

Segmentación y mejora de documentos

Artur Gil Torres

Visión Artificial

Objetivos y funcionalidades

En esta práctica se expondrá la implementación de un pequeño sistema que permita la segmentación de texto en imágenes de folios. Se partirá de un conjunto de 16 imágenes JPG que contienen 1 hoja cada una, hojas que podrán estar arrugadas, rotadas o manchadas de café o bolígrafo y se aplicarán distintas técnicas de visión artificial para extraer el texto contenido en ellas y reconstruirlo en una imagen en blanco. Para aproximar una solución se identifican distintos subproblemas a resolver:

- **Preprocesado.** Se incluye aquí un tratamiento robusto de la imagen que facilite la posterior detección o estimación de los bordes del folio en distintas circunstancias de luz, fondo o ruido de distinto tipo (arrugas, pelos, manchas...). Existen diversas opciones: filtrado morfológico, mejora de contraste, suavizado, umbralización...
- **Detección del folio.** Allanado el camino en el paso anterior se deberá obtener los contornos y esquinas del folio. Gracias a este paso se podrá reducir el problema a segmentar el texto de una imagen que sea exclusivamente una hoja de papel. Para obtener los contornos es crucial la obtención de las esquinas, ya que partiendo de que la hoja tiene forma trapezoidal con las esquinas detectadas se pueden determinar los límites de ésta. Esto se podrá realizar detectando una serie de líneas rectas que delimiten el folio y calculando sus intersecciones, un detector de esquinas (Harris, Shi-Tomasi, Susan...) o un detector de contornos como el de Satoshi Suzuki.
- **Segmentado del texto y eliminación de ruido.** Una vez se tenga la hoja de papel aislada se deberá tratar de extraer el texto contenida en ella, eliminando cualquier ruido y para distintas iluminaciones o contrastes. Para esto las opciones más claras son jugar con los colores y la umbralización.

Contents

1 Metodología general	1
2 Evaluación de resultados y posibles mejoras	4

1 Metodología general

Definidos los objetivos a alcanzar se comentarán ahora en detalle las técnicas empleadas en cada uno de los pasos citados anteriormente.

1. **Preprocesado.** Se ha decidido como

primer paso mejorar el contraste de la imagen para poder resaltar el folio sobre el fondo, sobre todo para aquellas imágenes en las que no sea tan obvia la diferencia entre el *objeto folio* y el *fondo*. Se ha empleado para ésto una ecualización de histograma adaptativa. La principal diferencia con la ecualización clásica es que la ecualización de histograma adaptativa divide la imagen en regiones o bloques más pequeños y aplica la ecualización de histograma a cada uno de ellos de forma local. Esto permite mejorar el contraste local en

áreas específicas de la imagen, especialmente en aquellas con variaciones de iluminación o regiones oscuras, sin afectar significativamente el resto de la imagen.

Se sigue con un suavizado de la imagen, ya que con la ecualización adaptativa se genera mucho ruido sal. Se ha decidido como buenas opciones un filtrado Gaussiano y uno bilateral, gracias a ellos se puede reducir gran parte de éste ruido sin comprometer los bordes del folio.

Luego se ha decide aplicar una umbralización local óptima. En esta técnica el umbral se calcula a partir de una media ponderada basada en una ventana gaussiana alrededor del píxel actual. Se probó a aplicar una umbralización global pero observados los malos resultados por las diferencias de luminosidad y regiones de contraste en puntos localizados de las imágenes, la umbralización local fue la elección final. Con esta umbralización y un posterior filtrado morfológico se consiguen resaltar todavía más los contornos del folio. Se aplica un cierre morfológico con el objetivo de cerrar y destacar contornos fragmentados.

En este punto se puede proceder con la detección de bordes, para ellos se aplica un Canny, y presuponiendo que en la mayoría de casos hemos resaltado los bordes, el comportamiento de éste, habiendo ajustado unos umbrales de histéresis superior e inferior deberá ser bueno. También se probó a delimitar los bordes determinando los gradientes G_x y G_y de las imágenes con Sobel, ponderando más el G_x para resaltar más los bordes verticales del folio, los bordes que más problemas daban con las diferencias de contraste y luz, pero finalmente decide usar tan sólo Canny por la gran introducción de ruido de Sobel en la imagen.

2. Segmentado del folio. Ahora que se dispone de los bordes de la imagen se debe decidir entre distintas estrategias para estimar en qué parte de la imagen está el folio. Se intentó extraer las líneas de Hough de la imagen, pero debido al preprocesado anterior se generaban demasiadas líneas sobre el ruido de la imagen, y por mucho que se filtrase por longitud para luego intentar estimar exactamente cuáles eran las del folio, se daba que muchas veces los bordes estaban constituidos por 3 o incluso más líneas cortas, haciendo imposible distinguirlas de manera sistemática de otras que representasen ruido. Se puede ver un ejemplo de esto en las imagen 1 y 2.

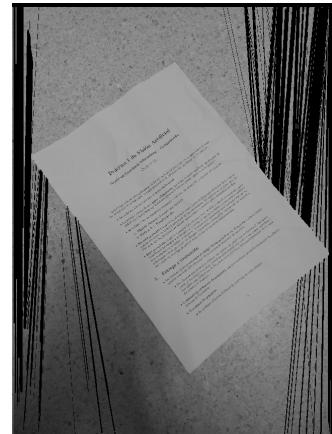


Figure 1: Problemas con Hough a).

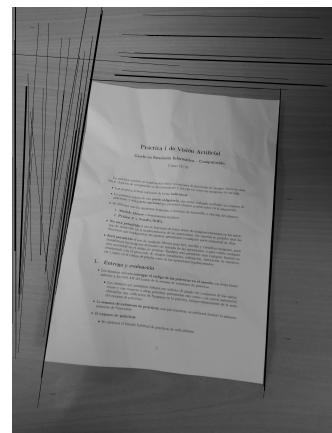


Figure 2: Problemas con Hough b).

También se intentó sin mucho éxito

extraer las esquinas de la hoja con el detector de Harris y el de Shi-Tomasi pero se encontraban demasiadas en el ruido de la imagen y resultó imposible limitarlas a las 4 necesarias. Además también se probó a aproximar el contorno del folio usando el modelo deformable Active Contours Without Edges por su robustez frente a bordes débiles o mal delimitados pero su alto coste computacional y el atascamiento en zonas de ruido pesaron para abandonar la idea. Además también era un problema complejo proporcionar un contorno inicial ajustado para cada imagen. Descartadas todas las opciones anteriores se decide extraer los límites del folio con la implementación de la librería OpenCV del algoritmo de Suzuki y Abe de detección de contornos. Teóricamente es una estrategia con buenos resultados para imágenes binarizadas y muy eficiente computacionalmente, que sigue los bordes cerrados y proporciona además una jerarquía basada en lo interno o externo que es un contorno. Se realizan diversos experimentos y se observa que se detectan contornos con forma trapezoidal que encierran los folios. Una vez se obtienen estos contornos resta filtrarlos para obtener el necesario, y para esto se miden sus áreas y se establece un umbral que permite seleccionar contornos suficientemente grandes como para encerrar un folio. Luego se intenta aproximar ese contorno de gran área a un trapezoide mediante el algoritmo de simplificación de polígonos de Douglas-Peucker. Este algoritmo reduce el número de vértices de una curva o polígono, manteniendo una forma aproximada que sigue siendo fiel a la estructura original. En la imagen 3 se puede ver el resultado de la detección de contornos de Suzuki-Abe.

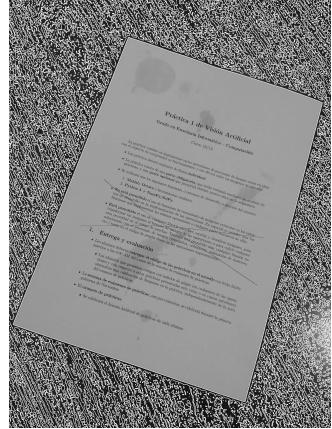


Figure 3: Contornos detectados

Ahora mismo, los contornos que hayan superado este filtrado ya se pueden considerar con gran certeza como candidatos a los contornos del folio, por lo que solo resta extraer sus vértices.

3. Segmentado del texto y eliminación de ruido. Para abordar este último paso se decide reducir la imagen al folio detectado y corregir su perspectiva. Esto se hace para que cualquier procesado posterior se aplique exclusivamente al área relevante, eliminando distracciones externas y garantizando que las dimensiones del folio sean consistentes. Para la corrección de la perspectiva se calcula primero una matriz de transformación de perspectiva (homografía) que luego se usa en interpolación y mapeo inverso para calcular los nuevos píxeles de la imagen transformada.

Por último se procesa la imagen con el objetivo de eliminar la mayor cantidad de ruido posible, a la vez que se intenta perder el menor texto. Para ello primero se ha decidido realizar una segmentación por color para quitar las manchas de bolígrafo, se asumirá que las manchas pueden ser azules o rojas. Para ello se extraerán los canales HSV de la imagen, que nos permiten trabajar de manera más efectiva con los colores, separando la tonalidad (Hue), la saturación (Satu-

ration) y el valor (Value) de cada píxel. Esto facilita identificar y aislar nuestros colores específicos. Mediante máscaras adaptadas a los rangos de tonalidad correspondientes, se detectan estas manchas y se sustituyen por blanco para eliminarlas del documento.

Para el ruido de las manchas de café, las diferencias de luminosidad y las posibles arrugas se empleará umbralización global, pero antes se suavizará un poco la imagen con un filtro Gaussiano. Se emplea en este caso umbralización global por ser eficaz para eliminar ruido de intensidad relativamente constante. Se ha probado también con una umbralización adaptativa pero los resultados no han sido demasiado buenos, como se puede ver en la imagen 4.

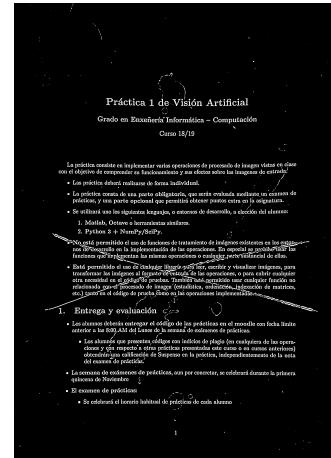


Figure 4: Manchas de café todavía muy visibles y ruido sal

Por último y para reconstruir el mayor número letras fragmentadas se emplará la operación morfológica de cierre de nuevo y se reconstruirá la imagen procesada en una imagen blanca.

2 Evaluación de resultados y posibles mejoras

Para validar los resultados del sistema se analizarán para las imágenes en las que se han detectado folios qué regiones de la hoja se han segmentado correctamente y cuáles se pierden. Se considerarán las siguientes regiones de la figura 5

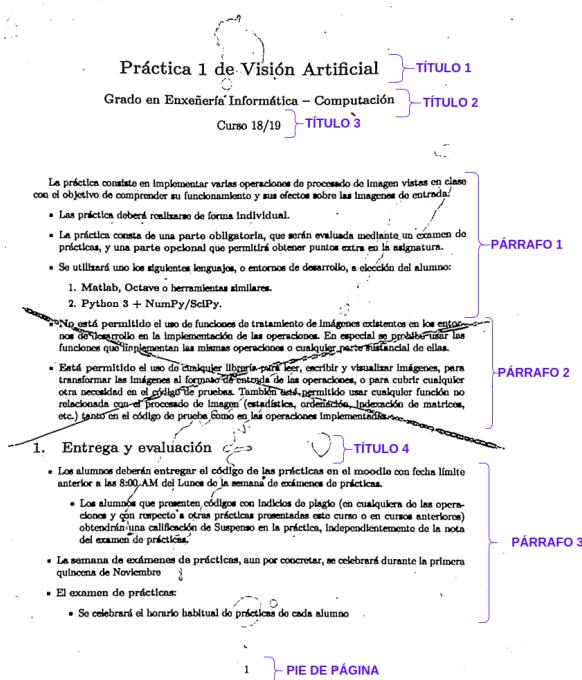


Figure 5: Secciones a considerar para evaluación de resultados

Con este sistema se detectan folios en 8/16 imágenes y los resultados son los de la tabla 1 (consultar resultados anexos). Se marca con un *SI* si el texto es legible y un *NO* si no lo es. Además en la tabla 2 se incluye un resumen de resultados de legibilidad para los folios detectados.

Folio	TÍTULO 1	TÍTULO 2	TÍTULO 3	PÁRRAGO 1	PÁRRAGO 2	TÍTULO 4	PÁRRAGO 3	PIE DE PÁGINA
5	SI	SI	SI	SI	SI	SI	SI	SI
6	SI	SI	NO	NO	NO	SI	NO	SI
7	SI	SI	SI	SI	SI	SI	SI	SI
10	SI	SI	NO	NO	NO	SI	NO	NO
11	SI*	SI*	NO*	NO*	NO*	SI*	NO*	NO*
12	SI	SI	NO	NO	NO	SI	NO	NO
13	SI	SI	NO	NO	SI	SI	NO	SI
14	SI	SI	SI	SI	SI	SI	SI	SI

Table 1: Elementos detectados

*. El folio se recupera rotado, por lo que se requiere un post-procesado manual para realizar esa evaluación.

Folio	Total elementos detectados correctamente	Precisión (%)
5	8/8	100.00
6	4/8	50.00
7	8/8	100.00
10	3/8	37.50
11	3/8	37.50
12	3/8	37.50
13	5/8	62.50
14	8/8	100.00
Promedio	5.25/8	65.62

Table 2: Resumen de resultados

En cuanto a las manchas y arrugas ignoradas. Se observa en los siguientes folios:

- **Folio 5.** Se consigue eliminar todas las manchas de café y la cruz de bolígrafo azul casi en su totalidad. Se observa que el texto en el que había café se ve más oscuro.
- **Folio 7.** Se consigue eliminar todas las marcas de bolígrafo rojo y prácticamente todas las marcas de bolígrafo azul. Se observan restos de la eliminación de estas últimas arriba a la derecha
- **Folio 11.** Se consigue eliminar todas las marcas de café. La cruz azul del medio no se elimina correctamente, se observan bastantes restos de la eliminación.
- **Folio 12.** Resultado prácticamente igual al anterior, además se ha incluido algo de ruido por estar el folio ligeramente doblado.
- **Folio 13.** Se han ignorado bastante bien las arrugas de la hoja, a excepción de una mancha que aparece arriba a la derecha.
- **Folio 14.** Resultado igual al anterior, pero no se observa ninguna mancha.

Se ha conseguido detectar tan sólo 8/16 folios, y en algunos de ellos existen muchas partes del documento no son legibles. Por otra parte se han conseguido eliminar las manchas y marcas de los documentos detectados prácticamente en su totalidad. Lo más mejorable es la detección del contorno de la hoja, se debe poder realizar un preprocesado más robusto frente a las distintas condiciones de luz y contraste con el fondo que facilite el detectado de los contornos. Además se ha observado que no se ha podido detectar ningún folio en imágenes con fondos claros, y es porque no se consigue encontrar un equilibrio entre el resultado de los contornos de las hojas y la introducción de ruido innecesario. Por último también se ha visto complejo segmentar las marcas de bolígrafo azul, sobre todo en las zonas en las que el color tenía más a un gris que a un azul.