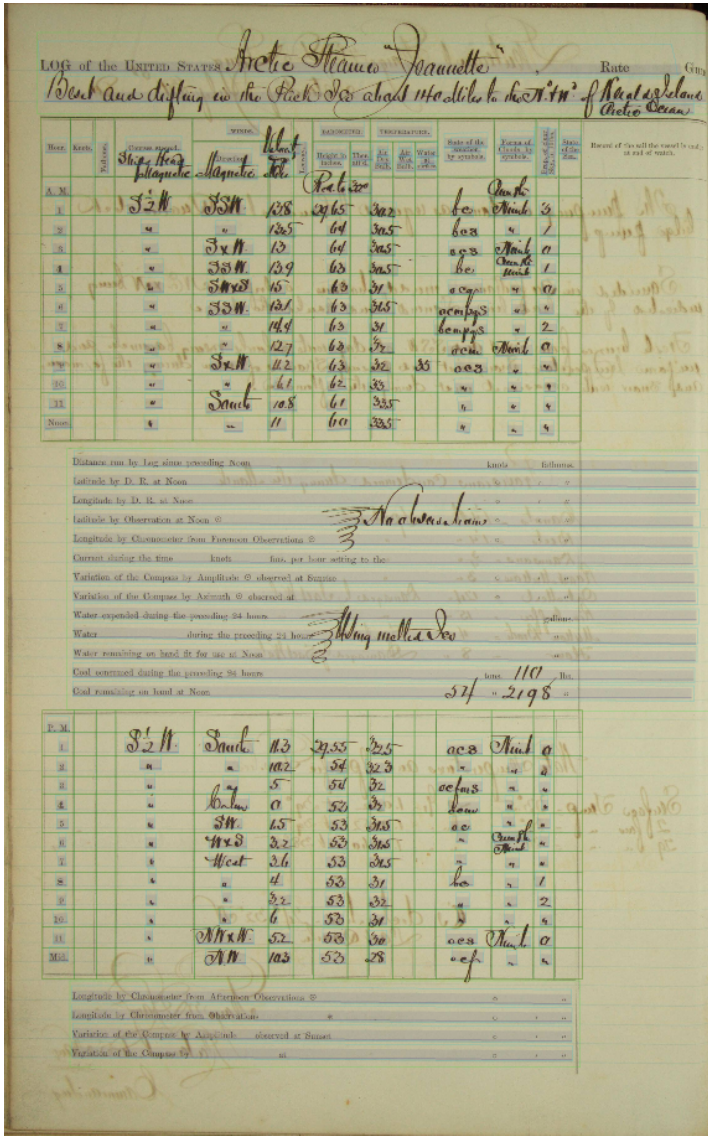
The HisClima Dataset: Annotations and partitions

- It has been endowed with two different types of annotations: layout analysis of each page to indicate blocks, columns, rows and lines; the transcription including relevant information.

- The 208 table pages were divided into three shuffled partitions aimed at performing experiments.

Number of:	Train	Validation	Test	Total
Pages	143	15	50	208
Lines	23 617	2 284	7 838	33 739
Running words	46 599	4 604	15 611	66 814
Lexicon	1 287	491	924	1 483
Character set size	76	76	76	76
Rel. Information	10 917	1 021	3 533	15 471

Technologies: Layout Analysis



LOG of the UNITED STATES Arctic Steamer "Jeannette", Rate Gun

Beset and drifting in the Pack Ice about 140 Miles to the N.t W. of Herald Island Arctic Ocean

Hour.	Knots.	Fathoms.	Courses steered. Ship's Head Magnetic	Winds. Direction Magnetic	Velocity Force	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.	Proportion of clear sky in 10ths.	State of the Sea.
A. M.							Red. to 32°								
1			S½W	SSW	13.8		29.65		30.2			bc	Cumulo Nimb	3	
2			"	"	13.5		64		30.5			bcs	"	1	
3			"	SxW	13		64		30.5			ocs	Nimb	0	
4			"	SSW	13.9		63		30.5			bc	Cumulo Nimb	1	
5			"	SWxS	15		63		31			ocq	"	0	
6			"	SSW	13.1		63		31.5			ocmpqs	"	"	
7			"	"	14.4		63		31			bcmpqs	"	2	
8			"	"	12.7		63		32			ocm	Nimb	0	
9			"	SxW	11.2		63		32		35	ocs	"	"	
10			"	"	6.1		62		33			"	"	"	
11			"	South	10.8		61		33.5			"	"	"	
Noon.			"	"	11		60		33.5			"	"	"	

Record of the sail the vessel is under at end of watch.

Distance run by Log since preceding Noon knots fathoms.

Latitude by D. R. at Noon

Longitude by D. R. at Noon

Latitude by Observation at Noon ⊙ No observations

Longitude by Chronometer from Forenoon Observations ⊙

Current during the time knots fms. per hour setting to the

Variation of the Compass by Amplitude ⊙ observed at Sunrise

Variation of the Compass by Azimuth ⊙ observed at

Water expended during the preceding 24 hours gallons.

Water during the preceding 24 hours Using melted Ice

Water remaining on hand fit for use at Noon

Coal consumed during the preceding 24 hours tons. 110 lbs.

Coal remaining on hand at Noon 51 " 2198 "

Hour.	Knots.	Fathoms.	Courses steered.	Winds. Direction	Velocity Force	Leeway.	Barometer. Height in inches.	Ther. att'd.	Air Dry Bulb.	Air Wet Bulb.	Water at surface.	State of the weather.	Forms of Clouds.	Proportion of clear sky.	State of the Sea.
P. M.															
1			S½W	South	11.3		29.55		32.5			ocs	Nimb	0	
2			"	"	10.2		54		32.3			"	"	0	
3			"	"	5		54		32			ocfms	"	"	
4			"	Calm	0		52		32			bcm	"	"	
5			"	SW	1.5		53		31.5			oc	"	"	
6			"	WxS	3.2		53		31.5			"	Cumulo Nimb	"	
7			"	West	3.6		53		31.5			"	"	"	
8			"	"	4		53		31			bc	"	1	
9			"	"	3.2		53		32			"	"	2	
10			"	"	6		53		31			"	"	"	
11			"	NWxW	5.2		53		30			ocs	Nimb	0	
Mid.			"	NW	10.3		53		28			ocf	"	"	

Longitude by Chronometer from Afternoon Observations ⊙

Longitude by Chronometer from Observations

Variation of the Compass by Amplitude observed at Sunset

Variation of the Compass by at

- The main document components are automatically detected.
- Technology based on neural networks.
- The page segmentation si considered as a pixel labelling problem.
- A *M-net* was defined as the main network and a *A-net* as adversarial one.
- The document analysis tool called P2PaLA has been used.

Technologies: Automatic Transcription Technology

- Let $\mathbf{x} = x_1 x_2 \ldots x_m$ be a handwritten text line image represented as a feature vector sequence, the HTR problem can be formulated as the problem of finding the most likely word sequence, $\widehat{\mathbf{w}} = \widehat{w}_1 \widehat{w}_2 \ldots \widehat{w}_l$:

$$\widehat{\mathbf{w}} = \arg\max_{\mathbf{w}} \Pr(\mathbf{x} \mid \mathbf{w}) \Pr(\mathbf{w})$$

- A CRNN is used for character optical modelling, $\Pr(\mathbf{x} \mid \mathbf{w})$.
- A character N-grams is used for language modelling, $\Pr(\mathbf{w})$.

Technologies: Information Extraction

- The semantic information related with every data is given by its position: columns have information about kind of data and rows about time.

- The retrieval proces is based on structured multi-word queries in the 1-best transcription of the detected lines.
  - Every column-heading word is retrieved.
  - Row-heading words are retrieved.
  - Every cells-content word is searched by the conbination of the corresponding column and row regions.

Experimental Framework: Results

- **Layout Analysis**

P	R	F1
0.91	0.72	0.8

Precision (P), recall (R) and F-measure (F1).

- **Automatic Transcription Technology**

	CER	WER
CRNN	2.8	5.2
CRNN + LM	2.7	4.4

Character/Word Error Rate (CER/WER).

- **Information Extraction**

	P	R	F1
Cell position	0.95	0.95	0.95
Line geometry	0.79	0.79	0.785

Precision (P), recall (R) and F-measure (F1)

Conclussions and Future Work

- A historic handwritten database compiled from a historical weather ship log is presented.
- Baseline results for state-of-the-art lines segmentation, recognition and information extraction approaches have been provided.
- The obtained results are encouraging.

Thanks for your attention!