UNIVERSIDAD AUTÓNOMA DEL ESTADO DE MÉXICO

UNIDAD ACADÉMICA PROFESIONAL TIANGUISTENCO

"Cálculo del índice de complejidad en documentos manuscritos para la segmentación de líneas de texto"

Tesis
Que Para Obtener el Grado de
Doctor en Ciencias de la Computación

Presenta
M.C.C. Miguel Ángel García Calderón

Tutor académico:
Dr. René Arnulfo García Hernández

Tutores adjuntos:
Dra. Yulia Ledeneva
Dr. Ángel Hernández Castañeda

TIANGUISTENCO, MÉX. Diciembre 2021

Resumen

Hasta el momento el conocimiento almacenado en los manuscritos antiguos no se ha utilizado en su totalidad debido a la falta de métodos robustos en el estado del arte para el reconocimiento de texto manuscrito.

La principal dificultad de los métodos para el reconocimiento de texto manuscrito es que se requiere que el texto se encuentre dividido en líneas. Además, los métodos para la Segmentación de Líneas de Texto (SLT) no han sido optimizados para procesar manuscritos antiguos.

La primera etapa de la SLT es la Localización de Líneas de Texto (LLT). En la SLT se han propuesto métodos que buscan los valores máximos locales en un histograma. El problema de estos métodos es que existen demasiados máximos locales y no es posible identificar cuáles conjuntos de máximos locales representan una línea de texto.

La segunda etapa de la SLT es la búsqueda de una ruta que permita separar las líneas de texto vecinas. Por un lado, el problema de los métodos actuales es que en algunos casos se realiza una búsqueda local de la ruta. Por otro lado, los métodos que realizan una búsqueda global de la ruta tienen problemas para encontrar una ruta entre trazos que se sobreponen.

Los problemas de las dos etapas conforman un valor de complejidad. La complejidad visual de un documento mansucrito antiguo para ser segmentado puede apreciarse por el humano experto, sin embargo, no existe en el estado del arte un método para calcular la complejidad.

En el estado del arte existen técnicas que permiten realizar una separación del cuerpo de letras y el espacio interlineal. Este trabajo se enfoca cuantificar la cantidad de información en el espacio interlineal para establecer un índice de complejidad. El índice de complejidad propuesto calcula la cantidad de información que aportan los trazos horizontales y verticales; además de la cantidad de información que aporta la tinta del documento y los valores del color del material de escritura.

Contenido

Página

Lista de figuras

Lista de tablas

CAPÍTULO 1.
Introducción

1.1 Antecedentes

Desde hace tiempo el hombre ha tenido la necesidad de comunicar, transmitir y almacenar sus necesidades, pensamientos y conocimiento. El conocimiento es almacenado para perdurar por generaciones (Rendón Rojas, 2005). El primer medio para transmitir el conocimiento de una generación a otra fue a través del lenguaje natural.

Desde que el hombre vivía en cavernas comenzó a almacenar el conocimiento mediante un lenguaje basado en dibujos. Con el paso de los siglos, las pinturas rupestres evolucionaron hasta la creación de las primeras formas de escritura en el año 3200 A.C (Baines et al., 2008). Al igual que las pinturas rupestres, la escritura cambia con el paso del tiempo; algunos lenguajes y estilos de escritura desaparecen y son tomados como base para nuevos estilos de escritura. La invención de la escritura ha permitido acumular y compartir el conocimiento.

Con el paso del tiempo se crearon y perfeccionaron moldes de trazos para replicar caracteres manuscritos (Bagley, 2004). Los primeros materiales utilizados para representar información de manera manuscrita fueron piedras, papiros, pergaminos, tablillas, cuero y papel. Cada material usado para escribir tiene diferentes propiedades de color y resistencia a la degradación (Trubek, 2017). En la figura 1.1 se muestran documentos escritos en diferentes materiales. Cada página tiene un fondo de color diferente debido a la degradación y al tipo de material (Bar-Yosef et al., 2009; Valizadeh & Kabir, 2012).

Figura 1.1. Ejemplo de documentos manuscritos con diferente fondo de color, materiales y épocas (Mauricio et al., 2016; *Voynich Manuscript, Beinecke MS 408, General Collection*, 1912).

Cada obra manuscrita antigua puede contener conocimiento histórico, teológico, cultural, literario y científico de una civilización o región (Gray, 1948; Győry Hory, 2008). Además, es natural el deterioro de los materiales de escritura debido al paso del tiempo. Por ello, es necesario proveer herramientas que faciliten la búsqueda y localización de información en documentos manuscritos antiguos y así, evitar dañar las obras durante su estudio (Mauricio et al., 2016).

Muchas bibliotecas y universidades del mundo están interesadas en dar acceso a documentos manuscritos mediante la indexación y creación de plataformas para la transcripción manual de documentos manuscritos con el objetivo de facilitar la búsqueda y recuperación de información (Causer &

Wallace, 2012; Mauricio et al., 2016). Un ejemplo de estas plataformas es el proyecto "Transcribed Bentham" en el que 1,009 hojas manuscritas fueron transcritas por 1,207 personas en un periodo de 6 meses (Causer & Wallace, 2012). Uno de los problemas de estas plataformas se presenta cuando una obra contiene diferentes estilos de escritura y diferentes lenguajes, por lo que se hace necesario que el humano conozca todos los lenguajes presentes en el documento.

Otro ejemplo es la piedra roseta que permitió la traducción de los jeroglíficos egipcios en el siglo XVII con más de un lenguaje (Medina Morán, 2011). En la Figura 1.2 se muestra un segmento de la imagen de la piedra roseta. Las primeras dos líneas contienen jeroglíficos egipcios y en la parte media escritura demótica y en la parte inferior griego antiguo.

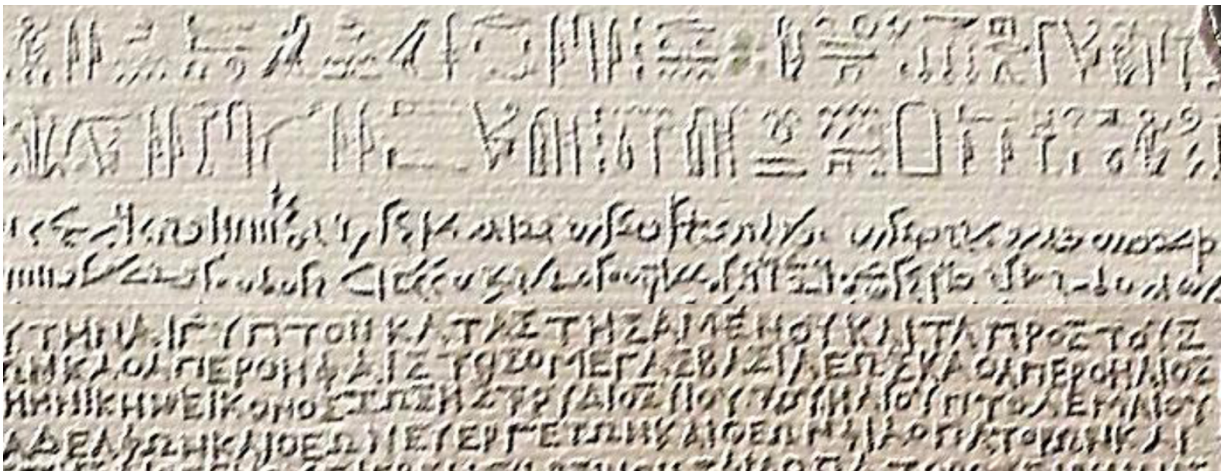


Figura 1.2. Ejemplo de documento que contiene más de un sólo lenguaje (Houston et al., 2004).

Para automatizar el proceso de indexación se debe considerar que los documentos pueden contener elementos que dificultan la transcripción, un ejemplo de esto es que cuando se cometía algún error de escritura el material no podía ser borrados ó reemplazado y se tenían dos opciones; tachar la palabra con el error o tachar el párrafo completo (Houston et al., 2004; Trubek, 2017; Wildgen, 2004). El texto tachado provoca que se conecten palabras completas de forma horizontal e incluso que se conecten líneas de texto vecinas, estos rayones hacen aún más complicada la Segmentación de Líneas

de Texto (SLT) incluso para el humano. En la figura 1.3 se muestra un ejemplo de documento con texto tachado.

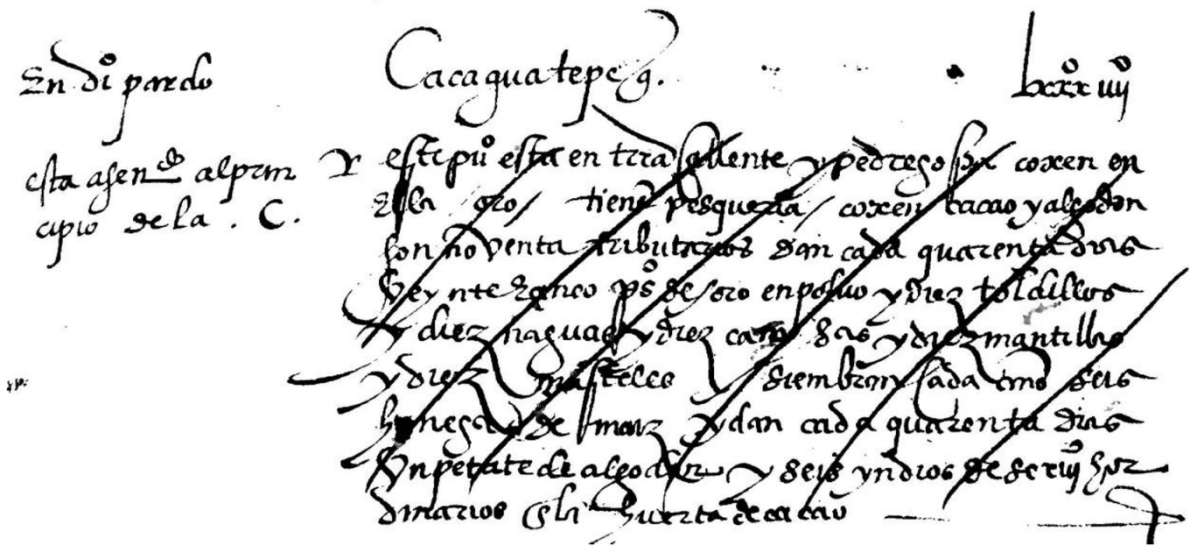


Figura 1.3. Ejemplo de documento manuscrito con errores de escritura que fueron tachados (García Castro, 2013).

Otra de las características presente en los documentos manuscritos antiguos es el color de tinta usada para trazar los caracteres. En la antigüedad se utilizaron pigmentos orgánicos para escribir documentos, posteriormente comenzaron a usar distintas mezclas de minerales, cada escribano o institución utilizaba diferentes pigmentos que permitieran identificar al autor del documento (Muñoz y Rivero, 1880). En la Figura 1.4 se muestra un ejemplo de documento que contiene mezclas de diferentes tintas y materiales.

Figura 1.4. Ejemplo de documento manuscrito escrito con diferentes materiales de tinta (Voynich Manuscript, Beinecke MS 408, General Collection, 1912).

La automatización de la indexación y la transcripción es extraordinariamente compleja en imágenes de documentos históricos porque tienen propiedades como degradación de la superficie de escritura, el material de la tinta, espacio interlineal variable y los trazos superpuestos de líneas de texto adyacentes.

Se sabe que la precisión del proceso de indexación y transcripción depende de qué tan bien se hayan localizado y segmentado las líneas del documento (Fischer et al., 2020; J. A. Sánchez et al., 2015; Mauricio et al., 2016; V. Romero et al., 2015). Por lo tanto, la Segmentación de Líneas de Texto (SLT) es una etapa crítica para otros sistemas.

La SLT contiene tanto localización como extracción. La Localización de Líneas de Texto (LLT) localiza patrones, mientras que la extracción de líneas de texto se busca una ruta dentro del espacio interlineal que permita separar dos líneas vecinas. En la figura 1.5 se muestran los elementos generales de las líneas de texto.

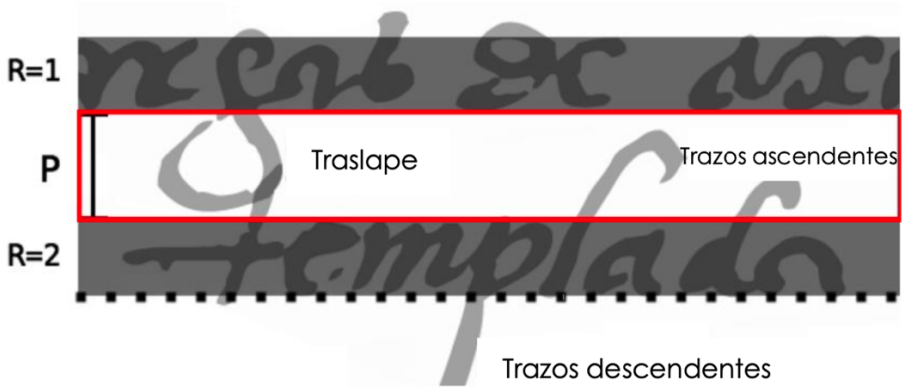


Figura 1.5. Elementos generales de las líneas de texto. El rectángulo marcado en color rojo (P) representa al espacio interlineal. $R1$ y $R2$ representan a los cuerpos de letras.

En el estado del arte se han identificado características de estilo independientes del lenguaje y época que influyen en la SLT como: líneas que se sobreponen, caracteres con trazos ascendentes o descendentes, tamaño de los caracteres, separación entre líneas de texto, separación entre palabras.

Al tener localizado el espacio interlineal es necesario realizar la búsqueda de una ruta que permita segmentar dos líneas de texto vecinas. El proceso tiene que lidiar con la cantidad de información que existe en el espacio interlineal. A mayor cantidad de información en el espacio interlineal, mayor será la complejidad para realizar la búsqueda de una ruta que permita segmentar dos líneas de texto vecinas.

Además, se ha identificado que el rendimiento de los métodos para la LLT depende de las variables presentes en el espacio interlineal como: cantidad de trazos ascendentes y trazos descendentes; qué tan estrecho es el espacio entre dos líneas de texto vecinas, trazos superpuestos y trazos coincidentes (Likforman-Sulem et al., 2006; Saabni et al., 2014).

Si se contara con una técnica para poder calcular la complejidad de un documento manuscrito se podría seleccionar al método óptimo para realizar la búsqueda de una ruta, que permita realzar la búsqueda con el menor costo computacional. En la figura 1.6 se muestran dos ejemplos de espacio interlineal con diferente nivel de complejidad visual.

En la figura 1.6a se muestra un documento con una complejidad visual baja. El método de Ptak (línea punteada) es adecuado para segmentar documentos con una complejidad baja (Ptak et al., 2017). La figura 1.6b muestra un documento con una complejidad mayor donde el método de Arvanitopoulus (Arvanitopoulos & Süsstrunk, 2014) supera al método de Ptak. El método de Arvanitopoulus (línea sólida) es más adecuado para documentos con una complejidad mayor, pero requiere una mayor cantidad de recursos computacionales en comparación del método de Ptak (Ptak et al., 2017) por lo que hace falta un criterio que permita seleccionar el método más adecuado para la SLT.

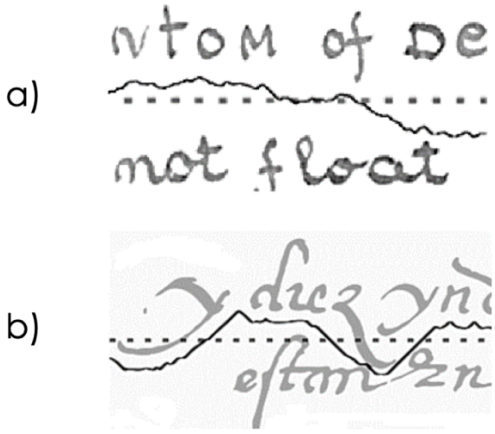


Figura 1.6. Búsqueda de una ruta que permita segmentar dos líneas de texto vecinas. La ruta representada con una línea punteada fue generada por un método heurístico (Ptak et al., 2017). La ruta representada como una línea continua fue generada por un método con una mayor carga computacional (Arvanitopoulos & Süsstrunk, 2014).

Hay evidencia de que la cantidad de información en el espacio interlineal influye directamente en la complejidad. Se puede observar en el estado del arte que se han realizado esfuerzos por eliminar o reducir la información en el espacio interlineal (Demır & Özkaya, 2020; Pearlsy & Sankar, 2020; Ptak et al., 2017).

Los métodos del estado del arte no mantienen el mismo comportamiento con documentos visualmente menos o más complejos, solo mantienen el

comportamiento con los documentos usados en su etapa de experimentación. Por lo tanto, los métodos del estado del arte no son aptos para ser usados en un escenario real (Demır & Özkaya, 2020; Likforman-Sulem et al., 2006; Saabni et al., 2014).

Se ha tratado de investigar por qué los métodos del estado del arte no mantienen su rendimiento cuando se usan para procesar colecciones de documentos nuevas o más antiguas. Arivazhagan (Arivazhagan et al., 2007) propone un índice de complejidad extrínseco basado en el resultado de la evaluación con datos generados por el humano y un método para la Segmentación de líneas de texto.

Los índices extrínsecos consideran información ajena al objeto de estudio (Lewis, 1983) y dificultan el proceso de comparación de objetos de la misma clase porque las propiedades extrínsecas varían. Por el contrario, los índices intrínsecos sólo consideran información contenida en el objeto de estudio (Sider, 1996). Dos objetos de estudio pueden compararse si tienen las mismas propiedades intrínsecas.

Después de un análisis del estado del arte no se encontró un índice de complejidad intrínseco. La creación de un índice de complejidad intrínseco que considere las características presentadas a lo largo de este capítulo sería ideal para documentos manuscritos antiguos porque no se requiere la creación de datos adicionales generados por el humano experto. Además, facilitaría la comparación de características entre documentos de diferentes lenguajes, estilos y épocas.

Un índice de complejidad intrínseco permitiría seleccionar el método más adecuado para realizar la Segmentación de Líneas de Texto de acuerdo con la complejidad del documento.

1.2 Planteamiento del problema

Después de un análisis de las características de los documentos manuscritos sabemos que no es posible seleccionar un método del estado del arte para su aplicación en un escenario real. Además, los rangos de complejidad en los

que puede trabajar cada método del estado del arte se desconocen por lo que no se pueden aprovechar las fortalezas de cada uno.

En este trabajo se resuelve el siguiente problema ¿cómo calcular el índice de complejidad intrínseco de un documento manuscrito para poder seleccionar el método del estado arte que realice de manera óptima la segmentación de líneas de texto?

1.3 Justificación o motivación

Es importante tener un ranking de complejidad en las colecciones de documentos y conocer los límites en donde puede trabajar cada método del estado del arte con el objetivo de enfocar los esfuerzos de investigación en diferentes rangos de complejidad. Actualmente, la mayoría de las investigaciones se enfocan en un solo lenguaje, las investigaciones multilenguaje se enfocan en documentos cuya complejidad visual es similar. Contar con un índice de complejidad intrínseco para la SLT permitiría ordenar las colecciones de documentos con lo cual se podría seleccionar el mejor método de acuerdo con la complejidad del documento a segmentar.

La indexación automática de documentos manuscritos es de gran ayuda a historiadores y paleógrafos para que puedan dedicar mayor tiempo al estudio del contenido de los documentos.

Al analizar el estado del arte sabemos que la tarea abordada en este trabajo no se ha resuelto (Fischer et al., 2020; Saabni et al., 2014; Zohrevand et al., 2019). Además, los esfuerzos actuales para medir la complejidad de un documento manuscrito miden la complejidad de forma extrínseca (Saabni et al., 2014), por lo tanto, siempre se requiere la intervención de un humano experto para generar datos de entrenamiento y datos de prueba.

Ya se cuenta con un análisis de las características intrínsecas de los documentos manuscritos por lo que se puede proponer un índice intrínseco considerando las características presentadas a lo largo de este capítulo.

1.4 Objetivo General

Proponer, implementar y desarrollar un índice de complejidad que permita organizar las colecciones de documentos manuscritos y métodos para la SLT para tener un criterio de selección del método óptimo a usar para la segmentación de líneas de texto manuscrito.

1.4.1 Objetivos particulares

- Proponer un índice de complejidad intrínseco independiente del lenguaje para la tarea de SLT que sea aplicable por página.

- Realizar un análisis de los métodos del estado del arte para identificar técnicas estándar.

- Implementar los métodos del estado del arte que hasta el momento tienen mejores resultados para la SLT y combinarlos en un método híbrido basado en el índice de complejidad que se propone en este trabajo.

1.5 Hipótesis

Existen técnicas que permiten realizar una separación del cuerpo de letras y el espacio interlineal. Por lo tanto, la hipótesis de este trabajo consiste en que si se calcula la cantidad de información que aportan los trazos horizontales y verticales; además de la cantidad de información que aporta la tinta del documento y los valores del color del material de escritura, entonces será posible calcular el índice de complejidad intrínseco de un documento manuscrito.

Cuando el humano realiza un análisis visual de un texto manuscrito no necesita comprender su contenido porque puede identificar que existen obstáculos que se tienen que evitar para realizar la segmentación. De esta manera, si se calcula la cantidad de información que aportan los trazos horizontales y verticales; en el espacio interlineal (espacio que hay entre dos cuerpos de letras) además de la cantidad de información que aporta la tinta del documento y los valores del color del material de escritura, entonces será posible calcular el índice de complejidad intrínseco de un documento manuscrito.

1.6 Estructura de la tesis

En el capítulo uno, correspondiente a la introducción, se describen las características que anteceden al problema abordado en este trabajo. También se describen los conceptos fundamentales necesarios para poder comprender el problema abordado en este trabajo. Además, se realiza una descripción del proceso actual que se sigue para poder realizar la segmentación de líneas de texto manuscrito. Se presenta la hipótesis que se comprobó en este trabajo. Al final de esta sección se describe el objetivo general y los objetivos particulares que se completaron con este trabajo.

En el capítulo dos, correspondiente al marco teórico, se presentan todos los conceptos necesarios para poder comprender el método propuesto.

El capítulo tres presenta el estado del arte actual para el problema abordado en este trabajo. En esta sección también se presenta una comparación de los métodos actuales dependiendo de su enfoque (métodos supervisados y métodos no supervisados). Además, se describen los corpus que se han usado actualmente para la tarea de segmentación de líneas de texto manuscrito.

En el capítulo cuatro está descrito el método propuesto basado en la hipótesis planteada en el capítulo uno. En este capítulo se presentan los algoritmos de cada una de las etapas del método propuesto en este trabajo. Además, se presenta el diagrama de componentes para describir el funcionamiento general del método propuesto.

En el capítulo cinco se describe la experimentación realizada para validar el índice de complejidad propuesto.

En el capítulo seis muestra la validación de nuestra hipótesis tomando como referencia los resultados de la etapa de experimentación. También se presentan nuevos objetivos para trabajo futuro.

CAPÍTULO 2.
Marco Teórico

En este capítulo se presentan los conceptos necesarios utilizados en el estado del arte, los conceptos básicos del procesamiento de imágenes usados a lo largo de esta investigación, así como los conceptos utilizados para proponer e implementar el método propuesto.

2.1 Imagen digital

Es una representación de objetos o escenas reales o imaginarios. La representación visual de la escena se realiza mediante una función bidimensional $f(x, y)$, donde (x) y (y) representan una coordenada en la función bidimensional, y se almacena un valor de brillo en cada coordenada durante el proceso de captura de imágenes. Cuando se almacena la imagen, finaliza el proceso de creación de la imagen digital (Gonzalez et al., 1996).

Las imágenes digitales se pueden generar de dos formas: usando dispositivos digitales como escáneres o cámaras digitales, o usando programas de computadora llamados editores de mapas de bits (González et al., 1996).

2.2 Píxel

Es el elemento básico que constituye una imagen digital. Cada píxel tiene un valor de brillo (Burger & Burge, 2008; González et al., 1996). El número total de píxeles se puede calcular multiplicando el ancho por la altura de la imagen (x, y), donde (x) es el número de columnas y (y) el número de filas.

El proceso de digitalización asigna un valor y una posición a cada píxel para formar una matriz bidimensional. La figura 2.1 muestra un ejemplo de representación de píxeles en una imagen digital.

La matriz generada cuyo origen es la esquina superior izquierda conforma un sistema de coordenadas (Martinsanz et al., 2007). Todas las coordenadas de la imagen tienen valores positivos, la posición en el eje (x) va de izquierda a derecha y la posición en el eje (y) va de arriba a abajo. La figura 2.1 proporciona una imagen con un sistema de coordenadas en la imagen digital. En una imagen rgb los valores mínimos y máximos se encuentran en el rango $[0 - 255]$.

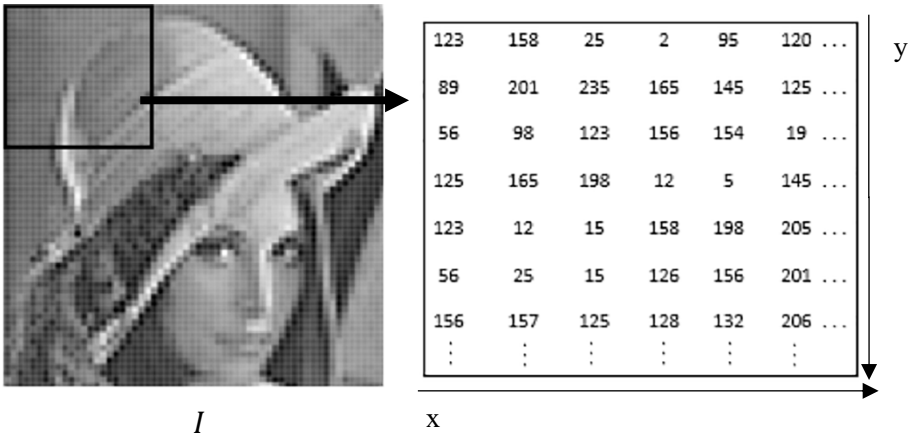


Figura 2.1. Representación matricial de una imagen digital (Gonzalez et al., 1996). Cada elemento de la matriz es un píxel.

2.3 Ruido

Los cambios en el proceso de impresión, el proceso de escaneo, los cambios debido a la antigüedad del documento, manipulación y la captura de la imagen producirán cambios aleatorios. En el procesamiento de imágenes

digitales, estos cambios se denominan ruido. El ruido se define como un conjunto de valores aleatorios que no corresponden con la realidad.

El ruido es generado por el dispositivo de captura de la imagen y el ruido aumenta la complejidad para el tratamiento de las imágenes digitales (Burger & Burge, 2008).

Cada dispositivo de captura agrega un patrón de ruido diferente, y en la Figura 2.2 se muestra un ejemplo de ruido generado por el dispositivo de captura.

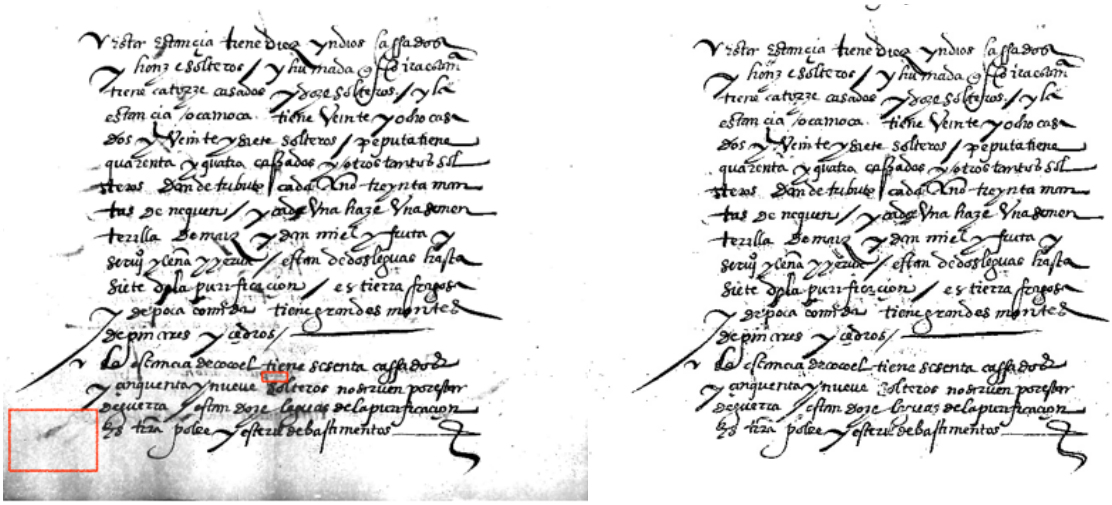


Figura 2.2. Ejemplo de ruido en imágenes digitales (García Castro, 2013). La imagen de la izquierda fue capturada con un escáner con lámpara gastada, que agregó y modificó las características originales del documento sin modificar su estructura. La imagen de la derecha muestra el mismo documento escaneado con el escáner en condiciones óptimas.

2.4 Procesamiento digital de imágenes

Son un conjunto de técnicas que permiten la adquisición, representación, mejora y procesamiento de imágenes digitales. El objetivo de estas técnicas es procesar imágenes digitales para facilitar el proceso de búsqueda de información (González et al., 1996).

2.5 Imágenes en escala de grises

Una imagen digital en escala de grises es un tipo de imagen digital en el cuál se tiene un solo canal de color (Gonzalez et al., 1996). En las imágenes digitales

binarias los píxeles sólo toman un valor, blanco o negro (0 o 1) (Burger & Burge, 2008). La Figura 2.3 muestra ejemplos de una imagen en color y una imagen en escala de grises.

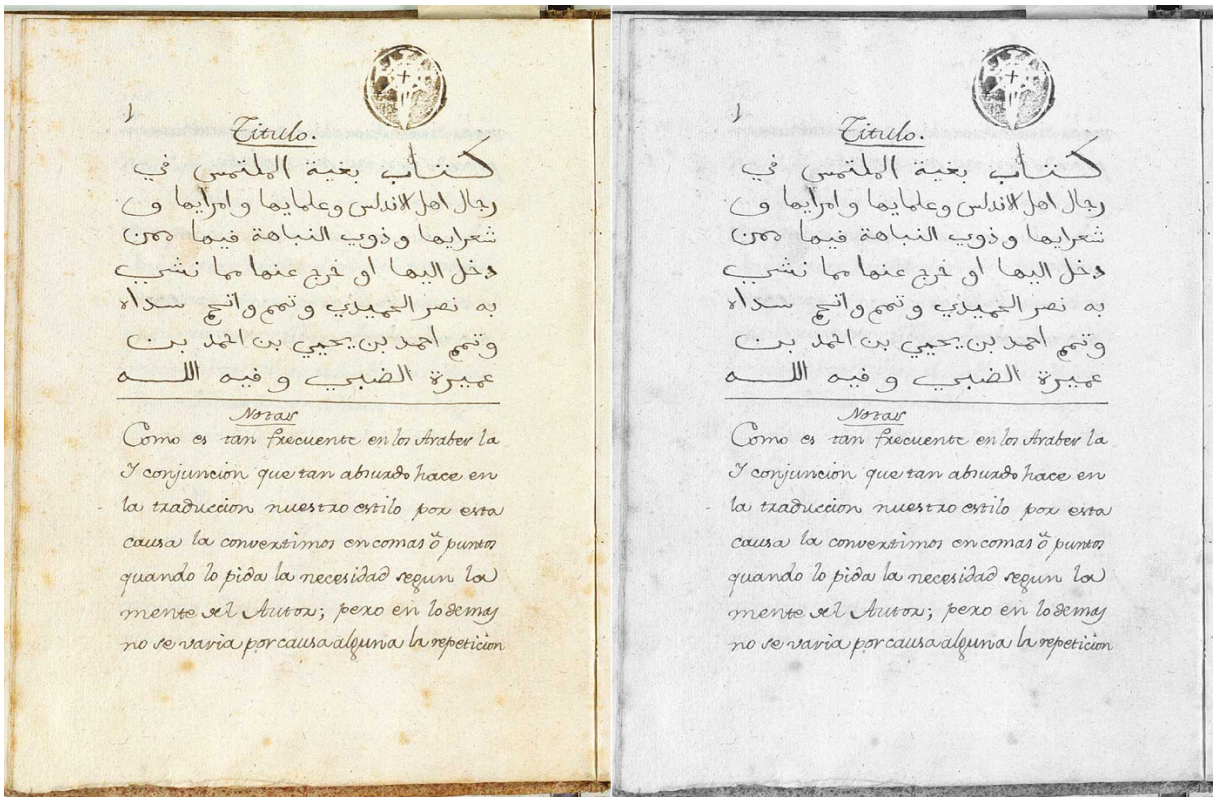


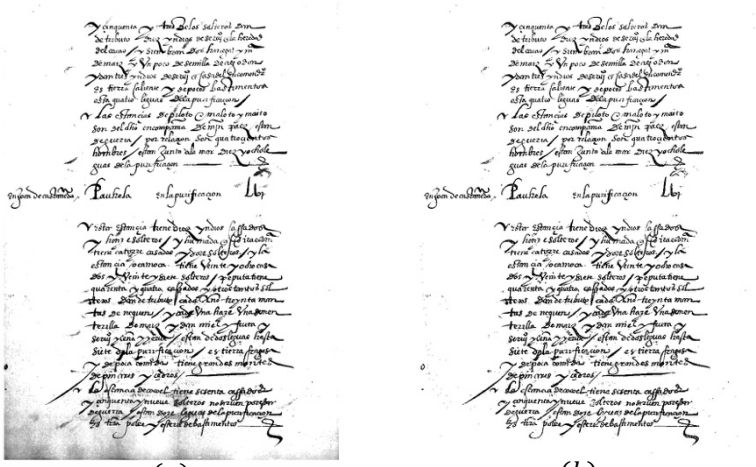
Figura 2.3. La imagen de la izquierda muestra el documento escaneado, conservando la configuración de color del documento (Gatos et al., 2009). La imagen de la derecha muestra un documento que se ha procesado para producir una imagen en escala de grises. La imagen de la derecha contiene una tercera parte de la información.

2.6 Imágenes binarias

Una imagen binaria es una imagen digital con solo 2 niveles de color, generalmente en blanco y negro. En la Figura 2.4 se muestra un ejemplo de la conversión de una imagen en escala de grises (a) a una imagen binaria (b). El proceso de binarizado permite separar el objeto del fondo de la imagen (O'Gorman et al., 2008).

A través del proceso de binarización, los valores de cada píxel se reducen a dos valores, generalmente 0 y 1 (blanco y negro). Utilizan un bit por píxel para

el almacenamiento y la codificación del color. Por lo general, uno de los colores se usa para el fondo y el otro color se usa para representar objetos (Burger & Burge, 2008). Los métodos de binarización se utilizan para reducir el ruido en imágenes de documentos antiguos (Peng et al., 2016).



(a) (b)

Figura 2.4. Conversión de una imagen en escala de grises (a) en una imagen binarizada (b) para reducir la cantidad de información a procesar (García Castro, 2013).

2.7 Operaciones geométricas

La geometría es una rama de las matemáticas que se puede utilizar en el procesamiento de imágenes digitales para modificar imágenes, como acercar, alejar, trasladar y rotar (González et al., 1996).

2.7.1 Traslación

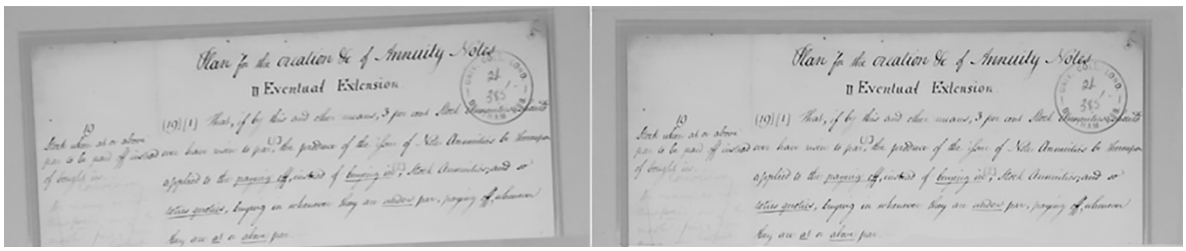
El operador de traducción realiza una transformación geométrica en la imagen. El proceso de traslación de imagen digital vuelve a dibujar la imagen (a) en la nueva posición de la imagen de salida (b). Es el proceso de cambiar la posición de la imagen sin cambiar su rotación o tamaño (Burger & Burge, 2008). Como se puede ver en la Figura 2.5, la imagen (a) se desplaza hacia la derecha y hacia arriba para generar la imagen (b).

(a) (b)

Figura 2.5. Traslación de una imagen digital 100 pixeles sobre el eje x y 50 píxeles sobre el eje y (Gonzalez et al., 1996).

2.7.2 Rotación

Es el proceso de rotar una imagen digital alrededor de un centro predefinido y en un ángulo predefinido (Burger & Burge, 2008). Un ejemplo de rotación de imagen se muestra en la Figura 2.6, que muestra una imagen de un documento escrito a mano escaneado en un ángulo oblicuo (a) y una imagen (a) con un ángulo corregido en una imagen (b).



a b

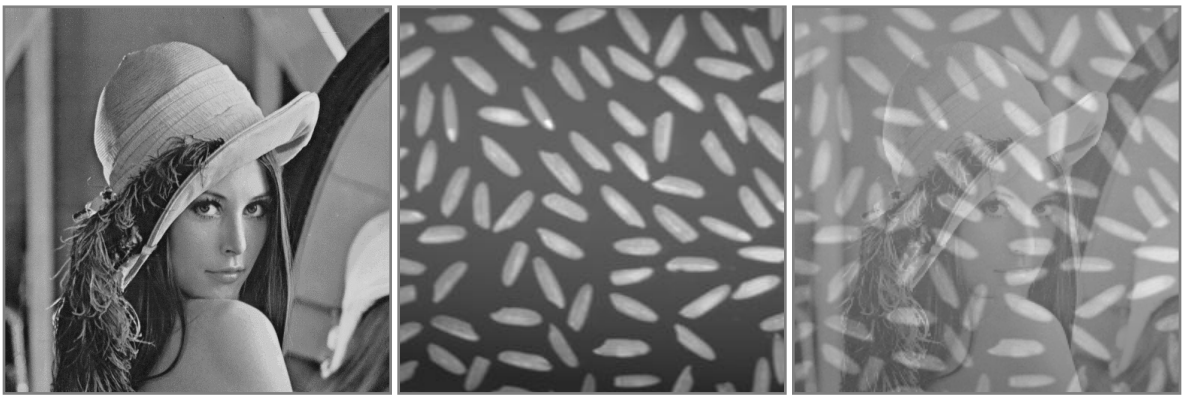
Figura 2.6. Rotación de una imagen digital. La imagen (a) muestra la imagen original y la imagen (b) muestra el resultado después de rotar la imagen (a) 3 grados (Mauricio et al., 2016).

2.8 Mezcla alfa

Es una técnica para mezclar dos imágenes (a) y (b) con un umbral de transparencia (Burger & Burge, 2008). Durante este proceso una imagen (a) es cubierta por una imagen (b), el valor de transparencia se controla con la variable $\propto$ de la siguiente forma

$$\mathrm{Alpha}(I, w, \alpha)^r = \mathrm{a}(u + w) + (1{-}\propto) \cdot \mathrm{b}(u + w)$$

donde $0 \leq \propto \leq 1, \propto = 0.5, u$ es la posición en el eje x y r es la cantidad de veces que se aplica la mezcla alfa.



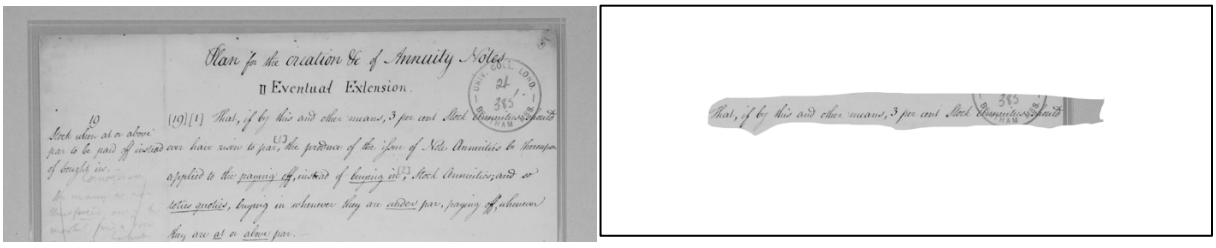
(a) (b) (c)

Figura 2.7. Ejemplo de procesamiento usando la técnica de mezcla alfa .

2.9 Segmentación

Es el proceso de extraer objetos de una imagen digital de forma manual o automática. El proceso de segmentación suele ser la etapa más delicada y difícil de cualquier sistema de procesamiento de imágenes (Gomez-Allende & Gómez-Allende, 1993; Gonzalez et al., 1996).

El proceso de segmentación se realiza con el propósito de simplificar la imagen para facilitar su análisis. En la imagen (a) de la Figura 2.9 se muestra un ejemplo de segmentación de imágenes. El proceso de segmentación tiene diferentes objetivos, en la imagen (b) de la Figura 2.8 se proporciona un ejemplo de segmentación manual de una línea de texto manuscrito.



a *b*

Figura 2.8. Ejemplo de segmentación de imágenes. La imagen *b* muestra el resultado de la segmentación manual de una línea de texto de la imagen *a* (Mauricio et al., 2016).

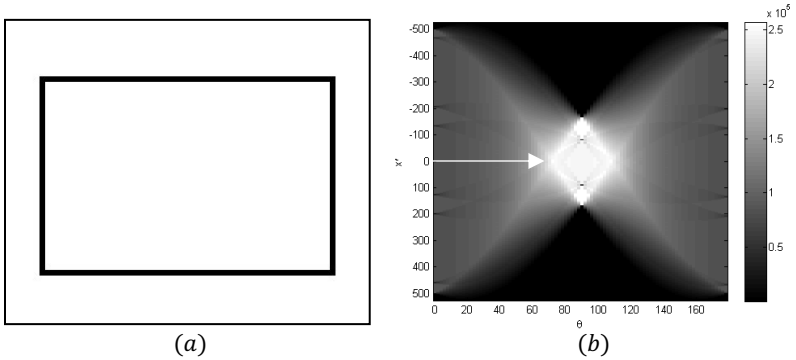
2.10 Transformada de Radon

Es una función utilizada en muchas disciplinas, incluida la cartografía por radar, las imágenes geofísicas y las captura imágenes médicas (Helgason, 1999). En el trabajo propuesto en (Helgason, 1999), se señala que la transformada de Radon consiste en mapear la función $f(x,y)$ con $(x,y) \in R2$ a una función $Rf(t,\theta)$ con $t \in R$ y $\theta \in (0,\pi)$ se define mediante la siguiente fórmula propuesta en (Helgason, 1999):

$$Rf(t,\theta) = f(x,y)\delta(t - xcos(\theta) + y\sin(\theta))dxdy$$

La transformada de Radon es una función que asigna un valor numérico a cada elemento de un grupo de líneas para permitir identificar el nivel de inclinación de las imágenes.

En la Figura 2.9a se muestra una figura con un ángulo de inclinación igual a 0. En la Figura 2.9b se muestra la trasformada de Radon del rectángulo, las intersecciones entre líneas indican el ángulo de cada imagen.



(a) (b)

Figura 2.9. Al analizar el resultado de la transformada de Radon (b) se puede apreciar que solo hay dos intersecciones entre las líneas las intersecciones se presentan en el valor cero, esto indica que la imagen (a) tiene un valor de inclinación igual a cero.

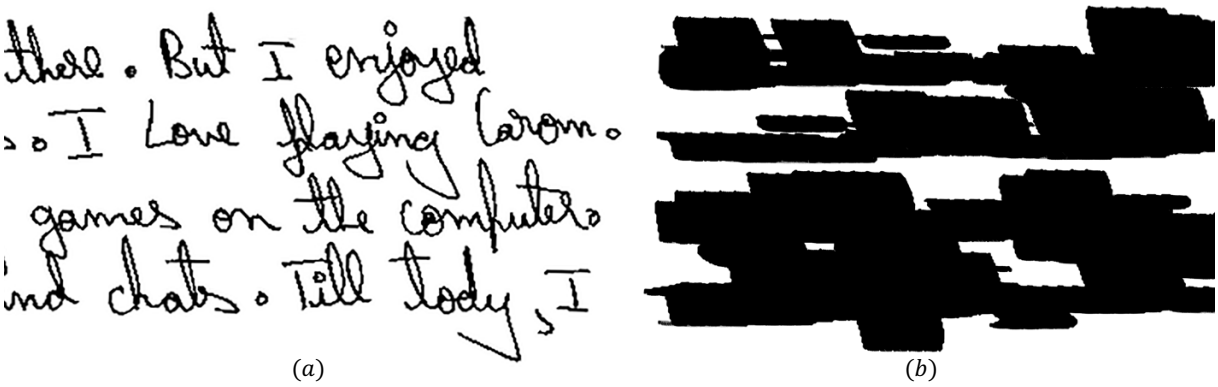
2.11 Mapa de energía

El mapa de energía es una técnica de caracterización de imágenes que permite rellenar los espacios en blanco entre caracteres y palabres (Du et al., 2009).

Esta técnica es utilizada en la etapa de preprocesamiento de documentos manuscritos (Kesiman et al., 2016; Koppula & Negi, 2014; Nicolaou & Gatos, 2009).

Se han propuesto métodos para la extracción de mapas de energía de documentos manuscritos, la técnica más simple de extracción de mapa de energía c

El objetivo de la extracción de un mapa de energía es eliminar los espacios horizontales vacíos entre caracteres (Kesiman et al., 2016). En la Figura 2.10 se muestra un ejemplo de la extracción del mapa de energía de una imagen digital que contiene texto manuscrito.

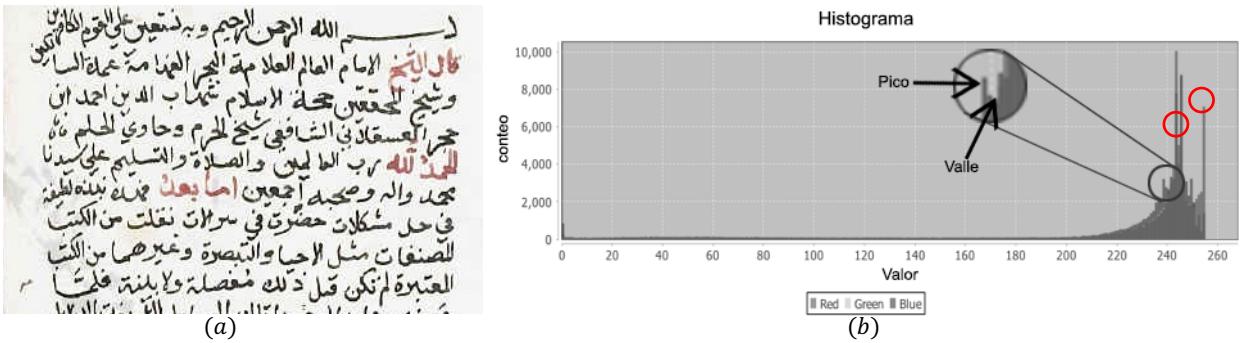


(a) (b)

Figura 2.10. Ejemplo de extracción del mapa de energía de una imagen que contiene texto (Gatos et al., 2009).

2.12 Histograma

Un histograma de una imagen digital es una representación de la frecuencia de los valores de intensidad presentes en una imagen (Gonzalez et al., 1996). En la Figura 2.11 se muestra el histograma de una imagen digital. Los elementos por los que está compuesto un histograma son los picos y valles. Un pico es un valor que está por encima de sus vecinos. Un valle es un valor que está por debajo de sus vecinos. En la imagen (b) se muestra encerrado en un círculo un pico y un valle del histograma.



(a) (b)

Figura 2.11. Histograma de color de una imagen. En la imagen (a) se muestra un ejemplo de documento manuscrito en idioma Árabe. En la imagen (b) se muestra el histograma del documento de ejemplo (Gatos et al., 2009).

2.13 Histograma de proyección horizontal

Un Histograma de Proyección Horizontal (HPH) es una presentación unidimensional de una imagen bidimensional. Los valores mostrados en el histograma de proyección vertical representan la densidad de la distribución de la información contenida en la imagen digital (O'Gorman et al., 2008; Ptak et al., 2017).

Los Histogramas de Proyección Vertical (HPV) son usados para calcular la inclinación de una imagen (Bagdanov & Kanai, 1998), reducción de ruido en imágenes de documentos digitales (A. Prachanucroa & S. Phongsuphap, 2013), identificación de autores en documentos manuscritos (Biswas & Das, 2012), entre otros. Para una imagen P de un documento con (x) pixeles de ancho y (y) pixeles de alto columnas se puede extraer con la Ecuación 2.1 el histograma de proyección vertical (Likforman-Sulem et al., 2006).

$$HPV(p) = \sum_{1 \le x \le m} f(x, y)$$

Ecuación 2.1 Ecuación para el cálcudo del histograma de proyección vertical en un documento de 'm' filas y 'n' columnas.

En la Figura 2.12 se muestra el histograma vertical de una imagen digital de un documento manuscrito. Este histograma es usado en el estado del arte para localizar la posición de cada línea buscando los picos del histograma.

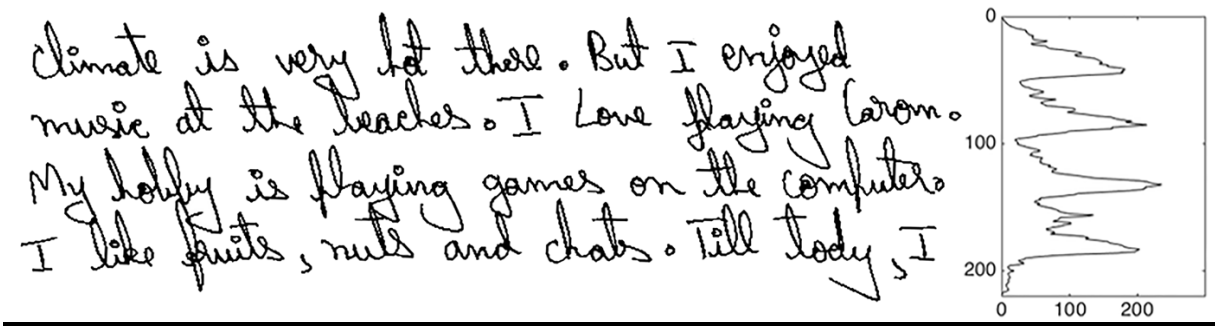


Figura 2.12. En la imagen del lado izquierdo se muestra un documento a analizar y en la imagen del lado derecho se proporciona el histograma de proyección horizontal del documento (Gatos et al., 2009).

2.14 Resumen del capítulo

A lo largo de este capítulo se han descrito los elementos básicos que conforman una imagen digital. Además, se han descrito los elementos que intervienen en la representación de las imágenes digitales. Se mencionan las variaciones que modifican el contenido de las imágenes. Por último, se presentan técnicas de representación unidimensional para el procesamiento de imágenes digitales.

CAPÍTULO 3.
Estado del Arte

En este capítulo se presenta un análisis de los métodos del estado del arte para la Localización de líneas de Texto (LLT) y Segmentación de Líneas de Texto (SLT) en documentos manuscritos.

3.1 Preprocesamiento

Diversos autores del estado del arte concuerdan en que el proceso de binarización, la corrección de inclinación y el filtrado para reducción de ruido son procesos fundamentales en la segmetación de líneas de texto (A. Prachanucroa & S. Phongsuphap, 2013; Arica & Yarman-Vural, 2001; B. Gatos et al., 2009; Bagdanov & Kanai, 1998; I. Pratikakis et al., 2016; Mauricio et al., 2016).

3.2 Métodos de abajo hacia arriba para la LLT

Los métodos de abajo hacia arriba para la Localización de Líneas de Texto agrupan pixeles o caracteres para crear patrones de líneas de texto(Koppula & Negi, 2014; Saabni et al., 2014; Valy et al., 2016).

Los métodos de abajo hacia arriba han sido diseñados para procesar documentos que contienen grupos de líneas de texto. La Figura 3.1 muestra un ejemplo de documento ideal para los métodos de abajo hacia arriba. Estos métodos no pueden procesar documentos con líneas de texto que se intersectan verticalmente.

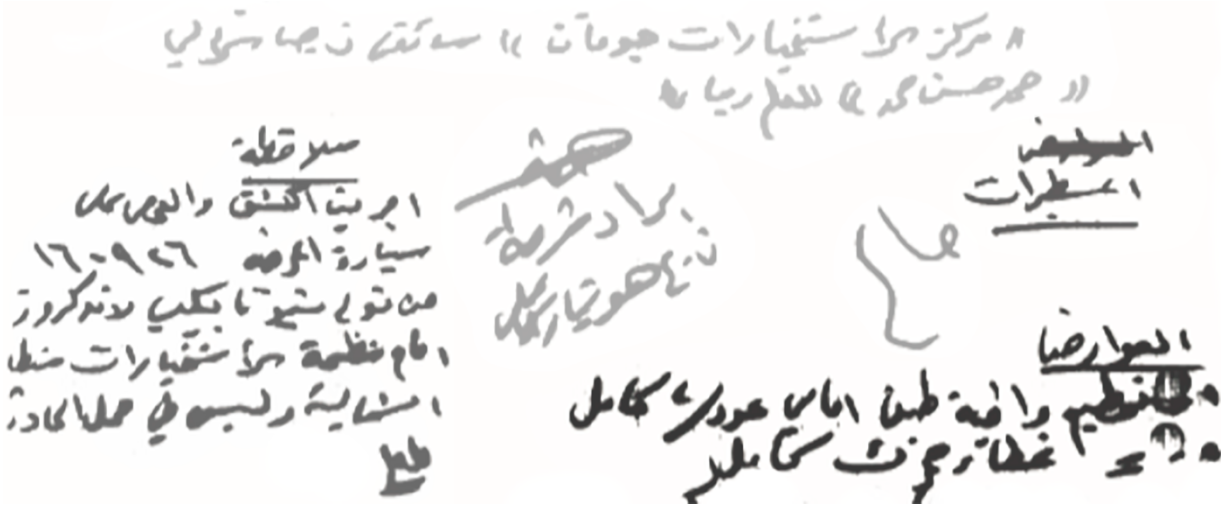


Figura 3.1. Documentos donde los métodos de arriba hacia abajo muestran un mejor rendimiento (Du et al., 2009).

3.3 Métodos de arriba hacia abajo para la Localización de Líneas de Texto (LLT)

Los métodos de arriba hacia abajo para la localización de líneas de texto utilizan extracción de Histograma de Proyección Horizontal (HPH), extracción de mapa de energía y aprendizaje supervisado.

3.3.1 Métodos basados en aprendizaje

Los métodos de esta categoría necesitan una muestra de entrenamiento para la LLT y una muestra de entrenamiento para generar un modelo que permita aprender a generar trazos para separar líneas de texto vecinas (Q. N. Vo & G. Lee, 2016). La desventaja de los métodos basados en aprendizaje es que el modelo generado tiende a ajustarse a la muestra de entrenamiento.

3.3.2 Métodos basados en el perfil de proyección horizontal

En caracteres impresos la extracción de un perfil de proyección horizotal ha demostrado tener el mejor resultado para la Localización de Líneas de Texto (Jaekyu Ha et al., 1995). Estos métodos no pueden ser aplicados directamente a textos manuscritos porque necesitan una separación clara entre cada línea de texto. Caracteres que se tocan y la inclinación de los documentos afectan el rendimiento de estos métodos.

Los métodos de esta categoría identifican los picos en un PPH para identificar una separación entre líneas de texto. Hasta el momento no ha sido posible aplicar esta técnica en documentos con líneas de texto que se intersectan verticalmente (Figura 3.2) porque todos tienen diferente anchura y longitud.

Los métodos de esta categoría necesitan el ajuste manual de un conjunto de umbrales definidos empíricamente y específicos para cada colección de documentos (Arvanitopoulos & Süsstrunk, 2014; Kesiman et al., 2016b; Ptak et al., 2017). Los métodos en este enfoque toman la ubicación de un pico, sin embargo, documentos con líneas de texto que se intersectan como el de la figura 1.6 presentan más de un pico por cada línea de texto. Por lo tanto, con una búsqueda de picos en un PPH se puede calcular que existen cinco líneas de texto manuscrito para el documento presentado en la figura 1.6

(Arvanitopoulos & Süsstrunk, 2014; Kesiman et al., 2016b; Ptak et al., 2017) propusieron un método basado en la extracción de PPH para estimar la posición de cada línea de texto identificando los valores de los máximos locales. Sin embargo, cuando las líneas de texto se superponen o se tocan, estos métodos no pueden estimar la posición del espacio interlineal (mínimo local) entre líneas de texto adyacentes. En el trabajo de (Peng et al., 2016), solo se llevó a cabo la etapa LLT. La aplicación de un método basado en PPH requiere que el texto esté representado horizontal, por lo que es imposible aplicarlo directamente al documento que se muestra en la Figura 3.1.

El método Arivazhagan se destaca porque tiene un menor costo computacional y se basa en métodos heurísticos (Arivazhagan et al., 2007). Este método puede dividir rápidamente líneas de texto en un documento

escrito a mano, pero requiere documentos con espacio interlineal libre de intersecciones, por lo tanto, no puede utilizarse para procesar documentos manuscritos antiguos.

La Figura 3.2 muestra dos ejemplos de documentos que se pueden segmentar utilizando el método propuesto por Arivazhagan (Arivazhagan et al., 2007).

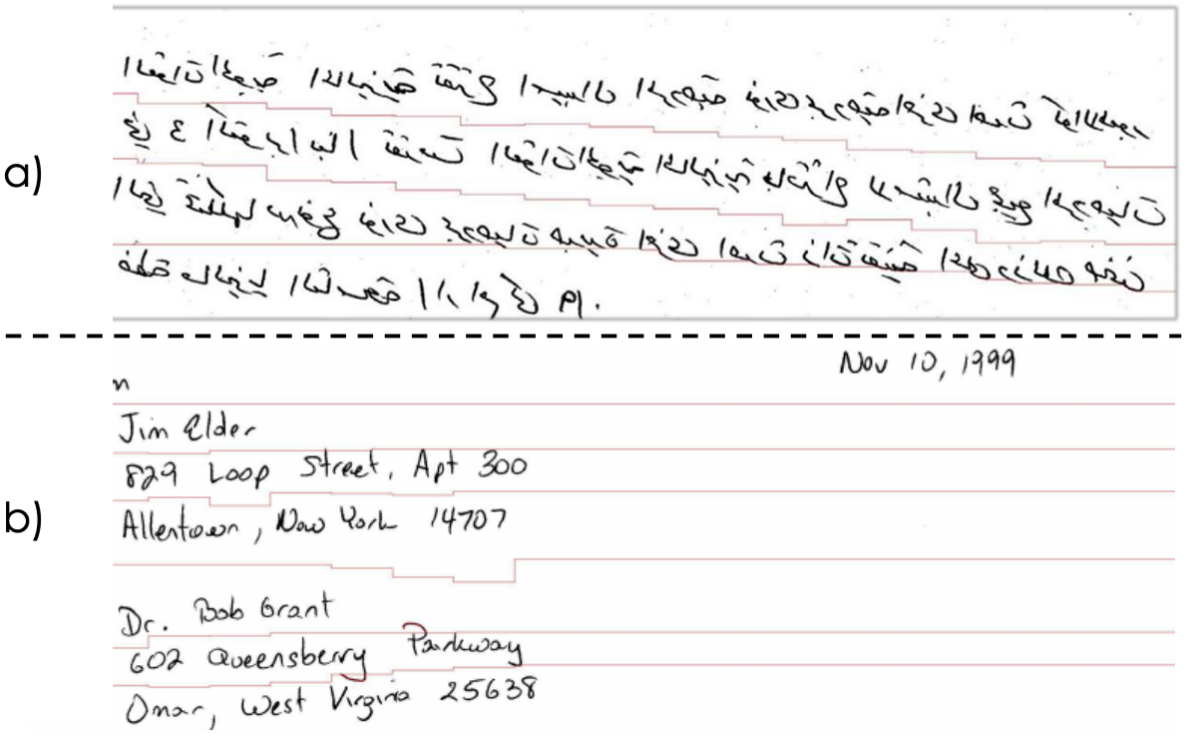


Figura 3.2. Ejemplo de documentos en donde el método propuesto por Arivazhagan puede ser aplicado (Arivazhagan et al., 2007). Para que este método tenga exito en la segmentación requiere la ausencia de traslapes entre líneas de texto adyacentes.

3.3.3 Métodos basados en la extracción de un mapa de energía

La extracción de un mapa de energía permite eliminar espacios en blanco entre palabras y caracteres (Du et al., 2009; Liwicki et al., 2007; Saabni et al., 2014); al extraer un PPH de un mapa de energía se puede apreciar una mayor diferencia entre máximos y mínimos locales en el PPH.

(Du et al., 2009; Kesiman et al., 2016; Liwicki et al., 2007; Nicolaou & Gatos, 2009; Saabni et al., 2014) presentaron trabajos en donde realizan una reducción de los espacios en blanco entre cada carácter y cada palabra con un método de extracción de mapa de energía basado en un operador gradiente (ME-gradiente) o una función específica para una colección de documentos (ME-F) (Du et al., 2009), etc. Para hacer esto, la imagen del documento es filtrada o trasladada sobre la imagen original para obtener un mapa de energía como el que se muestra en la Figura 3.3.

Después de aplicar este proceso, algunos de los trabajos del estado del arte se agrupan conjuntos de pixeles para encontrar los patrones de líneas de texto (Du et al., 2009). En los trabajos de (Arvanitopoulos & Süsstrunk, 2014; Ptak et al., 2017), se recomienda extraer el PPH del mapa de energía. Sin embargo, al aplicar los métodos actuales de extracción de mapas de energía en documentos con líneas de texto que se intersectan verticalmente, es imposible distinguir el espacio interlineal (Ver Figura 3.3).

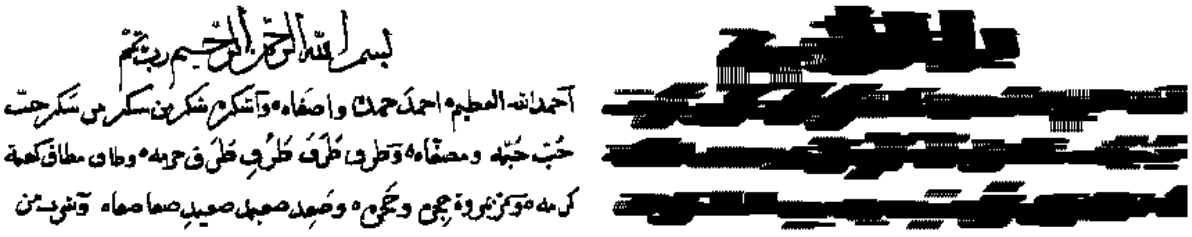


Figura 3.3. Mapa de energía generado usando el operador morfológico dilatación para manuscritos (Kesiman et al., 2016). En esta imagen es posible observar que los espacios vacíos entre cada caracter se cubren, esto dificulta la idetificación de la separación de las primeras dos líneas de texto.

Como puede ver en la Figura 3.3, los espacios para cada carácter y palabra han desaparecido, y los espacios entre las líneas de texto adyacentes también han desaparecido. Es importante mantener el espacio en blanco entre las líneas de texto adyacentes para facilitar la búsqueda de rutas que permitan dividir líneas de texto manuscrito. Estos métodos tienen problemas para separar documentos donde las líneas de texto adyacentes se intersectan verticalmente.

3.4 Métodos para la Búsqueda de una Ruta para Segmentar Líneas de Texto (BRSLT)

En el estado del arte para encontrar la ruta con la mayor cantidad de pixeles blancos se realiza una búsqueda local, este enfoque de búsqueda no garantiza encontrar la ruta óptima (Koppula & Negi, 2014; Nicolaou & Gatos, 2009; Peng et al., 2016). (Kesiman et al., 2016) proponen realizar una búsqueda local de la mejor ruta considerando el valor de una función de costo que considera la menor cantidad de píxeles negros.

En el método que se propone en (Arvanitopoulos & Süsstrunk, 2014) se usa una adaptación del método Seam Carving (Avidan & Shamir, 2007) para encontrar una ruta óptima. El objetivo de esta modificación es encontrar el Seam Carving o ruta que mejor separe dos líneas de texto manuscrito. Eventualmente, la ruta con el menor error o costo es la ruta deseada. Para evitar que este método se desvíe en un mínimo local es necesario usar un método de optimización global, esta técnica es discutida en (Liwicki et al., 2007; Saabni et al., 2014).

El método propuesto en (Arvanitopoulos & Süsstrunk, 2014), se utiliza una adaptación del método *Seam Carving* (Avidan & Shamir, 2007) para encontrar la ruta óptima. El propósito de esta modificación es encontrar la ruta que mejor separa las dos líneas de texto manuscrito. Al final, la ruta con el menor error o costo es la ruta deseada. Para evitar que este método se estanque en un mínimo local, es necesario utilizar un método de optimización global, esta mejora se discute en (Liwicki et al., 2007; Saabni et al., 2014) y se demostró que la optimización global permite obtener mejores resultados para la segmetación de líneas de texto manuscrito.

3.5 Corpus

En el estado del arte se han presentado corpus de documentos escritos en tablillas de palma en (Kesiman et al., 2016; Peng et al., 2016; Valy et al., 2016). En algunos corpus se tienen valores repetidos en todos los documentos, en algunos se tiene el mismo número de líneas (Peng et al., 2016; Valy et al., 2016). Algunos corpus sólo están compuestos por documentos del mismo idioma (Peng et al., 2016; Valy et al., 2016).

Hasta el momento no se ha realizado una comparación entre los métodos del estado del arte con cada una de las colecciones del estado del arte disponibles.

G.Greek16.- Colección de documentos manuscritos griegos propuesta en 2016 por Ptak et al. (Ptak et al., 2017). Esta colección contiene 60 documentos que fueron creados por 30 escritores usando el mismo texto. En la Figura 3.5 se muestran dos ejemplos de documentos que pertenecen a ese corpus.

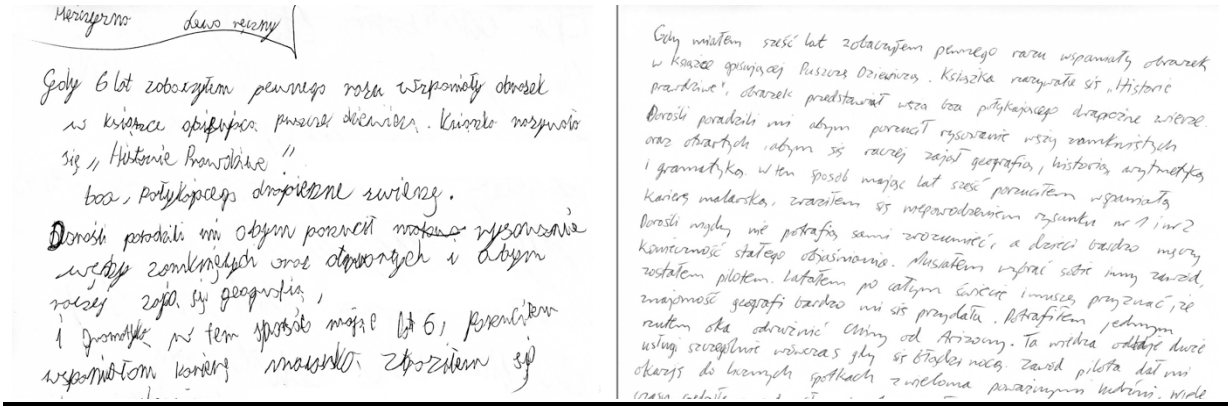


Figura 3.4. Ejemplo de documentos del corpus del corpus utilizado en (Ptak et al., 2017). En ninguna línea de los documentos se tienen caracteres que intersectan otras líneas verticalmente.

En el trabajo presentado en (Valy et al., 2016) se presenta un corpus compuesto por 134 líneas de texto agrupadas en 20 páginas de texto manuscrito antiguo en idioma Khmer. Este corpus no contiene líneas de texto conectadas verticalmente.

El corpus **M.Alaei11** contiene documentos escritos en Persa, Bangali, Oriya y Kannada con un total de 12,565 líneas de texto que se encuentran distribuidas en 707 imágenes de documentos escritos por 436 personas (Alaei et al., 2012)

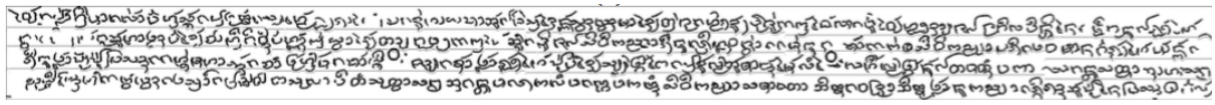


Figura 3.5. Ejemplo de documentos del corpus utilizado en (Peng et al., 2016). El método diseñado para este corpus no considera documentos con líneas de texto que se intersectan verticalmente.

M.Saabni14.- Colección manuscrita histórica multilenguaje usada en (Saabni et al., 2014) con 315 manuscritos históricos en cuatro idiomas: chino, inglés, español y árabe. Los documentos proceden de cuatro bibliotecas diferentes: Centro Juma Al-Majid de Cultura y Patrimonio, Centro Wadod de manuscritos, Universidad Americana de Beirut y Biblioteca del Congreso.

Figura 3.6. Ejemplo de documentos que pertenecen al corpus usado en los trabajos presentados en (Arvanitopoulos & Süsstrunk, 2014; Saabni et al., 2014). Los documentos de este corpus se han usado para evaluar el método propuesto en este trabajo.

M.ICDAR13.- Colección de documentos en varios idiomas de la Conferencia Internacional sobre Análisis y Reconocimiento de Documentos (ICDAR) en

2013. Utilizamos el conjunto de entrenamiento para la tarea TLS en tres idiomas, tres alfabetos y 150 documentos.

E.CLEF16.- En 2016, la tarea TLS formó parte de la conferencia ImageCLEF (Mauricio et al., 2016) centrada en la recuperación de información de documentos escritos a mano. Toda la colección de ImageCLEF está escrita en inglés por el filósofo Jeremy Bentham.

K.KhmerXX.- Las colecciones en idioma jemer propuestas por Valy (Valy et al., 2016) con 20 documentos del año 1900. En la Figura 3.7 se muestra un ejemplo de documentos de esta colección.

Figura 3.7. Ejemplo de documentos del corpus utilizado en (Valy et al., 2016). El método diseñado para este corpus no considera documentos con líneas de texto que se intersectan verticalmente.

S.VisitasXVI.- Colección histórica antigua española con 444 documentos del año 1548 (García Castro, 2013).

M.AmoXVII.- Fondo histórico plurilingüe con 20 documentos en español antiguo y náhuatl del año 1600.

En (Peng et al., 2016) se presenta un corpus en idioma Dai que contiene 1,050 líneas de texto manuscrito distribuidas en 290 documentos. Este corpus no contiene líneas de texto conectadas verticalmente. Todos los documentos del corpus contienen el mismo número de líneas. En la Figura 3.8 se proporcionan documentos del corpus usado para desarrollar el método propuesto en (Peng et al., 2016).

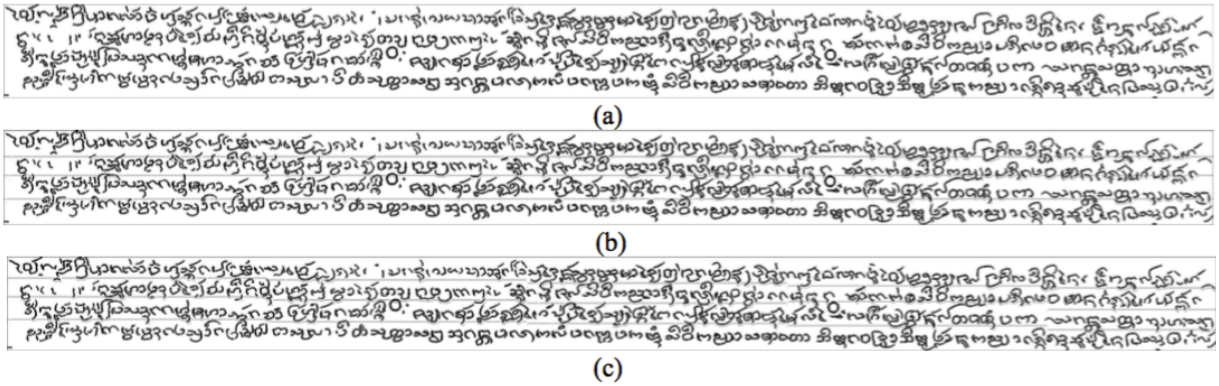


Figura 3.8. Ejemplo de documentos del corpus utilizado en (Peng et al., 2016). El método diseñado para este corpus no considera documentos con líneas de texto que se intersectan verticalmente.

Se ha establecido contacto con los autores de los trabajos en el caso de las colecciones privadas, pero no fue posible que compartieran el corpus debido a restricciones de derechos de autor.

3.6 Cálculo de complejidad

La mayoría de los métodos del estado del arte pasa la SLT requieren un ajuste de parámetros que se establece de manera empírica por el humano. Este ajuste manual de parámetros implica que existen conjuntos de configuraciones para cada nivel de complejidad.

En el estado del arte, es claro que la información en el espacio interlíneal dificulta la búsqueda de la ruta para segmentar líneas de texto (Arivazhagan et al., 2007; Arvanitopoulos & Süsstrunk, 2014; Likforman-Sulem et al., 2006; Saabni et al., 2014; Z. Shi & Venu Govindaraju, 2004) por eso es necesario tener una función de costo que orienta la búsqueda. En otras palabras, la complejidad de TLS radica en la cantidad de información que existe en el espacio interlineal. Para ello, Saabni (Saabni et al., 2014) propone el valor de precisión MatchScore como medida de complejidad.

MatchScore no mide la información que se encuentra en el espacio interlineal, MatchScore mide la similitud que existe en la región del cuadro delimitador de una línea de texto segmentada por el humano. Es decir, el índice de complejidad de Saabni (Saabni et al., 2014) es extrínseco y se centra más en la línea de texto que en el espacio interlineal.

Específicamente, el método Saabni usa el error del método Arivazhagan (Arivazhagan et al., 2007) cuando se evalúa con la precisión MatchScore. Esto es, dada una precisión de MatchScore del método Arivazhagan para una imagen I, el índice $Saabni - ECI$ se puede calcular como:

$$Saabni - ECI_{(i)} = \frac{1}{MatchScore\left(Arivazhagan_{(i)}\right)}$$

Ecuación 3.1 Cálculo del índice extrínseco de complejidad propuesto en (Saabni et al., 2014).

Un valor elevado de complejidad representa una baja precisión de MatchScore. Sin embargo, es importante considerar que la similitud se mide con el área de la línea segmentada por el humano, por lo tanto, no es posible automatizar la ejecución de este índice.

3.7 Resumen del capítulo

A lo largo de este capítulo se han presentado un análisis de las etapas del estado del arte para la segmentación de líneas de texto. Además, se describen las colecciones estándar del estado del arte para la SLT.

Posteriormente se han descrito los esfuerzos para reducir el impacto de la complejidad en la SLT.

Los métodos del estado del arte no mantienen el mismo comportamiento con documentos visualmente menos o más complejos. Los métodos del Estado del arte solo mantienen el comportamiento con los documentos usados en su etapa de experimentación, por lo tanto, los métodos del estado del arte no son aptos para ser usados en un escenario real (Likforman-Sulem et al., 2006; Saabni et al., 2014).

CAPÍTULO 4.
Método Propuesto

En este capítulo se presenta el método propuesto en esta tesis para el cálculo de complejidad de documentos manuscritos para ser segmentados.

Retomando los problemas planteados en este trabajo que consisten en: ¿cómo calcular la complejidad intrínseca de un documento manuscrito para poder seleccionar el método del estado arte que realice de manera óptima la segmentación de líneas de texto? La hipótesis de este trabajo consiste en que si se calcula la cantidad de información que aportan los trazos horizontales y verticales; además de la cantidad de información que aporta la tinta del documento y los valores del color del material de escritura, entonces será posible calcular el índice de complejidad intrínseco de un documento manuscrito.

En el capítulo dos se presentaron técnicas de procesamiento de imágenes y se describió como se puede realizar la extracción del Perfil de Proyección Horizontal (PPH) para obtener una distribución de la información de los trazos horizontales. También se presenta la extracción del Perfil de Proyección

Vertical (PPV) para obtener una distribución de la información de los trazos verticales.

Además, se presentó el concepto de histograma de color para obtener la distribución de color de una imagen digital.

En las siguientes subsecciones se describe la caracterización de documentos manuscritos con: PPH, PPV, un histograma de color de la tinta (HCT) y un histograma de color del fondo (HCF).

En la primera sección se presenta una descripción general de la metodología de solución. En las siguientes subsecciones se describe el método propuesto para la extracción del espacio interlineal. Posteriormente se describe el cálculo de complejidad considerando tres regiones del espacio interlineal.

4.1 Metodología general propuesta

Retomando el estado del arte para la segmentación de líneas de texto se proponen dos etapas. En la primera etapa se realiza la Localización de líneas de texto. En la segunda etapa se propone realizar el cálculo del índice de complejidad intrínseco.

Etapa 1. En esta etapa se aplica el método PEM-alpha para rellenar los espacios entre caracteres y palabras y para eliminar los trazos presentes en el espacio interlineal.

Etapa 2. En esta etapa se realiza la extracción del perfil de proyección horizontal (PPH) de la imagen resultante de la etapa 1.

Etapa 3. Se realiza una búsqueda de los valles en el perfil de proyección horizontal para obtener las coordenadas del espacio interlineal.

Etapa 4. El cálculo de complejidad se realiza tomando como entrada los espacios interlineales obtenidos en la etapa 3. El proceso de esta etapa se realiza tomando como entrada la imagen del documento original.

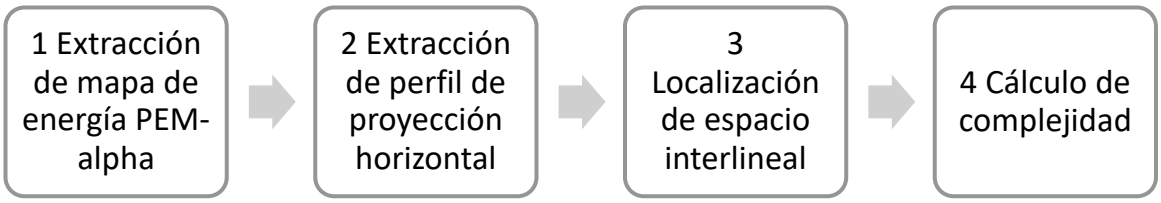


Figura 4.1. Secuencia de la metodología general propuesta.

4.1.1 Etapa 1: Extracción de mapa de energía PEM-alpha

En la figura 4.2 las regiones con mayor energía (regiones más oscuras) corresponden con el centro de las líneas de texto y las regiones con menos energía corresponden con los bordes superiores e inferiores de cada línea de texto. La Figura 4.2 muestra un ejemplo de regiones con mayor energía y menor energía.

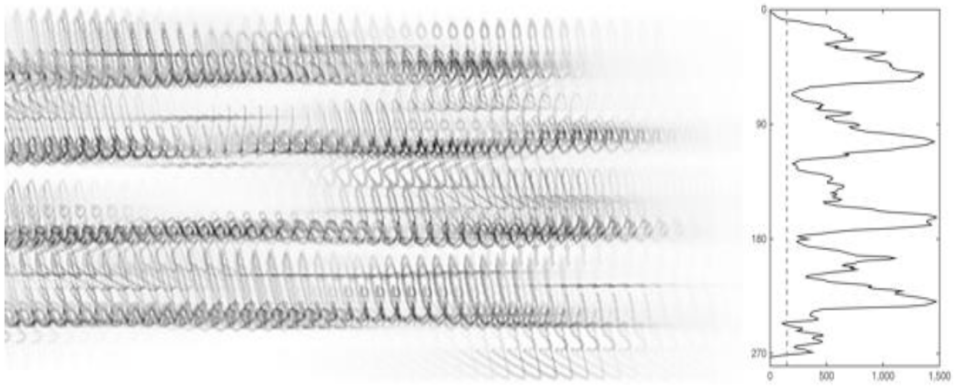


Figura 4.2. Ejemplo del ME-Alfa propuesto y el PPH.

La binarización del mapa de energía alfa (ME-Alfa) permite remover los píxeles con menos energía en comparación de la extracción directa del perfil de proyección horizontal (ver Figura 4.2). A partir de la extracción del PPH podemos mejorar la etapa de la LLT como puede verse en la Figura 4.2.

Figura 4.3. Ejemplo de PPH del ME-Alfa binarizado.

4.1.2 Etapa 2: Extracción de perfil de proyección horizontal

Los puntos de origen para los espacios interlineales se presentan en la figura 4.4.

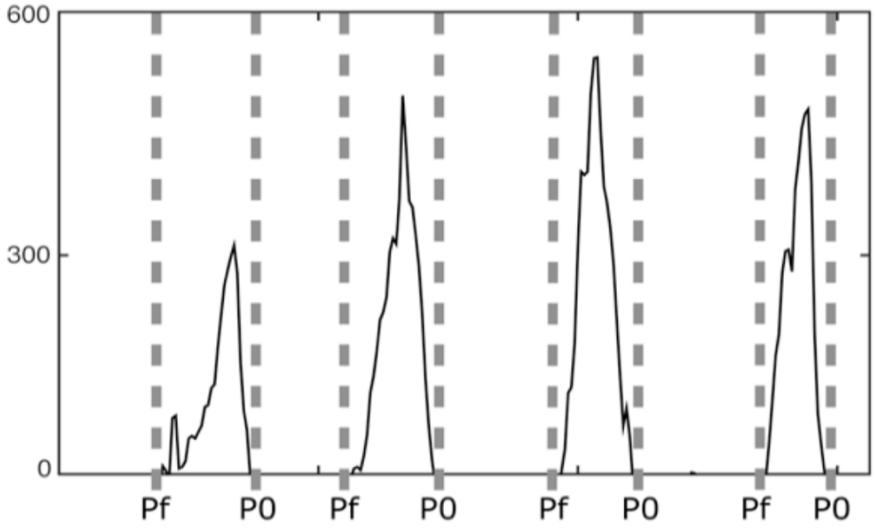


Figura 4.4. Perfil de proyección horizontal del ME-Alfa del documento E18 de la colección presentada en (Saabni et al., 2014).

Las coordenadas pf y $p0$ mostradas en la figura 4.4 se usan para extraer los espacios interlineales del manuscrito.

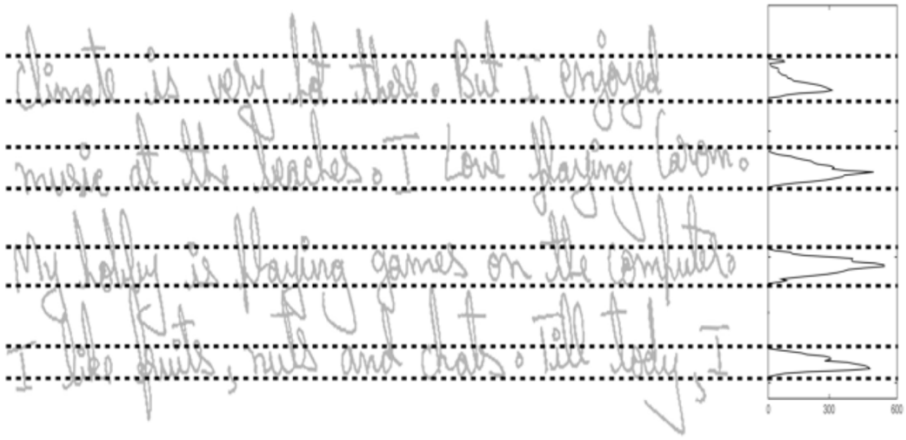


Figura 4.5. Espacios interlineales encontrados mediante la búsqueda de los valles en el PPH mostrado en la Figura 4.3. Imagen del documento E18 de la colección presentada en (Saabni et al., 2014).

Por medio de esta técnica, es posible generar un PPH donde la distancia entre los picos y valles sea mayor (véase Fig. 4.3) a diferencia del PPH mostrado en la Figura 4.1, dejando un salto muy grande entre los mínimos locales con el valor cero.

4.1.3 Etapa 3: Localización de espacio interlineal

Para determinar los puntos de origen de cada espacio interlineal se realiza una localización de los valles con una longitud mayor a un píxel en el PPH. La Figura 4.3 muestra el inicio y el fin de cada valle en el PPH. En la Figura 4.4 el valle inicial (P0) y el valle final (PF) encontrado por medio del ME-Alfa está dibujado sobre cada línea de texto correspondiente.

El resultado de esta etapa es el conjunto de espacios interlineales que contiene el documento. Después de esta etapa es posible realizar el cálculo de complejidad para cada espacio interlineal.

4.1.4 Etapa 4: Cálculo de complejidad

Con el objetivo de reducir la cantidad de recursos necesarios para calcular la complejidad se propone reducir el tamaño de la muestra. Por lo tanto, se propone extraer tres segmentos de cada espacio interlineal. Los tres

segmentos son usados como datos de entrada para el índice de complejidad intrínseco propuesto. Los tres segmentos se colocan de forma equidistante.

El índice de complejidad intrínseco para la segmentación de líneas de texto propuesto (TLS- ICI) se define como el promedio de la suma de cuatro subíndices normalizados. El primer subíndice mide la información de cantidad horizontal en el espacio interlineal mediante un Perfil de Proyección Horizontal (PPH). Perfil de Proyección Vertical (PPV), Histograma de Fondo de Color (HCF) y el Histograma de Tinta de Color (HCT).

El índice de complejidad intrínseca de segmentación de líneas de texto propuesto (TLS-ICI) se define como el promedio de la suma de los cuatro subíndices normalizados. El primer subíndice mide la cantidad de información horizontal en el espacio interlineal por un Perfil de Proyección Horizontal (PPH). El PPH permite medir la cantidad de rayones horizontales, tildes, trazos ascendentes y descendentes.

El segundo subíndice mide la cantidad de información vertical en el espacio interlineal por un PP vertical (PPV). El PPV permite medir la cantidad de rayones verticales, acentos, componentes superpuestos, trazos ascendentes y descendentes simples.

El tercer subíndice es un histograma de color del mapa de bits (HCF) en el espacio interlineal que mide la cantidad de información relacionada con el color (Gonzalez et al., 1996; Martinsanz et al., 2007). El HCF se basa en el hecho de que una imagen con un contraste bajo es más difícil de distinguir el fondo de la tinta.

El cuarto subíndice es un histograma de color de la tinta (HCT) en el espacio interlineal que mide la cantidad de información relacionada con el color de la tinta de escritura. Por lo tanto, dada la imagen del espacio interlíneal I con ancho w y alto h, el TLS-ICI se define como:

$$TLS - ICI_{(I)} = \frac{PPH_{(I)} + PPV_{(I)} + HCF_{(I)} + HCT_{(I)}}{4}$$

Ecuación 4.1 Función que describe el cálculo del promedio de los cuatro sub-índices propuestos.

Con el promedio de la suma de los cuatro índices se tiene el mismo nivel de importancia para cada subíndice. En la figura 3.9 se muestra una representación visual de los tres segmentos que se extraen en el espacio interlineal.

Dado el histograma horizontal I^H de la imagen I como un vector de tamaño h. El sub-índice HPP mide la entropía del histograma de proyección horizontal, de la siguiente manera:

$$PPH(I^H) = \sum_{i=1}^{h} -\frac{I_i^H}{w \cdot h} log2 \frac{I_i^H}{\mathrm{w} \cdot h}$$

Ecuación 4.2 Sub-índice propuesto para el cálculo de entropía en un histograma de proyección horizontal.

Dado el histograma vertical I^V de la imagen I como un vector de tamaño w. El sub-índice PPH mide la entropía del histograma de proyección vertical, de la siguiente manera:

$$PPV(I^V) = \sum_{i=1}^{w} -\frac{I_i^H}{\mathrm{w} \cdot h} log2 \frac{I_i^H}{\mathrm{w} \cdot h}$$

Ecuación 4.3 Sub-índice propuesto para el cálculo de entropía en un histograma de proyección vertical.

Dado el histograma de color I^C en escala de grises de la imagen I como un vector de tamaño 256. El HCB se define como la entropía de la información, de la siguiente manera:

$$HCF(I^C) = \sum_{i=1}^{256} -\frac{I_i^C}{w \cdot h} log2 \frac{I_i^C}{w \cdot h}$$

Ecuación 4.4 Sub-índice propuesto para el cálculo de entropía en un histograma de color.

El HCI se define como el porcentaje de tinta de la imagen I, y se calcula de la siguiente forma:

$$HCT(I) = \frac{\left|\left|\{I_{pixel} | I_{pixel} \in ink\}\right|\right|}{w \cdot h}$$

Ecuación 4.5 Sub-índice propuesto para el cálculo de entropía de los pixeles que representan la tinta con la que fue escrito el documento.

Dado un documento D con k imágenes de espacios interlineales I_k^D, el índice de complejidad intrínseco propuesto para D se define como el promedio de $TLS - ICI$ de sus espacios interlineales. Es decir:

$$TLS - ICI(D) = \frac{1}{k}\sum_{i=1}^{k} TLS - ICI(I_i^D)$$

Ecuación 4.6 Función que describe el cálculo del índice de complejidad propuesto $TLS - ICI$ para un documento manuscrito.

En la figura 4.6 se muestra una representación visual de los tres segmentos usados de cada espacio interlineal que se usan el cálculo del índice de complejidad con el índice $TLS - ICI$ propuesto.

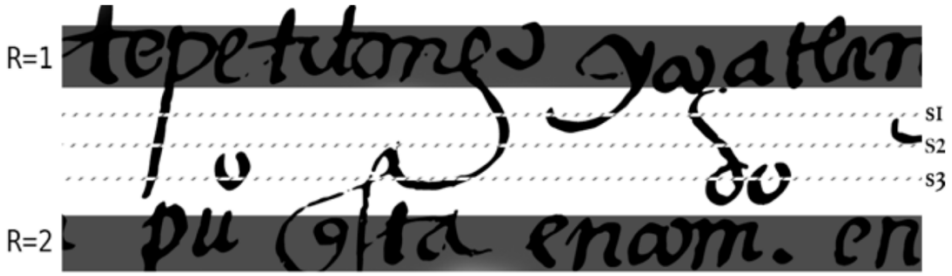


Figura 4.6. Representación visual de la extracción de tres segmentos del espacio interlineal usados para calcular la complejidad.

4.2 Resumen del capítulo

En este capítulo se ha presentado un índice de complejidad intrínseco para el cálculo de complejidad en documentos manuscritos. Se ha descrito cómo se realiza la localización del espacio Interlineal entre dos líneas de texto vecinas. Posteriormente se presentó el índice de complejidad propuesto que integra cuatro subíndices. Cada subíndice mide diferentes características únicas que se presentan en los documentos manuscritos.

CAPÍTULO 5.
Experimentación y resultados

A lo largo de este capítulo se describe la metodología y configuración utilizada para validar la hipótesis con el índice TLS-ICI que se propone en este trabajo. También se describen las colecciones de documentos estándar utilizadas. Además, se muestra el índice de complejidad obtenido con cada colección de documentos.

La hipótesis de este trabajo consiste en que si se calcula la cantidad de información que aportan los trazos horizontales y verticales; además de la cantidad de información que aporta la tinta del documento y los valores del color del material de escritura, entonces será posible calcular el índice de complejidad intrínseco de un documento manuscrito.

La sección de evaluación está dividida en dos secciones, en la primera sección se describen las colecciones de documentos utilizadas, la segunda sección contiene los índices de complejidad obtenidos para cada colección, los rangos de complejidad donde los métodos del estado del arte pueden realizar

la segmentación de líneas de texto y también se presenta la evaluación de un método híbrido construido con un ensamblre de métodos del estado del arte considerando los rangos de complejidad obtenidos en la etapa anterior de la experimentación.

5.1 Colecciones de documentos

Después de realizar una búsqueda exhaustiva encontramos un conjunto de colecciones de documentos estándar para la segmentación de líneas de texto manuscrito. Cada colección del estado del arte contiene texto con diferentes características de estilo y diferentes periodos de tiempo.

5.1.1 Descripción de las colecciones estándar utilizadas

El conjunto de colecciones de prueba está compuesto por ocho colecciones estándar. Se tienen cinco alfabetos diferentes, 12 lenguajes, 2,512 documentos y 58,138 líneas de texto. Todos los documentos presentados en el conjunto de datos contienen un conjunto de pruebas segmentada por humanos.

Tabla 5.1. Descripción general de las colecciones estándar usadas en la etapa de experimentación.

Colección	Documentos	Líneas	Lenguajes	Año
M.Alaei11	707	12,565	Persa, Bangali, Oriya y Kannada	2011
M.Saabni14	315	4,034	Árabe, chino, español e inglés	2009
M.ICDAR13	150	2,649	Griego, Inglés e Hindi	2013
E.CLEF16	796	20,234	Inglés	1838
K.KhmerXX	20	134	Khemer	1900
G.Greek16	60	1,514	Griego	2016
S.VisitasXVI	444	16,450	Español	1548
M.AmoXVII	20	558	Español y Nahuatl	1600
Total	2,512	58,138	12	-

5.2 Preprocesamiento

El método propuesto está diseñado para calcular la complejidad en imágenes en tres diferentes modelos de color; RGB, escala de grises y binario. Es por eso por lo que no se realiza ningún preprocesamiento sobre el modelo de color de las imágenes de entrada.

La mayoría de las imágenes tienen un ángulo de inclinación que se generan por error humano al momento de digitalizar los documentos. Es por eso que el único prepocesamiento aplicado es la corrección de inclinación con la implementación de la transformada de Radon (Helgason, 1999).

5.3 Determinación del TLS-ICI para las colecciones del estado del arte

En la figura 5.1 se presenta una evaluación promedio para cada documento de la colección utilizando el índice TLS-ICI propuesto. La figura 5.1 presenta el primer ranking de colecciones del estado del arte de acuerdo con su complejidad para ser segmentadas.

Como se puede observar, las colecciones en español antiguo (S.VisitasXVI y M.AmoXVII) presentan los índices de complejidad más altos. Además, podemos ver que a pesar de que la colección K.KhemerXX y G.Greek16 contienen documentos con lenguajes diferentes el valor obtenido con el índice TLS-ICI indica que las complejidades son similares, esto nos permite concluir que la complejidad no está asociada con el lenguaje de escritura (como algunos autores suponían) (Biswas & Das, 2012; Koppula & Negi, 2014; Likforman-Sulem et al., 2006; Valy et al., 2016).

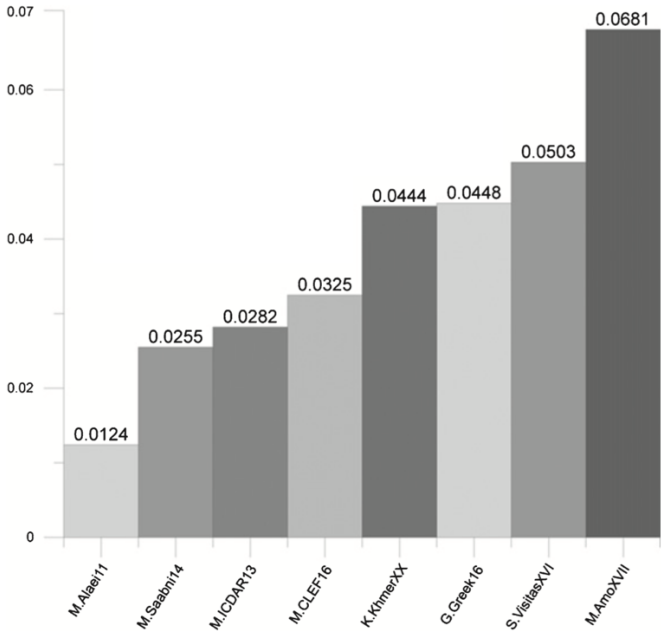


Figura 5.1. Ranking de las colecciones estándar del estado del arte de acuerdo con su complejidad. Las colecciones están ordenadas según su índice de complejidad intrínseco (TLS-ICI).

5.3 Determinación del TLS-ICI para los métodos del estado del arte

Con el objetivo de medir la complejidad máxima en la que cada método del estado del arte puede mantener una precisión de más de 95% al segmentar, se evaluó cada método de manera individual.

Está establecido que para que las líneas segmentadas tengan utilidad en la etapa posterior de transcripción automática se requiere que se haya obtenido una precisión mayor al 95% al evaluar el resultado de la segmenatación con la métrica *MatchScore* (Likforman-Sulem et al., 2006; Mauricio et al., 2016; V. Romero et al., 2015).

Para la evaluación individual de cada método del estado del arte se utilizó la métrica *MatchScore*. *MatchScore* fue presentada por Yanikoglu et. al. (Yanikoglu & Vincent, 1998) y está definida como el porcentaje de los píxeles en el fondo G_u que están cubiertos por R_u menos el porcentaje de los píxeles en el frente de R_u que se encuentran fuera de G_v. Hasta el momento esta métrica

se ha tomando como un estándar de evaluación en el estado del arte (Demır & Özkaya, 2020; Likforman-Sulem et al., 2006).

Dado G_u, el conjunto de todos los puntos de la región i de la muestra de entrenamiento, R_v el conjunto de todos los puntos de la región j resultante, $T(s)$ es una función que cuenta los elementos del conjunto s. El valor $MatchScore(u, v)$ representa el número de coincidencias de la región i de la muestra de entrenamiento G_u y la región j resultante de la siguiente forma:

$$MatchScore(u, v) = \frac{T(G_u \cap R_v)}{T(G_u \cup R_v)}$$

Ecuación 5.1 cálculo de coincidencias de una region de la muestra de entrenamiento contra una region generada con un método automático para la evaluación del rendimiento en la segmentación de líneas de texto.

Es importante mencionar que esta métrica de evaluación fue usada para medir el rendimiento de los métodos para la SLT en ICDAR 2007, ICDAR2010, ICDAR2013 en la competencia *Handwriting Segmentation* y se toma como medida de evaluación estándar (Arivazhagan et al., 2007; Likforman-Sulem et al., 2006; Mauricio et al., 2016; Saabni et al., 2014).

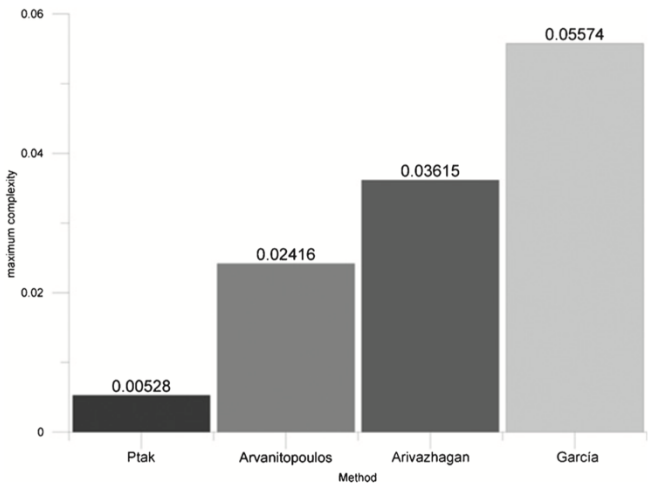


Figura 5.2. Representación visual de la extracción de tres segmentos del espacio interlineal usados para calcular la complejidad.

Los métodos de Ptak (Ptak et al., 2017), Arvanitopoulos (Arvanitopoulos & Süsstrunk, 2014), García (García-Calderón et al., 2018) y Arivazhagan (Arivazhagan et al., 2007) son seleccionados para la experimentación ya que estos métodos superan a otros. En particular, el método Arvanitopoulos (Arvanitopoulos & Süsstrunk, 2014) supera al método Saabni (Saabni et al., 2014) y utilizó la colección M.Saabni14 en su etapa de experimentación. Los métodos de Ptak (Ptak et al., 2017) y Arvanitopoulos (Arvanitopoulos & Süsstrunk, 2014) tienen un enfoque no-supervisado pero necesitan varios parámetros. En este caso, los parámetros definidos por los autores son utilizados en nuestra experimentación. Los métodos de Arivazhagan (Arivazhagan et al., 2007) y García (García-Calderón et al., 2018) son métodos automáticos no supervisados que no necesitan un ajuste manual de sus parámetros.

De acuerdo con los resultados del índice TLS-ICI propuesto nuestro método previamente publicado para la SLT (García-Calderón et al., 2018) puede manejar valores de complejidad más altos en comparación con los métodos de Ptak (Ptak et al., 2017), Arvanitopoulus (Arvanitopoulos & Süsstrunk, 2014) y Arivazhagan (Arivazhagan et al., 2007).

Una vez que se ha calculado el índice TLS-ICI para las colecciones del estado del arte y los métodos de la tarea SLT, se establece una relación que muestra el desempeño de los métodos en cada colección. En la Tabla 3 se muestra la relación de desempeño entre colecciones y métodos. Por ejemplo, el método Ptak obtiene un TLS-ICI de 0.00528 que no es suficiente para niguna de las colecciones estándar del estado del arte utilizadas.

Para la colección M.AmoXVVII no existe un método suficientemente robusto que pueda procesarla con una precisión mayor al 95%.

5.2 Un método híbrido basado en el índice TLS-ICI propuesto

Considerando que cada método SLT fue diseñado para un rango de complejidad (Likforman-Sulem et al., 2006; Ptak et al., 2017; Saabni et al., 2014; Valy et al., 2016), en esta sección se propone usar los valores de complejidad obtenidos con el índice TLS-ICI propuesto para realizar un ensamble de métodos del estado del arte. Llamanos este ensamble "método híbrido" y su

función es seleccionar el método TLS más apropiado por documento de acuerdo con el índice TLS-ICI propuesto, ver Fig.8.

Tabla 5.2. Representación de la relación Colección/Método genererada con los datos de la figura 5.1 y la figura 5.2. En esta representación podemos ver que en promedio, el método de Ptak no puede aplicarse para segmentar ninguna colección completa del estado del arte.

Colección/Método	Ptak	Arvanitopoulos	Arivazhagan	Garcia
M.Alaei11				
M.Saabni14				
E.CLEF16				
M.ICDAR13				
K.KhmerXX				
G.Greek16				
S.VisitasXVI				
M.AmoXVII				

El proceso general del método TLS Híbrido (MH) consiste en calcular la complejidad del documento a segmentar como primer paso y la selección de un método como segundo paso. El método híbrido se define como:

$$MH(D) = \begin{cases} Ptak(D) & 0 \leq TLS_{ICI(D)} < 0.00528 \\ Arvanitopoulus(D) & 0.00528 \leq TLS_{ICI(D)} < 0.02416 \\ Arivazhagan(D) & 0.02416 \leq TLS_{ICI(D)} < 0.03615 \\ Garcia(D) & 0.03615 \leq TLS_{ICI(D)} \end{cases}$$

Ecuación 5.2 Función que describe el proceso del método híbrido (MH) propuesto para seleccionar un método del estado del arte dado el índice de complejidad intrínseco obtenido para una imagen de un documento manuscrito antiguo.

5.3 Comparasión del método híbrido propuesto con los métodos del estado del arte

En una primera etapa de experimentación se planteó comparar los métodos del estado del arte contra el método híbrido propuesto. En la tabla 5.3 se muestra la comparación de exactitud.

Tabla 5.3. Comparación de exactitud del método propuesto (MH) contra los métodos de Arivazhagan , Ptak y Arvanitopoulos (Arivazhagan et al., 2007; Arvanitopoulos & Süsstrunk, 2014; Ptak et al., 2017).

Colección	Arivazhagan	Ptak	Arvanitopoulus	García	MH
M.Alaei11	70.1	65.5	69.2	95.1	95.1
M.Saabni14	92	89.2	84.5	96	96.3
E.CLEF16	62.3	41.3	53.1	95.3	95.3
M.ICDAR13	81	84.1	62.3	93.9	93.9
K.KhmerXX	33.6	0	7.3	93.5	93.5
G.Greek16	34.5	32.8	22.1	87.6	87.6
S.VisitasXVI	0	5.8	0	90.1	90.1
S.AmoXVII	0	3.6	5.1	84.1	84.1

Al analizar los resultados de la tabla 5.3 se puede llegar a la suposición de que el método híbrido sólo usa el método de García. En una segunda etapa de experimentación se eliminó el método de García dentro del método híbrido propuesto. El objetivo de la segunda etapa de experimentación fue comprobar que el método híbrido puede decidir qué método es el más apto para segmentar un document manuscrito de acuerdo con su complejidad.

Tabla 5.4. Comparación de exactitud del método propuesto contra el método de Arvanitopoulos (Arvanitopoulos & Süsstrunk, 2014).

Colección	Arivazhagan	Ptak	Arvanitopoulus	MH
M.Alaei11	70.1	65.5	69.2	71.9
M.Saabni14	92	89.2	84.5	92.5
E.CLEF16	62.3	41.3	53.1	65.6
M.ICDAR13	81	84.1	62.3	86.7
K.KhmerXX	33.6	0	7.3	33.6
G.Greek16	34.5	32.8	22.1	43.2
S.VisitasXVI	0	5.8	0	5.8
M.AmoXVII	0	3.6	5.1	5.3

Los resultados mostrados en la tabla 5.4 muestran que en todos los casos el método híbrido mantuvo o superó a los métodos del estado del arte cuando se ponen a prueba de manera individual.

CAPÍTULO 6.
Conclusiones y Trabajo Futuro

En este capítulo se presentan las conclusiones generadas a partir de la etapa de experimentación realizada en este trabajo. Este capítulo se encuentra dividido en tres subsecciones: conclusiones, aportaciones y trabajo futuro.

6.1. Conclusiones

La mayoría de los métodos TLS deben analizar las características de estilo de escritura para estimar de manera empírica la complejidad de un documento para ajustar los parámetros y aumentar su rendimiento.

En esta investigación se propuso un índice de complejidad intrínseco para la segmentación de líneas de texto (TLS-ICI) que estima rápidamente *a priori* la complejidad de segmentación de línea de texto de un par de líneas, un documento o una colección completa.

El índice propuesto TLS-ICI permite caracterizar los métodos TLS por su complejidad. En otras palabras, el TLS-ICI proporciona orden tanto a las colecciones de documentos como a los métodos del estado del arte. Además, en este documento se propone un Método Híbrido (MH) para la tarea TLS basado en TLS-ICI.

El TLS-ICI se puede utilizar para probar lo robusto que es un método TLS o para encontrar un método TLS adecuado para reducir el tiempo necesario para realizar la SLT. Según nuestra revisión, en este documento se presenta el primer índice de complejidad intrínseco y el primer método híbrido que combina un conjunto de métodos TLS del estado del arte.

Además, el método híbrido propuesto demuestra que TLS-ICI evalúa correctamente la complejidad del documento porque mantiene o supera a cualquier método de estado individual de TLS. Los resultados mostrados en la tabla 5.4 y 5.5 dan indicios de que es posible mejorar el método híbrido agregando más implementaciones de métodos del estado del arte.

Los resultados obtenidos muestran que la complejidad de los documentos para TLS no depende del lenguaje, sino de la cantidad de información presente en el espacio interlineal, por lo que las siguientes investigaciones deben tener en cuenta los rangos de complejidad de los documentos y no necesariamente el tipo de idiomas.

Se presentaron dos colecciones nuevas en español antiguo de los siglos XV y XVI. Estas colecciones históricas son un nuevo desafío para los métodos del estado del arte porque presentan los índices de complejidad más altos.

6.2. Trabajo futuro

En el futuro se espera analizar la correlación de los parámetros de los métodos del estado del arte con el índice de complejidad propuesto para estimar a priori sus parámetros. Además, esperamos agregar otros métodos TLS al método híbrido.

En el futuro es interesante hacer pruebas en documentos con una complejidad mayor, donde incluso para el humano es más difícil trazar una línea de corte. Con los resultados de esta investigación y trabajos previos se podría ensamblar un sistema para la indexación automática de líneas de texto manuscrito.

Es de nuestro interés realizar investigación para encontrar la forma de caracterizar documentos manuscritos con las complejidades más altas, analizar cómo el humano puede reconocer texto con espacio interlineal

pequeño y realizar experimentación para obtener propuestas de solución para segmentar documentos que hasta el momento ningún método del estado del arte puede segmentar correctamente.

Referencias

A. Prachanucroa & S. Phongsuphap. (2013). Marginal noise removal for scanned document images by projection profile based method. The 2013 10th International Joint Conference on Computer Science and Software Engineering (JCSSE), 17–20. https://doi.org/10.1109/JCSSE.2013.6567312

Alaei, A., Pal, U., & Nagabhushan, P. (2012). Dataset and ground truth for handwritten text in four different scripts. 26(4), 25. https://doi.org/DOI: 10.1142/S0218001412530011

Arica, N., & Yarman-Vural, F. T. (2001). An overview of character recognition focused on off-line handwriting. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 31(2), 216–233. https://doi.org/10.1109/5326.941845

Arivazhagan, Harish Srinivasan, & Sargur Srihari. (2007). A statistical approach to line segmentation in handwritten documents. 6500. https://doi.org/10.1117/12.704538

Arvanitopoulos, N., & Süsstrunk, S. (2014). Seam Carving for Text Line Extraction on Color and Grayscale Historical Manuscripts. 2014 14th International

Conference on Frontiers in Handwriting Recognition, 726–731. https://doi.org/10.1109/ICFHR.2014.127

Avidan, S., & Shamir, A. (2007). Seam Carving for Content-aware Image Resizing. ACM Trans. Graph., 26(3). https://doi.org/10.1145/1276377.1276390

B. Gatos, K. Ntirogiannis, & I. Pratikakis. (2009). ICDAR 2009 Document Image Binarization Contest (DIBCO 2009). 2009 10th International Conference on Document Analysis and Recognition, 1375–1382. https://doi.org/10.1109/ICDAR.2009.246

Bagdanov, A., & Kanai, J. (1998). Projection profile based skew estimation algorithm for JBIG compressed images. Proceedings of the Fourth International Conference on Document Analysis and Recognition, 1(1), 43–51. https://doi.org/10.1109/icdar.1997.619878

Bagley, R. W. (2004). Anyang Writing and the Origin of the Chinese Writing System. En S. D. Houston (Ed.), The first writing: Script invention as history and process (pp. 190–249). Cambridge University Press. https://contentstore.cla.co.uk//secure/link?id=5e2dff9d-5c36-e711-80c9-005056af4099

Baines, J., Bennet, J., & Houston, S. D. (2008). The disappearance of writing systems: Perspectives on literacy and communication. Equinox; /z-wcorg/.

Bar-Yosef, I., Hagbi, N., Kedem, K., & Dinstein, I. (2009). Line Segmentation for Degraded Handwritten Historical Documents. 2009 10th International Conference on Document Analysis and Recognition, 1161–1165. https://doi.org/10.1109/ICDAR.2009.191

Biswas, S., & Das, A. K. (2012). Writer Identification of Bangla Handwritings by Radon Transform Projection Profile. 2012 10th IAPR International Workshop on Document Analysis Systems, 215–219. https://doi.org/10.1109/DAS.2012.98

Burger, W., & Burge, M. J. (2008). Digital Image Processing: An Algorithmic Introduction Using Java. Springer-Verlag London.

Causer, T., & Wallace, V. (2012). Building a Volunteer Community: Results and Findings from Transcribe Bentham. Digital Humanities Quarterly, 6(2). http://www.digitalhumanities.org/dhq/vol/6/2/000125/000125.html

Demır, A. A., & Özkaya, U. (2020). A Semantic Segmentation Based Approach for Segmentation and Recognition of Touching and Overlapping Digits. 2020 28th Signal Processing and Communications Applications Conference (SIU), 1–4. https://doi.org/10.1109/SIU49456.2020.9302132

Du, X., Pan, W., & Bui, T. D. (2009). Text line segmentation in handwritten documents using Mumford-Shah model. Pattern Recognition, 42(12), 3136–3145. https://doi.org/10.1016/j.patcog.2008.12.021

Fischer, A., Liwicki, M., & Ingold, R. (2020). Handwritten Historical Document Analysis, Recognition, and Retrieval—State of the Art and Future Trends. WORLD SCIENTIFIC. https://doi.org/10.1142/11353

García Castro, R. (2013). Suma de visitas de pueblos de la Nueva España. Universidad Autónoma del Estado de México. http://ri.uaemex.mx/handle/123456789/33111

García-Calderón, M. Á., García-Hernández, R. A., & Ledeneva, Y. (2018). Unsupervised multi-language handwritten text line segmentation. 34(5), 2901–2911. https://doi.org/10.3233/JIFS-169476

Gatos, B., Stamatopoulos, N., & Louloudis, G. (2009). ICDAR 2009 Handwriting Segmentation Contest. 2009 10th International Conference on Document Analysis and Recognition, 1393–1397. https://doi.org/10.1109/ICDAR.2009.245

Gonzalez, R. C., Woods, R. E., Davue Rodríguez, F., & Rosso, L. (1996). Tratamiento digital de imágenes. Addison-Wesley ; Díaz de Santos; /z-wcorg/.

Gray, N. M. B. (1948). The Paleography of Latin Inscriptions in the Eighth, Ninth and Tenth Centuries in Italy. Macmillan. https://books.google.com.mx/books?id=xlGhOQAACAAJ

Győry Hory, H. (2008). Medicine in Ancient Egypt. En H. Selin (Ed.), Encyclopaedia of the History of Science, Technology, and Medicine in Non-Western Cultures (pp. 1508–1518). Springer Netherlands. https://doi.org/10.1007/978-1-4020-4425-0_9748

Helgason, S. (1999). The Radon Transform. Birkhäuser Boston. https://books.google.com.mx/books?id=tq3eStnBwIUC

Houston, S. D., Boudreau, V., & Houston, P. A. S. D. (2004). The First Writing: Script Invention as History and Process. Cambridge University Press. https://books.google.com.mx/books?id=jsWL_XJt-dMC

I. Pratikakis, K. Zagoris, G. Barlas, & B. Gatos. (2016). ICFHR2016 Handwritten Document Image Binarization Contest (H-DIBCO 2016). 2016 15th International Conference on Frontiers in Handwriting Recognition (ICFHR), 619–623. https://doi.org/10.1109/ICFHR.2016.0118

J. A. Sánchez, A. H. Toselli, V. Romero, & E. Vidal. (2015). ICDAR 2015 competition HTRtS: Handwritten Text Recognition on the tranScriptorium dataset. 2015 13th International Conference on Document Analysis and Recognition (ICDAR), 1166–1170. https://doi.org/10.1109/ICDAR.2015.7333944

Kesiman, M. W. A., Burie, J.-C., & Ogier, J.-M. (2016). A New Scheme for Text Line and Character Segmentation from Gray Scale Images of Palm Leaf Manuscript. 2016 15th International Conference on Frontiers in Handwriting Recognition (ICFHR), 325–330. https://doi.org/10.1109/icfhr.2016.0068

Koppula, V. K., & Negi, A. (2014). Segmentation of closely set and touching lines in handwritten document images using fringe maps. International Conference for Convergence for Technology-2014, 1–6. https://doi.org/10.1109/i2ct.2014.7092176

Lewis, D. (1983). Extrinsic Properties. Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition, 44(2), 197–200. JSTOR.

Likforman-Sulem, L., Zahour, A., & Taconet, B. (2006). Text line segmentation of historical documents: A survey. International Journal of Document Analysis and Recognition (IJDAR), 9(2–4), 123–138. https://doi.org/10.1007/s10032-006-0023-z

Liwicki, M., Indermuhle, E., & Bunke, H. (2007). On-Line Handwritten Text Line Detection Using Dynamic Programming. Ninth International Conference on Document Analysis and Recognition (ICDAR 2007), 447–451. https://doi.org/10.1109/icdar.2007.4378749

Martinsanz, G. P., PAJARES, G., & de la Cruz García, J. M. (2007). Visión por computador. Imágenes Digitales y Aplicaciones. 2a Edición. RA-MA S.A. Editorial y Publicaciones. https://books.google.com.mx/books?id=EQqsPgAACAAJ

Mauricio, V., Alejandro, T., Joan-Andreu, S., & Enrique, V. (2016). Overview of the ImageCLEF 2016 Handwritten Scanned Document Retrieval Task. 1609.

Medina Morán, S. (2011). ¿Un error en la piedra de rosetta? 30(1), 21–27.

Muñoz y Rivero, J. (1880). Manual de paleografia diplomatica española de los siglos XII al XVII : Método teórico-prático para aprender á leer los documentos españoles de los siglos XII al XVII. Madrid : Moreno y Rojas.

Nicolaou, A., & Gatos, B. (2009). Handwritten Text Line Segmentation by Shredding Text into its Lines. Proceedings of the 2009 10th International Conference on Document Analysis and Recognition, 626–630.

O'Gorman, L., Sammon, M. J., & Seul, M. (2008). Practical Algorithms for Image Analysis with CD-ROM. Cambridge University Press. https://books.google.com.mx/books?id=8dXkUPv2DGYC

Peng, G., Yu, P., Li, H., & He, L. (2016). Text line segmentation using Viterbi algorithm for the palm leaf manuscripts of Dai. 2016 International

Conference on Audio, Language and Image Processing (ICALIP), 336–340. https://doi.org/10.1109/icalip.2016.7846561

Ptak, R., Zygadlo, B., & Unold, O. (2017). Projection Based Text Line Segmentation with a Variable Threshold. International Journal of Applied Mathematics and Computer Science, 27(1), 195–206. https://doi.org/10.1515/amcs-2017-0014

Q. N. Vo & G. Lee. (2016). Dense prediction for text line segmentation in handwritten document images. 2016 IEEE International Conference on Image Processing (ICIP), 3264–3268. https://doi.org/10.1109/ICIP.2016.7532963

Rendón Rojas, M. Á. (2005). Relación entre los conceptos: Información, conocimiento y valor. Semejanzas y diferencias. Ci. Inf, 34(2), 52–61.

Saabni, R., Asi, A., & El-Sana, J. (2014). Text line extraction for historical document images. Pattern Recognition Letters, 35, 23–33. https://doi.org/10.1016/j.patrec.2013.07.007

Sider, T. (1996). Intrinsic properties. Philosophical Studies, 83(1), 1–27. https://doi.org/10.1007/BF00372433

Trubek, A. (2017). The History and Uncertain Future of Handwriting. Bloomsbury USA. https://books.google.com.mx/books?id=dj06MQAACAAJ

V. Romero, J. A. Sánchez, V. Bosch, K. Depuydt, & J. de Does. (2015). Influence of text line segmentation in Handwritten Text Recognition. 2015 13th International Conference on Document Analysis and Recognition (ICDAR), 536–540. https://doi.org/10.1109/ICDAR.2015.7333819

Valizadeh, M., & Kabir, E. (2012). Binarization of degraded document image based on feature space partitioning and classification. International Journal on Document Analysis and Recognition (IJDAR), 15(1), 57–69.

Valy, D., Verleysen, M., & Sok, K. (2016). Line Segmentation Approach for Ancient Palm Leaf Manuscripts Using Competitive Learning Algorithm. 2016 15th International Conference on Frontiers in Handwriting Recognition (ICFHR), 108–113. https://doi.org/10.1109/icfhr.2016.0032

Voynich Manuscript, Beinecke MS 408, General Collection. (1912). Yale University.

Wildgen, W. (2004). The Evolution of Human Language: Scenarios, Principles, and Cultural Dynamics. John Benjamins Pub. https://books.google.com.mx/books?id=VvSsQgAACAAJ

Yanikoglu, B. A., & Vincent, L. (1998). Pink Panther: A Complete Environment For Ground-Truthing And Benchmarking Document Page Segmentation. Pattern Recognition, 31(9), 1191–1204. https://doi.org/10.1016/S0031-3203(97)00137-4

Z. Shi & Venu Govindaraju. (2004). Line separation for complex document images using fuzzy runlength. First International Workshop on Document Image Analysis for Libraries, 2004. Proceedings., 306–312. https://doi.org/10.1109/DIAL.2004.1263259