

THE HISCLIMA DATABASE: HISTORICAL WEATHER LOGS FOR AUTOMATIC TRANSCRIPTION AND INFORMATION EXTRACTION

Verónica Romero
Universitat de València
veronica.romero@uv.es

Joan Andreu Sánchez
Universitat Politècnica de València
jandreu@prhlt.upv.es



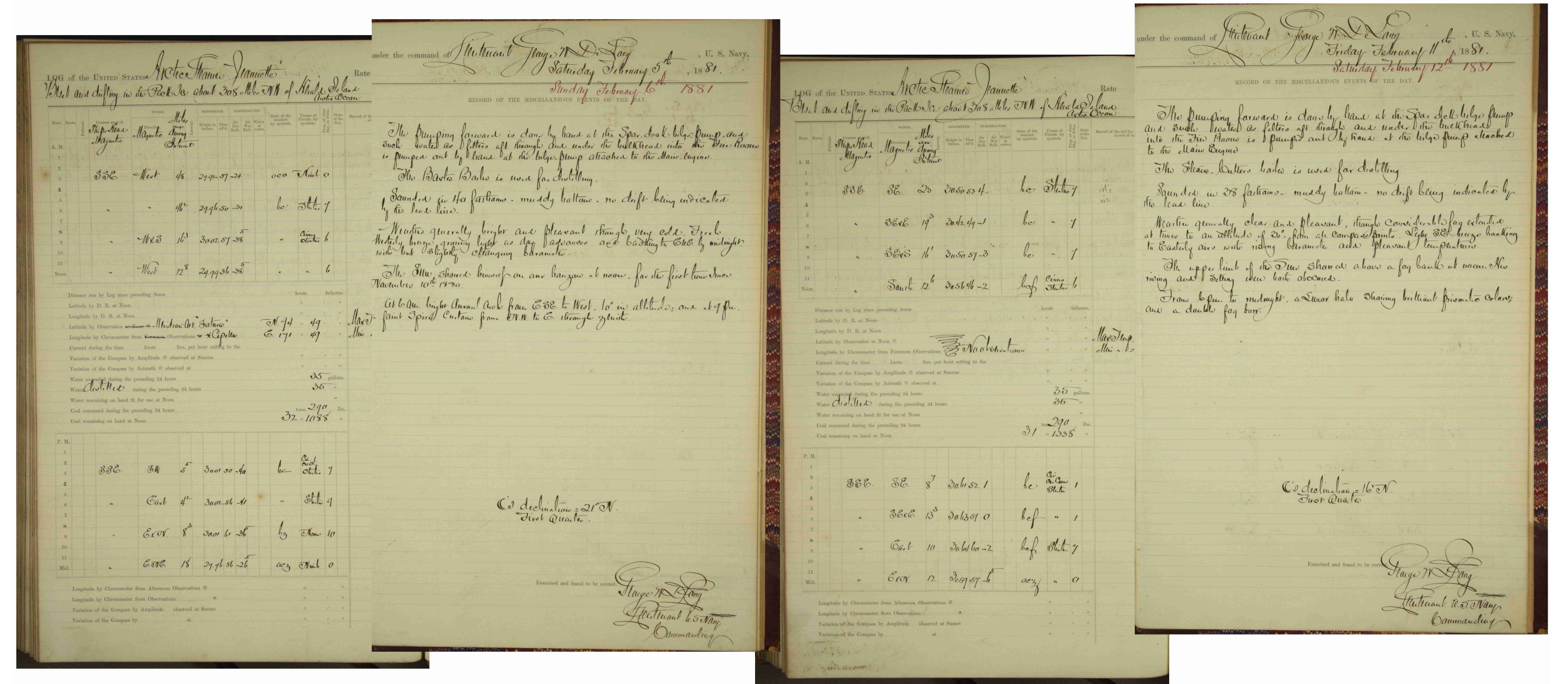
UNIVERSITAT
DE VALÈNCIA



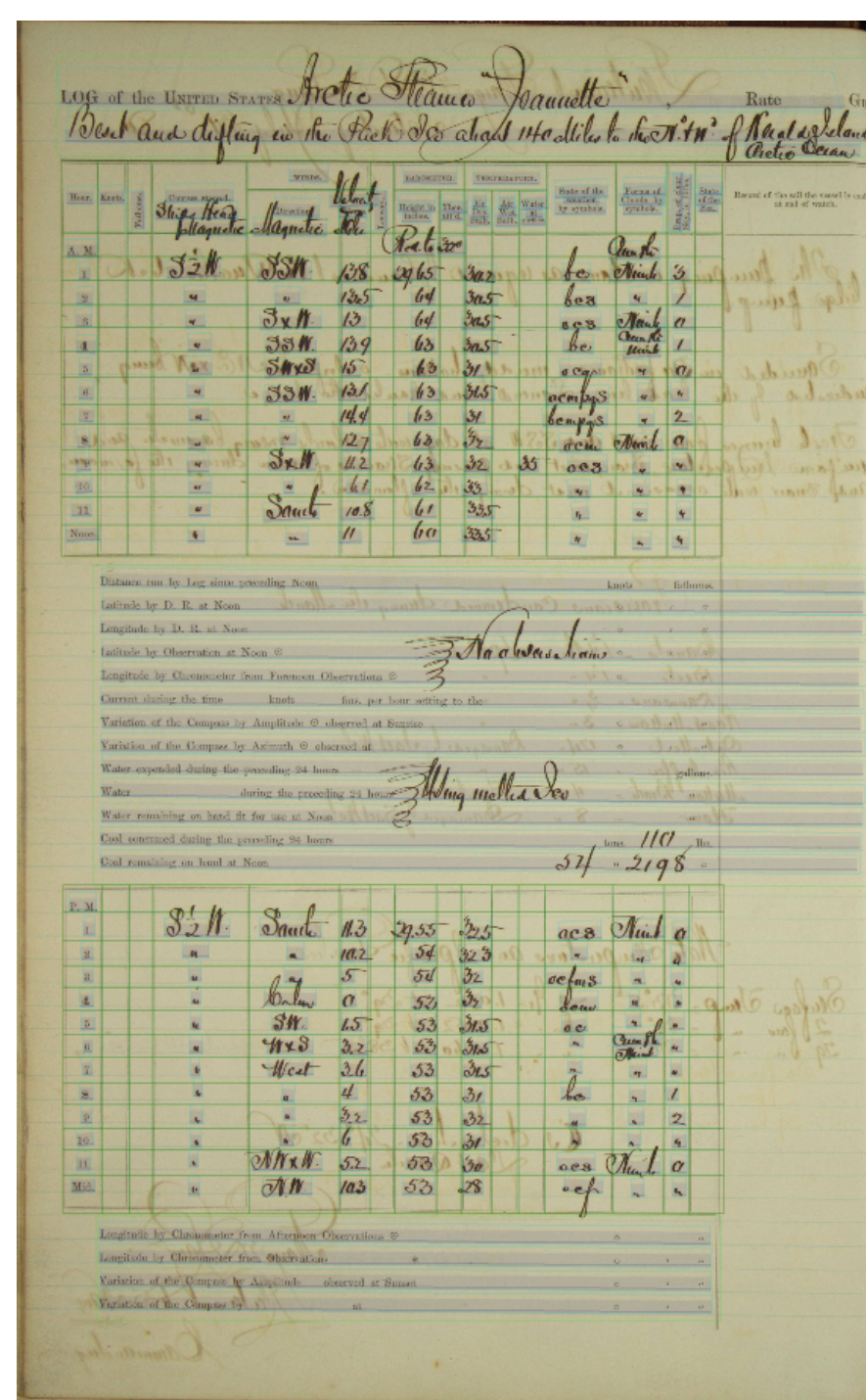
UNIVERSITAT
POLITÈCNICA
DE VALÈNCIA

INTRODUCTION

- The state of the art in document digitalization has increased the interest in preserving and providing access to handwriting historical documents.
- A particularly interesting and important type of historical documents are the ship log records, that were written daily when ships were sailing.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



- 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.
- The 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.

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

- The main document components are automatically detected.
- Technology based on neural networks.
- The page segmentation is considered as a pixel labelling problem.
- A *M-net* was defined as the main network and a *A-net* as the adversarial one.

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 process is based on structured multi-word queries in the 1-best transcription of the detected lines.

Automatic Transcription Technology

- Let x be a handwritten line image, the HTR problem can be formulated as the problem of finding the most likely word sequence, \hat{w} :

$$\hat{w} = \arg \max_w \Pr(w | x)$$

- A CRNN is used for character optical modelling.
- A character N -grams is used for language modelling.

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)

SUMMARY 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.

ACKNOWLEDGEMENTS

Work partially supported by the BBVA Foundation through the 2018–2019 Digital Humanities research grant “HisClima” and by the Generalitat Valenciana under the EU-FEDER grant IDIFEDER/2018/025.