

ICFHR 2014

ICFHR2014 Competition on Handwritten Text Recognition on tranScriptorium Datasets (HTRtS)

Joan A. Sánchez, Verónica Romero, Alejandro H. Toselli and Enrique Vidal

Pattern Recognition and Human Language Technology Research Center
Departamento de Sistemas Informáticos y Computación
Universitat Politècnica de València

{jandreu,vromero,ahector,evidal}@prhlt.upv.es



tranScriptorium



UNIVERSITAT
POLITÀCNICA
DE VALÈNCIA

September 4th, 2014

Problem description

To automatically transcribe a set of historical handwritten documents of the Bentham collection used in the *tranScriptorium* project

Challenges

- ▶ Document Image Analysis Challenges
 - ▶ different layouts and geometry
 - ▶ noisy and bleed-through images
 - ▶ marginal notes
 - ▶ faded writing
 - ▶ stamps
 - ▶ skewed images
 - ▶ slanted lines and with lines with different slope in the same page
 - ▶ inter-line sentences
- ▶ Handwritten Text Recognition Challenges
 - ▶ several hands
 - ▶ many crossed-out and hyphenated words
 - ▶ portions of the collection are written in French
 - ▶ Bentham occasionally used Latin and ancient Greek in his writings

Data set description and available data

Basic statistics

Number of:	Training	Validation	Test	Total
Pages	350	50	33	433
Lines	9,198	1,415	860	11,473

Available data:

- ▶ The original images of all the training and validation pages
- ▶ The PAGE file corresponding to each page image
- ▶ The preprocessed and extracted line images for all the lines of the training and validation sets in grayscale
- ▶ The corresponding transcripts of each of these lines

that they are either complicated with Injuries operating

on a different Subject, viz: Person; or else take a

Ground-Truth in PAGE format

CT_Test_PAGE-2013

File Mode Options Help

0 / 0 paragraph [101] k-y 387_001.jpg

Whitehall 14 Oct. 1799

116L
322

Mr King having laid before me a Letter from
Mr Long of the 27th Division, by command of
your Lordships, to be apprized of the number of
sails, which the Panglossian proposed to be erected to
St. Beatham, is intended to accommodate
examined the Acts of the 19th & 34th of his present
Majesty, relative to the binding of Convictive Houses
for confining said employings in said labour Persons

Transcription

Plain text

Unicode

Mr King having laid before me a Letter from

Competition Protocol

- ▶ Evaluation based on Word Error Rate (WER)
- ▶ The participants submitted their transcript results on the test set
- ▶ A baseline system based on hidden Markov models n-gram models was provided
- ▶ The participants were allowed to improve this baseline by changing any step in the recognition process
- ▶ Two tracks were planned in this competition:
 - ▶ Restricted track: participants were allowed to use just the data provided by the organisers for training and tuning their systems
 - ▶ Unrestricted track: participants were allowed to use any data of their choice

Participants and obtained results

- ▶ Artificial Intelligence and Image Analysis (**A2IA**)
- ▶ Computational Intelligence Technology Laboratory (**CITlab**) at University of Rostock, Institute of Mathematics, in collaboration with PLANET intelligent systems GmbH
- ▶ Spoken Language Processing Group at Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur (**LIMSI**)

Best Word Error Rate and Character Error Rate (WER/CER) obtained by the participants on each track.

	Restricted track	Unrestricted track
A2IA	-	8.6 / 2.9
CITlab	14.6 / 5.0	-
LIMSI	15.0 / 5.5	11.0 / 3.9

The winners

Restricted track: **CITlab**

Unrestricted track: **A2IA**

Congratulations !!