



STATE

Assisted Text Transcription System

David Llorens, Andrés Marzal, Vicente Palazón, Federico Prat, Juan Miguel Vilar (UJI, Spain) María José Castro, Salvador España, Joan Pastor, Francisco Zamora (UPV, Spain)

The people



Automatic? Transcription

Optical Character Recognition (OCR) systems perform poorly:

- On damaged documents.
- On ancient documents.
- On handwritten documents.
- On text with non-estandar characters or fonts.
- It is always needed to supervise and correct the automatic transcription.

aquella palabra, e den lo por fechor que puede fer justificado, segu dicho es. Mas el querelloso no le deue matar: e filo mata, deue fer dado por enemigo de los sus paristes: e pechar el omezillo. Y esto es por razóq no gelo dio el alcalde por enemigo, segun dize enel titulo de los omezillos: en la ley que comiença. Si aquelenel. &c. E todo otro hombre que matare su enemi go, maguer que lo aya desafiado con dere

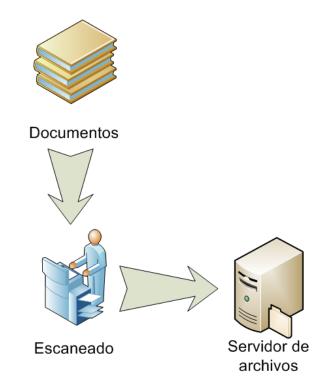
Del estylo.

fuero, dar lo an por enemigo delos tes, e que falga dela villa, e del term fobre efto es a faber, que quando manera de defendimiento fe comi demandar la muerte fegun el fuero que todo lo que dize eneste fuero de fazer, e de juzgar despues del de miento, q esso fe ha de guarda, e den e juzgar: e no puede mudar la qrella demanda en orra manera fi no forur

Fo

" He Donall finds it easies to take it out of mothers, childrens and cick people than to take on this verst industrie," He Born commented icity. " let us have a full inquiry into the cost of drugs and the pharmacantical industry." The health of children today or set much to the welfare freet scheme. It was main tained during the was. New in conditions of Tary offluence it seemed it could not be carried or.

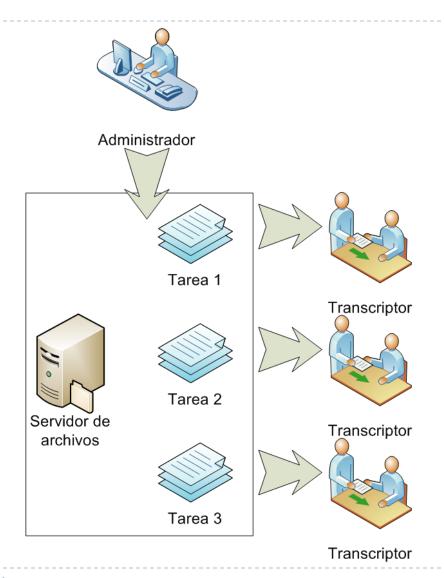
Documentary resources digitization



Documents are scanned...

 ...and the obtained images are uploaded to a file server.

Documentary resources digitization



The administrator

divides the documents in groups of pages (tasks).

Each task is assigned to a human transcriber.

His dutie is to obtain a text file from the original document.

Documentary resources digitization



- Each document is transcribed with an OCR sytem.
- The resulting text contains many errors.
- Text editing: the transcriber must manually correct the text.
- It is the most time
 consuming stage, where it is
 easy to introduce errors.
- Finally, the correct transcription is uploaded to the server.
- Now, the digital information can be exploited in many ways.

How is text editing performed?

aquella palabra, e den lo por fechor que puede fer justificado, segui dicho es. Mas el querelloso no le deue matar: e filo mata, deue fer dado por enemigo de los sus parietes: e pechar el omezillo. Y esto es por razoq no gelo dio el alcalde por enemigo, segun dize enel titulo delos omezillos: en la ley que comiença. Si aquelenel. &c. E todo otro hombre que matare su enemi go, maguer que lo ava desafiado con dere

Del estylo.

fuero, dar lo an por enemigo delos parien tes, e que falga dela villa, e del termino. E fobre efto es a faber, que quando enefta manera de defendimiento fe comiença a demandar la muerte fegun el fuero viejo: que todo lo que dize enefte fuero q fe ha de fazer, e de juzgar defpues del defafiamiento, q effo fe ha de guarda, e demadar, e juzgar: e no puede mudar la qrella, nila demanda en orra manera fi no fegun lo co

Fol.7

aquella palabra , e den lo por fechor que .puede serj u [st]i[si]cado,seg[un] dicho es.Mas el quer.elloso no le deue matar: e si lo mata, deue ser dado por enemigo de los sus pari[en]tes:e pechar el omezillo. [ss] e[st]o es por raz[on][que] no gelo dio el alcalde por enemigo, segun dize enel titulo delos omezillos: en .l.a ley que comienca. bi aquel enel. ac. .E todo otro hombre que matare su enemi go, m.uguer guelo aya desa[si]ado con dere Comparing the page with its transcription is very uncomfortable and difficult.

The human expert must

- jump continuously from the image to the text,
- find each error,
- move with the mouse or keyboard,
- and **type** the correction.

Ancient documents

Problems get worse with ancient documents

Stains, handwritten annotations on the margins, bad conditions, cracks, patches, disappearing ink in some regions, and so on:

Text extraction is difficult for OCR systems.

Special fonts or ancient graphemes:

OCR systems do not recognize them.

Non-modern orthography or syntax and archaic language: Language models of OCR systems are not suitable.

Humidity

stains e de al rey el querellofo, o el acufado que diga que quiere fer oydo e librado por el,fi efto dixere ante que el pleyto fea contestado ante los alcaldes del lugar: en tonce suyo es de oyr e de librar estas cosas fobredichas, o puede los embiar el rey fi ron fechoseftermalas ane la libra fa el fuero delos

Ancient chos. Pero fi e graphemes gun los fuere do tales fech

destos en algunos destos fechos de muer-

vegan por querella ante el regedeuen fer

embiados a que los libren fus alcaldes de

efta querella

biudas e de

las villas do tales Echos acacícieron por la querella dela amino quebratado, maguer - 1- l'heros fi querellare al rev: Archaic language

- Ley. xey. Que el que no fu injuria, o delos fuyos no deue fer recebi do a aculacion, fi no fe obliga ala pena del Talion.

onas.

Ialgunoviene diziédo al alcalde, que fulano hombre q es ay enel lugar, que hizo algu malfecho que no arañe a el. vale el reftimonio dela muger. Si fe quifiere obligar a acufarle, e obligar a la pena que el otro deue auer, fi no gelo probare, deuelo oyr el alcalde:mas en otra guila no lodeue oyr: faluo fi moftrare carta,o alguna otra cofa que hizieffe alguna fe al alcalde, porque se ouiesse de mouer contra el aculado.

Ley.xciij. Como el marido no pue de matar al vno de los adulteros e dexar alotro.

Neltitulo delos adulterios, en la primeraley dize alsi. Si muger calada fa ze adulterio, ambos sean en el poder del marido: e faga dellos lo o quifiere e de lo o an, afsi que no pueda matar el vno dellos, e dexar al otro. Sobre eftas palabras fi acae sce q fe vaya el vno, e prenden al otro: y el

Las leves

rarcs

Orthography?

Hyphenation?

ena:

prefo es vécido d'adulterio por juyzio dar gelo an los al aldes en poder del marido:v el marido deuclo tener, mas no lo deuc matar fata q aya el otro y le veça por juy-zio, vor de los mate ambos fi quifiere.

Hayacumbry

Handwritten

annotations

, xciiij. Que eleriuanos an de quifiere estos pleytos a los alcaldes do fue te de los presos sueltos sobre fianças e de fus pleytos.

Os pleytos delos que estan presos, e delos g fueron enfiados ante el alcal de,o de elcreuir la fiadur

rey que escriue concl al tos de los fobredichos a te,o de tolimiento de miembro,o de echa efercuir el eferiuano de miento de tierra, mas ay otra pena de diconclel alguazil en cafa uerrey. nero,o de al. Entoce tales pleytos, maguer

Ley. xcv. Que manera terna el calde fi el acufado no viene a responde la acufacion.

Trofi, fi alguno acufa a otro que le quemo fus cafas,o quele mato fu pa

c, o fobre otra cofa defguifada echo:y el alcalde lo fizo empla alos plazos que el fuero manne,entonce deue el alcalde fa-

per del techo de que querello, fi fue fecho: masno a de faber quien lo fizo. E fi fallare que tal fecho es fecho, entoce lo deue dar porfechor. A pt det na

Ley.xcvj. En que calos, e quando

Obre la ley que comiéça, toda muger, que es en el titulo delos teftimonios: es a faber, que pueden las mugeres fer refeebidas en teftimonio fobre las cofas, quier fea ceuiles, quier criminales q fe fazen en tal lugar, q no es razo ni aguifado de fer, y hobres colas mugeres Forrafi, fi ferescibe las m

védidas, y en la

zer las mugere

Handwritten taunderlines ma

leficios q acael uese por su dicho de mugeres en testimonio. E otrofi en la pesquisa que le faze de los yerros fechos de noche en yerra, fi ellas dan testimonio de vista, juzgen lo por prucua. E otrofi faze los fus dichos prefum pcion para poder tormentar, mas en aque llos lugares do es cierto que el fecho fue

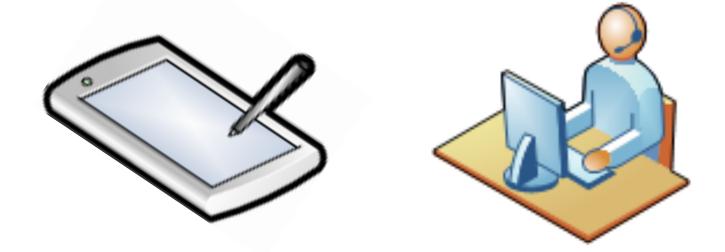
Standard OCR systems just do not work!!!!!

Our aims

- Our transcription system integrates advanced and accurate tools
 - for image processing,
 - to detect page layout,
 - for text recognition.
- But also... it assists the human expert to
 - partially automatize the text editing process, and
 - attend to usability in order to ease the interactive text editing process.
- And the recognition system includes adaptive learning: it learns from samples of each new task.
- With all of this: we aim to drastically reduce the time devoted to text editing.

Multimodal interaction

- Mouse, keyboard, and stylus.
- Voice



STATE: An Assisted Text Transcription System

dere de provarlo que dize.E si este demăder e de prouar lo que dize.È [si] e[st]e dem[an]dadorno quisiere responder,de dadorno qui[si]ere resoonder, deue deram Pararlos bienes del deudor.M parar los bienes del deudor. re fueffe el principal deudor, primerole te rue[ss]e el pr incipal deudor [si] presen nnimono

StateTA: Line Transcriber

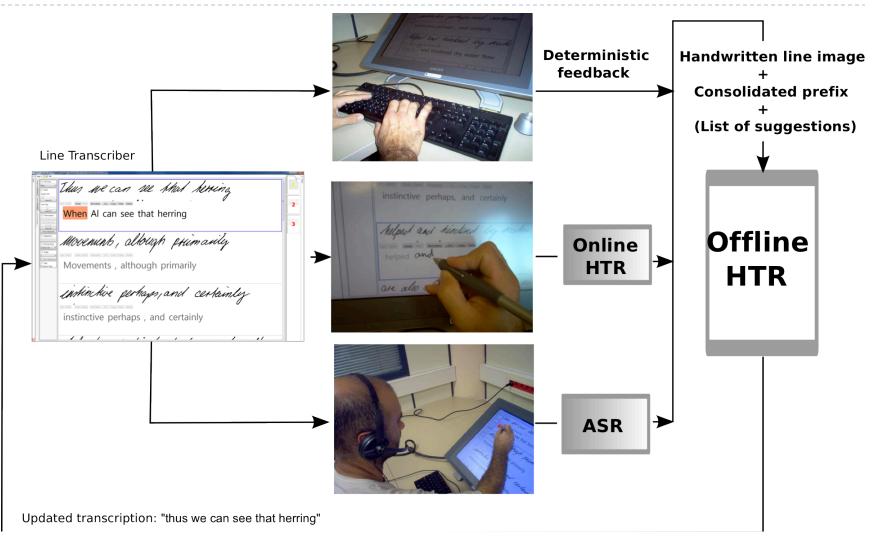
✓ It allows to
 edit text line by
 line

✓ Comfortable
 interaction

✓ Keyboard,stylus or speech

dere de provarlo que de lor e de provar la	ze.E fielte demä- que dize.E
der [si] e[st]e dadorno quiserer dadorno qui[si dadorno qui[si resoonder, deue	lere deraudor.M. Heldeudor.M.
Parar los ba parar los ba [si] presen [si] presen te fueffe el p te rue[ss	incipal deute incipal deu

User-interaction cycle with STATE via keyboard, stylus, or voice

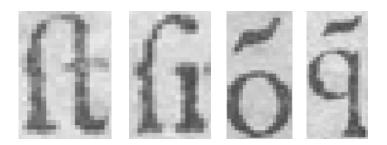


But... Does STATE increase the productivity? System evaluation

der e de provar lo que dize. E si este demădador no quisiere responder, deue desamparar los bienes del deudor. Mas si presente suesse el principal deudor, primero le deue demandar a su deudor la deuda que el deue en juyzio, o si el deudor otros bienes touiesse que cumpliessen al su deudo del demandador, saluo si los bienes que de manda fuessen senada deuda mente obligados a esta deuda.

Ley.iiijComo no puede hóbre tomar los bienes de su deudora otro que los tenga en su poder por si mismo.!

Maguer es derecho, q ha poder de to mar los bienes de fu deudor aquela de auer el deudo por el obligamien to a q fe obligo: maguer passen los bienes a otro en su poder, por qual manera quiere q passen. Pero de costúbre se guarda assien casa delrev: q fi passan los bienes a otro q Task: a Spanish book dated in 1569, printed in ancient font, with archaic lexicon and syntax, hundreds of abbreviations, and no consistent rule about word separation.



System evaluation: Experiment

Time saving when using STATE with respect to using only a text editor.

System	Time for each page
Editor only	between 27 and 35 minutes
STATE	around 12 minutes

System evaluation

Task: unrestricted handwritten task from the IAM database: collection of forms written by different people.

He said these concerned Mr. Meavers alleged anocistion with organisations blacklided by the Jovernement innediately Mr. Kennedy rushed a letter to benator Robertson anying the Federal Burean of Movestigation had reported on Mr. Meaver.

"Of course you must count about two hundred for legal charges and stamp duties, maybe less, depending on the price of the house, and whether it has been regis tered. I take it you have a mortgage lined " Hr. Dowell finds It easies to take it out of mothers, childrens and sick people than to take on this vest industrie," Hr. Bown commented icity. " Let us have a full inquiry into the cost of drugs and the pharmacentical industry." The health of children today owed much to the welfare food scheme. It was main tained during the was. Now in conditions of Tory offluence it seemed it could not be carried on.

System evaluation: Experiment

Time saving when using STATE with respect to using only a text editor.

CER (Character Error Rate)	Editor only	STATE
CER < 15	100	49
15 ≤ CER < 30	100	69
30 ≤ CER	100	95

To summarize

- STATE is a practical solution to the assisted transcription problem for difficult ancient printed or handwritten documents.
- It offers a multimodal interface: mouse, keyboard, stylus and voice.
- It can be connected to different recognition engines (at the moment: NN, HMM, HMM/ANN).
- It can be easily adapted to new documents: it learns from samples obtained from the documents to be transcribed.



