

Autor: Miguel Ángel Rodrigo Lisbona

Tutor: Daniel Limón Marruedo

Reconocimiento de Sellos en “Archivo Histórico de Osborne”

Trabajo de Fin de Grado

Grado en Ingeniería de Tecnologías Industriales

Dep. Teoría de la Señal y Comunicaciones

Escuela Técnica Superior de Ingeniería

Universidad de Sevilla

Sevilla, 2013

Proyecto Fin de Carrera

Ingeniería de Telecomunicación

**Reconocimiento de Sellos en “Archivo Histórico de Osborne”**

Autor:

Miguel Ángel Rodrigo Lisbona

Tutor:

Daniel Limón Marruedo

Profesor titular

Dep. de Ingeniería de Sistemas y Automática

Escuela Técnica Superior de Ingeniería

Universidad de Sevilla

Sevilla, 2016

Trabajo de Fin de Grado: Reconocimiento de Sellos en “Archivo Histórico de Osborne”

|  |  |
| --- | --- |
| Autor: | Miguel Ángel Rodrigo Lisbona |
| Tutor: | Daniel Limón Marruedo |

El tribunal nombrado para juzgar el Proyecto arriba indicado, compuesto por los siguientes miembros:

Presidente:

Vocales:

Secretario:

Acuerdan otorgarle la calificación de:

Sevilla, 2016

El Secretario del Tribunal

*A mi familia*

*A mis maestros*

Agradecimientos

Blabla...

*Miguel Ángel Rodrigo Lisbona*

*Sevilla, 2013*

Resumen

La Fundación Osborne posee un archivo de documentos relacionados con la empresa que datan desde principios de 1800. Previas investigaciones han permitido descubrir información sobre la vida de Tolkien. Se sabe además, que existe correspondencia con casas reales o el Vaticano. También existe correspondencia de personajes públicos como Washington Irving. Los inicios de la empresa están estrechamente relacionados con otros personajes históricos como Fernán Caballero, cuñada del fundador, o “Frasquita” Larrea, su suegra. Por ello, podría haber información importante sobre dichos personajes que podría ser de valor histórico.

Para extraer dicha información, se ha optado por desarrollar una web donde se suben los archivos escaneados junto con su transcripción. Además de ello, se quiere desarrollar un sistema que realice dichas transcripciones de manera autónoma mediante detección de los caracteres manuscritos.

En este trabajo se describe el desarrollo de un sistema que reconozca los escudos en las cartas para clasificar su autor. Además, el sistema debe calcular la posición de dichos escudos con el objetivo de eliminarlos. Esto es necesario ya que los sistemas de detección de caracteres necesitan que los elementos que no son texto sean eliminados del documento antes de empezar cualquier análisis.

Abstract

Fundación Osborne owns an archive of documents related to the company dated up to early 1800. Previous investigations on the archive have revealed information about the live of Tolkien. Moreover, it is know that the archive has correspondence with royal houses as well as The Vatican. There are also letters of historical characters such as Fernán Caballero – founder’s sister-in-law – or “Frasquita” Larrea – his mother-in-law. For that reason, it is reasonable to think that there might be some historical value on such documents.

In order to retrieve that information, there is a website in development where scanned documents will be uploaded with their respective transcription and metadata. Those transcriptions are currently being typed manually. A system to make those transcriptions automated is being developed too by recognizing handwritten characters on the documents.

The purpose of this work is to describe the development of a system which recognizes seals on letters to automatically classify them by author. Moreover, the system calculates the position of such seal to delete it from the document. This is a need because character recognition system needs every non-text element removed from the image before any analysis is performed.

Índice

Agradecimientos ix

Resumen xi

Abstract xiii

Índice xiv

Índice de Tablas xv

Índice de Figuras xvii

Notación xix

1 Introducción 1

Referencias 11

Índice de Conceptos 13

Glosario 15

# **Índice de Tablas**

**No se encuentran elementos de tabla de ilustraciones.**

# **Índice de Figuras**

**No se encuentran elementos de tabla de ilustraciones.**

Notación

|  |  |
| --- | --- |
| A\* | Conjugado |
| c.t.p. | En casi todos los puntos |
| c.q.d. | Como queríamos demostrar |
| ∎ | Como queríamos demostrar |
| e.o.c. | En cualquier otro caso |
| e | número e |
| IRe | Parte real |
| IIm | Parte imaginaria |
| sen | Función seno |
| tg | Función tangente |
| arctg | Función arco tangente |
| sen | Función seno |
| sin*xy* | Función seno de *x* elevado a *y* |
| cos*xy* | Función coseno de *x* elevado a *y* |
| Sa | Función sampling |
| sgn | Función signo |
| rect | Función rectángulo |
| Sinc | Función sinc |
| ∂y ∂x  *x*◦ | Derivada parcial de *y* respecto  Notación de grado, *x* grados. |
| Pr(*A*) | Probabilidad del suceso *A* |
| SNR | Signal-to-noise ratio |
| MSE | Minimum square error |
| : | Tal que |
| < | Menor o igual |
| > | Mayor o igual |
| \ | Backslash |
| ⇔ | Si y sólo si |

# Introducción

Individual commitment to a group effort, that is what makes a team work, a company work, a society work, a civilization work.

- Vince Lombardi -

L

a Fundación Osborne “es el vehículo canalizador de la Responsabilidad Social Corporativa del Grupo Osborne”. Posee dos fines fundamentales: “la preservación y difusión del patrimonio histórico de la compañía y la promoción de la formación y el emprendimiento en las áreas de influencia del Grupo Osborne.” (citar web fundación Osborne)

La Fundación SEPI, entre otras actividades, “promueve y gestiona un extenso Programa de Becas de Inserción Laboral que cada año facilita la formación práctica en las empresas a más de 300 jóvenes titulados.” (citar web fundación sepi)

La Fundación Telefónica tiene por misión “mejorar las oportunidades de desarrollo de las personas a través de proyectos educativos, sociales y culturales, adaptados a los retos del mundo digital.” (citar web fundación tele)

Las tres colaboran en un proyecto denominado Lab Osborne. El objetivo de dicho proyecto es la digitalización del archivo histórico que posee la empresa. Dicho archivo contiene, por un lado, facturas, notas de pago y demás documentación de carácter administrativo y, por otro, correspondencia entre miembros de la familia y con otros personajes públicos desde la época de la fundación de la empresa en 1772. Entre estos personajes se encuentran Washington Irving o J.R.R. Tolkien así como miembros de distintas casas reales o El Vaticano.

Esta correspondencia podría tener valor histórico. De hecho, existe un libro sobre la vida de Tolkien donde una gran cantidad de datos han sido extraídos de investigaciones en este archivo.

No sólo eso, las cartas de miembros de la familia también podrían contener información importante ya que algunos estos miembros en los orígenes de la empresa fueron personajes públicos. Por ejemplo, la suegra de Thomas Osborne Mann, fundador de la empresa, fue Francisca Javiera Ruiz de Larrea y Aherán, **Frasquita Larrea**. Esta señora fue una escritora gaditana impulsora de la tertulia literaria que se cree que tuvo una gran influencia en la Constitución de 1812. Su cuñada, Cecilia Böhl de Faber y Larrea, **Fernán Caballero**.

Gran parte de la documentación que se posee tiene un contenido desconocido de ahí nace la motivación por digitalizarlo, ya que se podría analizar e investigar con mayor agilidad. Para ello, se desarrolla una página web donde se muestran los documentos escaneados junto con su transcripción. Esta transcripción es necesaria por un doble motivo: por un lado, los documentos tienen en su mayoría una caligrafía de muy difícil reconociminto lo cual ralentiza enormemente la lectura e interpretación. Por otro lado, si se quiere realizar algún tipo de búsqueda en función de alguna palabra que forme parte del texto, sería imposible sin transcripción previa.

Dicha transcripción se realiza a mano, pero es una tarea lenta y tediosa. Por ese motivo se quiere desarrollar, como parte del proyecto, un sistema visión artificial que reconozca el contenido manuscrito de los documentos y genere una transcripción de manera automática. Esto puede realizarse por distintos métodos, por ejemplo, mediante reconocimiento óptico de caracteres (OCR). Esto consiste en seccionar de alguna manera el texto en caracteres para más tarde utilizar algún método para reconocer a qué carácter corresponde cada una de las secciones de imagen previamente obtenidas. (citar artículo donde se ha hecho incluyendo precisión) El proceso de seccionamiento es costoso e impreciso (citar artículo) y puede dar caracteres divididos en dos regiones o varios caracteres en una misma región, lo que al final lleva a un reconocimiento erróneo de dichos caracteres. Por ello, otra forma de hacerlo es directamente reconocer palabras completas. Esto elimina el paso de seccionamiento de los caracteres lo que supone un gran avance, ya que el seccionamiento de palabras es mucho más simple y preciso generalmente (citar). Este segundo enfoque presenta, sin embargo, una gran desventaja. El método más utilizado para reconocer los caracteres o palabras consiste en entrenar un sistema mostrándole un gran número de caracteres o palabras, según corresponda. En el caso de las palabras, es evidente que existe una variedad mucho más elevada que de caracteres, por ello para obtener un reconocimiento razonablemente preciso es necesario una cantidad de texto de entrenamiento mucho mayor. Independientemente de cuál de las dos sea la solución adoptada, ambas requieren una serie de pasos previos.

1. Primero necesitamos una **binarización** que convierta las imágenes escaneadas en una matriz binaria. Deseablemente este sistema eliminará todos los píxeles de fondo del papel, incluyendo posibles manchas o texto transparentado perteneciente a la cara opuesta del papel. Este paso incluye posibles transformaciones morfológicas o filtrados que fueran necesarios con el objetivo de eliminar ruido procedente del uso de un papel heterogéneo, de manchas, de oscurecimiento de posibles pliegues, texto transparentado, etc…Una binarización ideal no elimina nada de los elementos gráficos del documento, incluyendo firmas, fechas, sellos, además de texto, pero eliminará todos los elementos que no contengan información.
2. Posteriormente, procedemos a **segmentar** el documento, localizando qué píxeles dentro de la imagen corresponden a texto manuscrito, cuáles a texto impreso, a sellos o firmas.
3. Una vez que sabemos qué parte del documento corresponde a texto, se procede a **dividirlo por líneas**. Esta división es sencilla, precisa y computacionalmente rápida. El método más utilizado consiste en sumar cuántos píxeles de texto hay en cada una de las filas de la imagen binarizada. Ni siquiera es necesario calcular mínimos locales, aquellas filas donde el número de píxeles es menor que un cierto umbral, corresponden con los huecos entre líneas de texto.
4. El siguiente paso es obtener las coordenadas que **delimitan cada palabra**. La forma más extendida para resolver este problema se denomina *etiquetado de componentes conexas* que se explicará en detalle más adelante. Una vez que tenemos las coordenadas de cada palabra, podemos extraer las secciones de la imagen original que las contienen para entrenar al sistema que reconocerá dichas palabras.

Existen otros enfoques alternativos para realizar este procedimiento, en (cita) se propone un método alternativo que se salta la fase de binarización y trabaja directamente con imágenes en escala de grises. Este procedimiento no presenta ventajas notables para el caso de imágenes de alta resolución. Las imágenes con las que trabajamos tienen una resolución de 3525x2810 píxeles. Por ese motivo, se ha preferido atenerse al método tradicional, ya que es más sencillo y no presenta desventajas notables. También existen métodos para eliminar la fase de segmentación. En (cita) Aouadi N. y Kacem A. proponen un método basado *GHT* (*Generalized Hough Transform*). Este método no requiere segmentación y, previo entrenamiento, consigue un sistema que detecta el texto a partir de parametrizaciones aprendidas. Aparentemente, este sistema consigue muy buenos resultados, sin embargo, requiere un entrenamiento. Dado que el método pretende evitar una segmentación automática errónea, dicho entrenamiento debe realizarse a partir de muestra de texto segmentada manualmente. Eso es precisamente lo que se quiere evitar aquí, por lo tanto este algoritmo no nos es válido.

Por todo ello, se ha optado por seguir un planteamiento más ortodoxo descrito con anterioridad e ilustrado en la figura a continuación.

Este trabajo pretende describir el desarrollo de un sistema de segmentación que detecte los sellos en los textos. Esto tiene un doble propósito:

1. Por un lado y como ya se ha descrito anteriormente, es necesario conocer qué partes de la imagen escaneada del documento corresponden o no a texto.
2. Por otro lado, conocer qué píxeles en la imagen corresponden a un sello permite clasificar dicho sello con el objetivo de asociar de manera automática un autor al documento en cuestión. La mayoría de los documentos corresponden a cartas y buena parte de ellos contienen un sello en la parte superior que corresponde al remitente de dicha carta.

# Algoritmos disponibles para seccionamiento de sellos

Everything is theoretically impossible, until it is done.

- Robert A. Heilein -

E

xisten múltiples maneras de resolver este problema. En (cita) se utiliza un enfoque frecuencial para diferenciar entre regiones gráficas (sellos, imágenes…) y texto. Pruebas con ese algoritmo demuestra no ser suficiente ya que algunos de los sellos presentes en los documentos agregan componentes de alta frecuencia igual que el texto, por lo tanto, eliminarían el sello de la misma manera que eliminan el texto.

Otro procedimiento probado, consiste en localizar formas geométricas mediante HT ya que muchos sellos presentan algún tipo de circunferencia o rectas de dimensiones de los que se presentan en el texto. Las pruebas no muestran buenos resultados ya que, en lugar de figuras geométricas de gran tamaño, se detectan múltiples círculos y rectas de pequeñas dimensiones.

Se ha probado, además, extraer las características mediante SURF de un sello modelo y comparar el número de emparejamientos en los documentos; si este número es suficientemente alto, se considera que dicho sello ha sido encontrado en el documento. Este método ofrece mucho mejores resultados, sin embargo, presenta el inconveniente de necesitar una muestra de cada uno de los distintos sellos que puedan aparecer.

El último intento consiste en aislar los elementos que aparecen en el documento y almacenarlos de manera independiente para más tarde aplicar una serie de criterios heurísticos para determinar si dicha región corresponde o no a un sello. Esto se hace mediante *etiquetado de componentes conexas* (o *connected component labeling*). Este algoritmo asigna una etiqueta identificativa única para cada región conexa de píxeles. El problema de este método es la existencia de texto muy cercano al sello o incluso encima de éste. En ese caso el algoritmo detecta como una región tanto al sello como a ese texto. Cuando esto ocurre, los filtros heurísticos fallan ya que el conjunto de sello más texto no cumple las mismas condiciones que si solamente estuviera el sello.

A continuación se detalla cada procedimiento en mayor profundidad además de los resultados obtenidos en cada caso, utilizados para decidir qué solución adoptar.

## Enfoque frecuencial para seccionamiento de texto y sellos.

En (cita) se desarrolla un método para separar texto de elementos gráficos en documentos escaneados. La idea se basa en el hecho de que el texto introduce componentes armónicas de alta frecuencia a la imagen. Por lo tanto, un filtro de paso bajo eliminaría ese texto. En el algoritmo, se utiliza un filtro de Gauss para eliminar las componentes de alta frecuencia. El resultado es una imagen con los elementos de texto eliminados casi en su totalidad mientras que los elementos gráficos (en el artículo incluye también logos) permanecerían en la imagen, idealmente, en su totalidad. El algoritmo expuesto es el siguiente:

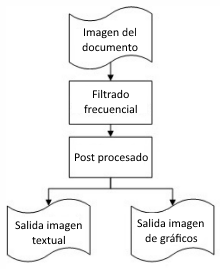


Figura X. Algoritmo espectral de seccionamiento.

El algoritmo original presenta pasos adicionales para separar distintos tipos de elementos gráficos, pero para nuestro problema, son pasos innecesarios.

### Filtrado frecuencial

El filtrado frecuencial es el paso más importante, y es donde se elimina la mayor parte del texto. Es una evidencia conocida que el texto contribuye a componentes espectrales de muy alta frecuencia en imágenes de documentos (cita). Por ello, se convoluciona con un kernel gausiano como filtrado de paso bajo y se utiliza un post procesado para eliminar restos residuales de texto. El algoritmo puede verse en la figura X.

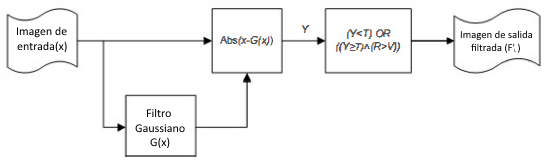


Figura X. Filtrado frecuencial.

El proceso consiste en:

1. Se toma la imagen original convertida a escala de grises, X.
2. Se le aplica un filtrado gaussiano, G(X).
3. Se calcula la imagen la imagen de **paso alto** Y, restándole G(X) a la original.
4. Un filtrado selectivo en frecuencias sencillo sería entonces

Donde I es la imagen original de dimensiones NxM, 1≤ i ≤ N y 1≤ j ≤ M. T es el umbral de frecuencias. Los valores que superan un cierto valor en la imagen filtrada Y se consideran de alta frecuencia, los que no, de baja frecuencia. Por último, Bk es el valor del color considerado de fondo. En el artículo utilizan una media de los colores de fondo del papel, obtenidos sumando los píxeles de la imagen original I, que han quedado como fondo en una binarización de la imagen en escala de grises X. En este trabajo se ha considerado innecesario ya que no se pretende reconstruir los huecos dejados por los elementos de eliminados del documento, y se ha tomado Bk como blanco directamente.

Hay que destacar que un filtrado solamente en frecuencias no es apto para elementos cromáticos cuyos tonos cambien de manera abrupta, ya que también introducen componentes de alta frecuencia y dicho filtro los eliminaría. Para solventar este problema, se conservan también aquellos elementos que tengan una cantidad de color suficientemente alta, esto se consigue cuantificar mediante la *cromaticidad*.

**Definición** cromaticidad: Se entiende por cromaticidad de un tono de color a la medida de cómo de alejado se encuentre ese color de tonos grises, independientemente de su luminancia, es decir, de cómo de claro u oscuro sea.

En un espacio de colores YCrCb, la cromaticidad R se cuantifica como

Por tanto, dada nuestra imagen original, se define la matriz de cromaticidad aquella que contiene para cada píxel el valor de su cromaticidad.

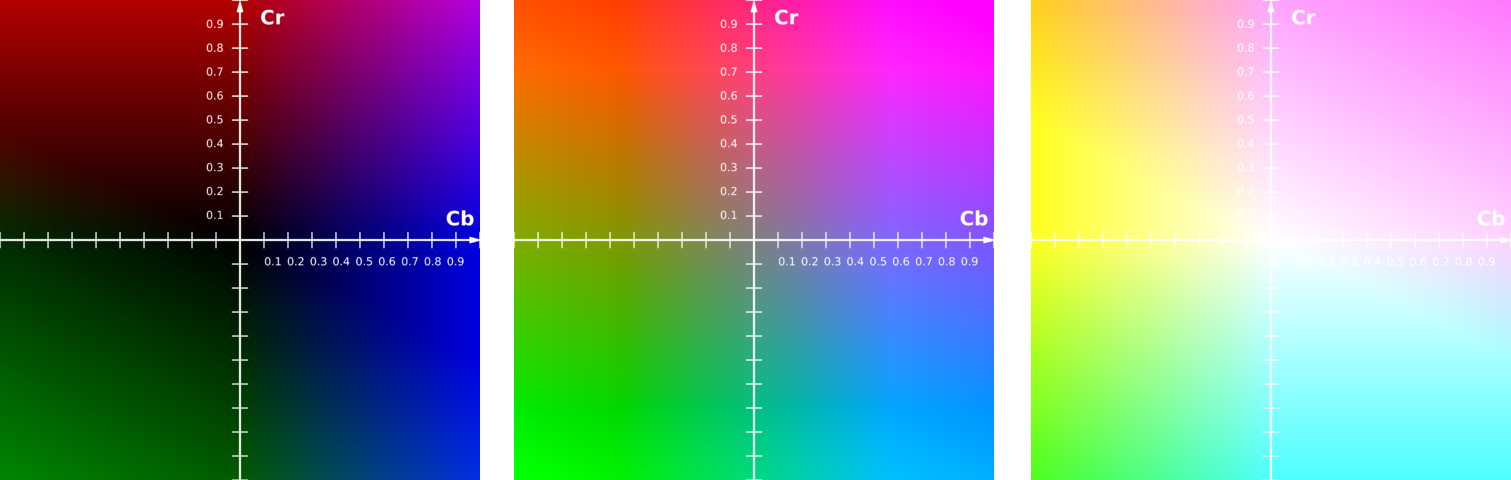


Figura x. Espacios de color YCrCb para valores de luma (Y) de 0, 0.5 y 1 respectivamente.

En la figura anterior puede verse como aquellos colores con una menor R, es decir, aquellos más cercanos al origen de coordenadas, poseen tonos más grises.

Volviendo al problema anterior, queremos evitar que al filtrar componentes de alta frecuencia eliminemos elementos gráficos que posean cambios bruscos en los colores. Para ello, imponemos un umbral V. Aquellos píxeles cuya cromaticidad sea superior a dicho umbral serán conservados en la imagen filtrada. Esto no supone ningún problema para la eliminación de texto ya que los caracteres tienen generalmente una cromaticidad muy baja.

Los resultados obtenidos mediante este procedimiento son los siguientes.



Figura X. Ejemplo de correcto funcionamiento del filtrado frecuencial.

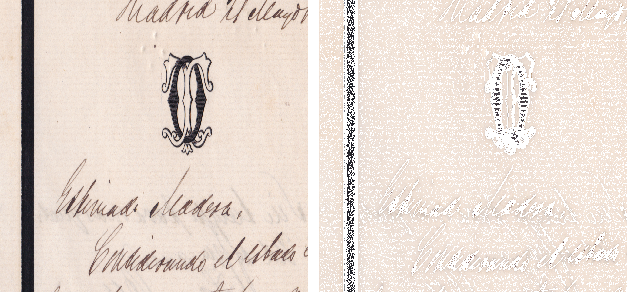


Figura X. Ejemplo en el cual el filtrado frecuencial falla.

Como puede verse, los resultados en el primer caso son verdaderamente alentadores. Sin embargo, en el segundo caso el filtrado frecuencial elimina el sello además del texto. Esto se debe a que el sello posee claras diferencias tonales entre colores de muy baja cromaticidad. Por tanto, la condición de cromaticidad incluida para evitar la eliminación de elementos gráficos con variaciones tonales importantes, aquí no consigue salvar el sello.

Por todo eso, se procederá a probar otro método.

## Seccionamiento de texto y sellos mediante extracción de características

Definición Extracción de Características: En el contexto de tratamiento de imágenes, proceso mediante el cual, a partir de una imagen se obtiene un conjunto de datos (características), ligados a algunos píxeles específicos de la imagen (*keypoints)*. Estos datos deben ser informativos y no redundantes. En ocasiones las características pueden tener una representación física tangible, como esquinas. En otras en cambio, es un cálculo matemático sin una interpretación evidente.

El objetivo de la extracción de características es encontrar objetos que aparezcan en distintas imágenes, a pesar de presentar transformaciones homográficas o algún otro tipo de cambio menor, como cubrimiento parcial del objeto, deterioro, etc. Esto se hace partiendo de un modelo del objeto a encontrar, se realiza una extracción de características, luego se hace lo mismo en la imagen. Más tarde se comparan las características encontradas en la imagen y en el objeto modelo y aquellas que se parezcan lo suficiente suponen una correspondencia o *match*. Aquella zona de la imagen que dé suficientes matches, contiene al objeto buscado.

El proceso normal en una extracción de características es:

1. Detección de puntos clave o *keypoints*: según el algoritmo utilizado, estos puntos tienen una definición matemática diferente. En general, se busca encontrar puntos que tienen alguna propiedad matemática que los hace suficientemente diferentes al resto de puntos próximos.
2. Descripción de dichos puntos. A cada punto se le calcula un valor identificativo único denominado **descriptor**. En función del algoritmo escogido este descriptor puede ser un conjunto de vectores o un valor binario.
3. Matching. Consiste en la búsqueda de correspondencias de estos puntos clave en función de los descriptores. Para cada punto se le busca el punto de la imagen cuyo descriptor se parece más. Este “parecerse más” puede variar también según el algoritmo: mínima distancia euclídea o mínima distancia Manhattan son los criterios más comunes, aunque no los únicos.
4. Desechado de matches erróneos. Se toma algún tipo de criterio para intentar eliminar matches erróneos en el paso anterior. El criterio más común es el descrito por D. Lowe desarrollado para utilizarlo en su algoritmo SIFT, pero extendido hoy en día casi como criterio estándar de eliminación de falsos matches. Consiste en, dado un keypoint, calcular sus dos mejores matches. Si el primero y el segundo se parecen demasiado (un 70% o más es el valor utilizado por él, aunque este valor es configurable a medida para satisfacer las necesidades de cada problema), entonces es un falso match. Si el primer y el segundo candidato son suficientemente diferentes, se considera válido. Según su artículo, este método elimina un alto porcentaje de falsos positivos, eliminando un número muy bajo de positivos verdaderos. (cita).

(INCLUIR AQUÍR TESTS DE ALGORITMOS DE DETECCIÓN DE ROS) y por esto usamos SURF…

Para el caso en el que el sello usado para comparar encaja con el que se muestra en el documento, se obtienen más de 200 correspondencias (típicamente, entre poco más de 200 hasta unas 350, según el documento) como puede apreciarse en la imagen. Muchos de ellos son falsos positivos, pero en su mayoría son buenas.

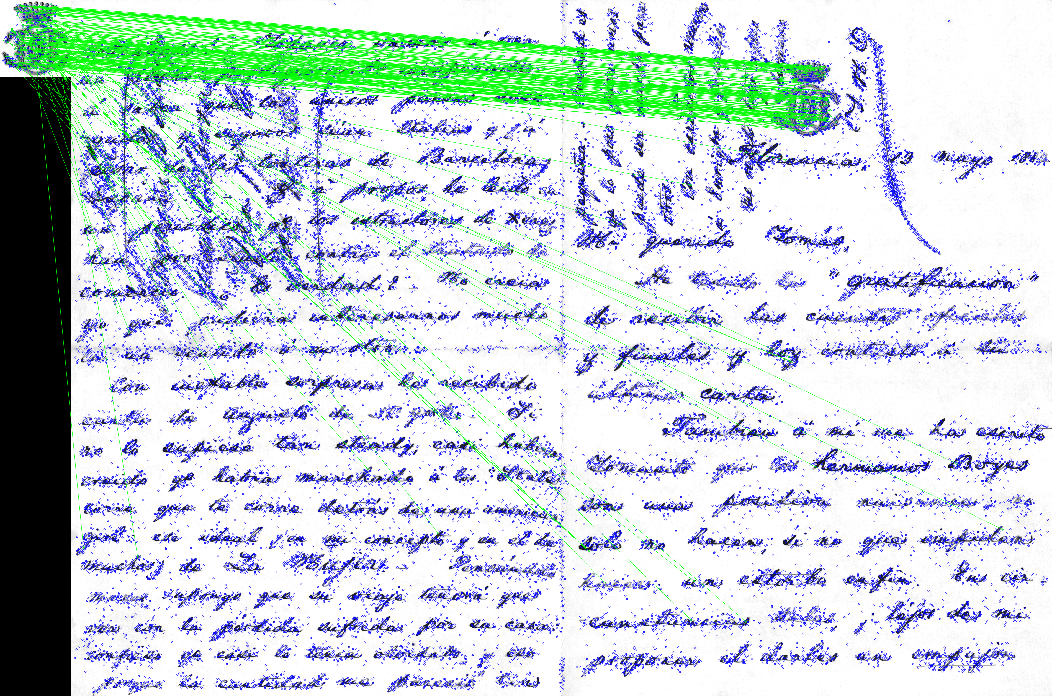


Figura X. Sello localizado en el documento mediante SURF

En cambio, cuando la muestra y el sello del documento no coinciden, el número de correspondencias encontradas, es de poco más de 100. Pueden parecer muchas pero hay que tener en cuenta que el método de Lowe para discriminar falsos positivos sólo es útil cuando el objeto buscado se encuentra en la imagen. En caso de que el número de falsos positivos cuando el sello buscado no está presente fuera demasiado grande podríamos calcular la distancia media entre descriptores emparejados en el caso de emparejamientos acertados (es decir, cuánto se parecen cuando el emparejamiento es correcto). Aquellos descriptores cuya distancia sea muy superior a un umbral prefijado (digamos por ejemplo, un 30% más que la media calculada), podrían ser descartados. Sin embargo, ya que el número de falsos positivos disminuye lo suficiente y que además los emparejamientos erróneos se encuentran distribuidos por toda la imagen, es fácil determinar cuándo el sello se encuentra en el documento y cuándo no.

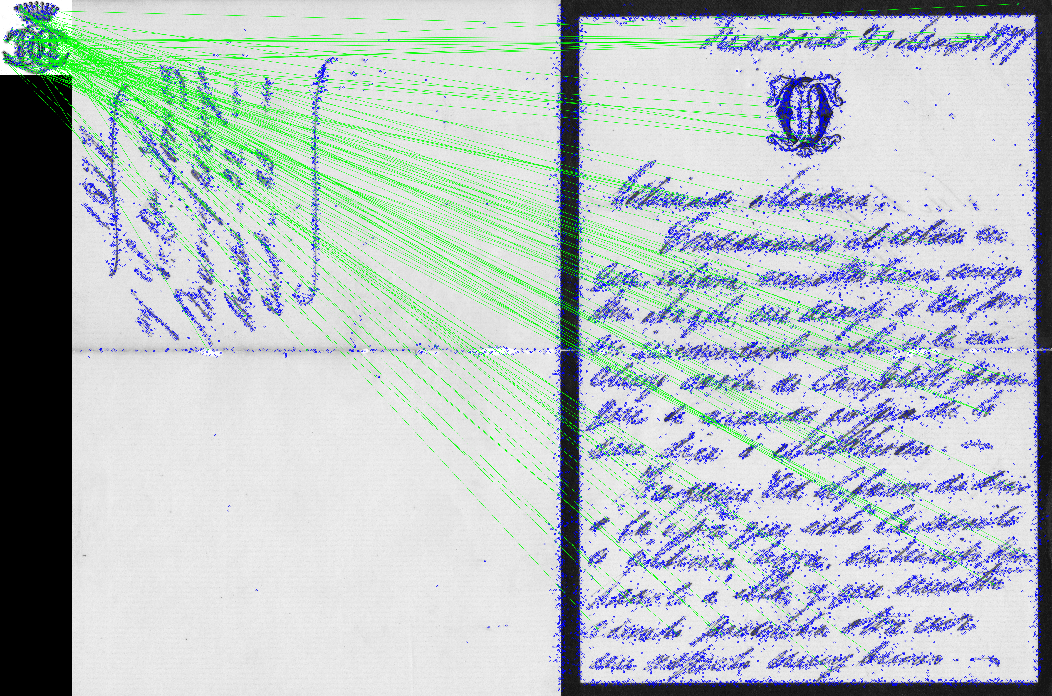


Figura X. Sello buscado mediante SURF pero no localizado en el documento.

(Después de explicar cada método). Debido a que este sistema es imperfecto y que el enfoque de la extracción de características presenta muy buenos resultados, se ha optado por **aplicar extracción de características** **y alimentar** ese sistema con muestras de sellos obtenidas **mediante etiquetado de componentes conexas y eliminación heurística**.

# Referencias

|  |  |
| --- | --- |
| [1] | Autor, «Este es el ejemplo de una cita,» *Tesis Doctoral,* vol. 2, nº 13, 2012. |
| [2] | O. Autor, «Otra cita distinta,» *revista,* p. 12, 2001. |

# **Índice de Conceptos**

conceptos 9

# **Glosario**

ISO: International Organization for Standardization 4

UNE: Una Norma Española 4