Artificial Neural Networks based techniques for Ancient Documents processing

Estudiante de Doctorado: Joan Pastor Pellicer
Director de Tesis: María José Castro Bleda

Introducción

Las bibliotecas y Archivos Históricos están digitalizando sus colecciones de textos. La mayoría están escaneando los documentos y publicando las imágenes resultantes sin sus correspondientes transcripciones. Esto limita seriamente las posibilidades de explotación de dichos documentos. Cuando una transcripción es absolutamente necesaria, ésta se realiza manualmente mediante un experto humano, lo cual resulta caro y además se trata de una tarea sujeta a errores. Disponer de herramientas como son los sistemas de reconocimiento de documentos manuscritos e impresos ayudaría a preservar, estudiar y publicar el legado cultural de estos documentos. Incluso con estas mejoras, los resultados de un sistema automático de reconocimiento no sería perfecto. Para obtener transcripciones de una determinada calidad, requiere la intervención de expertos para revisar y corregir la salida obtenida del sistema de reconocimiento.
Aunque la transcripción del texto es el principal objetivo, se necesitan de ciertas etapas previas (conocidas como preproceso) para obtener una transcripción acurada para la imagen digitalizada. Limpieza y binarización (si se precisan) son las primeras etapas del "pipeline" del sistema de reconocimiento.

Los documentos antiguos presentan un serie de degradaciones, manchas, tinta del reverso, y otros artefactos debido al paso del tiempo y las condiciones de conservación de estos. Estas degradaciones no pueden ser tratadas satisfactoriamente con técnicas tradicionales. Además, se precisan de técnicas más complejas y elaboradas, incluso, la supervisión de un experto. Una vez se tienen las imágenes limpias, se pasa a detectar las partes principales de la imagen: aquellas que contienen líneas de texto y otras partes como imágenes o figuras, decoraciones, versales, etc. También es importante, además, detectar las relaciones entre estos elementos. Estas etapas de preproceso son cruciales para el rendimiento final del sistema de transcripción, puesto que un error en fases tempranas se propagará por el resto de fases.

Por dichas razones, el trabajo de Tesis se centrará en las fases de preproceso con el fin de mejorar los sistemas y herramientas actuales de transcripción.

Objetivos

Los principales objetivos del proyecto de tesis son:
- Desarrollar modulos y metodos avanzados y precisos para el preproceso y reconocimiento de

textos manuscritos.
- Proveer de un sistema que incluya estos motores de preproceso y búsqueda y que permita la transcripción interactiva de texto manuscrito complejo y documentos antiguos.
- Desarrollar un corpus propio basado en documentos escaneados de los siglos XVII y XVIII del *Arxiu Històric de València*, y su correspondiente ground-truth para obtener la transcripción y otra información relevante.

Estado Actual

Se ha desarrollado satisfactoriamente un sistema de limpieza de imágenes antiguas deterioradas y con diversa suciedad basado en redes neuronales. El sistema ha sido evaluado con diferentes imágenes de ediciones de Digital Image Binarization Contest, obteniendo resultados mejores que otras técnicas basadas en redes neuronales. A su vez, esté método se está utilizando con muy buenos resultados para la limpieza de documentos antiguos como fase previa a la tarea de detección de líneas y layout. Se está desarrollando un método novedoso de detección de líneas basado en el cálculo de forma supervisada de puntos de interés en imágenes de textos manuscritos.

Artificial Neural Networks based techniques for Ancient Documents processing.

Joan Pastor-Pellicer

Departamento de Sistemas Informáticos y Computación
Universidad Politécnica de Valencia
Spain

Valencia, June 2014

1. What I am working on?
   - Image Cleaning and enhacement
   - Interest points and layout analysis

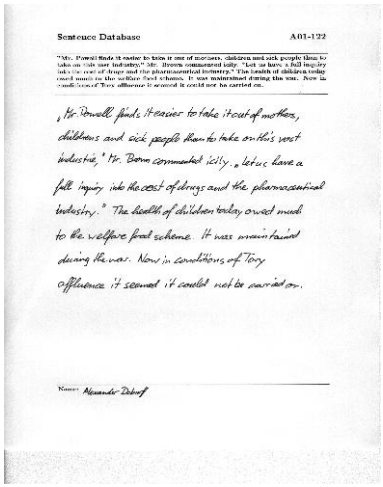
My Thesis

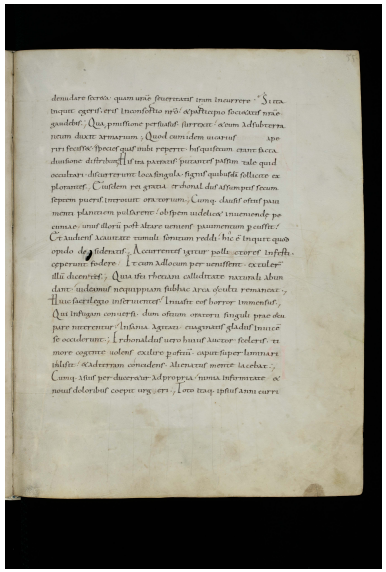
Artificial Neural Networks based techniques for Ancient Documents processing.

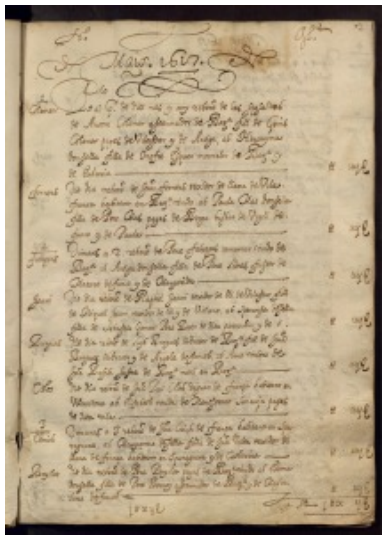
- Image cleaning and enhancement: use of a neural filter to estimate the probability of ink for each pixel, given a window of the original image centered at the pixel to be cleaned.
- Detection of interesting points: extract local maxima and local minima from the vertical contour of strokes in order to classify them by means of connectionist classifiers.
  - Text line Normalization.
  - Text Line Extraction and layout analysis.
  - Word spotting and Query by Example on handwritten documents.

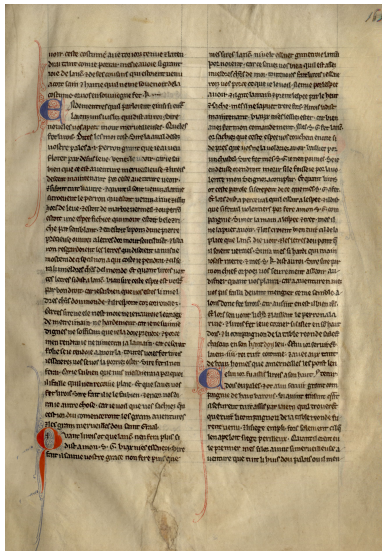
Handwritting ancient documents

- Handwritten Ancient documents are dirtier and more degraded than modern handwritting text.









- Cleaning methods have an important impact on posterior stages.

Image cleaning and enhacement using Neural Networks

- Work fine with similar data.
- Ink value can be used as a probability (or score) shown in a grayscale value.
- Smoothing on the letter borders (Ambiguity).

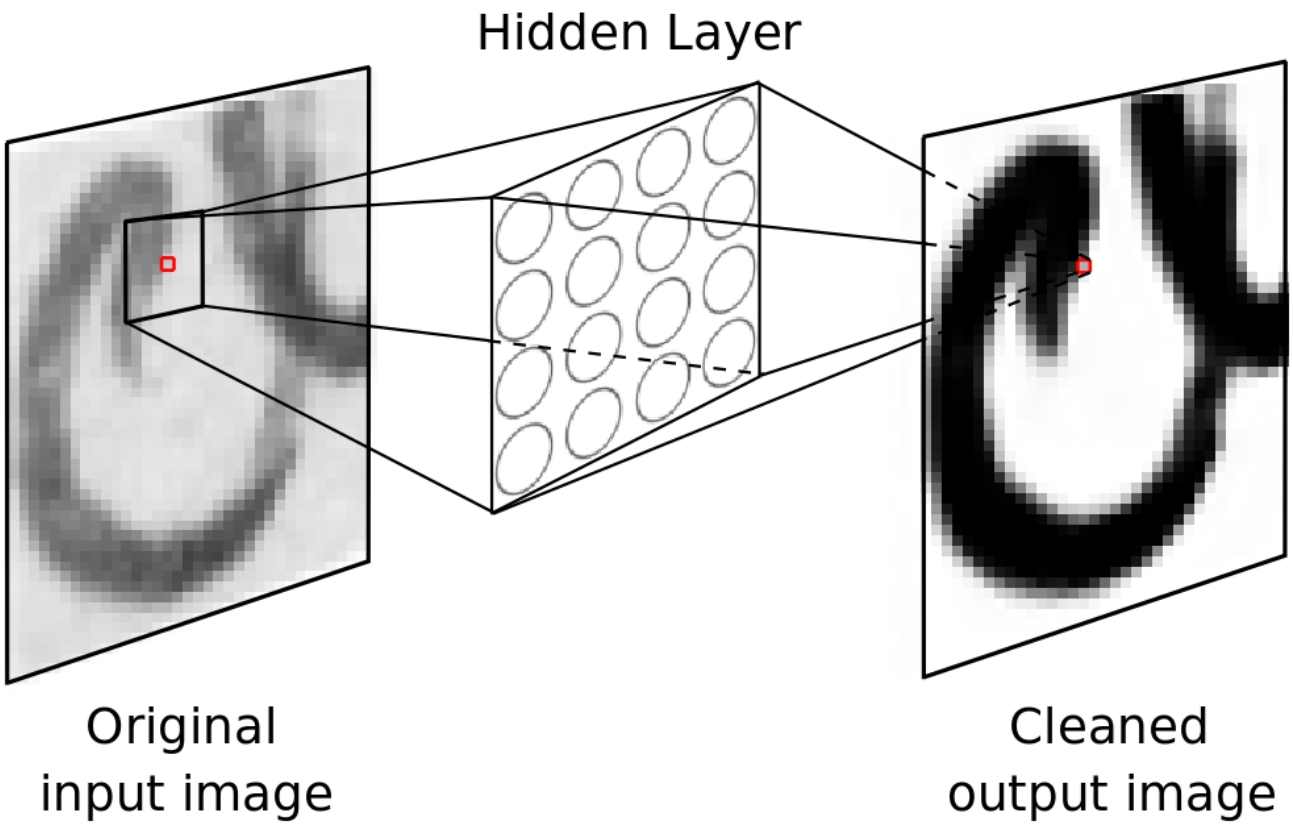
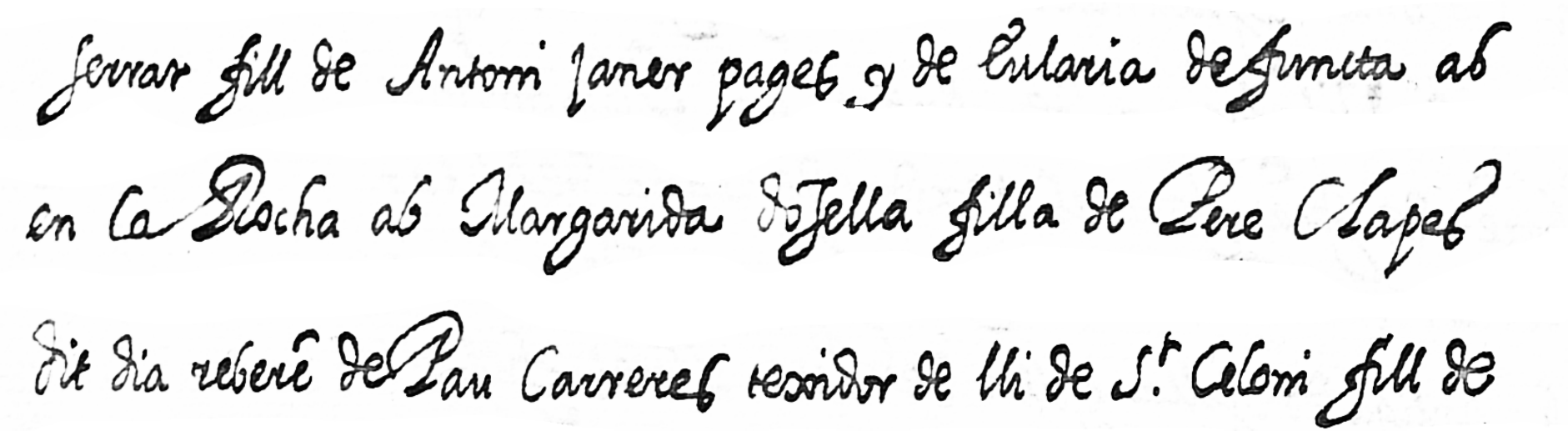


Image Cleaning Neural networks

- Improvements
  - Adding features: median filter, histograms.
  - Artificial Noise: gaussian perturbation, salt pepper mask.
  - Deeplearning: extract features from raw image, i.e. edges.
- Drawbacks
  - Slow. Slidding window. (Convolutional ANN are very slow)
  - Supervised images. (Bootstrap)
  - Evaluation (pixel accuracy,F-measure, text line accuracy, layout measures, WER...)
  - Very unbalanced data (5% 15% black pixels)

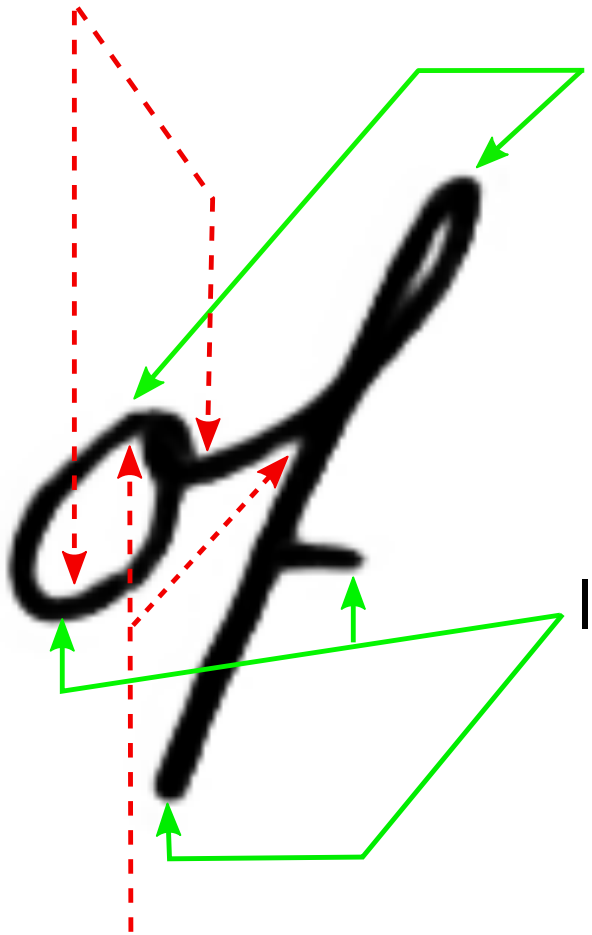
Cleaned lines



Interest Points

- **Types of local extrema**

local minima of upper contours



local maxima of upper contours
- *ascenders*
- *upper baseline*
- *not belong to reference lines*

local minima of lower contours
- *lower baseline*
- *descenders*
- *not belong to reference lines*

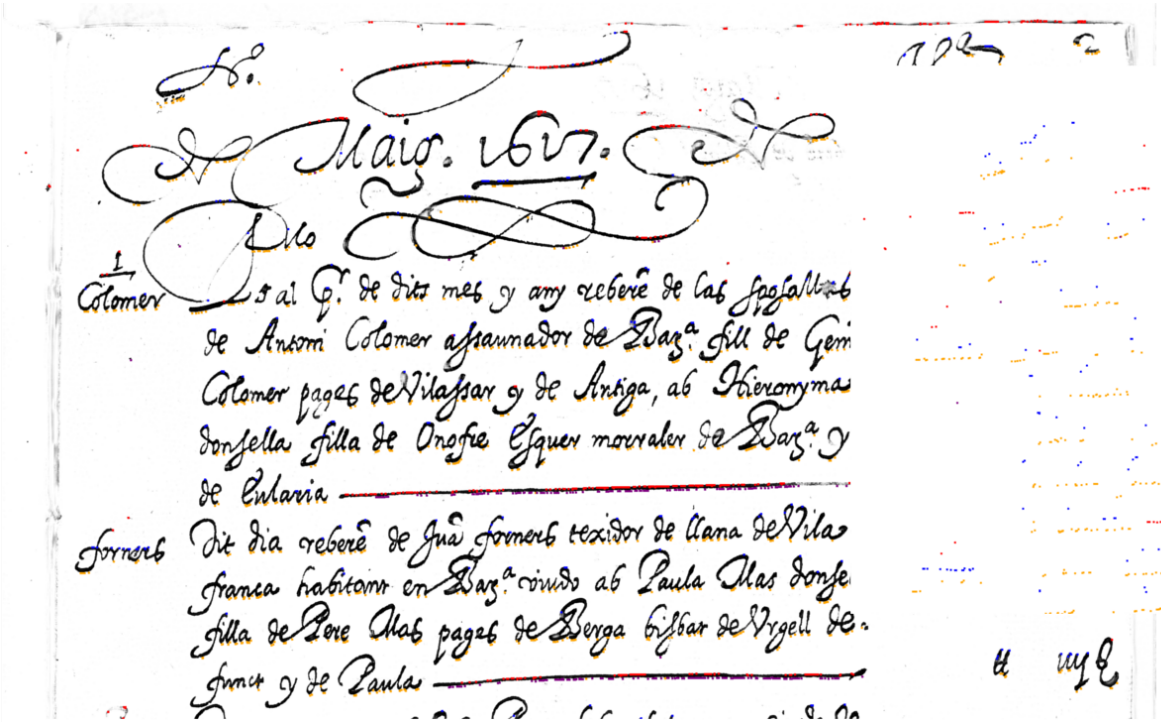
local maxima of lower contours

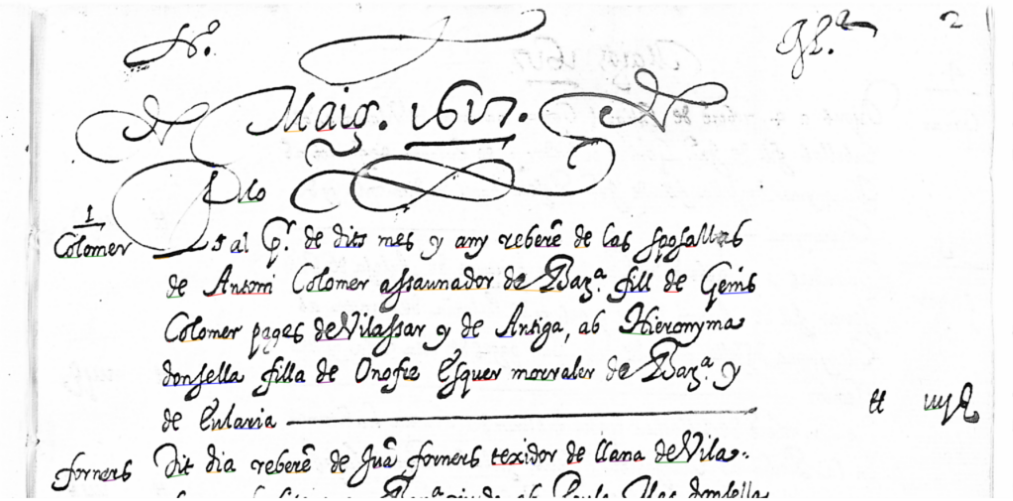
Interest Points

- Classification is computed by means of geometric transformation (fish eye) and a ANN.

- Interest points can be applied to an entire page. It detects the main points of the line.
- Join this points in order to get the lines.
- Improvements
  - Add a new class to separate body area pixels than noise.
  - Convolutional ANN. Fish eye distorsion is slow.

Using Interest Points at page level





Thank you for your attention.