Organización memoria de TFG

Reconocimiento de Sellos en Archivo Histórico de Osborne

Había pensado la siguiente organización, todo muy breve para que lo veas rápido. Si es necesario, te puedo detallar los puntos que sean (o todos) más en profundidad.

1. Explicar en qué consiste el proyecto en su conjunto (desarrollar métodos de OCR para digitalizar el archivo de manera autónoma), la Fundación Osborne así por encima…etc
2. Explicar por qué es importante extraer los sellos de los documentos escaneados.
   1. Para que el sistema de OCR sepa dónde está el texto y dónde no
   2. Para usar esos sellos para una clasificación autónoma del origen de los documentos
3. Explicar qué necesitamos exactamente
   1. Primero obtener una base de datos de todos los distintos sellos que existen.
   2. Utilizar esa base de datos para, mediante extracción de características, saber sus coordenadas en el documento y a qué sello corresponde.
4. Hablar sobre el estado del arte del tema
   1. Hay muchos artículos que utilizan transformada de Hough.
   2. Otro muy interesante utiliza un filtrado de paso bajo mediante transformación Gaussiana
   3. Etc…
5. Primero, vamos a generar la base de datos de sellos. Para realizar esto de manera autónoma tenemos que localizar dónde hay sellos. Esto parece contradictorio, ya que elaborar la base de datos tiene por objetivo precisamente, localizarlos y clasificarlos. En realidad no es contradictorio ya que este primer paso es imperfecto y falla cuando el sello tiene texto encima, que son bastantes ocasiones. Además, para realizar la base de datos de sellos, no necesitamos registrar todas las veces que aparezca el sello. El objetivo es obtener al menos una muestra de cada sello. En la práctica uno de ellos no se consigue registrar automáticamente, por lo que es necesario añadirlo de manera manual.
6. Después, se eliminan sellos repetidos. También se elimina un falso positivo de sello que aparece.
7. Luego, se etiquetan manualmente esos sellos (se les da alguna denominación).
8. Después, se extraen las características de los sellos mediante SURF, y se almacenan los puntos clave detectados junto con sus respectivos descriptores en una base de datos.
9. Una vez que tenemos ya toda esa información, se cargan los puntos almacenados en la base de datos y sus descriptores, y se buscan en el documento. Aquel sello que tenga un mayor número de puntos similares encontrados en el documento, será el que aparece en él. Si el número de puntos similares encontrados de cualquier sello en el documento es menor que un umbral, se considera que no se ha encontrado ningún sello y que el documento no contiene ninguno.

A grandes rasgos esto es lo que hace el algoritmo, con algunos matices más específicos para resolver algunos problemas.