Universidad de Cuenca

Facultad de Ingeniería

Carrera de Ingeniería en Ciencias de la Computación

Analizar y aplicar técnicas de tratamiento de imágenes de periódicos antiguos del Ecuador para mejoras en el proceso de reconocimiento de textos (OCR).

Trabajo de titulación previo a la obtención del título de Ingeniero en Ciencias de la Computación

Autores:

Kevin Ismael Ochoa Arevalo

Lucia Carolina Quituisaca Suconota

Director:

Victor Hugo Saquicela Galarza

ORCID: 0000-0002-2438-9220

Cuenca, Ecuador

2023-07-26

Resumen

En el mundo, se están llevando a cabo proyectos de digitalización de documentos históricos con el objetivo de preservar la información contenida en ellos. Muchos de estos proyectos utilizan el Reconocimiento Óptico de Caracteres (OCR, por sus siglas en inglés). Sin embargo, actualmente no existen proyectos de este tipo en Ecuador. Durante el proceso de digitalización, surgen desafíos que afectan la calidad de la información obtenida mediante OCR, debido a problemas relacionados directamente con la imagen, como manchas, dobleces, iluminación, entre otros. Por lo tanto, es necesario buscar soluciones para contrarrestar estos problemas y obtener una mejor calidad de información.

En este trabajo de investigación se propone analizar técnicas de procesamiento de imágenes para mejorar los procesos de OCR con imágenes de periódicos antiguos del Ecuador. Se lleva a cabo un proceso de comparación y análisis de los datos obtenidos del OCR, centrándose en la cantidad de palabras correctamente reconocidas en las imágenes que fueron tratadas y no tratadas, con el objetivo de identificar mejoras en los resultados. Las técnicas de procesamiento, para facilitar el análisis, se dividen en tres grupos: técnicas tradicionales, técnicas de segmentación y técnicas de super resolución.

Los resultados demuestran que los procesos de super resolución, en particular la técnica LAPSRN, presentan una mejora significativa en los resultados del OCR. Estos hallazgos tienen importantes implicaciones para el campo de la preservación y acceso a la información histórica en Ecuador.

Palabras clave: digitalización, tradicional, segmentación, super resolución, información histórica

Abstract

Around the world, projects are being carried out to digitize historical documents with the aim of preserving the information contained in them. Many of these projects use Optical Character Recognition (OCR). However, there are currently no such projects in Ecuador. During the digitization process, challenges arise that affect the quality of the information obtained through OCR, due to problems directly related to the image, such as stains, folds, lighting, among others. Therefore, it is necessary to find solutions to counteract these problems and obtain a better quality of information.
In this research work we propose to analyze image processing techniques to improve OCR processes with images of old newspapers from Ecuador. A process of comparison and analysis of the data obtained from OCR is carried out, focusing on the number of words correctly recognized in the images that were treated and untreated, with the objective of identifying improvements in the results. The processing techniques, for ease of analysis, are divided into three groups: traditional techniques, segmentation techniques and super-resolution techniques.
The results demonstrate that super-resolution processes, in particular the LAPSRN technique, show a significant improvement in OCR results. These findings have important implications for the field of preservation and access to historical information in Ecuador.

Keywords: digitization, traditional, segmentation, super Resolution, historical Information

Índice de contenidos

Índice de figuras

Índice de tablas

Agradecimientos

En primer lugar, quiero expresar mi profundo agradecimiento a Dios por todas sus bendiciones, en especial por haberme dado una familia tan maravillosa que con su amor incondicional y apoyo constante han sido fundamentales en este largo recorrido de mi vida universitaria.

A mis padres, les debo un agradecimiento sincero, su dedicación, sacrificio y ejemplo han sido la fuente de inspiración que me ha impulsado a alcanzar mis metas. Siempre han estado a mi lado, brindándome su apoyo incondicional en cada momento importante de mi vida, y como no, extender mi agradecimiento a mis amigos que con su amistad y apoyo han ayudado de gran manera en mi vida académica y personal, ellos han sido testigos de mis logros y me han brindado su aliento en momentos difíciles, gracias por compartir esta travesía conmigo.

Kevin Ismael Ochoa Arévalo

Agradecimientos

Quiero expresar mi profundo agradecimiento a mis padres por su inmenso apoyo y por brindarme todas las oportunidades necesarias para llegar hasta este momento. Además, quiero agradecer a mis hermanos, Micaela, Sebastián y Ronald, quienes han sido una fuente constante de apoyo, motivación y cariño en cada paso que he dado.

Quiero expresar mi más sincero agradecimiento al director de mi tesis, Victor Saquicela. Sus enseñanzas, consejos y guía han sido fundamentales para el desarrollo de este trabajo de titulación. Sin su apoyo, este proyecto no habría sido posible.

No puedo dejar de agradecer a mis amigos, quienes han estado a mi lado a lo largo de esta travesía. Sus palabras de aliento, consejos y compañía han sido de gran valor para mí.

Lucia Carolina Quituisaca Suconota

1. Introducción

En este capítulo introductorio se presenta una visión general del contexto, motivación, planteamiento y solución del problema, además de la metodología utilizada en este trabajo de titulación, así también los objetivos que se desean cumplir.

1.1. Contexto

En el pasado, la información era transmitida de forma oral entre las personas, posteriormente con el avance de la ciencia y tecnología se comenzó a registrar la información en medios escritos, tales como periódicos, libros, inclusive pergaminos, en la actualidad mucha información valiosa e importante continúa en papel, lo cual representa un riesgo de pérdida, puesto que los medios escritos físicos se encuentran en constante deterioro, esto debido a las condiciones de almacenamiento, humedad, entre otros.

Dicho riesgo alimenta la idea de que solo la información que está digitalizada o disponible en línea de forma instantánea es relevante, según indica [5]. Uno de los escritos físicos que contiene una cantidad considerable de información, sobre todo histórica y que se encuentra en deterioro constante, es la prensa antigua, por lo cual, se está comenzando a realizar un proceso de digitalización, Neudecker y Antonacopoulos en [6], indican que la mayor parte del contenido digitalizado de periódicos tiene la forma de páginas escaneadas sin texto asociado, es decir, son imágenes puras, de las cuales no es posible obtener información almacenable o utilizable, tomando en cuenta a [7], que dicen que al manipular los documentos deteriorados para realizar un procesos de digitalización a menudo se pierden parte de ellos, lo cual conlleva a necesitar un proceso de restauración de documentos, o realizar un post procesado de los textos luego de haber sido digitalizados, con el objetivo de obtener la mayor cantidad de información posible. También hacen hincapié en que no todos los documentos se pueden digitalizar, pero al hacerlo, estos documentos antiguos pueden ayudar a entender las políticas, cultura y la sociedad en general [8], además de ello pueden mostrar acontecimientos históricos de los cuales no haya evidencia digital.

1.2. Motivación

En la vida diaria, social y científica, la combinación de nuevas tecnologías con las tecnologías antiguas es esencial para transmitir y almacenar información de manera efectiva. Esta combinación enriquece las opciones de acceso a la información, ya sea referente al pasado, presente o futuro. La tecnología permite una mayor accesibilidad a la información que, de lo contrario, sería difícil de manejar debido a su fragilidad y peligro de desaparición, como los documentos históricos. Por lo tanto, en el ámbito académico e investigativo, preservar y perpetuar la historia es una responsabilidad importante. Para lograr esto, se utilizan tecnologías de digitalización adaptadas a las necesidades de la sociedad actual, asegurando así la conservación, visualización y acceso a estos documentos.

Tener los acontecimientos históricos al alcance, es una tarea relevante, además que se puede tener un gran impacto dentro de la sociedad, debido a que permitiría

conocer más a fondo las raíces de un pueblo, además de conocer información sobre el clima, economía, entre otros temas importantes. Teniendo en cuenta la importancia de salvaguardar la información histórica, es necesario tener presente que los medios de almacenamiento no han sido los adecuados, siendo el papel la opción más común, el mismo que al paso del tiempo sufre deterioros y manchas considerables, dificultando la lectura al ojo humano, por ende existen muchas técnicas y nuevas tecnologías que permiten analizar cada uno de los píxeles de las imágenes, el mismo análisis permite la obtención de los diferentes textos, aunque las manchas, causan mucho ruido dificultando a la tecnología la detección clara de información, por ende es necesario tomar las imágenes y darle un tratamiento con técnicas adecuadas para ayudar a mejorar el análisis y por ende entregar menos ruido, además de textos más claros y comprensibles.

1.3. Planteamiento del problema

Los procesos OCR enfrentan varios problemas, entre los cuales se mencionan en [9]: el tipo de letra, caracteres perdidos, rotaciones y manchas en los textos, incluso el ancho de cada una de las letras. Todos ellos relacionados con la imagen ingresada, dando a entender así que la calidad de la información obtenida depende de la calidad de la imagen que ingresa, por ello es importante un tratamiento adecuado a la imagen previo al proceso OCR.

En el procesamiento de imágenes es muy difícil identificar las técnicas adecuadas para cada tipo de imágenes como indican en [10], puesto que existen una gran cantidad de técnicas, por ende, es necesario identificar, seleccionar y dar a conocer las técnicas que entreguen una mejor calidad de información cuando se aplique el proceso OCR. Es fundamental dar un tratamiento adecuado a las imágenes, para poder realizar procesos de digitalización, debido a que muchas de las veces una imagen con mucho ruido imposibilita hacer uso de motores OCR, tal y como menciona en [11], la prensa antigua contiene pérdida de fragmentos de texto, iluminación no uniforme, además de manchas, lo cual hace que el texto obtenido obtenga mucho ruido, pudiendo llegar a hacer que sea prácticamente ilegible.

1.4. Solución propuesta

En este trabajo de investigación se plantea explorar, probar, presentar y aplicar técnicas de tratamiento de imágenes que permitan obtener buenos resultados al realizar los procesos OCR sobre imágenes de periódicos antiguos, entendiéndose por buenos resultados una información clara, y lo más fidelizada a la imagen original. Se tomará en cuenta que cada imagen necesita un tratamiento diferente, todo esto con el objetivo de probar la hipótesis “al pasar una imagen previamente procesada o tratada por un proceso OCR entregará mejores resultados que una imagen sin tratar”.

Para la presente investigación se hará uso de imágenes de la prensa histórica del Ecuador que se encuentran en la Casa de la Cultura Ecuatoriana[1] que datan entre 1860 y 1920, dichas imágenes contienen mucha información valiosa que no puede ser utilizada en formato texto, ni almacenada, además al aplicar un proceso OCR

[1] http://repositorio.casadelacultura.gob.ec//

sobre las mismas, entregan mucho ruido, por lo cual se convierte en un escenario ideal para ser tomado como caso de estudio.

1.5. Objetivos

En esta sección se presentan los diferentes objetivos que se quieren alcanzar en el presente trabajo de titulación.

1.5.1. Objetivo general

Analizar técnicas de tratamiento de imágenes para mejorar los procesos OCR en periódicos antiguos digitalizados que permitan obtener una mejor calidad de información.

1.5.2. Objetivos específicos

1. Explorar técnicas de tratamiento de imágenes en procesos OCR.
2. Evaluar las técnicas encontradas.
3. Seleccionar las técnicas de tratamiento de imágenes que mejores resultados entreguen para el tratamiento de periódicos antiguos digitalizados.

1.6. Metodología

El proceso metodológico que se va a seguir es el propuesto por Wohlin [12], el mismo que consta de las siguientes fases: Identificación del problema, formulación del problema, estudio del estado de arte y una solución candidata, con la cual se quiere demostrar que el tratamiento de imágenes ayuda a la obtención de mejores resultados en procesos OCR. En la evaluación de cada una de las técnicas se realizarán comparaciones y análisis de los datos de entre los resultados del proceso OCR, se realizará un análisis de la precisión de las palabras de las imágenes, tanto procesadas como sin procesar, esto para identificar si existe o no la mejora en los resultados. Para obtener la precisión de las palabras, Olson y Berry [13] plantean la siguiente ecuación donde se compara el resultado del proceso OCR con el archivo original, contando las palabras correctas e incorrectas.

$$Presicion\,de\,las\,palabras\,(\,\%) = \frac{(\sum Palabras\ totales - \sum Palabras\ incorrectas)}{\sum Palabras\ totales} x100 \tag{1.1}$$

Los resultados que se desean obtener, son las técnicas de tratamiento de imágenes para prensa histórica que permita que un proceso OCR pueda entregar la información más fidelizada al documento físico original, esto para simplificar la elección de las técnicas al momento de dar tratamiento a imágenes de periódicos antiguos previo a procesos OCR.

1.7. Estructura del trabajo de titulación

El presente trabajo de titulación se estructura de la siguiente manera:

En el capítulo 2 se presenta el marco conceptual y trabajos relacionados en el cual se encuentran conceptos generales sobre inteligencia artificial y afines, así co-

mo la descripción de las técnicas de tratamiento de imágenes que van a ser probadas en los siguientes capítulos.

En el capítulo 3 se presenta una descripción detallada del procedimiento utilizado para evaluar diversas técnicas. Se exponen los pasos seguidos durante la investigación, destacando la aplicación de técnicas sobre las imágenes, la implementación del proceso OCR y la realización de la evaluación correspondiente.

En el capítulo 4, se presenta la conclusión basada en los resultados obtenidos, donde se indica si se han alcanzado los objetivos, junto con una propuesta para futuros trabajos de investigación.

2. Marco conceptual y Trabajos relacionados

En este capítulo, se incluyen y explican los principales conceptos con el objetivo de facilitar la lectura y comprensión del trabajo. Al proporcionar una explicación clara y concisa de estos conceptos, se busca asegurar que los lectores tengan una base sólida y una comprensión adecuada de los términos utilizados a lo largo del documento.

2.1. Inteligencia Artificial

Según [14] la inteligencia artificial es una rama multidisciplinaria de la ciencia que va tomando gran importancia en el avance de la misma, el objetivo principal de esta rama es lograr que una computadora pueda emular a un ser humano, siendo capaz de tomar decisiones por sí sola y seguir instrucciones.

2.1.1. Aprendizaje máquina

El aprendizaje máquina, según [15] se define como un proceso automatizado que extrae patrones de los datos. Para construir los modelos utilizados en las aplicaciones de análisis predictivo de datos, se utiliza el aprendizaje automático supervisado, es decir, aprenden automáticamente un modelo de la relación entre un conjunto de características descriptivas y una característica objetivo basándose en un conjunto de ejemplos históricos, o instancias. A medida que se abordan problemas más complejos como el reconocimiento de objetos y análisis de texto, los datos se vuelven extremadamente de alta dimensión. Para abordar esta complejidad, como menciona [16] la investigación reciente en aprendizaje automático ha intentado construir modelos que se asemejen a las estructuras usadas por el cerebro, es decir, cuando se aprenden cosas nuevas en la vida cotidiana desde muy temprana edad. Un ejemplo propuesto es cuando se enseña a un niño a reconocer un perro, no le enseñaron a reconocer un perro al medir su nariz o la forma de su cuerpo, sino que aprendió a reconocerlo al ver varios ejemplos y ser corregido cuando cometía errores.

2.1.2. Aprendizaje profundo

El aprendizaje profundo es una parte de un campo más específico del aprendizaje automático, se basa en la idea de aprender a partir de ejemplos. Es decir, en vez de enseñar a una computadora una lista masiva de reglas para resolver un problema, el aprendizaje profundo utiliza redes neuronales artificiales profundas, que consisten en múltiples capas de procesamiento de datos, para aprender de forma automática a partir de una gran cantidad de conjunto de datos para mejorar su capacidad para realizar tareas específicas [16].

2.1.3. Redes neuronales convolucionales

En [17] se menciona que las redes neuronales convolucionales (CNN) son una categoría específica de redes neuronales profundas que emplean la operación de convolución para procesar los datos de entrada. El incremento de su uso en los últimos tiempos se debe al desafío de reconocimiento visual a gran escala de ImageNet [18] de 2012. A diferencia de las redes profundamente conectadas, las CNN sobresalen en el procesamiento de datos con una organización espacial o de rejilla

(series de tiempo, imágenes, vídeos, etc.) y al mismo tiempo disminuyen el número de parámetros entrenables debido a sus propiedades de intercambio de peso.

2.2. Procesamiento y análisis de imágenes

El procesamiento y análisis de imágenes es un proceso que permite identificar partes importantes de una imagen, es decir, identificar los aspectos más relevantes para rescatarlos y posteriormente poder utilizarlos. En la representación digital de una imagen en 2D, se utiliza una estructura matricial compuesta por n filas y m columnas. Cada unidad elemental en esta matriz se denomina píxel, y corresponde a la intersección específica entre una fila y una columna [19]. La información visual de la imagen se codifica en los valores numéricos asignados a cada píxel. Dependiendo del tipo de imagen que se considere, los valores pueden representar la intensidad lumínica o el color [20].

La mayoría de las imágenes de gráficos por computadora utilizan el espacio de color rojo, verde y azul (RGB). Este se basa en la mezcla de tres luces primarias: rojo, verde y azul. Al atenuar estas luces desde apagado (valor de píxel 0) hasta encendido (valor de píxel 1), se puede crear cualquier color que se pueda visualizar en un monitor. Los niveles de los colores RGB se especifican en un número entero entre 0 y 255, lo que equivale a 256 posibles niveles para cada componente [19]. Por lo general, al adquirir imágenes a través de sensores modernos, estas pueden estar contaminadas por una variedad de fuentes de ruido. Por ruidos se entiende como sombreado o falta de enfoque [20].

2.3. Reconocimiento Óptico de Caracteres (OCR)

Reconocimiento Óptico de Caracteres (OCR), es un software que permite reconocer el texto de imágenes, es decir, convertir el texto de imágenes a cadenas de texto completamente utilizables, permitiendo realizar búsquedas [21]. Cada OCR tiene diferentes especificaciones, como por ejemplo, el tipo de imágenes que recibe (PNG, JPG,TIFF,etc),algunos siendo capaces de recibir hasta PDF. En [21] identifica algunas de las posibles aplicaciones que se puede dar a dicha tecnología, como almacenamiento de facturas, Captchas, creación de bibliotecas digitales, reconocimiento de eventos históricos, etc. Una de las herramientas más utilizadas es Tesseract, el cual es un motor de reconocimiento óptico de caracteres(OCR), cumple con la función de convertir imágenes de texto en texto utilizable. Según la documentación oficial de Tesseract “se desarrolló en Hewlett-Packard Laboratories Bristol UK y en Hewlett- Packard Co, Greeley Colorado USA entre 1985 y 1994, desde 2005 el código es abierto por HP y desde 2006 hasta 2018 fue desarrollado por Google”[22], desde Tesseract 4 es posible utilizar redes neuronales de corto-largo plazo (LTSM), las mismas que son un tipo de Red neuronal recurrente (RNN), capaz de aprender dependencias a largo plazo [23], es decir que es capaz de capturar y comprender las relaciones entre elementos distantes o que no guarden características en común.

2.4. Técnicas de tratamiento de imágenes

Se presentan las definiciones de las diferentes técnicas de tratamiento de imágenes probadas en este trabajo de titulación.

2.4.1. Escala de Grises

Por lo general los Motores OCR funcionan mejor con imágenes a escala de grises, aunque Tesseract puede trabajar con imágenes a color. Este proceso consiste en convertir los colores de las imágenes en blanco, gris y negro según el grado de luz, mientras más luz haya en la imagen, el color va a tender a blanco y por lo contrario, mientras menos luz, el color tenderá a negro [24]. Según [25] el convertir adecuadamente una imagen a escala de grises ayuda a obtener de mejor manera las características de dichas imágenes, lo cual resulta beneficioso en un proceso OCR, ya que se evalúa píxel a píxel para reconocer las diferentes letras.

2.4.2. Binarización

La binarización es una técnica de tratamiento de imágenes, cuyo objetivo es convertir una imagen en escala de grises a una imagen a blanco y negro [26], lo cual elimina las diferencias de intensidad de luz, permitiendo la mejor identificación de objetos dentro de la misma. Al ser la iluminación el principal aspecto y problema a tomar en cuenta en ésta técnica, se han propuesto varias técnicas de binarización, las cuales intentan solucionar dicho problema.

2.4.2.1. Binarización Simple

Es una de las técnicas más básicas en el tratamiento de imágenes, utilizada para realizar el agrupamiento de píxeles [27], ya que convierte una imagen en escala de grises a una imagen en blanco y negro, todo ello con el objetivo de lograr identificar objetos dentro de la misma. Este método se basa en los umbrales, que no es más que determinar un valor donde la intensidad del tono de grises cambia, tomando en cuenta el valor del objeto que se desea rescatar con respecto al resto de la imagen [28], lo cual permite diferenciar el objeto a rescatar del resto de la imagen, y luego mediante una función determinar si el valor del píxel supera el valor del umbral, lo coloca de un color (por lo general blancos) y el resto de otro color (por lo general negro), dando como resultado una imagen en dos colores claramente diferenciados.

2.4.2.2. Binarización de Otsu

Esta técnica de binarización hace un análisis previo a la imagen, obteniendo el umbral de forma automática, basado en una varianza global [29], es decir se toma todos los píxeles de una imagen y se analiza la dispersión de sus valores, lo cual muestra como los grises varían dentro de la imagen, entonces con dicho análisis se marcan la diferencia entre áreas oscuras y claras, lo cual entrega el valor del umbral a ser utilizado en cada una de las imágenes. Ésta técnica es similar a la anterior, teniendo el mismo fundamento de funcionalidad, además de tener el mismo objetivo, pero difieren en que la técnica de Binarización simple, tiene un mismo valor de umbral fijo para todas las imágenes, mientras que Otsu realiza un cálculo previo según la dispersión de tonos.

2.4.2.3. Binarización Adaptativa

La binarización adaptativa se puede decir que es una combinación de los dos métodos anteriores, funciona con valores umbrales, los mismos que va calculando según los valores de los píxeles aledaños a píxel que se desea umbralizar. [30].

Con ésta técnica el valor del umbral varía en cada una de las zonas de la imagen, diferenciando de mejor manera los objetos de interés, del resto de la imagen.

2.4.3. Filtro Mediano

El filtro mediano analiza cada uno de los píxeles de una imagen, va cambiando el valor de cada uno de los píxeles, según el valor promedio de los píxeles adyacentes [31], este filtro por lo general es utilizado para eliminar ruidos normalmente llamados "de sal y pimienta", que no son más que puntos dispersos en la imagen debido a problemas de iluminación o de la calidad de la imagen, también sirve para quitar los valores pico tanto superiores como inferiores, y dejar la imagen en un tono uniforme.

2.4.4. Filtro de Gauss

Según [32], el filtro de Gauss o Filtro Gaussiano, es una técnica de tratamiento de imágenes que tiene como objetivo suavizar partes de la imagen donde la intensidad del píxel sea homogéneo, con ésta técnica se logra eliminar los objetos más pequeños de la imagen, dejando los principales objetos claramente diferenciados del resto de la imagen, en [33] se menciona que en relación al filtro mediano, produce una imagen más suave, aunque con menos nitidez.

2.4.5. Contraste

El contraste es definido como la diferencia de iluminación entre el entorno y un objeto, por lo general se toman zonas brillantes y oscuras [34], esta técnica sirve para hacer que la imagen sea más amigable a vista humana, también es usada para destacar zonas de la imagen, haciendo posible rescatar objetos.

2.4.6. Técnicas de super resolución

El objetivo de los algoritmos de super resolución (SR) es generar imágenes de alta resolución (High-Resolution HR) a partir de una entrada de imagen de baja resolución (Low-Resolution LR) con un enfoque basado en aprendizaje profundo [35][36][2][1][3].

2.4.6.1. Fast Super-Resolution Convolutional Neural Networks (FSRCNN)

La Red Neuronal Convolucional de Super Resolución (SRCNN), propuesto por[36] tiene varias características que lo hacen atractivo. En primer lugar, su estructura fue diseñada con la intención de ser sencilla, pero también ofrece una precisión superior en comparación con otros métodos. Tiene tres etapas, la primera etapa se basa en la extracción de características importantes de la imagen de baja resolución. Estas capas convolucionales aprenden filtros para capturar patrones y detalles específicos de la imagen. En la siguiente etapa se aplica una capa de mapeo no lineal, esta capa mejora la capacidad del modelo para comprender las complejas relaciones entre las características extraídas. En la última etapa, se utiliza una capa de convolución para convertir las características mapeadas en una imagen de alta resolución. Para producir una imagen de mayor resolución, esta capa de convolución reconstruye las texturas y los detalles de alta frecuencia. El SRCNN se entrena con datos de pares de imágenes de baja y alta resolución. El modelo aprende a mapear efectivamente las características de baja resolución a las características de alta resolución durante el entrenamiento, lo que permite generar imágenes de alta calidad. La función de

minimización de pérdidas compara las imágenes generadas por el modelo con imágenes reales de alta resolución para facilitar el entrenamiento.
El enfoque FSRCNN[2] se creó con el objetivo de mejorar la velocidad de SRCNN mientras se mantiene la calidad de restauración. Como resultado, una de las principales ventajas del FSRCNN es su rapidez. Al tener una arquitectura eficiente, puede acelerarse hasta cuarenta veces, produciendo un rendimiento de SR satisfactorio y un tiempo de ejecución superior. Esto se debe al hecho de que el algoritmo FSRCNN utiliza una combinación de capas de convolución y sub-píxel para extraer características, aumentar la resolución y refinar la imagen generada. En las capas sub-píxeles se utilizan para reconstruir los detalles perdidos y aumentar la resolución de la imagen. Se puede observar la comparación visual en la Figura 2.1.

2.4.6.2. Efficient Sub-Pixel Convolutional Neural Network (ESPCN)

El ESPCN[2] se caracteriza por su eficiencia computacional y su capacidad para generar resultados de super resolución en tiempo real. El funcionamiento del algoritmo se describe en los siguientes pasos: en la etapa principal se utilizan capas de convolución para extraer características significativas de la imagen de baja resolución. Estas capas convolucionales se encargan de aprender filtros que capturan detalles y patrones relevantes en la imagen; la siguiente etapa se aplica una capa de convolución conocida como convolución sub-píxel. Esta capa está diseñada para aumentar la resolución espacial de la imagen. Funciona mediante la convolución de cada píxel de entrada con los filtros aprendidos y luego reorganizando los píxeles para aumentar la resolución de la imagen, la última etapa, se aplica otra capa de convolución para refinar la imagen de alta resolución generada. Esta capa adicional ayuda a mejorar la calidad visual de la imagen y a eliminar cualquier artefacto no deseado.
Una diferencia destacada es la eficiencia computacional. ESPCN es conocido por ser más eficiente en términos de cálculo y tiempo de ejecución debido a su enfoque basado en la convolución sub-píxel. La convolución sub-píxel permite aumentar la resolución directamente en una sola etapa, evitando la necesidad de capas adicionales o de Sub-muestreo. La Figura 2.1 muestra una representación visual.

2.4.6.3. Layered Recursive Super-Resolution Network (LAPSRN)

El algoritmo de super resolución LAPSRN[3] se caracteriza por su enfoque de reconstrucción en capas y su capacidad para capturar detalles finos en las imágenes de salida. El funcionamiento del algoritmo LAPSRN se puede describir en los siguientes pasos: primero el procesamiento de la imagen de baja resolución se somete a un preprocesamiento para ajustar su tamaño y formato; el segundo paso es la reconstrucción en capas, en este utiliza una estructura de red en capas, donde cada capa tiene su propia función de super resolución. La primera capa se encarga de aumentar la resolución de la imagen inicial de baja resolución. Luego, las capas subsiguientes toman la imagen de salida de la capa anterior y la refinan aún más. Este proceso se repite en varias capas, lo que permite una reconstrucción en cascada de la imagen de baja resolución a una imagen de alta resolución. El tercer paso son las conexiones recursivas, es decir, además de las conexiones entre capas, LAPSRN emplea conexiones recursivas que retroalimentan la información de las capas anteriores a las capas posteriores. Esto permite que la información

de alta resolución se propague y se refine a medida que pasa a través de las capas sucesivas, lo que ayuda a capturar detalles finos y preservar la coherencia de la imagen; finalmente, el entrenamiento, se realiza usando un conjunto de datos de pares de imágenes de baja y alta resolución. Durante este proceso, el modelo aprende a mapear eficazmente las características de baja resolución a características de alta resolución mediante la minimización de una función de pérdida que compara las imágenes generadas por el modelo con las imágenes de alta resolución reales. La comparación visual se puede ver en la Figura 2.1.

Figura 2.1: Ejemplos super resolución a) “Butterfly” FSRCNN factor de aumento de escala 3 [1], b) “Monarch” ESPCN factor de aumento escala 3 [2], c) “MukoukizuNo-Cho” LAPSRN factor de aumento escala 4[3].El PSNR es una medida utilizada para evaluar la calidad de una imagen, se expresa en decibelios (db).

2.4.7. Técnicas de segmentación de página

Los componentes fundamentales de un documento se encuentran en las líneas de texto individuales y los bloques de texto[37]. Para el análisis de documentos, se recurre a la segmentación, la cual consiste en subdividir el área de un documento en secciones o bloques que contienen texto [38]. Históricamente, los métodos tradicionales de segmentación de texto utilizaban algoritmos de visión por computadora sencillos. No obstante, en la actualidad, los métodos convencionales han sido superados por los enfoques basados en redes neuronales profundas. Previo a la ejecución de cualquier tarea de procesamiento de imágenes de documentos, resulta crucial implementar tanto el análisis de diseño como la segmentación de páginas. Estas etapas preliminares son necesarias para asegurar un procesamiento

adecuado y eficiente de los documentos [37].

2.4.7.1. Newspaper Navigator

Newspaper Navigator [4] es uno de los modelos disponibles de Layout Parser [39] que se ha creado para analizar y extraer contenido visual de una gran colección de 16,358,041 páginas de periódicos históricos que se encuentra en Chronicling America. Para realizar esta tarea, se empleó un modelo de detección de objetos previamente entrenado utilizando anotaciones específicas obtenidas de las páginas de Çhronicling Americaçorrespondientes a periódicos de la Primera Guerra Mundial. Este proceso ha producido un modelo de reconocimiento de contenido visual que puede reconocer una variedad de tipos de contenido que se encuentra en las imágenes que se extrajeron. Algunos de los tipos de contenido que este modelo puede reconocer son:

- Fotografía
- Ilustración
- Mapa
- Cómics/Dibujos animados
- Caricatura editorial
- Titular
- Anuncio

Basándonos en el modelo Newspaper Navigator, se ha tomado la decisión de emplearlo como método de segmentación para dividir distintas secciones de los periódicos. En la Figura 2.2 se puede observar la detección de algunos tipos de contenido.

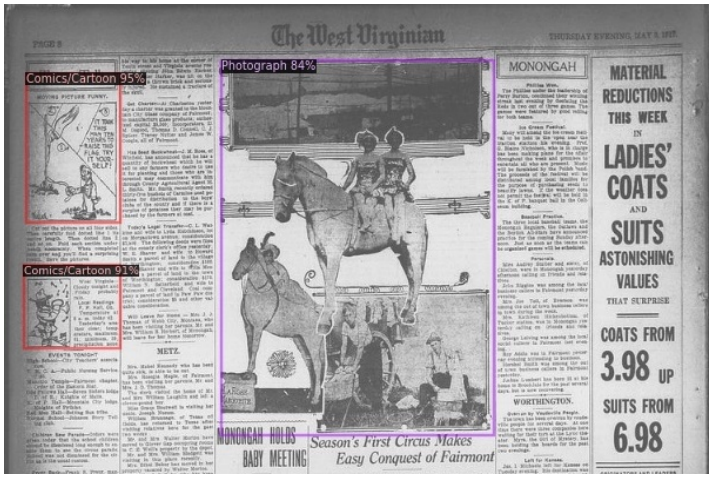

Figura 2.2: Identificación de contenido según el modelo Newspaper Navigator [4].

2.5. Trabajos relacionados

En ésta sección se va ha presentar los trabajos que más relación guardan con el tratamiento de imágenes previo a procesos OCR, así mismo, se presentan algunos

de los diferentes proyectos de digitalización con procesos OCR que se han llevado a cabo alrededor del mundo.

2.5.1. Tratamiento de imágenes

La metodología de preprocesamiento de [38] propuesta para la recuperación de textos, consta de varios pasos, los cuales son: escaneo de la imagen; binarización; eliminación del ruido; detección y corrección de sesgos para finalmente realizar la segmentación de página, partiendo de la binarización, cuyo objetivo es tener una imagen en dos colores (Blanco y Negro), el siguiente paso es eliminar el ruido, el cual es un proceso muy importante para mejorar la calidad de una imagen, en dicho estudio se recomienda utilizar el filtro mediano, el mismo que utiliza una suavización de la imagen. El tercer paso es la detección y revisión de sesgos, par lo cual hay que seguir seis pasos: aplicar el filtro Wiener [40]; detectar el cuadro delimitador; aplicar el algoritmo de smearing [41]; detección de bordes; el aplicar la transformación de Radón [42] y el paso final es la revisión de inclinación.

Para el preprocesamiento de imágenes, en [43] indican la necesidad de convertir las imágenes en escala de grises esto mediante el uso del método de umbral adaptativo local, el siguiente paso es la rotación automática de imágenes, el otro paso es la segmentación, ya que mencionan que se puede obtener una mejora en la precisión del proceso OCR. Al igual que [44] menciona que la imagen de entrada debe pasar de escala de grises a RGB, el proceso de binarización se debe realizar mediante el método de Otsu y para la eliminación de ruido se debe utilizar el filtro mediano.

En la revisión sistemática de [45], indican que el preprocesamiento es fundamental para obtener tasas de reconocimiento más altas, en dicha revisión sistemática se inicia con la binarización, posteriormente se aplican operaciones de filtrado de imágenes espaciales, otro paso es la umbralización, puesto que aísla datos de partes innecesarias de la imagen, el siguiente paso es la eliminación de ruido y finalmente la detección o corrección de inclinación, También se menciona que la segmentación es un paso importante, comenzando por la segmentación de página, lo cual ayuda a aislar contenido del resto de la imagen, los siguientes pasos son la segmentación de caracteres, la normalización del tamaño de la imagen y finalmente el procesamiento morfológico, es decir eliminar ciertos píxeles durante la umbralización para que la imagen sea más fácil de procesar.

La construcción de un sistema eficiente para documentos históricos propuesto por [37], el cual consiste en una secuencia de pasos. En primer lugar, se realiza el procesamiento de la página de entrada, seguido por al aplicación del método de binarización para obtener una imagen binaria. En este caso, se sugiere el uso del método de Otsu [46] para llevar a cabo la binarización. El siguiente paso es la segmentación de bloques. Para ello, se emplean redes convolucionales que son capaces de predecir una máscara que indica las posiciones de las regiones de texto en la página, el mismo que está basado en el modelo U-Net [47].

La evaluación realizada por [48] indica la necesidad de un método robusto de súper resolución debido a que las imágenes de baja resolución obtenidas por una cámara afectan la precisión al obtener el texto al pasar por el OCR. Por lo cual

analiza 6 distintos métodos de súper resolución con el fin de evaluar el desempeño, puesto que son métodos que han demostrado ser eficientes en imágenes. Utiliza tres métricas de evaluación, como la relación pico señal a ruido, índice de similitud de estructura y la precisión de reconocimiento óptico de caracteres. Obteniendo como resultados que las redes con capas de reconstrucción de múltiples escalas funcionan mejor que una capa de una sola escala.

En base a la revisión de la literatura realizada, se puede evidenciar que existen varias técnicas que son comúnmente utilizadas para el tratamiento de imágenes previo a procesos OCR, por lo cual en este trabajo de titulación se van a probar las técnicas de: escala de grises; binarización, tanto de Otsu como Adaptativa; filtro mediano; filtro gaussiano, además de algoritmos de segmentación y de mejora de la resolución de las imágenes.

2.5.2. Reconocimiento Óptico de Caracteres (OCR)

En varios países se han llevado a cabo procesos de digitalización de documentos históricos, como por ejemplo en Finlandia [49], los mismos toman documentos históricos, los pasan por un proceso de escaneado, además por el software ABBY Fine Reader OCR obteniendo un 81 % de efectividad[49], lo cual indica que el 19 % de la información presente en dichos documentos se ha perdido, esto se puede deber a varios factores relacionados directamente con la imagen, los cuales pueden ser distorsiones, el tamaño, estilo y color de la fuente, iluminación, resolución entre otros [50], aunque también depende del proceso OCR utilizado para la digitalización, luego en [49] se menciona que se realizó el entrenamiento de Tesseract con redes neuronales, y se volvieron a pasar los documentos históricos Finlandeses, dicho proceso entrega una mejora de resultados de hasta un 9 % que los datos obtenidos por Abby Fine Reader.

Otro proyecto de digitalización se llevó a cabo en Uruguay con el objetivo de obtener eventos climáticos [51], en dicho paper toma documentos de prensa Uruguaya del siglo XIX, los convierte en imágenes, lleva a cabo un procesamiento OCR (Abby Fine Reader) para finalmente realizar clustering de los eventos encontrados en el texto, sin embargo, se hace referencia a la presencia de una gran cantidad de interferencias en los documentos antiguos, lo que resulta en la pérdida de información y limita la obtención de datos significativos, por consiguiente, en las pruebas realizadas toma únicamente periódicos actuales.

En Barcelona [52] también realizaron digitalización de prensa, se toman los periódicos desde 1933 hasta 1939 y se pasa por un proceso OCR para permitir efectuar búsquedas en las imágenes de periódicos, aunque toma únicamente recortes, y sin pasar por ninguna técnica de tratamiento previa. Si bien existen varios trabajos de digitalización de documentos históricos, en muchos de los casos, hay gran pérdida de información, debido a que la imagen que se está tomando tiene un nivel de degradación bastante alto, lo cual genera ruido en los textos obtenidos, provocando pérdida de información, inclusive llegando a hacer que no sea óptimo utilizar OCR en la digitalización.

Tal y como se menciona, en el mundo se están ejecutando varios proyectos de digitalización de archivos [53], aunque muchas de las veces solo se realiza un pro-

ceso de escaneado, cuyos resultados no entregan un texto sobre el cual se pueda extraer información y muchos otros únicamente presentan un texto resultado de un proceso OCR simple [53] lo cual extrae una cierta cantidad de información pero contiene mucho ruido. Al tener una gran variedad de motores OCR es importante buscar el que entrega mejores resultados, en algunas de las comparativas como [54], se toman varios software OCR como Google Docs [1], Tesseract[2], Abby Fine Reader[3] y Transym [3], los cuales se someten a pruebas de precisión, y rendimiento, esto para determinar el OCR que mejor resultados entrega, dando como resultado que Abby Fine Reader es el mejor software de reconocimiento de caracteres, seguido por Google Docs y Tesseract, destacando Tesseract como software de código abierto, además que se puede utilizar en varios lenguajes de programación.

En otro estudio comparativo [55] en donde se toman archivos médicos para digitalización, se pone a prueba los softwares de PyOCR[4], PyTesseract y Tesse OCR, todos ellos de código abierto,luego de realizar algunas pruebas se concluye que PyTesseract es el mejor software para digitalizar documentos, esto debido a que es capaz de identificar las diferentes zonas del documento, además de que es el primer OCR en ser capaz de manejar imágenes a color, esto debido a que internamente realiza binarización de la imagen antes de extraer los textos, además de ser capaz de realizar un segmentación interna.

Finalmente [56] hace un análisis comparativo bastante completo, distinguiendo entre softwares OCR de pago, de código abierto y de OCRs que son prestados como servicio, se toman varios softwares entre los que figuran Abby Fine Reader, Transym, Readiris[5], Adobe Acrobat[6], Omni Page[7], Microsoft Office Document Imaging[8] , entre los de pago, mientras que entre los de código abierto están Tesseract, Ocrad[9], OCRopus[10], GOCR[11],Cuneiform [12] y Calamari[13], entre los que son brindados como un servicio están, Abby Web Service, Google Docs, Free-Online OCR[14] y Online OCR, dicha investigación entrega como resultado el mejor software de cada categoría, y entre los mejores figuran Abby Fine Reader (en categoría de pago y categoría de servicio), Tesseract (en código abierto), también se menciona que Tesseract es capaz de funcionar con Redes neuronales del tipo LSTM (redes recurrentes), lo cual lo hace más atractivo,debido a que se puede volver a entrenar la

[1]https://support.google.com/drive/answer/176692?hl=es-419&co=GENIE.Platform %3DDesktop

[2]https://tesseract-ocr.github.io/tessdoc/

[3]https://www.abbyy.com/ocr-sdk/features/documentation/

[3]https://github.com/Transym

[4]https://pypi.org/project/pyocr/

[5]https://www.irislink.com/ES-EC/c2251/Readiris-PDF–Funciones-de-escaneado-y-OCR.aspx

[6]https://www.adobe.com/mx/acrobat/how-to/ocr-software-convert-pdf-to-text.html

[7]https://docs.uipath.com/es/activities/other/latest/user-guide/omnipage-ocr

[8]https://support.microsoft.com/es-es/topic/instalar-modi-para-utilizarlo-con-microsoft-office-2010-4fbd3076-6d01-9cb7-c574-3bbabc9eead9

[9]https://www.gnu.org/software/ocrad/manual/ocrad$_manual.html$

[10]https://github.com/ocropus/ocropy

[11]https://docs.oracle.com/cd/E88353$_0$1$/html/E37839/gocr-1.html$

[12]https://github.com/cdli-gh/Cuneiform-OCR

[13]https://calamari-ocr.readthedocs.io/en/latest/

[14]http://www.free-online-ocr.com#: :text=Free %20Online %20OCR %20is %20a,install %20anything %20on %20your %20computer.

red con datos propios, para buscar mejores resultados en un proyecto en específico, dando la idea que aparte de los problemas que la imagen entrega, también depende de la elección del OCR que se utiliza.

Tomando en cuenta los documentos revisados se puede evidenciar que en Ecuador no se ha llevado a cabo un proceso adecuado de digitalización de prensa antigua, ya que los documentos existentes son imágenes dentro de un en formato PDF, siendo inutilizable para llevar a cabo un análisis de acontecimientos históricos, por lo cual, en este trabajo de titulación, además de probar las técnicas de tratamiento de imágenes mencionadas en la sección anterior, se va a utilizar Tesseract como motor OCR, debido a que en las diferentes comparativas muestran que es uno de los mejores, además de permitir utilizar redes neuronales para la detección de textos.

3. Proceso de análisis y evaluación de técnicas de tratamiento de imágenes

En este capítulo se describe el proceso de evaluación de diversas técnicas de tratamiento de imágenes previo a un proceso OCR, con el objetivo de identificar aquellas que contribuyan a mejorar la efectividad del reconocimiento óptico de caracteres y permitan recuperar la mayor cantidad de texto posible de las imágenes de periódicos antiguos.
La Figura 3.1, ilustra el proceso usado para realizar la evaluación de las diferentes técnicas de tratamiento de imágenes para procesos OCR. El proceso inicia con la selección de una muestra específica. En una parte del proceso, se realizó la transcripción manual de los periódicos, mientras que en otra parte los documentos PDF se transformaron en imágenes PNG, se aplicaron técnicas de procesamiento de imágenes, posteriormente se aplicó el proceso OCR. Finalmente, se realiza una evaluación en base a la comparación de palabras. A continuación se detalla cada uno de los pasos.

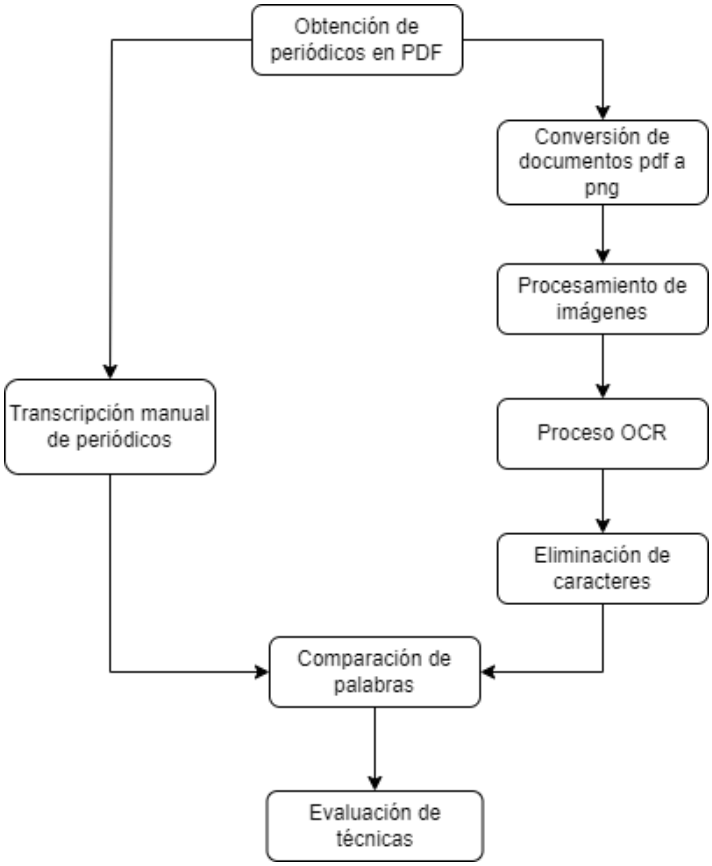

Figura 3.1: Diagrama de los procesos realizados en el presente estudio

3.1. Obtención de periódicos en PDF

A continuación, se detalla el proceso mediante el cual se obtuvo la muestra utilizada en este estudio. La muestra seleccionada representa un conjunto significativo de la población, en este caso, los 15679 periódicos antiguos ecuatorianos. Además,

se proporcionarán características relevantes de esta muestra para una mejor comprensión de su composición y representatividad.

3.1.1. Selección de la muestra

Al revisar los periódicos denominados Prensa Antigua[1] de la Casa de la Cultura Ecuatoriana, se obtuvieron un total de 15679 periódicos de entre los años 1860 y 1920, los mismos que se tomaron como la población total de este trabajo de titulación. Para establecer la muestra de una población finita basada en la fórmula 3.1, presentada en [57].

$$n = \frac{Z^2 \cdot N \cdot p \cdot q}{e^2 \left(N - 1\right) + \left(Z^2 \cdot p \cdot q\right)} \tag{3.1}$$

En donde:
n: tamaño muestral
Z: nivel de confianza
e: error estimado máximo aceptado
p: probabilidad de éxito
q; probabilidad de fracaso
Se ha fijado un nivel de confianza del 95 % (Z=1.96) con un error de estimación del 5 %, además, la probabilidad de éxito con un valor de 0,5, esta decisión se basa en que no se tiene información previa de la probabilidad de éxito del evento estudiado, por lo tanto, la probabilidad de fracaso es del q=(1-p)=0,5 asumiendo que la probabilidad de éxito es igual a la proporción de fracasos así asegurando que no haya sesgos en la estimación. Con un tamaño poblacional de 15679 periódicos obtenidos de la Casa de la Cultura.

$$n = \frac{1{,}96^2 \cdot 15679 \cdot 0{,}5 \cdot 0{,}5}{0{,}05^2 \left(15679 - 1\right) + \left(1{,}96^2 \cdot 0{,}5 \cdot 0{,}5\right)} = 375 \tag{3.2}$$

A pesar de que se obtuvo un total de 375 periódicos como resultado de la ecuación 3.2, la muestra consistió en una selección de 50 periódicos en PDF, estos se eligieron minuciosamente para asegurar que se abarcaran todos los tipos de periódicos y se tuviera en cuenta la variedad de ruidos presentes en los documentos seleccionados, con el fin de brindar una comprensión general del conjunto de datos e identificar patrones o tendencias iniciales. Aunque no se procesó la muestra completa de 375 periódicos, los datos analizados proporcionan información relevante y permiten extraer conclusiones significativas dentro de los límites establecidos. La elección de procesar una muestra parcial en lugar de toda la muestra estuvo motivada por las limitaciones de los recursos computacionales disponibles, los cuales no eran suficientes para procesar toda la muestra en el tiempo disponible, además es necesario llevar a cabo la transcripción de los documentos para realizar la evaluación, tal como se explica en la sección 3.2 y este proceso requiere una considerable inversión de tiempo. Por consiguiente, se decidió procesar una muestra más reducida que pudiera procesarse dentro de los límites de tiempo, de los recursos computacionales y humanos disponibles.

[1] http://repositorio.casadelacultura.gob.ec/handle/34000/1534

3.1.2. Descripción de la muestra

Es importante resaltar que se cuenta con 198 páginas de periódicos de los 50 seleccionados. Cada periódico normalmente consta de cuatro páginas, con la excepción de un periódico que tiene solo dos páginas. Cabe indicar que los periódicos analizados presentan una variedad de características, las cuales se describen detalladamente a continuación:

- La estructura común de los periódicos incluye una página principal que contiene un encabezado con el nombre de la editorial, la fecha de publicación y el contenido organizado en columnas.
- Las siguiente páginas, también presentan un encabezado pequeño con el nombre de la editorial, contenido en columnas y un posible pie de página, como se ilustra en la Figura 3.2.
- La cantidad de columnas varía dependiendo de la editorial y el tipo de contenido, generalmente oscila entre dos y seis columnas.
- Por lo general, la última página del periódico se destina principalmente a la publicación de anuncios.

Figura 3.2: Ejemplos de la estructura de distintos periódicos.

Es importante destacar que algunos periódicos muestran signos de deterioro causado por el paso del tiempo, incluso presentan manchas debido a las condiciones de almacenamiento inadecuado, como la humedad y otros factores, los mismos que se pueden distinguir en color marrón amarillento, mientras que otros presentan traspaso de tinta o letras de otras páginas, además de sellos cada uno de estos casos se pueden observar en la Figura 3.3.

Figura 3.3: Ejemplos de manchas, deterioro, sellos y traspaso de tinta en distintos periódicos.

Además, algunos ejemplares pueden tener páginas arrugadas, lo que puede causar la pérdida de letras o palabras. Los problemas de escaneo, como capturas inexactas o dobleces en el periódico, pueden causar pérdida de información o ruido visual al digitalizar archivos. Asimismo, es común encontrar fragmentos rotos en las muestras, estos ruidos se pueden visualizar en la Figura 3.4. Cabe aclarar que hay periódicos que se encuentran en perfectas condiciones.

Figura 3.4: Ejemplos de dobleces, escaneo inexacto y fragmentos rotos presentes en distintos periódicos.

3.2. Transcripción manual de periódicos

Una vez seleccionada la muestra, es necesario obtener el texto tal y como se encuentra en el periódico, esto para tener un texto con el cual comparar el resultado obtenido con el OCR, para lo cual hay que realizar una transcripción manual de los periódicos, debido a que el ser humano es el único ser capaz de identificar los diferentes textos, inclusive si estos contienen manchas, diferencias de iluminación o algún tipo de ruido que dificulte la lectura.
Debido a que la muestra contiene 50 periódicos, y en total 198 páginas, fue necesario solicitar la colaboración de los alumnos de las materias de Probabilidad y Estadística e Inteligencia Artificial de la carrera de computación de la Universidad de Cuenca, además de gente cercana a los autores. El proceso seguido fue el siguiente: primero se agregaron en una unidad de Google Drive[2] todos los documentos seleccionados en formato PDF; luego a cada una de las personas que colaboraron, se le asignó un documento almacenado en Drive, y se solicitó que el mismo deba ser transcrito lo más fiel posible al PDF tomando el formato de las columnas como factor determinante, además manteniendo los caracteres de punto, coma y guión medio, esto se debe a que algunos periódicos, debido a su formato particular, utilizan guiones medios para separar las palabras y el resto de caracteres especiales muchas veces aparecen como ruido por ende se ha decidido no tomarlos en cuenta para el transcrito. Además, se conservaron las palabras con faltas de ortografía, errores gramaticales y errores de escritura, es decir, se transcribió tal como se presenta en el texto original. Es posible apreciar un fragmento de la transcripción realizada en la Figura 3.5


BAHIA de CARAQUEZ

LA CAMARA COMERCIAL

El notable desarrollo que ha alcanzado el comercio de Bahía en los últimos años, y las necesidades diarias para ensanchar las transacciones mercantiles y fijar la base en los cálculos, hizo indispensable la instalación de la Cámara comercial que ha entrado al segundo año de su existencia, con éxito siempre creciente para las negociaciones, supuesto que el comercio no es otra cosa sino una manifestación evidente de esas actividades que impulsan á la humanidad á la transformación gradual que impele á los individuos hacia el progreso señalado por la Providencia.

No es para estas líneas patentizar el método por el cual la fuerza económica de los valores en acción realizan la misión civilizadora, ya por la traslación de los bienes de las manos del productor á las del consumidor, ya por medio del milagroso agente del crédito que junto á otros factores forman el flujo y reflujo de las operaciones comerciales.

BAHIA de CARAQUEZ
LA CAMARA COMERCIAL
El notable desarrollo que ha alcanzado el comercio de Bahía en los últimos años, y las necesidades diarias para ensanchar las transacciones mercantiles y fijar la base en los cálculos, hizo indispensable la instalación de la Cámara comercial que ha entrado al segundo año de su existencia, con éxito siempre creciente para las negociaciones, supuesto que el comercio no es otra cosa sino una manifestacion evidente de esas actividades que impulsan á la humanidad a la transformación gradual que Impele a los individuos hacia el progreso señalado por la Providencia.
No es para estas líneas patentizar el método por el cual la fuerza económica de los valores en acción realizan la misión civilizadora, ya por la tralación de los bienes de las manos del productor a las del consumidor, ya por medio del milagroso agente del crédito que junto a otros factores forman el flujo y reflujo de las operaciones comerciales.

Figura 3.5: Formato de la transcripción manual del periódico.

Los participantes generaron el documento transcrito en archivo TXT, y lo subieron a un espacio en Google Drive para el correcto almacenamiento y posterior análisis, además entregaron el número de palabras que han transcrito, lo cual resulta de vital importancia para llevar el proceso de evaluación. El resultado de la transcripción manual, permite definir una línea base que posteriormente será utilizada en

[2]https://www.google.com/intl/es/drive/

el proceso de evaluación, el documento detallado del experimento se encuentra en el Anexo A.

3.3. Conversión de documentos PDF a PNG

Debido a que la mayoría de los formatos de imágenes RGB utilizan ocho bits para cada uno de los canales de color rojo, verde y azul, esto equivale a tres mega bytes de información sin comprimir para una imagen de un millón de píxeles, por lo que para reducir los requisitos de almacenamiento, la mayoría de los formatos de imagen permiten algún tipo de compresión sin perdida o con perdida [19]. Por lo que se ha seleccionado el formato PNG, debido a que es un formato sin pérdida con un gran conjunto de herramientas de administración de código abierto.
La conversión se realizó mediante código en lenguaje python, en donde primero se identifican todos los archivos PDF, ingresándolos al programa, posteriormente se convierte el archivo PDF en una lista de imágenes, una por cada página, las mismas que son almacenadas en una carpeta generada por cada documento PDF leído,los nombres de dichas imágenes se generan basándose en el nombre del documento PDF inicial con un guión bajo seguido del número de página, como se ilustra en la Figura 3.6 (para facilitar el proceso cada periódico se nombró con números del 1 al 50). Cada imagen resultante tiene una resolución de 500 píxeles por pulgada (dpi), debido a que se considera relativamente alta y suele ser suficiente para capturar detalles finos y mantener una buena calidad en la imagen resultante.

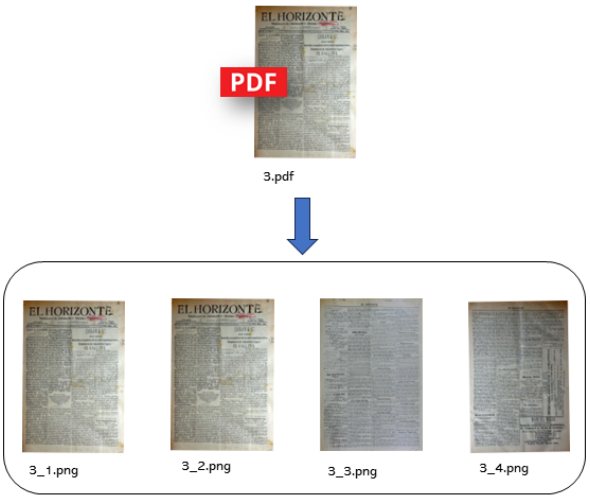

Figura 3.6: Ejemplificación del paso de formato PDF a PNG.

3.4. Procesamiento de imágenes

Luego de haber obtenido las imágenes a partir de los documentos PDF, es necesario evaluar cada una de las técnicas de tratamiento de imágenes, las mismas que han sido divididas en tres grupos: técnicas tradicionales; las de super resolución y segmentación, la división se realizó con el objetivo de poder entender y evaluar de mejor manera las diferentes técnicas, esta división se puede observar en la Figura 3.7, donde además se muestran las diferentes combinaciones que se han ido realizando con el fin de obtener una mejor efectividad en los procesos OCR.

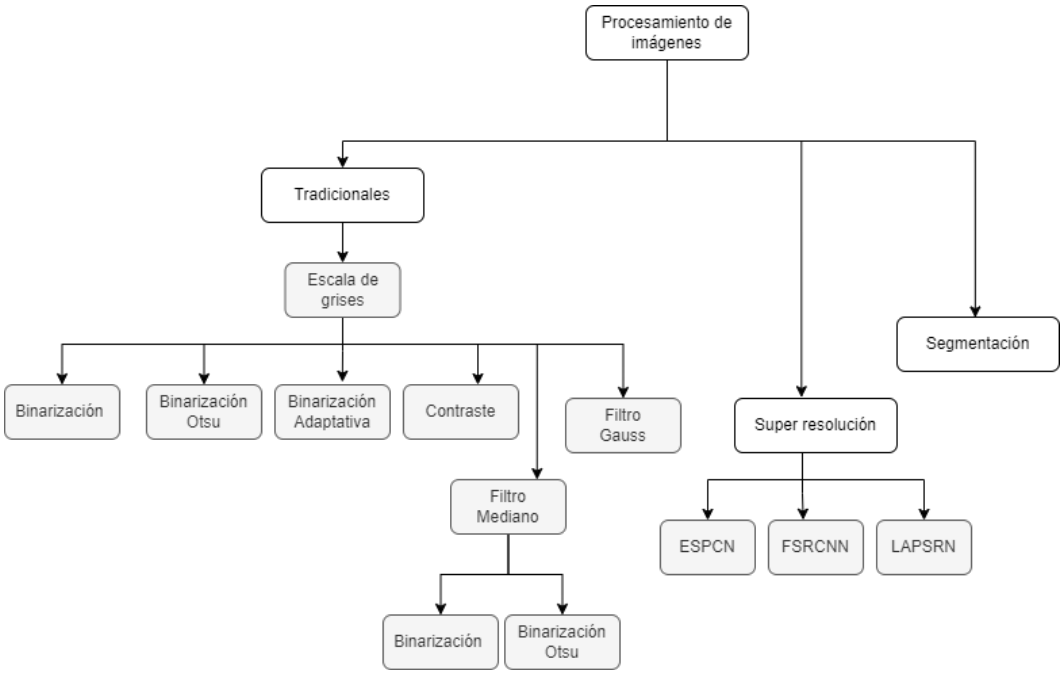

Figura 3.7: Diagrama de las distintas técnicas que se aplicaron a las imágenes

A continuación se explica como se ha aplicado cada una de las técnicas, junto con un ejemplo para ilustrar los cambios que sufre la imagen al someterse a dicho tratamiento, en la Figura 3.8 se muestra la imagen sin tratar que va a ser tomada como ejemplo.

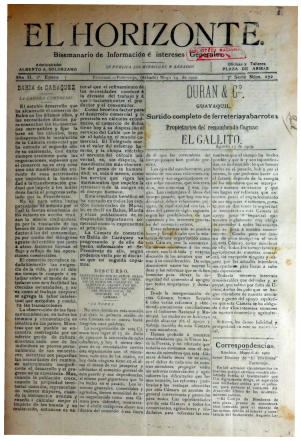

Figura 3.8: Imagen de página de periódico sin tratar a tomar como ejemplo

3.4.1. Técnicas Tradicionales

Para generar los códigos de este trabajo, en general, se ha utilizado la librería OpenCv en Python, ya que es una librería de código abierto, capaz de hacer análisis y tratamientos tanto de imágenes como de vídeos, siendo ideal para visión por computador puesto que permite la detección y seguimiento de objetos, además de permitir reconstrucción de imágenes en 3D entre otras[58], haciéndola útil para aplicar las diferentes técnicas tradicionales de tratamiento de imágenes.

Es importante mencionar que para aplicar cualquiera de las técnicas que se van a presentar hay que convertir la imagen a escala de grises.

a) Escala de Grises

En la Figura 3.9 se puede observar el resultado de la aplicación de la técnica escala de grises.

Figura 3.9: Imagen sin tratar e imagen en escala de grises

Visualmente se logra apreciar el cambio que sufre la imagen, pasando de un tono amarillento a varias tonalidades de gris, siendo más oscuro en zonas donde menos luz contiene la imagen original, ésta técnica es clave para la aplicación de las siguiente técnicas, ya que reciben como entrada un imagen en escala de grises.

b) Binarización Simple

Para aplicar esta técnica previamente se necesita tener la imagen en escala de grises, por ende se usa la técnica anterior, luego se define un valor de umbral medio recomendado, es decir el valor de 127. Todos y cada uno de los píxeles que tengan ese valor o un valor superior serán blancos, y aquellos que sean inferiores se le aplicará el valor 255 que es el correspondiente al color negro, se evalúa cada píxel para realizar una clasificación entre blancos y negros, formando una imagen completamente en blanco y negro, este proceso se puede observar en la figura 3.10, donde se convierte la imagen inicial a escala de grises, y posteriormente se aplica binarización, dando como resultado la tercera imagen.

Figura 3.10: Imagen sin tratar, en escala de grises y aplicada binarización simple

En la Figura 3.10 se puede observar una imagen completamente en blanco y negro, eliminando por completo los tonos de gris. Esto facilita la legibilidad del documento, pero también muestra manchas mucho más pronunciadas en color negro. Como resultado, se pueden perder secciones de texto en áreas donde existen manchas o donde la imagen es mas oscura.

c) Binarización de Otsu

Al ser una técnica donde se obtiene el valor de umbral según el análisis del valor de los píxeles aledaños al píxel que se quiere binarizar, no es necesario entregar un valor de umbral fijo, en la Figura 3.11 se muestra el resultado de aplicar ésta técnica.

Figura 3.11: Imagen sin tratar, en escala de grises e imagen aplicada binarización de Otsu

Tal y como se observa en la Figura 3.11 ,ésta técnica calcula el valor de umbral

en cada parte de la imagen, produciendo manchas bastante pronunciadas, haciendo que se pierda toda la información de la zona más oscura de la imagen.

d) Binarización Adaptativa

En ésta técnica, al igual que en la anterior, no es necesario entregar un valor de umbral, ya que este es calculado según los valores obtenidos mediante el análisis de todos los píxeles de la imagen en conjunto, lo cual entrega un valor umbral fijo para cada una de las imágenes, en la Figura 3.12 se muestra el resultado de aplicar ésta técnica.

Figura 3.12: Imagen sin tratar, escala de grises e imagen aplicada binarización adaptativa

Como se observa en la imagen, con este método se genera mucho ruido fuera de los textos, produciendo los puntos en el fondo, pero también se logran distinguir los textos, aunque se vuelve una imagen con el mismo grado de iluminación y de esa forma elimina las manchas grandes que los otros métodos producían.

e) Filtro Mediano

Con el objetivo de generar una imagen más suave y libre de ruido, se aplica ésta técnica con una apertura o kernel de 5x5 píxeles, es decir que va ha tomar áreas cuadradas alrededor del píxel analizado de 5 píxeles por lado, de donde se obtendrá el promedio y se aplicará al píxel ha ser tratado, dando como resultado la Figura 3.13.

Figura 3.13: Imagen sin tratar, escala de grises e imagen aplicada filtro mediano

En la Figura 3.13 se observa una imagen en escala de grises, con textos bastante desenfocados, esto es debido a que la imagen es suavizada, y al calcular el nuevo valor del píxel se toma en cuenta los que están al alrededor del mismo, por ende se producen tonos más uniformes dando un efecto de desenfoque del texto.

Combinación con filtro mediano

Con el objetivo de hacer que la imagen obtenida con el filtro mediano, se pueda "enfocar", es decir que se pueda distinguir claramente entre el fondo y las letras, se va a aplicar binarización.

i) **Filtro Mediano y Binarización Simple**
A la imagen obtenida aplicando el filtro mediano, se le aplica binarización simple, usando las mismas funciones aplicadas anteriormente.

Figura 3.14: Imagen sin tratar,en escala de grises e imagen aplicada filtro mediano y binarización simple

En la Figura 3.14 se puede ver como la distinción de luz en la imagen original, sumada a que en el filtro mediano se suavizó la imagen produciendo colores muy similares. En toda la imagen se pierde mucha información, tanto en zonas oscuras (mancha negra), como en zonas claras donde se observa que se ha perdido mucho texto, esto se puede deber a que se sigue utilizando el valor umbral de 127.

ii) Filtro Mediano y Binarización de Otsu

También se ha aplicado la binarización de Otsu, la cual calcula el valor de umbral automáticamente según la zona de la imagen.

Figura 3.15: Imagen sin tratar e imagen aplicada filtro mediano y binarización de Otsu

La imagen de la Figura 3.15 muestra una gran cantidad de ruido en la zona oscura (mancha negra), y en el resto del documento se observa texto casi ilegible debido a que existe demasiada iluminación, provocando que se pierda gran cantidad de información, por lo cual se puede llegar a deducir que el OCR no tendrá una buena efectividad aplicando ésta técnica

f) Filtro de Gauss

El filtro de Gauss o filtro Gaussiano al igual que el filtro mediano tienen como objetivo suavizar la imagen, se toma un área de 5x5 alrededor de un píxel, es decir se tomará un cuadrado de 5 píxeles por lado, y se hará el análisis y cálculo, para obtener el nuevo valor por el cual reemplazar el valor del píxel analizado.

Figura 3.16: Imagen sin tratar, en escala de grises e imagen aplicada filtro de Gauss

En la Figura 3.16 se observa una imagen en escala de grises, más suave, es decir sin mucho ruido, además de que el texto es completamente legible, por ende no pierde información de forma visual.

g) Aumento de Contraste

Para este caso en particular, se utilizó la librería “Python Imaging Library” (PIL) que como su nombre lo dice, es una biblioteca de procesamiento de imágenes, capaz de trabajar con diferentes formatos de imagen vectoriales tales como PNG, JPEG,etc.. Después de leer las imágenes es necesario convertirlas a un formato PIL, para luego mediante una función aumentar el contraste de la misma, aunque cabe mencionar que también se mejoran parámetros como el brillo y la nitidez, para aplicar dicha técnica es necesario entregar un factor de mejora, en este caso 2.0, para tener una mejora significativa en el contraste.

Figura 3.17: Imagen sin tratar, en escala de grises e imagen aplicada aumento de contraste

Como se observa en la Figura 3.17, se obtiene una imagen con un alto contraste, identificando claramente los textos de la imagen, en las áreas oscuras se puede apreciar que el texto es mucho más legible que en las zonas claras, debido al contraste de luz como tal.

3.4.2. Super resolución

Para llevar a cabo el proceso de super resolución, se optó por utilizar OpenCV debido a su amplia gama de algoritmos de aprendizaje profundo diseñados para escalar imágenes. En este estudio, en particular, se seleccionó una escala de 4, porque según la investigación de [48], ofrece resultados superiores para la aplicación de OCR, aunque con una diferencia mínima en comparación con la escala 2, lo que llevó a la elección de esta escala. Es necesario indicar la ubicación del archivo que del modelo pre-entrenado para la super resolución, en este caso, se utilizaron los modelos entrenados en TensorFlow[3] con una escala de 4x, como ESPCN[4], FSRCNN[5] y LAPSRN[6]. Dado el tamaño considerable de las imágenes empleadas en este estudio, se utilizó la biblioteca "patchify" [59], esta biblioteca permite dividir las imágenes en fragmentos más pequeños, conocidos como parches, facilitando el procesamiento de cada parche aplicando la super resolución correspondiente. Finalmente, los parches procesados se fusionaron para crear una imagen de salida completa y se guardó en un archivo. Dado que los recursos computacionales eran limitados, este método permitió el procesamiento en bloques más pequeños en lugar de procesar la imagen en su totalidad. Los resultados se pueden apreciar en la Figura 3.18.

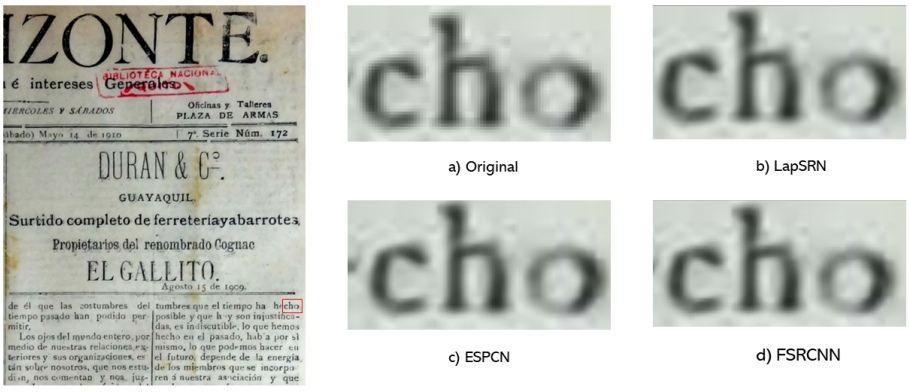

Figura 3.18: Comparación visual de las técnicas de super resolución

A partir de los resultados obtenidos, se puede observar en la Figura 3.18, la aplicación de técnicas de super resolución presentan mejores resultados en comparación con la imagen original. Específicamente, se destacan las técnicas FSRCNN Y LAPSRN en términos de mejora en la calidad de la imagen, logrando una resolución superior en comparación con el enfoque original.

[3]https://www.tensorflow.org/?hl=es-419
[4]https://github.com/fannymonori/TF-ESPCN
[5]https://github.com/Saafke/FSRCNN_Tensorflow
[6]https://github.com/fannymonori/TF-LapSRN

3.4.3. Segmentación

Para este paso fue necesario instalar “layout Parser torchvision”[7]. A partir de esto, se determinará la configuración del modelo que se utilizará. En este caso, se empleó la configuración del modelo “NewspaperNavigator” [4] que se encuentra en la plataforma en Model Zoo[8], ya que proporcionan modelos pre-entrenados en diferentes conjuntos de datos. Al analizar con diferentes umbrales de confianza, se evidenció que al aplicar un valor de 0,3 dio mejores resultados para reconocer bloques de texto, al igual que al aplicar escala de grises y filtro mediano.
El paso siguiente implica el reconocimiento de los bloques de texto. En base a cada bloque identificado se genera una nueva imagen, las dimensiones de cada bloque se registran en un arreglo, con el fin de generar una máscara para cada bloque segmentado sobre la imagen original permitiendo obtener la parte no segmentada como una nueva imagen, esto con el objetivo de preservar toda la información. Este proceso se puede observar con más detalle en la Figura 3.19.

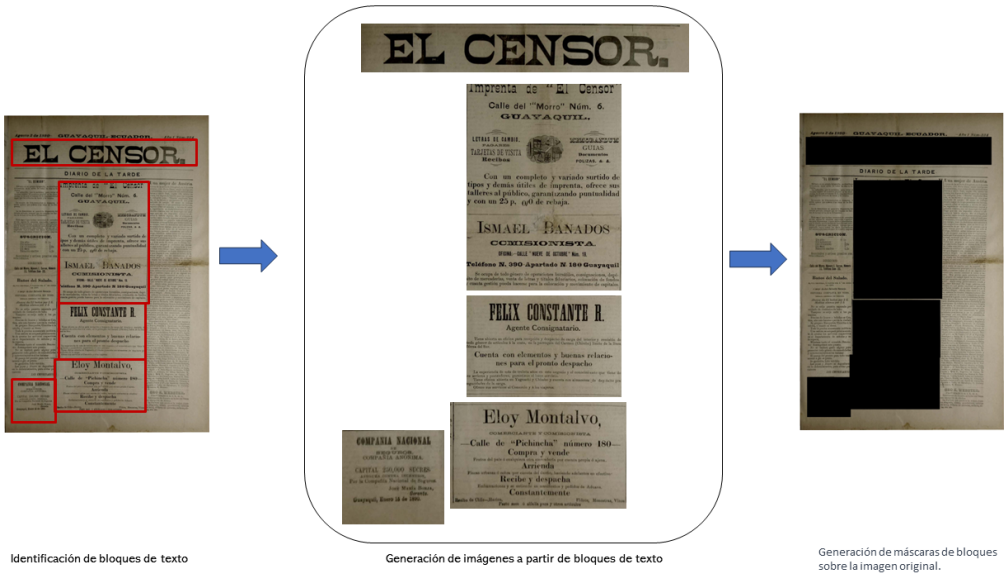

Figura 3.19: Visualización del proceso de segmentación en el periódico de la editorial “EL CENSOR”.

Es relevante destacar el resultado obtenido a partir del análisis del periódico base 3.8, donde se identifica una gran parte de la imagen del periódico como un bloque de texto, lo que resulta en una máscara que cubre casi la totalidad del periódico. Además, se identifica una cuadro de texto que genera una repetición de información en la misma área, como se muestra en la Figura 3.20. Estos resultados pueden ser atribuidos a la diversidad de formatos presentes en los periódicos analizados.

[7]https://layout-parser.readthedocs.io/en/stable/notes/installation.html
[8]https://layout-parser.readthedocs.io/en/latest/notes/modelzoo.html

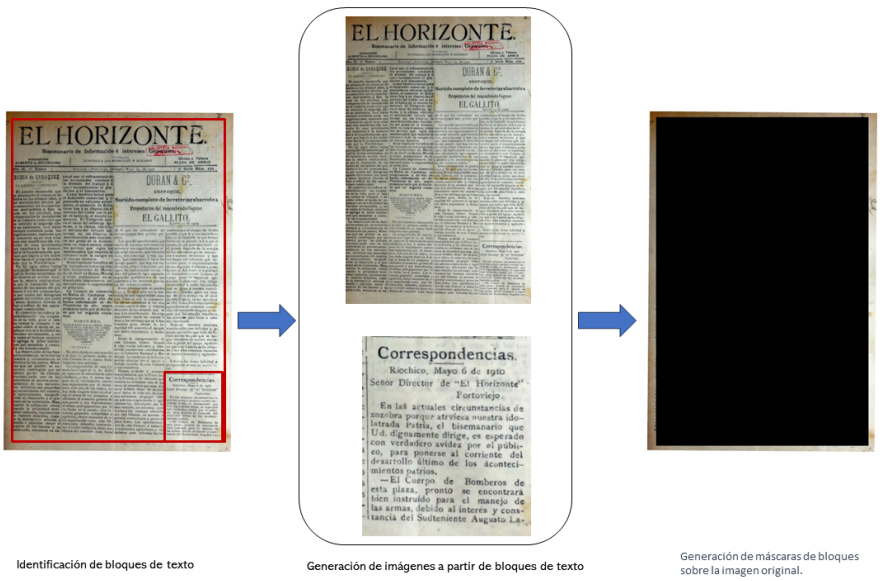

Figura 3.20: Visualización del proceso de segmentación, cuando reconoce todo como un bloque

3.5. Proceso OCR

Se utilizó Tesseract como motor OCR, el cual puede ser utilizado tanto en su modo OCR estándar como en su modo de Redes Neuronales, en este trabajo se realizaron pruebas con ambas modalidades con el objetivo de comparar los resultados obtenidos por cada uno y finalmente destacar las fortalezas y debilidades.
Para sacar el máximo provecho a dicho motor, hay que entender y configurar los diferentes parámetros que Tesseract recibe, entre los principales que menciona la página oficial. [60]

- Cadena “Config” : permite realizar la configuración del OCR en una sola línea, se usa para definir las rutas de los resultados del modelo pre entrenado, la forma de segmentar y el separador de página.
- Motor OCR (OEM): permite escoger entre 0 para Legacy u OCR Tradicional y 1 para usar Redes Neuronales, si se deja sin parámetro por lo general se va a fijar en el “config”, y en caso de tener un enlace a una red utilizará redes, caso contrario utilizará el motor Legacy.
- Idioma: Teseract por defecto utiliza el inglés como idioma, pero entrega la posibilidad de instalar o descargar los diferentes idiomas, para este proyecto se instaló el lenguaje español, cabe mencionar que permite usar varios idiomas a la vez.
- Modo de Segmentación de Página (PSM): es uno de los parámetros más importantes a configurar, debido a que depende de esto como va ha leer el OCR el documento, en total existen 13 formas de segmentar una página, pero la recomendada para imágenes con varias columnas, es el PSM 3.

Al tener los documentos transcritos en documentos TXT, es necesario que el resultado obtenido mediante el OCR también se encuentre en el mismo formato, por

lo cual, el proceso OCR deberá entregar como resultado una cadena de texto, el mismo que será almacenado en un archivo en formato TXT.

3.5.1. Tesseract Estándar (OCR1)

En este modo, Tesseract utiliza algoritmos tradicionales y técnicas de reconocimiento de patrones para llevar a cabo el proceso OCR. La configuración utilizada para este caso fue la siguiente:

- Modo de segmentación de página (psm) : 3 (imágenes con columnas)
- Separador de página: “” (ninguno)
- Idioma: español (spa)
- Tiempo de respuesta: 0 (ilimitado)
- Modo de salida: texto

No se especifica el OEM, debido a que por defecto utiliza OCR de modo predeterminado, cabe recalcar que se hicieron varias pruebas en la configuración, pero la descrita es la que lee mejor el documento en columnas, además de que no da errores en ejecución(ej. exceder el tiempo de respuesta).

3.5.2. Tesseract con Redes Neuronales (OCR2)

Para utilizar Tesseract con redes neuronales, es necesario descargar un archivo “traineddata”, este archivo es el resultado de un entrenamiento previo de Tesseract y es fundamental para lograr un reconocimiento preciso de palabras y caracteres. En este caso, se descargó el archivo “traineddata” de la versión “best”, la cual, de acuerdo a la documentación oficial [61] ofrece un mayor número de palabras reconocidas y menos presencia de ruido en los resultados. Sin embargo, es importante tener en cuenta que esta versión puede requerir más tiempo de ejecución, además de la versión “best”, existe también la opción de utilizar la versión “fast”, que se caracteriza por tener una ejecución más rápida, pero puede presentar más ruido en los textos reconocidos y es posible que no logre reconocer ciertos textos. Es fundamental tener en cuenta que la configuración de Tesseract debe especificar correctamente la ubicación del archivo “traineddata” descargado, para que pueda utilizar los datos de entrenamiento durante el proceso de reconocimiento óptico de caracteres (OCR), en este caso la configuración es la siguiente:

- Modo de segmentación de página (psm) : 3 (imágenes con columnas)
- Separador de página: “” (ninguno)
- Dirección del Tessdata: ubicación del archivo “best tessdata spa” (especificado en el “config”)
- Idioma: español (spa)
- Tiempo de respuesta: 0 (ilimitado)
- Modo de salida: texto

3.6. Eliminación de caracteres

Teniendo en cuenta que las imágenes estaban en mal estado y no tenían una buena calidad, se decidió realizar una depuración de los caracteres identificados debido a la presencia de ruido y diversas irregularidades que surgieron después de utilizar el proceso de Reconocimiento Óptico de Caracteres (OCR). Por lo que, se priorizó mantener los caracteres más frecuentes en el texto de los periódicos analizados en lugar de aquellos que presentaban agrupaciones o dispersión inapropiada. En consecuencia, se decidió conservar únicamente los caracteres de punto, coma y guión medio, ya que son muy comunes y se encuentran presentes en el contenido textual de varios periódicos.

3.7. Comparación de palabras

Debido a que se tiene los documentos transcritos de forma manual en formato TXT, lo cual ha sido tomado como un “archivo ideal al que se quiere llegar” (se asume que no existen errores en la transcripción), además se tiene todos y cada uno de los TXT obtenidos mediante los OCR, es necesario saber que tan efectivo es el OCR cuando se aplica una determinada técnica de tratamiento a una imagen. Para llevar a cabo esta evaluación, es necesario obtener una métrica de efectividad, para lo cual se debe realizar una comparativa de textos y obtener el número de palabras en común entre el texto “ideal” y los resultados obtenidos. Esto se ha realizado de dos maneras, aunque la primera es solo un primer acercamiento, y por ende finalmente no fue tomado como parte de la evaluación.

3.7.1. Uso de Difflib

Para llevar a cabo este proceso, se empleó la biblioteca difflib [62], la cual se utiliza para la comparación de secuencias. El procedimiento consistió en los siguientes pasos: en primer lugar, se procedió a la lectura del contenido de archivo de transcripción original y el archivo generado mediante unos de los procesos previamente mencionados, esto se realizó con todas las técnicas, incluyendo el resultado del OCR sin aplicar ninguna técnica. A continuación, se usó la función re.findall() junto con expresiones regulares para extraer las palabras presentes en cada uno de los archivos, almacenando los resultados en listas separadas. Seguidamente, se determinaron las palabras comunes entre ambas listas haciendo uso del operador de intersección “&”, y se almacenaron en un conjunto. Por último, se contó la cantidad de palabras usuales obtenidas. Este tipo de comparación de palabras es útil en tareas de análisis de texto, especialmente cuando se compara un texto original y su versión generada mediante OCR. Cabe mencionar que después de hacer pruebas, se descartó esta opción, ya que reconoce patrones de texto, mas no compara las palabras como tal, es decir si se busca por ejemplo la palabra “sol” y en el texto aparece la palabra “aerosol”, será tomado como acierto, lo cual no es correcto, por ende fue descartado, y se buscó el siguiente método que se describe a continuación.

3.7.2. Comparativa directa

En este proceso, se tienen dos documentos TXT: uno es el texto transcrito manualmente y el otro es el texto obtenido mediante OCR. El objetivo es determinar cuántas palabras son comunes entre ambos textos para evaluar la efectividad del OCR.

Primero, se lee cada documento y se utiliza la función split de Python para dividir el texto en palabras. Cada palabra se almacena en una lista correspondiente al documento al que pertenece, a continuación, se realiza un recorrido por la lista de palabras del texto obtenido mediante OCR. Para cada palabra, se verifica si está presente en la lista de palabras del texto transcrito. Si la palabra está en la lista transcrita, se agrega a una nueva lista que contiene las palabras comunes entre ambos textos, además, una vez que una palabra de la lista del OCR se ha encontrado en la lista transcrita, se elimina de esa lista. Esto se hace porque una palabra en el texto OCR idealmente solo debería aparecer la misma cantidad de veces que en la lista transcrita, ya que se considera como el texto de referencia sin errores.
Una vez que se ha recorrido toda la lista de palabras obtenida mediante OCR y se han verificado si están presentes en la lista transcrita, se obtiene la longitud de la lista de palabras comunes. Esto da el número de palabras que son iguales en ambos textos lo cual es de vital importancia para evaluar la efectividad del OCR con cada una de las técnicas.
En resumen, se divide cada texto en palabras, se compara si las palabras del texto OCR están presentes en el texto transcrito, se crea la lista de palabras comunes y se calcula la cantidad total de palabras comunes.

3.8. Evaluación de técnicas

Como ya se ha mencionado anteriormente, para facilitar el estudio y análisis, se ha dividido las técnicas en tres grupos: Tradicionales, o técnicas que son usadas comúnmente para OCR; las de Segmentación, es decir que de una imagen se obtienen varias imágenes más pequeñas; y finalmente las de super resolución, en las que se aumenta la resolución de la imagen, para que de esa forma puedan textos u objetos ser más reconocibles.

Se mostrará una tabla para cada grupo, que contendrá las diferentes técnicas en la primera fila, en la segunda fila se presentarán los promedios de efectividad del OCR al aplicar cada una de las técnicas en los 50 periódicos, dicho valor se calculará utilizando la fórmula presentada en la ecuación1.1, lo cual dará un porcentaje de efectividad para cada periódico después de aplicar alguna de las técnicas de tratamiento. Cabe mencionar que las palabras totales son el número de palabras existentes en el documento transcrito, mientras que las palabras incorrectas se calculan restando del total de palabras, las palabras en común obtenidas previamente. En la tercera fila se mostrará un porcentaje que indica si hubo una mejora, representado por un valor positivo, o un empeoramiento, representado por un valor negativo, en comparación con los resultados promedio obtenidos al utilizar el OCR con las imágenes sin ningún tratamiento.

En resumen, dicha tabla contendrá en la primera fila las técnicas probadas, en la segunda fila presentará los promedios de efectividad del OCR al aplicar las técnicas, y en la tercera fila mostrará el porcentaje de mejora o empeoramiento con respecto a un periódico sin tratamiento.

3.8.1. OCR Estándar (OCR1)

Los resultados obtenidos por el OCR 1 u OCR estándar, con las distintas técnicas de procesamiento de imágenes sin redes neuronales, y dando como entrada un

periódico completo sin tratamiento alguno, de entre los 50 periódicos procesados sin ningún tratamiento, el OCR 1 tiene una efectividad promedio de 50.95 %, recuperando así más de la mitad del texto presente en los periódicos, cabe aclarar que hay periódicos que entregan una efectividad sin tratamiento de hasta el 88.25 %, sin embargo también hay periódicos que tienen efectividad de hasta 17.33 %, entonces los resultados en gran medida dependen del estado de los periódicos, además de como tienen distribuidos los textos, en este caso el periódico que menos efectividad entregó es un periódico de cuatro páginas, con letras pequeñas y borrosas, además de que contienen tablas, gráficos e inclusive fragmentos de texto en distintas orientaciones, mientras que el que entrega 88.25 % es un periódico que únicamente tiene 2 páginas, cada una con 3 columnas, con textos grandes y legibles en su totalidad, la tabla completa de resultados del OCR 1 se puede observar en el Anexo B.

3.8.1.1. Técnicas Tradicionales

Evaluando la efectividad del OCR 1 con algunas de las técnicas tradicionales, se tienen los resultados presentes en la Tabla 3.1, en donde se destaca que escala de grises mejora hasta en un 1.88 %, con una efectividad promedio de 52.83 % y que la técnica de contraste tiene la misma efectividad promedio que cuando se envía la imagen sin tratamiento previo, además se puede observar que cuando se mezclan filtro mediano con binarización simple baja la efectividad promedio hasta en 43.35 %, teniendo una efectividad solo del 7.59 %, lo cual indica que la combinación de éstas técnicas no ayudan al tratamiento de imágenes en procesos OCR.

Técnica	Sin Tratamiento	Escala de grises	Contraste	Binarización	Otsu	Adaptativa	Gauss	Mediano	Mediana y Otsu	Mediana y binarización
Resultado	50.95 %	52.83 %	50.95 %	39.31 %	43.50 %	8.035 %	14.75 %	14.75 %	10.82 %	7.59 %
Diferencia con ”sin tratamiento”	-	+1.88 %	0 %	-11.64 %	-7.44 %	-42.91 %	-36.20 %	-36.20 %	-40.12 %	-43.35 %

Tabla 3.1: Comparación entre el resultado obtenido por OCR1 utilizando técnicas tradicionales y el resultado sin procesar la imagen.

3.8.1.2. Super resolución

En la Tabla 3.2, se observan los resultados obtenidos al aplicar las diferentes técnicas de super resolución, las mismas que mejoran la efectividad promedio del OCR 1 con respecto a un periódico sin tratar hasta en un 17.39 %, cabe mencionar que todas las técnicas provocan una mejora bastante significativa en porcentaje de resultados, la mejor es la técnica LAPSRN, con una efectividad promedio de entre los 50 periódicos tomados como prueba de 68.34 %, seguida de la técnica ESPCN con una efectividad promedio de 67.55 %, y finalmente de FSRCNN que entrega una efectividad promedio de 64.93 %, en conclusión se puede decir que todas las técnicas de super resolución aplicadas en este trabajo de titulación ayudan a mejorar los resultados en procesos OCR, con un OCR sin redes neuronales.

Técnica	Sin Tratamiento	LAPSRN	ESPCN	FSRCNN
Resultado	50.95 %	68.34 %	67.55 %	64.93 %
Diferencia con ”sin tratamiento”	-	+17.39 %	+16.59 %	+13.98 %

Tabla 3.2: Comparación entre el resultado obtenido por OCR1 utilizando técnicas de super resolución y el resultado sin procesar la imagen.

3.8.1.3. Segmentación

El objetivo de ésta técnica es dividir una página del periódico en imágenes más pequeñas, lo cual facilitará el reconocimiento de texto, pero al ser un primer acercamiento, el algoritmo probado no segmenta toda la imagen, sino solo obtiene algunos fragmentos, aunque ayuda al proceso OCR a tener una mejor efectividad en el reconocimiento de textos, ésta técnica mejora hasta en un 13.84 % a la efectividad que tiene el OCR 1 con un periódico sin tratar, en la Tabla 3.3, se puede observar que ésta técnica da un promedio de efectividad de 64.79 %, aunque como se mencionó anteriormente no se puede tomar como resultado definitivo ya que no segmenta correctamente toda la imagen.

Técnica	Sin Tratamiento	Segmentación
Resultado	50.95 %	64.79 %
Diferencia con ”sin tratamiento”	-	13.84 %

Tabla 3.3: Comparación entre el resultado obtenido por OCR1 utilizando la técnica de segmentación y el resultado sin procesar la imagen.

3.8.2. OCR con redes Neuronales y Tessdata Best (OCR2)

Se muestran los resultados obtenidos con el OCR 2, aplicando redes neuronales junto con el Tessdata Best.
Este tipo de OCR con redes neuronales, tiene una efectividad promedio de entre los 50 periódicos escogidos sin aplicar ningún tratamiento de 51.06 %, teniendo una efectividad mínima de hasta 15.88 % en un periódico con 4 páginas llenas de texto, imágenes, además de texto borroso, y con orientaciones variadas, y una efectividad máxima de 87.19 % con un periódico de 2 páginas que contienen únicamente texto en 3 columnas, además de que es completamente legible.

3.8.2.1. Técnicas Tradicionales

Tomando el OCR 2, junto con algunas técnicas tradicionales, o técnicas usadas para procesos OCR, se obtienen los resultados observados en la Tabla 3.4, De los que la única técnica tradicional que mejora la efectividad del OCR 2 con respecto a un periódico sin tratamiento, es la técnica de escala de grises con una efectividad

promedio de 53.12 %, mejorando así, hasta un 2.17 % a un periódico que no tenga tratamiento alguno, para este OCR, la peor técnica es la de Binarización Adaptativa, con una efectividad promedio de 8.72 %, empeorando los resultados hasta en un 42.23 %.

Técnica	Sin Tratamiento	Escala de grises	Contraste	Binarización	Otsu	Adaptativa	Gauss	Mediano	Mediana y Otsu	Mediana y binarización
Resultado	51.06 %	53.12 %	50.48 %	40.35 %	44.01 %	8.72 %	25.71 %	15.03 %	11.21 %	9.07 %
Diferencia con ”sin tratamiento”	-	+2.17 %	-0.47 %	-10.60 %	-6.93 %	-42.23 %	-25.24 %	-35.92 %	-39.74 %	-41.88 %

Tabla 3.4: Comparación entre el resultado obtenido por OCR 2 utilizando técnicas tradicionales y el resultado sin procesar la imagen.

3.8.2.2. Super resolución

En la tabla 3.5, se puede observar los resultados promedio obtenidos con cada una de las técnicas junto con el OCR 2, en el cual se pueden observar los siguientes resultados: la mejor técnica es LAPSRN con 68.34 % de efectividad promedio, mejorando hasta en un 17.18 % al promedio de los periódicos sin tratar, seguido de ESPCN con un 66.72 % de efectividad y por último la técnica FSRCNN con un 65.01 % de efectividad promedio, mejorando hasta en un 13.94 % la efectividad promedio de un periódico sin tratamiento. En conclusión, al igual que en el OCR 1, se puede decir que todas las técnicas de super resolución ayudan a mejorar la efectividad de los procesos OCR con redes neuronales.

Técnica	Sin Tratamiento	LAPSRN	ESPCN	FSRCNN
Resultado	51.06 %	68.34 %	66.72 %	65.01 %
Diferencia con ”sin tratamiento”	-	+17.18 %	+15.66 %	+13.94 %

Tabla 3.5: Comparación entre el resultado obtenido por OCR 2 utilizando técnicas de super resolución y el resultado sin procesar la imagen.

3.8.2.3. Segmentación

En la Tabla 3.6, se muestra el resultado obtenido aplicando la segmentación junto al OCR 2, como se mencionó anteriormente, dicha técnica segmenta la imagen ingresada, pero muchas partes no logra identificar los bloques de texto, por ende dicha técnica está sujeta a mejoras en un futuro, y los resultados no son del todo concluyentes, aunque se puede observar que si presenta una mejora considerable de hasta un 13.56 %, teniendo una efectividad promedio de 64.63 %.

Técnica	Sin Tratamiento	Segmentación
Resultado	51.06 %	64.63 %
Diferencia con ”sin tratamiento”	-	+13.56 %

Tabla 3.6: Comparación entre el resultado obtenido por OCR2 utilizando la técnica de segmentación y el resultado sin procesar la imagen.

3.8.3. Análisis de resultados

En esta sección se condensa un análisis de los resultados obtenidos en las secciones anteriores, en la columna “Mejora” de la Tabla 3.7 y la Tabla 3.8 se encuentran los porcentajes de mejoras con respecto a realizar el OCR sobre una imagen sin ningún tratamiento, el mismo que se obtiene aplicando una diferencia entre los porcentajes de los distintos procesos y el porcentaje de no haber realizado ningún proceso. Se ha destacado con color rojo el porcentaje más alto de mejora, y con color azul el segundo porcentaje más alto.
Como se puede apreciar en la Tabla 3.7, el proceso con el mejor resultado de mejora al realizar con el OCR 1 es el de super resolución LAPSRN con un 17,39 % de mejora y el segundo es el de ESPCN con un porcentaje del 16,59 %.
Tras examinar los resultados y analizar los documentos que presentaron porcentajes de similitud más bajos con el texto transcrito (consultar Anexo C), se identificaron ciertos aspectos comunes que influyeron en estos resultados. Dichos documentos exhiben características como texto borroso o con una escasa cantidad de tinta, así como una notable presencia de manchas, incluyendo manchas de gran tamaño, y secciones en las que la tinta se traspasaba de una página a otra.

Técnicas usadas		Proceso OCR1	Mejora
Sin Tratamiento		50,9529841	
TRADICIONAL	**Escala de Grises**	52,83648387	1,883499773
	Binarizacion	39,31147486	-11,64150924
	Otsu	43,50487797	-7,448106129
	Adaptativa	8,03540881	-42,91757529
	Mediano	14,75244331	-36,20054079
	Gauss	14,75244331	-36,20054079
	Contraste	50,9529841	0
	Mediano Binarizacion	7,597006598	-43,3559775
	Mediano Otsu	10,82596952	-40,12701458
SEGMENTACIÓN		64,79478635	13,84180225
SUPER RESOLUCIÓN	**FSRCNN**	64,93320556	13,98022146
	ESPCN	67,55190623	16,59892213
	LAPSRN	68,34826291	17,39527881

Tabla 3.7: Comparación entre todos los resultados obtenidos con OCR1

Se evidencia en la Tabla 3.8 que con el OCR2, el proceso con el mejor porcentaje de mejora es el de super resolución LAPSRN con un 17,29 % de mejora con respecto a un periódico sin tratamiento, seguida por la técnica ESPCN con un porcentaje de mejora del 15,779 %.

Técnicas usadas		Proceso OCR1	Mejora
Sin Tratamiento		51,067684	
TRADICIONAL	Escala de Grises	53,12519739	2,172213284
	Binarizacion	40,3516351	-10,601349
	Otsu	44,01457943	-6,938404676
	Adaptativa	8,721739294	-42,23124481
	Mediano	15,03304092	-35,91994318
	Gauss	25,71353125	-25,23945285
	Contraste	50,48022732	-0,4727567802
	Mediano Binarizacion	9,072546556	-41,88043755
	Mediano Otsu	11,210541	-39,7424431
SEGMENTACIÓN		64,63676245	13,68377835
SUPER RESOLUCIÓN	FSRCNN	65,00863404	14,05564994
	ESPCN	66,72588985	15,77290575
	LAPSRN	68,25017959	17,29719549

Tabla 3.8: Comparación entre todos los resultados obtenidos con OCR2

3.8.4. Mejores técnicas y OCR

En base a lo analizado en las secciones anteriores se puede decir que las mejores técnicas de tratamiento de imágenes para procesos OCR son las de super resolución, en específico la técnica LAPSRN junto con el OCR 1, sin redes neuronales, obteniendo hasta un 68.34 % de efectividad promedio de entre los 50 periódicos seleccionados.
En cuanto al OCR, se puede observar en las Tablas 3.7 y 3.8, que el OCR 2, con imágenes sin tratamiento da un mejor resultado que en el OCR 1, pero al tratar la imagen, hay algunas técnicas que funcionan mejor en el OCR 1, mientras que otras mejoran en el OCR 2, pero los resultados más altos se han obtenido en el OCR 1 junto con la técnica de Super resolución LAPSRN, viéndose así el OCR 2 con la misma técnica superado con un 0.09 %, lo cual no es muy significativo, por ende con ésta muestra y este estudio no se puede decir claramente cuál de los dos OCRs funciona de mejor manera, ya que presentan resultados muy similares en ambos casos.

3.8.5. Amenazas a la validez de resultados

Los resultados pueden verse afectados directamente por la muestra seleccionada, al tener la población tan grande y haber seleccionado una muestra tan pequeña, debido al tiempo y recursos que se tuvieron al momento de realizar la investigación, puede que no se hayan tomado todos los casos presentes. También cabe mencionar que las técnicas aplicadas, por lo general, se desarrollaron con la librería OpenCV, pero también las mismas técnicas se pueden aplicar con otras librerías y hasta con otros lenguajes, trabajando de una manera diferente a la imagen ingresada, lo cual podría dar una imagen tratada de forma diferente y entregar mejores o peores resultados a los obtenidos. Además, una de las principales amenazas que podría modificar los resultados de gran manera, sería la eliminación de caracteres especiales, los mismos que representaban ruido en el texto obtenido mediante OCR, en caso de no eliminar dichos caracteres, el texto sería diferente, por ende se podría

tener otros resultados.
En este trabajo se realizó una prueba con el fin de obtener otras métricas, en donde se tomaba el número de palabras del texto en crudo, es decir, sin eliminar ningún tipo de caracteres, y ver cuánto porcentaje representaba con respecto al total de palabras del texto transcrito, en muchos casos se daba un porcentaje mayor al 100 %, por ende ésta no podía ser tomada como una métrica válida, esto debido a que por el ruido en las imágenes muchas veces el OCR divide a una palabra, o contiene muchos caracteres especiales, dando un número de palabras mayor que el transcrito. Finalmente, los transcritos, podrían contener errores de codificación, así como falta de letras, palabras mal escritas, o quizá párrafos repetidos o faltantes, lo cual afectaría directamente a los resultados. Esto no se puede validar al 100 %, debido a que hay 50 archivos TXT, con un promedio de 8284 palabras, representando mucho tiempo de revisión y corrección. Aunque se han dado casos de documentos TXT donde no tienen la estructura correcta, o textos duplicados y faltantes que si se han corregido, esto se pudo detectar mediante una inspección rápida, además de que número de palabras reportadas variaba mucho con respecto al número promedio de palabras obtenidas mediante los procesos OCR.

4. Conclusiones y Trabajos Futuros

En este capítulo se presentan las conclusiones obtenidas a partir de los resultados, además verificar el cumplimiento de los objetivos del proyecto presentados al inicio, así también se muestran los trabajos futuros que podrían ayudar a mejorar los resultados en los procesos OCR.

4.1. Conclusiones

En este trabajo de titulación se ha logrado revisar, identificar y evaluar las diferentes técnicas utilizadas en el tratamiento de imágenes con procesos OCR, las mismas que son: escala de grises; binarización simple; binarización de Otsu; Binarización Adaptativa; Filtro de Gauss y Filtro Mediano, además de las técnicas de super resolución y la de segmentación, dando como resultados que las mejores técnicas de tratamiento de imágenes para procesos OCR son las de super resolución, en específico la técnica LAPSRN, además de entre las técnicas tradicionales, la mejor fue la técnica de escala de grises, cumpliendo así con los objetivos planteados, además se llega a la conclusión que aplicar las técnicas de segmentación y de super resolución, previo a un proceso OCR ayuda a obtener mejores resultados.

4.2. Trabajos Futuros

Con el objetivo de mejorar los resultados presentados en este proyecto de titulación, se propone los siguiente trabajos futuros:

- Mejorar el algoritmo de segmentación, de modo que se pueda detectar cada bloque de texto de forma correcta y tome como una nueva imagen. Se recomienda la creación de modelos específicamente para la segmentación de periódicos ecuatorianos, ya que sus diseños difieren de los diseños de periódicos antiguos internacionales.
- Aplicar las técnicas tradicionales en las imágenes obtenidas con las técnicas de segmentación, ya que al tener imágenes más pequeñas permite detectar de mejor manera los textos, además dichas técnicas se pueden aplicar a las imágenes de super resolución.
- Probar combinaciones de técnicas de tal manera que eliminen el ruido, aumenten la resolución y permitan reconocer bloques de textos para realizar la posterior segmentación.
- Es importante también, aplicar algoritmos de similaridad semántica y sintáctica, para de esa forma poder evaluar si los textos tienen sentido Y si realmente coinciden con el periódico original.
- También sería importante realizar un proceso de corrección de los textos obtenidos mediante OCR, lo cual ayudaría a facilitar la lectura y análisis de los mismos.
- Se sugiere llevar a cabo pruebas con diferentes escalas de modelos preentrenados de super resolución disponibles. Debido a las limitaciones en los

recursos disponibles, esta exploración no pudo ser realizada en el presente estudio.

- Finalmente se recomienda entrenar Tesseract con los periódicos antiguos ecuatorianos, con el objetivo de generar un tessdata propio y obtener mejores resultados.

Referencias

[1] C. Dong, C. C. Loy, y X. Tang, “Accelerating the super-resolution convolutional neural network,” 2016.

[2] W. Shi, J. Caballero, F. Huszár, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, y Z. Wang, “Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network,” 2016.

[3] W.-S. Lai, J.-B. Huang, N. Ahuja, y M.-H. Yang, “Fast and accurate image super-resolution with deep laplacian pyramid networks,” 2018.

[4] B. C. G. Lee, J. Mears, E. Jakeway, M. Ferriter, C. Adams, N. Yarasavage, D. Thomas, K. Zwaard, y D. S. Weld. (2020) The newspaper navigator data-set: Extracting and analyzing visual content from 16 million historic newspaper pages in chronicling america.

[5] H. Wijfjes, “Digital humanities and media history: A challenge for historical news-paper research,” *Tijdschrift voor Mediageschiedenis*, vol. 20, num. 1, p. 4, 2017.

[6] C. Neudecker y A. Antonacopoulos, “Making europe’s historical newspapers searchable,” in *2016 12th IAPR Workshop on Document Analysis Systems (DAS)*, pp. 405–410.

[7] M. Florensa Flix y A. Rossell Badia, “Digitalizar para no restaurar o restaurar para poder digitalizar,” accepted: 2019-07-08T11:26:57Z. [En línea]. Disponible: https://diposit.ub.edu/dspace/handle/2445/136663

[8] L. M. Vilches-Blázquez, D. Comesaña, y L. d. J. Arrieta Moreno, “Construcción de una red de ontologías sobre eventos meteorológicos a partir de periódicos históricos,” *Transinformação*, vol. 32, p. e180077, 2020. [En línea]. Disponible: https://www.scielo.br/j/tinf/a/NFQC4DMHGP5BDxK7dcqZRnk/?lang=es

[9] S. S. Ballesteros Estrada, G. Morales Romero, y P. A. Cedillo Pérez. (2012) Los problemas de identificación de caracteres OCR para la recuperación de texto en el libro antiguo: un análisis de caso en el fondo antiguo de la biblioteca central, UNAM | biblioteca universitaria. [En línea]. Disponible: https://bibliotecauniversitaria.dgb.unam.mx/rbu/article/view/39

[10] D. V. Gómez Trejos y A. Guerrero Guzmán, “Estudio y Análisis de Técnicas para Procesamiento Digital de Imágenes,” 2016.

[11] D. Comesaña y L. M. Vilches Blazquez, “Un estudio de la prensa latinoamericana entre los siglos XIX y XX con un enfoque en eventos meteorológicos,” num. 156, pp. 29–59, 2019, publisher: Instituto Panamericano de Geografía e Historia Section: Revista de Historia de América. [En línea]. Disponible: https://dialnet.unirioja.es/servlet/articulo?codigo=7985716

[12] C. Wohlin, P. Runeson, M. Höst, M. C. Ohlsson, B. Regnell, y A. Wesslén, *Experimentation in Software Engineering*. Springer, 2012. [En línea]. Disponible: http://link.springer.com/10.1007/978-3-642-29044-2

[13] L. Olson y V. Berry, “The code4lib journal – digitization decisions: Comparing OCR software for librarian and archivist use,” 2021-09-22. [En línea]. Disponible: https://journal.code4lib.org/articles/16132

[14] J. C. Ponce Gallegos, A. Torres Soto, F. S. Quezada Aguilera, A. Silva Sprock, E. U. Martínez Flor, A. Casali, E. Scheihing, Y. J. Túpac Valdivia, M. D. Torres Soto, F. J. Ornelas Zapata, J. A. Hernández, C. Zavala, N. Vakhnia, y O. Pedreño, *Inteligencia Artificial*. Iniciativa Latinoamericana de Libros de Texto Abiertos (LATIn), 2014, accepted: 2020-03-01T18:09:56Z. [En línea]. Disponible: http://rephip.unr.edu.ar/xmlui/handle/2133/17686

[15] J. D. Kelleher, B. M. Namee, y A. D’Arcy, *Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies*, 2015-07-24.

[16] N. Budoma, *Fundamentals of Deep Learning*, 2017-06.

[17] M. Vakalopoulou, S. Christodoulidis, N. Burgos, O. Colliot, y V. Lepetit, “Deep learning: basics and convolutional neural networks (CNN),” 2023. [En línea]. Disponible: https://hal.science/hal-03957224

[18] A. Krizhevsky, I. Sutskever, y G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in *Advances in Neural Information Processing Systems*, F. Pereira, C. Burges, L. Bottou, y K. Weinberger, Eds., vol. 25. Curran Associates, Inc., 2012. [En línea]. Disponible: https://proceedings.neurips.cc/paper_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf

[19] S. Marschner y P. Shirley, “Fundamentals of computer graphics,” 2015.

[20] I. Young, J. Gerbrands, L. Van Vliet, C.-d. Bibliotheek, D. Haag, Y. Theodore, G. Jacob, V. Vliet, y L. Jozef, “Fundamentals of image processing,” 07 2004.

[21] A. Singh, K. Bacchuwar, y A. Bhasin, “A Survey of OCR Applications,” *International Journal of Machine Learning and Computing*, pp. 314–318, 2012. [En línea]. Disponible: http://www.ijmlc.org/show-31-230-1.html

[22] “Tesseract OCR,” May 2023, original-date: 2014-08-12T18:04:59Z. [En línea]. Disponible: https://github.com/tesseract-ocr/tesseract

[23] A. M. Mañas, *Capítulo 8 Métodos basados en Deep Learning | Notas sobre pronóstico del flujo de tráfico en la ciudad de Madrid*, Jun. 2019. [En línea]. Disponible: https://bookdown.org/amanas/traficomadrid/m%C3%A9todos-basados-en-deep-learning.html

[24] J. Madrid, “Escala de grises,” 2018. [En línea]. Disponible: http://glosario.ldr.webs.upv.es/postout/3716/escala-de-grises

[25] J. A. Cortes Osorio, W. A. Urueña, y J. A. Mendoza Vargas, “Técnicas alternativas para la conversión de imágenes a color a escala de grises en el tratamiento digital de imágenes,” *Scientia et Technica*, vol. 1, num. 47,

pp. 207–212, 2011, publisher: Universidad Tecnológica de Pereira Section: Scientia et Technica. [En línea]. Disponible: https://dialnet.unirioja.es/servlet/articulo?codigo=4526322

[26] C. R. Zúñiga, “Maestro en Ciencias en Ingeniería Eléctrica,” Ago. 2021.

[27] E. Molina, J. Diaz, H. Hidalgo-Silva, y E. Chávez, “Algoritmos de Binarización Robusta de Imágenes con Iluminación No Uniforme,” *Revista Iberoamericana de Automática e Informática industrial*, vol. 15, num. 3, p. 252, Jun. 2018. [En línea]. Disponible: https://polipapers.upv.es/index.php/RIAI/article/view/8847

[28] R. Ochoa-Montiel, C. Sánchez-López, V. H. Carbajal-Gómez, y E. Juárez-Guerra, “Segmentation of Microscopic Images with NSGA-II,” *Computación y Sistemas*, vol. 22, num. 2, Jul. 2018. [En línea]. Disponible: http://www.cys.cic.ipn.mx/ojs/index.php/CyS/article/view/2944

[29] W. A. Mustafa y M. M. M. Abdul Kader, “Binarization of Document Images: A Comprehensive Review,” *Journal of Physics: Conference Series*, vol. 1019, p. 012023, Jun. 2018. [En línea]. Disponible: https://iopscience.iop.org/article/10.1088/1742-6596/1019/1/012023

[30] J. Uriarte Barragán, “Enhanced Local Adaptative Binarization,” 2021, accepted: 2021-07-21T11:30:45Z. [En línea]. Disponible: https://academica-e.unavarra.es/xmlui/handle/2454/40239

[31] A. Peña-Peñate, L. G. S. Rojas, y R. A. Núñez, “Módulo de filtrado y segmentación de imágenes médicas digitales para el proyecto Vismedic,” vol. 10, num. 1, 2016.

[32] J. F. Valencia-Murillo, D. A. Poveda-Sendales, y D. F. Valencia-Vargas, “Evaluación del impacto del preprocesamiento de imágenes en la segmentación del iris,” *TecnoLógicas*, vol. 17, num. 33, pp. 31–41, Jul. 2014, publisher: Instituto Tecnológico Metropolitano - ITM. [En línea]. Disponible: http://www.scielo.org.co/scielo.php?script=sci_abstract&pid=S0123-77992014000200004&lng=en&nrm=iso&tlng=es

[33] R. C. Gonzalez y R. E. Woods, *Digital image processing*. New York, NY: Pearson, 2018.

[34] J. C. M. Román, “MEJORA DE CONTRASTE UTILIZANDO MORFOLOGÍA MATEMÁTICA MULTIESCALA PARA IMÁGENES EN ESCALA DE GRISES E IMÁGENES EN COLOR,” Ago. 2017.

[35] K. Nasrollahi y T. B. Moeslund, “Super-resolution: a comprehensive survey,” vol. 25, num. 6, pp. 1423–1468, 2014-08-01. [En línea]. Disponible: https://doi.org/10.1007/s00138-014-0623-4

[36] C. Dong, C. C. Loy, K. He, y X. Tang, “Image super-resolution using deep convolutional networks,” 2015-07-31. [En línea]. Disponible: http://arxiv.org/abs/1501.00092

[37] J. Martínek, L. Lenc, y P. Král, “Building an efficient OCR system for historical documents with little training data,” vol. 32, num. 23, pp. 17 209–17 227, 2020-12-01. [En línea]. Disponible: https://doi.org/10.1007/s00521-020-04910-x

[38] M. Ramanan, A. Ramanan, y E. Charles, “A preprocessing method for printed tamil documents: Skew correction and textual classification,” in *2015 IEEE Seventh International Conference on Intelligent Computing and Information Systems (ICICIS)*, pp. 495–500.

[39] “Layout-parser/layout-parser,” original-date: 2020-06-10T20:22:54Z. [En línea]. Disponible: https://github.com/Layout-Parser/layout-parser

[40] L. Rey Vega y H. Rey, “Wiener filtering,” in *A Rapid Introduction to Adaptive Filtering*, ser. SpringerBriefs in Electrical and Computer Engineering, L. R. Vega y H. Rey, Eds. Springer, 2013, pp. 7–17. [En línea]. Disponible: https://doi.org/10.1007/978-3-642-30299-2_2

[41] N. Priyadharshini y V. Ms, “Document segmentation and region classification using multilayer perceptron,” 2013.

[42] W. Menke, “Chapter 11 - continuous inverse theory and tomography,” in *Geophysical Data Analysis (Fourth Edition)*, fourth edition ed., W. Menke, Ed. Academic Press, 2018, pp. 223–248. [En línea]. Disponible: https://www.sciencedirect.com/science/article/pii/B9780128135556000113

[43] Z. Z. Aung y C. M. M. Maung, “Myanmar optical character recognition using block definition and featured approach,” in *2017 3rd International Conference on Science in Information Technology (ICSITech)*, pp. 313–318.

[44] Y. L. Khomba Khuman, H. Mamata Devi, T. Romen Singh, y N. Ajith Singh, “Graphics separation system for printed document images,” in *2020 International Conference on Computer Communication and Informatics (ICCCI)*, pp. 1–4, ISSN: 2329-7190.

[45] R. Mittal y A. Garg, “Text extraction using OCR: A systematic review,” in *2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA)*, pp. 357–362.

[46] N. Otsu, “A threshold selection method from gray-level histograms,” vol. 9, num. 1, pp. 62–66, 1979-01, conference Name: IEEE Transactions on Systems, Man, and Cybernetics.

[47] O. Ronneberger, P. Fischer, y T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, ser. Lecture Notes in Computer Science, N. Navab, J. Hornegger, W. M. Wells, y A. F. Frangi, Eds. Springer International Publishing, pp. 234–241.

[48] R. Wongso, F. A. Luwinda, y Williem, “Evaluation of deep super resolution methods for textual images,” vol. 135, pp. 331–337, 2018-01-01. [En línea]. Disponible: https://www.sciencedirect.com/science/article/pii/S1877050918314704

[49] K. Kettunen y M. Koistinen, “Open Source Tesseract in Re-OCR of Finnish Fraktur from 19th and Early 20th Century Newspapers and Journals – Collected Notes on Quality Improvement,” 2019.

[50] D. Kumar y R. Singh, “A comparative Analysis of Feature Extraction Algorithms and Deep Learning Techniques for Detection from Natural Images,” in *2019 4th International Conference on Information Systems and Computer Networks (ISCON)*. Mathura, India: IEEE, Nov. 2019, pp. 483–487. [En línea]. Disponible: https://ieeexplore.ieee.org/document/9036279/

[51] P. Anzorena, M. Laguarda, y B. Olivera, “Extracción de eventos en prensa escrita Uruguaya del siglo XIX,” 2018.

[52] E. J. Morales y M. V. Meya, “Los recortes de prensa del Archivo Histórico de la Universidad de Barcelona. Propuesta de descripción y digitalización para su difusión,” 2016.

[53] S. Pletschacher, C. Clausner, y A. Antonacopoulos, “Europeana Newspapers OCR Workflow Evaluation,” in *Proceedings of the 3rd International Workshop on Historical Document Imaging and Processing*. Gammarth Tunisia: ACM, Ago. 2015, pp. 39–46. [En línea]. Disponible: https://dl.acm.org/doi/10.1145/2809544.2809554

[54] A. P. Tafti, A. Baghaie, M. Assefi, H. R. Arabnia, Z. Yu, y P. Peissig, “OCR as a Service: An Experimental Evaluation of Google Docs OCR, Tesseract, ABBYY FineReader, and Transym,” in *Advances in Visual Computing*, G. Bebis, R. Boyle, B. Parvin, D. Koracin, F. Porikli, S. Skaff, A. Entezari, J. Min, D. Iwai, A. Sadagic, C. Scheidegger, y T. Isenberg, Eds. Cham: Springer International Publishing, 2016, vol. 10072, pp. 735–746, series Title: Lecture Notes in Computer Science. [En línea]. Disponible: http://link.springer.com/10.1007/978-3-319-50835-1_66

[55] M. R. M. Ribeiro, D. Julio, V. Abelha, A. Abelha, y J. Machado, “A Comparative Study of Optical Character Recognition in Health Information System,” in *2019 International Conference in Engineering Applications (ICEA)*. Sao Miguel, Portugal: IEEE, Jul. 2019, pp. 1–5. [En línea]. Disponible: https://ieeexplore.ieee.org/document/8883448/

[56] P. Jain, D. K. Taneja, y D. H. Taneja, “Which OCR toolset is good and why,” 2021.

[57] M. R. Spiegel y L. J. Stephens, *Estadística*. México, D.F.: McGraw-Hill, 2005.

[58] “OpenCV: Introduction,” Dic. 2022. [En línea]. Disponible: https://docs.opencv.org/4.7.0/d1/dfb/intro.html

[59] “patchify: A library that helps you split image into small, overlappable patches, and merge patches back into the original image.” 2021. [En línea]. Disponible: https://github.com/dovahcrow/patchify.py

[60] “Command Line Usage.” [En línea]. Disponible: https://tesseract-ocr.github.io/tessdoc/Command-Line-Usage.html

[61] “Traineddata Files for Version 4.00 +,” 2017. [En línea]. Disponible: https://tesseract-ocr.github.io/tessdoc/Data-Files.html

[62] (2023) difflib — funciones auxiliares para calcular deltas. [En línea]. Disponible: https://docs.python.org/3/library/difflib.html

Anexos

Anexo A. Transcripción de periódicos

Experimento Propuesto

Tema tesis:

Analizar y aplicar técnicas de tratamiento de imágenes de periódicos antiguos del Ecuador para mejoras en el proceso de reconocimiento de textos (OCR).

Autores:
- Kevin Ismael Ochoa Arévalo
- Lucìa Carolina Quituisaca Suconota

Objetivo:
Transcribir el peridico a un archivo de texto, para realizar una evaluación de la eficiencia de los procesos OCR y las técnicas de tratamiento de las imágenes.

Proceso:

1. Asignación de Documentos PDF de periódicos:
    a. Los periódicos están nombrados del 1 al 50, se entregará uno a cada estudiante.

    b. Los periódicos se encuentran en:
        i. https://drive.google.com/drive/folders/1yJ7sWnwiS7LrvzHtcHUcPx6xtLrg5yqu?usp=sharing

    c. Poner su nombre en este documento, para que no se repitan las elecciones:
        i. https://docs.google.com/spreadsheets/d/1isU_gB41TVHcOhIPlTiPgWMtALxsprgfZqiy2ydMFUs/edit?usp=sharing

2. Transcribir el texto de los periódicos en un archivo .txt, de **forma manual**
    a. Consideraciones a tomar en cuenta:
        i. Solo tomar en cuenta puntos, comas y guiones medios, ignorar el resto de caracteres especiales (comillas, guiones bajos, punto y coma, barras, etc)

        ii. Transcribir siguiendo el orden el periódico, en caso de tener columnas se transcribe cada columna una debajo de otra, iniciando por la izquierda.

        iii. No utilizar ningún OCR o programas en línea para transcribir, ya que se necesita tener el texto lo más fiel posible al original.

Anexo B. Tabla de resultados con OCR1

Resultados OCR1

Periodico	Transcrito	Sin Tratamiento	Escala de Grises	Binarizacion	Otsu	Adaptativa	Mediano	Gauss	Contraste	Binarizacion	Mediano Otsu	Segmentacion	FSRCNN	ESPCN	LAPSRN	promedio por periodico
1	5301	83.19185059	85.58762498	82.61273345	5.470665197	43.7087342	83.19185059	17.03452179	32.33352198	93.66157329	85.55857197	90.69987795	90.22852882	65.75309241		
10	7670	43.31160365	45.72359844	23.01173403	35.77574957	6.818774446	11.74202868	11.74202868	83.31160365	9.061277705	9.845546284	66.80573564	63.82138201	63.53246416	63.53324641	35.34363941
12	10490	30.30505343	32.65967588	22.04957102	22.09310391	14.38513869	3.384175405	3.384175405	30.30505343	3.346043851	2.917063877	51.96377502	51.96377502	53.53716584		27.76424435
13	11415	19.7021463	20.95488392	17.43220193	15.78725736	6.520718353	4.564169952	4.564169952	19.7021463	3.576724415	3.576724415	49.5926412	49.5926412	46.03886407		21.74284945
14	9053	52.94377554	59.61559704	48.89987076	46.87948746	6.716005744	52.01590633	52.01590633	52.94377554	38.70540152	35.09333962	58.79181845	54.40185574	56.35700873	58.55517508	48.13794954
15	9363	41.38440671	63.58000413	47.22845242	55.93229748	6.386841824	14.66410339	14.66410339	41.98440671	11.95129766	10.92598526	66.02586642	66.02586642	76.12943569	77.94510307	43.47860243
16	8673	52.71532342	51.59128099	29.22864061	36.26196241	6.145509051	5.090155655	5.090155655	52.71532342	1.464314539	3.090862117	75.11818287	73.44632768	72.28179407	72.08578447	38.38898971
17	5137	75.93926416	75.76400643	67.29608721	65.87502433	0.214132762	12.94529881	12.94529881	75.93926416	7.57253134	10.08370644	73.91473623	73.3502044	81.17578353	81.95444812	51.06827334
18	4808	77.03826955	77.30826225	60.9400983	65.87502433	0.145590822	28.68136439	28.68136439	77.03826955	14.80865225	20.98359691	83.40266223	77.1472629	81.7803606	81.3435901	55.36308353
19	7722	73.86687387	74.0999741	63.07951308	62.89821229	4.817404817	21.10852111	21.10852111	73.86687387	14.07664408	15.83766584	79.73223973	79.11162911	80.1289303	80.5102305	53.17460317
2	7301	48.04821357	52.10245172	42.36405972	42.10382139	10.75195179	11.1354609	11.1354609	48.04821257	4.424405	7.597813998	63.92275031	68.1824076	70.3191347	74.01725791	39.60807717
20	11858	48.40613932	51.52539568	39.32366335	38.16832518	9.124641592	5.9959521	5.9959521	48.40613932	6.594703997	4.655085175	71.64783269	69.35766571	72.3983804	72.39883084	38.35264821
21	14753	30.80051515	32.18328476	21.08723649	25.02541856	9.462482207	4.57534069	4.57534069	30.80051515	3.951728629	59.28285772	58.1034365	57.89498407	59.20829562		28.69101684
22	9159	60.86390051	58.8738239	50.86799869	50.74789824	8.155912217	15.50387597	15.50387597	60.86390051	7.631837537	11.8353532	60.24675183	69.7783608	59.91440114	76.3074575	44.62042471
23	4843	27.42102003	21.9492504	25.10843088	17.5715305	20.50381995	5.360573198	5.360573198	27.42102003	5.360573198	4.665329011	45.48833368	37.7658472	53.3535129	55.0278752	23.21164867
24	12111	33.91132029	34.0269768	25.21674511	25.14243425	4.632152589	5.300966064	5.300966064	33.91132029	3.822970853	4.191031286	66.9061184	69.3190752	69.3419257		32.1049766
25	14067	28.98983436	28.74813393	19.27205516	18.9877071	15.12049478	2.381460155	2.381460155	28.98983436	2.089997867	1.677685363	15.98066397	47.50124405	46.9325714	50.3163421	22.09781759
26	8745	57.01543739	65.1572327	51.01200686	54.33966264	3.293310463	8.975299584	8.975299584	57.01543739	13.2075417	76.71812464	73.10463122	73.08176101	72.9445974		42.15840408
27	4089	62.16678895	57.25161616	37.95549034	43.53142578	6.627537295	8.975299584	8.975299584	62.16678895	3.350452433	7.092189582	70.09048667	71.8268523	75.6908797	74.5165996	41.92850006
28	9007	59.73132008	63.38403464	46.58598868	52.79232768	6.693047408	8.993005544	8.993005544	59.73132008	7.349839014	66.1152437	65.52681248	67.1366714	69.3904708		41.92850006
29	7841	67.03226629	67.86124219	54.08748884	59.99234791	23.31353289	21.6554011	21.6554011	67.03226629	10.86596097	17.40849381	62.41550823	75.5153878	78.293585	79.2628491	50.46912811
3	7574	56.36387642	68.25943547	51.09585424	56.86559282	5.439662002	16.00212149	16.00212149	56.36387642	12.27884869	11.81674148	75.75197613	66.39952697	73.7877866	74.4124636	46.81567207
30	4986	44.64500602	42.35860409	36.82310469	36.14119535	11.61251504	9.386281588	9.386281588	44.64500602	3.548993832	6.297633373	66.39552697	52.1460088	57.4007202	57.9823508	31.65408667
31	6976	30.46794572	32.20491312	23.26243259	25.74895147	5.557219892	6.7855003	6.7855003	30.46734572	4.358897543	5.092869982	46.9592704	49.2959856	50.2995859	53.9310964	23.37283645
32	8091	44.13545915	43.3444568	21.2458287	34.4580398	16.56161167	16.68520578	16.68520578	44.13545915	10.64145347	39.78494624	59.9678653	70.2632547	66.6543075		34.68227484
33	13725	22.30965392	20.95446266	14.47723133	15.58469945	5.129326047	3.409836066	3.409836066	22.30965392	2.593806922	35.21311475	45.6466304	39.0455374	49.527686		20.11345303
34	10214	21.49011161	20.41315841	12.91836793	10.43665557	1.869983277	1.869983277	1.869983277	21.49011161	2.300763658	1.351086744	39.48502056	24.2801899	35.5198766	37.0756158	17.45992895
35	9085	63.35531095	63.31315355	50.82003302	51.09521189	4.446890479	15.33296643	15.33296643	63.35531095	8.266373143	9.981474959	66.68134287	67.3087509	76.532746	77.3582288	45.25434389
36	6679	60.98218296	60.06887259	32.0257526	45.11154564	7.995208864	14.19374158	14.19374158	60.98218296	4.252133553	10.7950292	42.41652942	73.3642769	76.4186255	77.137296	
37	4923	19.45967906	22.14097095	21.30814544	17.89559212	7.434491164	3.798496852	3.798496852	19.45967906	5.484460695	3.087548243	54.47897623	46.7601056	47.4710542	46.7194999	22.80694118
38	10247	17.33190202	19.1373048	12.55977359	13.39433981	6.538499073	3.308285352	3.308285352	17.33190202	2.24455938	2.586127268	52.70810969	45.7890114	44.4422579	46.0328852	20.62228286
39	7577	21.23531741	32.65144516	20.91856935	24.79873101	5.397914742	5.173551538	5.173551538	21.23531741	3.853767982	4.315692226	64.98614227	59.3506649	64.9464877	66.2267881	28.69028282
4	10466	41.6704013	55.943057	34.32065737	46.91318617	6.51633862	15.25893169	15.25893169	41.63004013	10.81597554	10.98796102	76.81062488	72.4441047	75.0812153	74.842346	41.32104583
40	7088	69.25620767	51.49545853	55.48814898	9.001128668	19.65239454	19.65239454	19.65239454	69.28617738	8.767286662	14.22123896	79.43002257	70.2313768	76.5378038	76.573	41.32104583
41	6742	30.70305547	30.03559775	30.03559775	26.92079502	32.21595966	5.547315337	5.547315337	30.70305547	6.333432216	4.47958972	25.2447345	70.5132008	44.7196775	64.7285679	
42	5410	74.82439926	75.63770795	52.58780037	63.51210479	2.92051756	27.09796673	27.09796673	74.82439926	5.545286506	18.52125693	71.16451017	81.8299455	82.070403	82.6425323	
43	7072	61.26979638	65.55428664	38.36255656	49.9010181	6.207579186	12.74038462	12.74038462	61.26979638	3.563348416	10.50622172	75.83427602	77.2824434	79.3410635	79.6321493	45.30037977
44	1703	88.25601879	87.84479945	45.33176747	78.09747504	1.664157369	67.41045214	67.41045214	88.25601879	45.97768644	84.85026424	85.7890114	44.4422579	46.032885	46.0328852	52.87894745
45	8047	45.39579368	51.01127998	43.69331085	44.19038151	3.119474848	14.27861315	14.27861315	45.39579368	13.17261091	11.48254008	77.15918976	70.3889064	73.2943954	73.6796326	41.46782274
46	6229	77.52448226	76.856248	47.88890673	70.26810082	10.93377399	27.51645529	27.51645529	77.52448226	5.233568445	19.7	81.9557004	76.1619947	81.6021834	82.4586928	54.93773364
47	4657	71.48378705	58.92205282	43.11786705	58.92205282	11.31623685	14.58020185	14.58020185	71.48378705	4.230191111	11.99016053	82.24178566	77.8827574	79.7938971	79.3874116	49.30672722
48	14768	34.45964247	38.3193911	28.04712893	27.95910076	6.554712893	6.839111593	6.839111593	34.45964247	5.599945829	4.820806501	65.69609967	66.7050379	68.3301735	67.5582304	33.01394908
49	10476	47.05994654	60.91050572	37.94387171	50.31500573	14.08934708	8.648339061	8.648339061	47.05994654	9.938998855	7.035117911	74.33180603	70.887657	47.2346153	73.91.17984	36.83632479
5	3762	70.4678326	53.26953748	52.92397661	47.2089142	0	20.36150984	20.36150984	47.06783626	28.3625731	14.54013822	79.02711324	52.5518343	60.2604997	57.7352472	44.82418167
50	10310	61.72647915	63.90826538	50.98933075	52.51212415	8.069835112	15.72259942	15.72259942	61.72647915	11.24151309	11.29970902	76.19786515	72.1959292	78.1993629	78.1073036	47.23456745
6	9374	75.91209729	77.75762748	66.24706635	64.96629281	8.555579262	9.36532409	9.36532409	75.91209729	4.608491572	6.953408577	85.39577555	87.5400437	87.8387079	88.1053791	53.46967745
7	13039	26.65848608	27.14165197	23.85919166	20.41567605	10.14645363	5.031060664	5.031060664	26.65848608	3.545216504	3.527877905	50.1725925	54.1605954	60.3803972	59.6784216	26.85852604
8	4412	72.16681777	70.98821396	61.5593835	62.51133273	0.158658204	27.51586582	27.51586582	72.16681777	12.26201269	18.7	79.80507706	70.7162284	75.7	77.3345416	
9	4779	81.41870684	82.17202151	50.91023227	71.31198996	0.753295685	34.31680268	34.31680268	81.41870684	4.331450094	23.62418916	75.76898933	74.3670239	79.2843691	79.8911906	55.27755358
Promedio		**50,9529461**	**52,83648387**	**39,31147486**	**43,50487787**	**8,03450881**	**14,78244331**	**14,78244331**	**50,9529461**	**7,597006598**	**10,82586962**	**64,79476535**	**64,93320656**	**67,85190823**	**68,34082291**	

Anexo C. Tabla de resultados con OCR2

Resultados OCR2

Periodico	Transcrito	Sin Tratamiento	Escala de Grises	Binarizacion	Otsu	Adaptativa	Mediano	Gauss	Contraste	ano Binarizacion	Mediano Otsu	Segmentacion	FSRCNN	ESPCN	LAPSRN		promedio por periodico
1	5301	82,75797019	85,54989625	82,28636106	83,21071496	5,300886625	44,10488587	72,45802679	83,90869647	17,03452179	31,37143935	93,60498019	85,77626863	85,26693077	89,15299		**67,27032635**
10	7670	47,90091265	47,90091265	23,32464146	36,83181226	7,014341591	12,54237288	19,64797914	45,14993481	10,44328553	10,61277705	68,17470665	64,27640156	66,66232073	66,40156454		**37,63456882**
11	10490	34,96663489	32,84080076	23,21258341	22,18303146	14,38512869	3,517635844	4,671115348	30,36224976	3,698760724	3,012392755	66,00571973	52,46901811	53,99428027	53,68922784		**28,50061283**
12	11415	21,44546649	21,44546649	18,239159	16,74989049	6,727989488	4,581690758	5,519053876	20,06132282	3,407796759	3,5304424	51,61629435	47,21857205	47,81427946	49,65396408		**22,71509918**
13	9053	56,15817961	60,06848558	49,93924666	47,36551419	5,909643212	52,45774881	55,10880371	54,53440848	40,42858721	36,63978792	58,82028057	54,81056004	57,96973379	56,78780515		**49,07134178**
14	12710	36,34146341	39,83477577	28,30841857	31,62863887	7,317073171	6,15263572	9,35483871	40,69236821	4,6184107	5,066876475	66,9472856	61,18017309	66,30212431	67,64752164		**33,6709003**
15	9363	54,80081171	64,17814803	48,87322439	56,97960056	6,482964862	15,30492364	33,91007156	42,26209548	12,12218306	11,49204315	80,2627363	66,55986329	77,22952045	71,25921179		**45,83695702**
16	8673	53,76455667	53,76455667	30,9350859	39,20212153	5,684307621	5,868788193	13,31719128	53,55701603	1,475844575	5,107805834	75,1297129	74,57627119	75,4525539	73,65386833		**40,10640576**
17	5137	69,61261437	75,18006619	66,34222309	65,91395756	0,2725326066	14,03542924	57,09558108	75,66673156	7,630912984	10,45357212	73,27233794	73,46700409	80,41658556	80,72805139		**53,5776857**
18	4808	73,3985025	77,66222962	60,2953411	65,78618968	0,1039933444	29,90848586	55,63643927	76,70549085	15,93178037	21,96339434	83,11148087	77,05906822	81,46838602	80,03327787		**57,07600428**
19	7722	70,92722093	74,7992748	64,33566434	63,66226366	4,623154623	22,49417249	38,00828801	74,30717431	15,7990158	18,05231805	79,96632997	79,52602953	78,19217819	77,67417767		**54,45480445**
2	7301	41,09026161	52,62292837	43,14477469	42,35036296	11,12176414	12,01205314	16,16216957	48,59608273	5,163676209	9,012464046	65,00479386	69,30557458	69,62059992	73,16805917		**39,88396893**
20	11858	46,96407489	52,47933884	41,31388092	40,45370214	8,559622196	5,709225839	10,70163603	48,86152808	7,834373419	4,93337831	73,78141339	68,80586946	71,20931017	70,91415078		**39,46582175**
21	14753	32,50864231	32,96278723	21,50749	25,87948214	9,564156443	4,365213855	6,012336474	31,67491358	4,548227479	3,728055311	59,62855013	59,43875822	59,19474005	58,58469464		**29,25700342**
22	9159	63,41303636	59,19860247	51,54492849	52,01441205	8,144993995	15,59122175	31,16060705	60,97827274	9,040288241	13,59318703	59,83185937	69,84386942	74,95359755	72,68260727		**45,85653456**
23	4843	22,79578773	21,39169936	25,50072269	17,15878588	20,00825934	4,749122445	5,905430518	27,73074541	4,728474086	3,964484823	45,86000413	38,24075986	53,33470989	53,83027049		**24,65708976**
24	12111	32,80488812	33,96911898	25,5222525	26,45528858	4,55783998	4,772520849	6,291800842	33,78746594	3,550491289	4,062422591	67,21988275	68,07860623	42,12699199	69,79605317		**30,21397313**
25	14067	29,70071799	28,57041302	20,36681595	19,68436767	15,41195706	2,338807137	4,698940783	29,31684083	1,962038814	1,727447217	16,20103789	48,55335182	47,33063198	49,51304471		**22,52688663**
26	8745	63,47627216	65,90051458	52,12121212	55,42595769	3,739279588	17,58719268	26,31217839	57,31275014	12,77301315	14,96855346	78,33047456	74,61406518	72,99028016	75,4488279		**47,92861227**
27	4089	57,54463194	57,54463194	38,20004891	46,51504035	6,260699438	9,366593299	20,22499389	61,99559795	3,20371729	6,554169724	69,62582539	70,89752996	73,88114453	74,15015896		**42,56891311**
28	9007	56,01199067	63,07316532	47,48528922	53,24747419	6,306206284	10,09215055	24,33662707	58,5766626	5,14044632	8,304651937	66,09303875	65,75996447	68,30243144	68,46896858		**42,94279053**
29	7841	67,30008927	68,65195766	53,98546104	60,68103558	23,04552991	22,17829359	43,04297921	68,24384645	10,48335671	17,71457722	62,91289376	76,40607065	78,67618926	79,67096034		**52,35666005**
3	7574	68,36546079	68,12780565	52,13889622	57,31449696	4,792711909	16,33218907	29,21837866	56,15262741	13,1370478	12,83337734	77,17190388	67,38843412	73,07895432	71,56060206		**47,68663473**
30	4986	42,29843562	41,73686322	37,98636181	36,04091456	11,17127958	9,526674689	25,09025271	44,12354593	3,590052146	6,357801845	28,84075411	51,20336943	58,50381067	54,79342158		**32,23310985**
31	6676	33,58298382	32,74415818	23,44218095	26,25823847	5,467345716	6,411024566	15,44337927	31,11144398	3,924505692	5,017974835	48,29239065	52,20191732	51,6327142	51,57279808		**27,65021827**
32	8091	34,13669509	43,9747868	21,13459399	35,96588802	17,05598813	16,64812755	47,1511556	44,93882091	0,8157211717	10,61673464	40,45235447	60,74650847	67,48238784	67,3340749		**36,31813126**
33	13725	21,63934426	21,63934426	15,21311475	16,42258652	4,765027322	3,191256831	5,449908925	22,81967213	1,872495446	2,630236794	35,70127505	47,60655738	41,75591985	49,60291439		**20,73640385**
34	10214	15,88995496	19,65929117	14,90111612	13,16820051	10,64225573	1,723125122	3,270021539	21,78382612	2,095163501	1,282553358	39,2794204	23,90836107	34,79537889	36,32269434		**17,05152592**
35	9085	59,55971381	62,49862411	52,16290589	51,26031921	4,413869015	16,77490369	24,6119978	62,97193176	10,2916896	13,57182168	67,14364337	66,7473858	74,99174463	75,31095212		**45,87939303**
36	6679	58,15241803	58,15241803	61,01212756	32,504866	46,18954933	7,441233718	16,75400509	26,27638868	61,3864351	3,847881419	12,69651145	42,77586465	73,22952538	75,59514897		**41,14388381**
37	4923	24,17225269	22,50660167	21,40970953	18,09872029	7,353239894	4,04225066	7,434491164	19,96749949	5,301645338	2,884420069	55,75868373	48,03981312	45,88665448	43,28661385		**23,29589971**
38	10247	21,8600566	19,15682639	12,65736313	13,52591002	6,47994535	2,976480921	4,898994828	16,99033864	2,020103445	2,459256368	52,4738948	45,24251	41,18278521	44,87166976		**20,48543825**
39	7577	30,85653953	33,74686551	21,31450442	25,37943777	5,160353702	5,397914742	8,407021249	21,5652633	4,606044609	4,434472746	65,46126435	60,51207602	62,06942062	58,87554441		**29,12762307**
4	10466	54,4142939	55,43665202	35,65832219	48,05083126	6,392126887	16,51060577	26,44754443	41,26695968	12,2205236	12,18230461	77,39346455	73,76266004	74,89012039	75,01433212		**43,54576725**
40	7088	53,99266366	66,93002257	51,22742664	55,91139955	8,450902935	19,96331828	33,14051919	66,78893905	8,211060948	14,46106095	78,38600451	69,751693	73,58916479	74,01241535		**48,2011851**
41	6742	32,72026105	30,24325126	34,7967962	26,89113023	30,21358647	4,50904776	6,867398398	30,83654702	6,170275883	4,242064669	23,96914862	65,21803619	50,71195491	56,73390685		**28,86595754**
42	5410	74,71349353	75,56377079	51,34935305	62,5323475	2,92051756	28,8909427	54,15896488	74,2883549	5,61922366	19,70425139	70,72088725	81,92236599	82,08872458	82,45841035		**54,78082915**
43	7072	66,84106335	66,84106335	38,54638009	51,68269231	6,447963801	13,27771493	19,1459276	61,79298643	3,068438914	11,36877828	76,32918552	77,61595023	78,66233032	79,83597285		**46,53260343**
44	1703	87,19906048	86,84674105	45,91896653	79,21315326	1,820317087	69,64180857	84,4979448	87,31650029	2,701115678	48,38520258	84,38050499	75,57251908	79,50675279	85,37874339		**65,59852361**
45	8047	52,50403877	52,50403877	44,52591028	45,30881074	3,268298745	13,7691065	21,84665093	45,90530633	13,63240959	11,28370821	77,51957251	70,85870511	74,0648689	74,37554368		**42,9547835**
46	6229	77,34788891	79,70781827	47,31096484	71,56847006	11,51067587	27,43618558	42,94429282	77,39605073	4,687750843	18,75100337	83,06309199	80,43024563	82,35671857	83,06309199		**56,25530353**
47	4657	63,32402834	68,82112948	43,59029418	59,60919047	11,31629805	15,58943526	23,53446425	71,61262615	3,435688211	12,62615418	83,31543912	79,10672107	80,05153532	78,8275714		**49,62575539**
48	14768	35,2248104	39,53819068	29,8347779	29,24566631	7,123510293	7,455308776	8,132448537	35,59046587	6,940682557	5,335861322	66,99620802	68,81771398	69,50162514	68,1947454		**34,13800108**
49	10476	54,37189767	61,38793433	38,63115693	52,51050019	14,20389462	8,886979763	14,52844597	47,69950363	11,96067201	7,521954945	74,56090111	71,34402444	48,3008782	72,56586483		**41,3196149**
5	3762	48,61775651	52,60499734	50,9835194	46,1456672	0	20,38809144	36,49654439	70,28176502	28,70813397	14,72620946	77,53854333	51,9404572	55,7682084	55,63530037		**43,55965672**
50	10310	63,77303589	64,99515034	51,43549952	53,24927255	8,031037827	16,8671193	29,22405432	62,69641125	12,56062076	12,46362755	78,77788555	79,22405432	80,17458778	79,68962173		**49,5115699**
6	9374	77,26690847	77,26690847	66,42841903	65,51098784	8,779603158	10,52912311	30,3819074	75,02666951	4,768508641	8,726264135	85,23575848	88,28675059	87,41199061	88,36142522		**55,28437319**
7	13039	32,40279163	28,22302324	25,7995245	22,67045019	10,72168111	4,762635171	6,342510929	27,35639236	4,110744689	4,103075389	51,9441675	56,64544827	60,52611397	59,299026		**28,20768464**
8	4412	69,65095195	71,78150499	61,62737987	61,15140526	0,2039891206	27,71985494	48,64007253	72,12148685	11,42339075	19,19764279	79,46509519	71,03354488	72,00815956	76,11060743		**53,00964901**
9	4779	82,82067378	84,0343168	49,75936388	74,15777359	0,6486712701	35,25842226	52,04017577	83,05084746	3,347980749	23,64511404	77,56852898	75,66436493	79,68194183	80,91650973		**57,32819179**
Promedio	**8284,54**	**51,067684**	**53,12519739**	**40,3516351**	**44,01457943**	**8,721739294**	**15,03304092**	**25,71353125**	**50,48022732**	**9,072546556**	**11,210541**	**64,63676245**	**65,00863404**	**66,72588985**	**68,25017959**		