# Call for participation
# Handwritten Historical Document Recognition
# The RODRIGO database


## Summary

Lack of data for training statistical models is one of the main difficulties that researchers has to face when transcribing old handwritten documents. This competition aims to evaluate different methods to extract information from a real and complete historical document when few data is available. Thus, two different scenarios are proposed to participants on the same database: *handwritten text recognition* and *word spotting*.
The proposed document collection is the RODRIGO database, which has recently been released and is fully annotated. The RODRIGO database corresponds to a manuscript from 1545 entitled *"Historia de España del arçobispo Don Rodrigo",* and it is completely written in old Castilian (Spanish) by a single author. It is a 853-page bound volume describing chronicles from the Spanish history. Most pages only contains a single text block of nearly calligraphic handwriting on well-separated lines.
The first competition in the RODRIGO database takes place at the 12th International Conference on Frontiers in Handwriting Recognition (ICFHR2010), November 16-18, 2010, Kolkata, India.


## Description

Researchers, aiming to participate in this contest, have to send an e-mail with their names and their affiliation to ws-icfhr10@iti.upv.es or htr-icfhr10@iti.upv.es, to register in the handwriting recognition scenarios or the word spotting, respectively. After registration, free access to the data will be provided.
In all scenarios evaluation, participants will have to provide a zip file, containing: the binaries for executing the recognition, any auxiliary files required in the execution, and a README file describing the system requirements. Note that results reproducibility is one of the goals sought by the contest organizers. Usage of external data is forbidden, and further enquiries about the training process could be made.

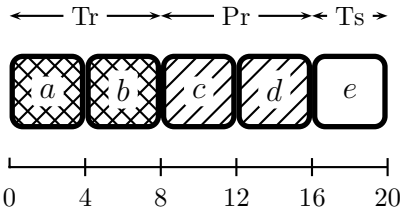


Figure 1: Block distribution in RODRIGO and its correspondent each set ($Tr, Pr$ and $Ts$). The database is split in 4 blocks ($a, b, c, d$ and $e$) of $4K$ lines, each.

As shown in Figure 1, the RODRIGO database is split into 5 consecutive blocks, of $4K$ lines each: $a$, $b$, $c$, $d$ and $e$. Blocks $a$ and $b$ correspond to the training set $Tr$; blocks $c$ and $d$ form the partially annotated set $Pr$; and block $e$ corresponds to the test block $Ts$, which will not be released to the participants.

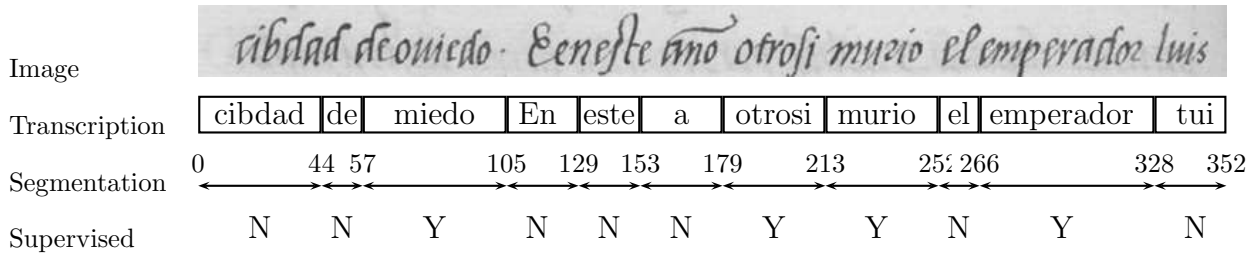


Figure 2: Example of a partially supervised text line image. The "segmentation" row indicates its word level segmentation (in pixels), and the "supervised" row indicates whether the correspondent word (on the "transcription" row) is supervised or not.

### Scenario 1: Handwriting recognition

In automatic transcription of old text documents, the lack of enough representative data is the main difficulty to train competitive systems. A possible strategy to overcome this problem, is to improve the system with partially supervised transcriptions. In this contest edition, two experiments are proposed to participants. The first one is the classical handwriting recognition experiment, in which systems are trained on a fully annotated database subset. In the second one, a new subset, which is partially annotated, is added to the previous experiment training subset. In both cases, the training data provided consists in: text line images, text line transcriptions and its segmentation (at word level) in pixels, and a flag indicating which words has been supervised. Note that, the line segmentation process was performed automatically and may contains errors. Figure 2 shows an example of the partially annotated training data. Finally, the proposed experiments performance will be evaluated in the $Ts$ set in terms of WER (word error rate).

### Scenario 2: Word spotting

A second strategy for overcoming the problem of lacking enough annotated data for training is by the means of word spotting methods. In this scenario, the same partitions, as for the handwriting recognition scenario, are provided to participants. Thus, $Tr$ and $Ts$ will be respectively used for training and testing word spotting systems, but $Pr$ will not be used. In addition, the training data provided consists of: gray-scale images, a lexicon of query words, and, for each image, a text file containing the list of words appearing in the image and their respective bounding boxes (Figure 3 shows two samples of the query word "Rei").

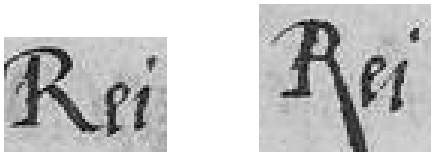

Figure 3: Samples of the query word "Rei"

Finally, the word spotting system have to return, for each processed page image, a text file listing the spotted words and their bounding boxes following the same format that the text files provided for the training data. Evaluation of the system will be also done in the $Ts$ set by the Precision and Recall measures.

## Important dates

- Release of the training the database for all scenarios: 24th march
- System submission deadline: 15th may

## Contact information

For further information or question, please contact ws-icfhr10@iti.upv.es or htr-icfhr10@iti.upv.es or consult the web http://prhlt.iti.es/w/contest-icfhr2010/.

Nicolas Serrano, Oriol Ramos Terrades and Alfons Juan
PRHLT group
Instituto Tecnológico de Informática
Univ. Politécnica de Valencia,
Cami de Vera s/n, 46022, Valencia, Spain