ICFHR 2014

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

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

Pattern Recognition and Human Language Technology Research Center Departamento de Sistemas Informticos y Computacin Universitat Politècnica de València

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







September 4th, 2014

Problem description

To automatically transcribe a set of historical handwritten documents of the Bentham collection used in the tran Scriptorium project

Challenges

- Document Image Analysis Challenges
 - different layouts and geometry
 - noisy and bleed-through images
 - marginal notes
 - fainted 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

Samples

parties white one Quai Sind shall have " here for place , if any one says yes . Le appoint Day & hour: carlier or ultrior time. of applies for, are granted by him or required , if Quari shead he int applied for the provisional Deenes, and provisional, become absolute. elit. S. The Quan Unial is performed at our titing or i more as une pay at dues titles sight is gove through at one sitting, when the whole time cupligable at that me bitting has been consumed hopere the suit has thus here 10 ale whe by Delinting Darces. est 9. ab di sceapitulation examination In service is seemable, that was not exhibited in the course of the original examination Art 10. At this Quese Unial way be re. anguest tim and a Calcunate line tengers. gind a confidence, the question of law elit. H. The inevental oreasion & on which a bearing before a Quesi Jury haspla and those on which alipeal for due is misden

f de James <u>Ample Descend Injerie</u> Bar de gaartie of an ange at the same the de ange gaartie of an ange at the same the de ange gaartie of an anter dy constraints of an gaartie at an at the anter dy constraints of an and the ange of the instance of a start to the part of the start del and the descendent of the same and the start del and the same to a start to the part of the start del and the same to a start to the part of the start del and the same to a start to the part of the start del and the same to a start to the part of the start del and the same to a start to the part of the start del and the same to a start to the part and and the same to an attart to part to a start to the of the same to an attart to part to a start to the same and the same to an attart to part to a start to the same and the same to an attart to part to a start to the same and the same to an attart to part to a start to a the same the same to a same to an attart to part to a start to a start to a start to a start to a same approximation of the same to a start of the same to a same to an attart to part to a start to a start to a start to a start to a same to an attart to a start to a same to an attart to a start to a same to an attart to a start to a start

ble grænskilge flærereger I græn intendiere uneverslegte på den binn in ti anne er ble lege svillend stepninge binn if villen trænst ense insalgedelig intendieres<mark>ter</mark> i græne intendiere noverte ægsine, hinn of his bligte er ble substate of binn ærer binnet

Whitelate a Col. 1799. My Logi Mor Then having laid before me a 16 to Long of the 27 all Prairies, by common your Settings, to be appresed of the cumber nite , shad the Parophis property to be or At Muthow , is intended to accommidtate . examined the chete of the 19 & 34" of his & Majuly, relation with hailing of Constantian for confining and complaying in hard labor consisted of transfortables and other lines . such Prostration from should be went fin as receptorles for such tronsportable Couriels word facts of the register Consider cannot from the true of their receiving section to appedients many der for their being transports With respect to this portable accurber of ath amoule. is refer who can be adopted; then by taking qualit number of the form last years of North any gold jurbally more theme county balance

all prototly non the constrate laws the

of decoreces Extennations

hashed when as likely to enfore in consequence of the other of any of the process of mature), it is a catematy and, formishes models of justification .

Main Just.

 $\begin{aligned} & \int dx & \sin 2 x \int dx & \sin 2$

any and the minder of the sector of the provided fatter inner and an and a sector of the sector and a sector of the helenging to the of mean is near the sector the **constant** of the

Course VI ... Deference la Intherity.

Al a faceta come degree of the come degree of the cont the come that with a segment of the content of the cont

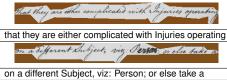
Data set description and available data

Basic statistics

Number of:	Training	Validation	Test	Total
Pages	350	50	33	
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



Ground-Truth in PAGE format

X - D GT_Tool_PAGE-2013					
File Mode Options Help					
: 0 /0	🗕 🔶 🛛 paragraph	; r101	ж <i>» к</i> -	387_001.jpg	
Transpin	the start the	7 alt, Jun 5 to approve Carrent con 1 unterest of the of the 19" to the built o confloging	before one to	flor redet (2) Shared connet	
Plain text		Unicode			
		Mr King hav	ing 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 (CITIab) at University of Rostock, Institute of Mathematics, in collaboration with PLANET intelligent systems GmbH
- Spoken Language Processing Group at Laboratoire dInformatique pour la Mécanique et les Sciences de IIngénieur (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
CITIab	14.6 / 5.0	-
LIMSI	15.0 / 5.5	11.0 / 3.9

Restricted track: CITlab Unrestricted track: A2IA

Congratulations !!