Universidad de las Ciencias Informáticas

Facultad 1

Solución de Software para la Digitalización de Documentos

Trabajo de Diploma para optar por el Título de Ingeniero en Ciencias Informáticas.

Autores: Caridad Odil Reyes Duconger

Frank Rosales Muñoz

Tutor: MSc. Alexeis Companioni Guerra

Co-Tutor: MSc. Siovel Rodríguez Morales

La Habana, junio del 2011

Declaración de autoría

Declaramos ser autores de la presente tesis y reconocemos a la UCI los derechos patrimoniales de la misma, con carácter exclusivo.

Para que así conste firmo la presente a los ____ días del mes de _____ del año _____.

_____ _____

Caridad Odil Reyes Duconger Frank Rosales Muñoz

_____ _____

SubDirector de Formación. CENIA. SubDirector de Investigación y Postgrado. CENIA.

MSc. Alexeis Companioni Guerra MSc. Siovel Rodríguez Morales

Resumen

Con el avance progresivo de la informática y las comunicaciones en el mundo, surgieron cambios transcendentales en cuanto al tratamiento de la información. Esto constituye un elemento relevante en el funcionamiento de instituciones y organizaciones. La gestión documental para las empresas, figura como un problema fundamental, lo cual se ve representado en gastos para locales e infraestructuras, con el objetivo de garantizar el estado de conservación de los documentos. A esto se suma el tiempo dedicado a la organización, la búsqueda de documentos, duplicados, gastos de fotocopias, entre otros. Encontrar hoy en día una solución que convierta grandes volúmenes de documentación en información controlable y perfectamente localizable es imprescindible, si se desea alcanzar la rentabilidad, la facilidad de difusión y la conservación de los documentos.

Actualmente en el departamento de Gestión Documental y Archivística de la Facultad 1 en la Universidad de las Ciencias Informáticas, se está desarrollando un sistema que se encargue del proceso íntegro de la gestión documental. Dicho sistema emplea el gestor de contenidos empresariales (ECM) Alfresco como repositorio de contenidos para almacenar los documentos digitalizados, pero no permite establecer búsquedas sobre estos, pues carece de un proceso capaz de extraer la información útil de los mismos; por tal motivo se hizo necesario diseñar un módulo para la digitalización de documentos con el objetivo de facilitar la extracción de la información contenida en los documentos digitalizados almacenados en el ECM Alfresco, para de esta forma hacer un mejor uso del contenido del documento procesado y poder realizar búsquedas sobre los mismos.

Mediante el desarrollo de un prototipo funcional de la solución propuesta realizado en Matlab, queda sustentada la factibilidad de desarrollar la solución de software que se propone para mejorar los procesos de recuperación de información desde el ECM Alfresco. Comprobando la veracidad y eficiencia de los algoritmos seleccionados a través de las diferentes fases por la cual transitó la presente investigación.

Palabras claves: gestión documental, digitalización, ECM Alfresco.

Índice General

Introducción

Desde tiempos remotos el hombre se ha preocupado por mantener guardadas sus vivencias y recuerdos. De esta manera surgieron diferentes medios para grabar estos conocimientos, tales como: el arte rupestre, los jeroglíficos, la escritura cuneiforme, el uso de la tinta y el papiro. Posteriormente con el descubrimiento del papel la humanidad obtuvo una nueva forma para expresar y hacer perdurar sus pensamientos; siendo estas las bases para la confección de los documentos, libros y artículos que conocemos en nuestros días.

Con la llegada de las ciencias informáticas, surge la inquietud de almacenar toda esta información en buen estado y en la menor cantidad de espacio posible, intercambiar documentos e información con terceros sin necesidad de llevarle la misma en formato duro, entiéndase el propio libro que usted compró en una librería y se lo desea mostrar a su amigo, por solo citar un ejemplo práctico. Con esta idea surgen los procesos de digitalización de documentos; con el fin de poder contar con cualquier documento, libro, revista, entre otros de forma digitalizada y poder compartirlo a través de la red o mediante cualquier dispositivo de almacenamiento.

Actualmente el departamento de Gestión Documental y Archivística se enfoca en desarrollar un sistema (eXcriba[1]), el cual se encargue del proceso íntegro de la gestión documental que se lleva a cabo en cada una de las empresas e instituciones del país. Dicho sistema emplea el ECM Alfresco como repositorio de contenidos para almacenar los documentos digitalizados, pero no permite establecer búsquedas sobre estos, pues carece de un proceso capaz de extraer la información útil de los mismos. Por tal motivo es necesario encontrar una vía que permita obtener la información

[1]eXcriba: Gestor de documentos administrativos, basado en Alfresco.

contenida en los documentos digitalizados, con el objetivo de hacer un mejor uso del contenido y poder realizar búsquedas sobre los mismos.

Ante tal situación se plantea la siguiente **pregunta científica:** ¿Cómo mejorar el proceso de digitalización de documentos de modo que se facilite la recuperación de los mismos desde el ECM Alfresco?

Para desarrollar esta investigación es necesario realizar un estudio de los procesos que se llevan a cabo a la hora de realizar la obtención de la información contenida dentro de los documentos. Para enmarcar los límites de la misma se define como **objeto de estudio:** el proceso de digitalización de un documento, delimitando el **campo de acción:** a los métodos para la obtención de la información contenida dentro de los documentos digitalizados almacenados en el ECM Alfresco.

La presente investigación tiene como **objetivo general:** diseñar un módulo para la digitalización de documentos que facilite la extracción de la información contenida en los documentos digitalizados almacenados en el ECM Alfresco.

Con el propósito de dar cumplimiento al mismo se trazan los siguientes **objetivos específicos:**

- Analizar las soluciones de digitalización de documentos existentes en la actualidad.
- Seleccionar las tecnologías, herramientas y metodología necesaria para el diseño del módulo.
- Implementar un prototipo funcional de la propuesta de solución en ambiente Matlab.
- Validar los resultados obtenidos mediante los casos de prueba de aceptación.

El trabajo de diploma queda sustentado en la siguiente **idea a defender:** El diseño de un módulo para la digitalización de documentos que incorpore técnicas de procesamiento digital de imágenes, clasificación y extracción de resúmenes, posibilitará extraer y representar de manera compacta la información contenida en los documentos digitalizados.

Con el propósito de dar validez a lo anteriormente planteado se formulan las siguientes **tareas de investigación:**

- Revisión y análisis de la bibliografía existente relacionada con las aplicaciones de digitalización de documentos.
- Selección de las técnicas y algoritmos de mejora de imágenes más apropiadas para su aplicación al tratamiento de documentos digitalizados.
- Profundización en el conocimiento de las aplicaciones para el reconocimiento óptico de caracteres (OCR), documentando de modo particular las limitaciones actuales de este tipo de aplicaciones.
- Realización de un estudio relacionado con los algoritmos y técnicas de segmentación de imágenes.
- Elección de los algoritmos y técnicas de clasificación supervisada y no supervisada.
- Exploración en la bibliografía existente relacionada con los algoritmos y técnicas de extracción de resúmenes de textos, para seleccionar la más factible.
- Elección de la metodología de desarrollo y las tecnologías adecuadas para el desarrollo de la solución.
- Realización del diseño del módulo a implementar.
- Implementación de los algoritmos en matlab de la solución propuesta.
- Ejecución de los casos de prueba de aceptación para asegurar la calidad del resultado.

Para la realización de este trabajo de diploma se utilizaron diferentes métodos científicos, ellos constituyen un *"Conjunto de reglas que señalan el procedimiento para llevar a cabo una investigación"*[1]. Este conjunto de reglas parten de principios claros, razonables e incuestionables, que servirán para dar validez a las reglas del método científico.

Métodos teóricos a utilizar:

Analítico-Sintético: para la elaboración del presente trabajo de diploma se dividirá el objeto de estudio en conceptos que serán examinados por separado, estudiándose rigurosamente cada uno de ellos de manera independiente, y confrontando el criterio de disímiles autores, y las correspondencias entre ellos. Además se realizará un análisis de las diferentes aplicaciones existentes en el ámbito nacional e internacional para la digitalización de documentos. Igualmente se analizarán las diferentes herramientas, algoritmos y técnicas necesarias para elaborar el diseño de este módulo.

Modelado: para una mejor comprensión del proceso de digitalización, se modelará dicho proceso de manera general, obteniendo una panorámica del flujograma de actividades inherentes al mismo.

El presente informe se compone de una introducción, tres capítulos, conclusiones, recomendaciones así como las referencias bibliográficas, bibliografía, anexos y un glosario de términos, donde se expone y da cumplimiento de forma progresiva a la totalidad de los objetivos planteados en el trabajo:

Capítulo #1: "Fundamentación teórica". El objetivo de este capítulo es abordar conceptos generales y básicos que permiten comprender temas relacionados con la digitalización de documentos. Además se exponen las herramientas, algoritmos, tecnologías y metodología imprescindible para la realización del diseño de la propuesta de solución.

Capítulo #2: "Planificación y diseño del módulo de digitalización". Este capítulo tiene como principal objetivo planificar y diseñar la solución propuesta haciendo uso de la metodología ágil SXP. Para ello se especifican los principales artefactos generados en las primeras fases; se definen las actividades de las cuales se obtienen los artefactos relacionados con la concepción inicial del sistema, los requisitos, el diseño y las tareas a realizar.

Capítulo #3: "Validación de la propuesta de solución". El objetivo de este capítulo es dar validez y vericidad a la propuesta de solución. Para ello se formularán y aplicarán los casos de pruebas de aceptación.

Capítulo 1

Fundamentación Teórica

En el presente capítulo se detallan los resultados obtenidos de la investigación realizada acerca del proceso de digitalización de documentos y las soluciones de software existentes en la actualidad que responden a este proceso. Además se precisan los basamentos matemáticos que rigen el procesamiento digital de imágenes, como parte de los sistemas de digitalización y el reconocimiento de patrones dentro del proceso de digitalización. Por otro lado se seleccionan las herramientas, algoritmos, tecnologías y metodología necesaria para el correcto desarrollo de la propuesta de solución.

1.1. Principales razones para digitalizar.

Hoy en día las tecnologías de la información y las comunicaciones ocupan un lugar destacado con el fin de brindar servicios eficientes y de elevada calidad. Uno de ellos es la digitalización de documentos, el cual es un proceso vital para el tratamiento archivístico. En la misma medida en que se desarrollan no solo las herramientas de software para la digitalización, sino, también los dispositivos ópticos asociados (escáneres, robots dedicados, etc) y los dispositivos de almacenamiento, se favorece la recuperación de espacio libre en archivos. Además se facilita la consulta e integración de los documentos a las redes informáticas, los gestores documentales y los flujos de trabajo a fin de alargar su vida útil.

1.1.1. Ventajas de la digitalización de documentos y el trabajo con documentos digitales.

Entre las principales ventajas de la digitalización de los documentos y la manipulación posterior de los mismos en este formato podemos mencionar:

- Eliminación de gastos relacionados con la impresión (tinta, papel, etc.).
- Ahorro de tiempo (no hay que esperar hasta que se impriman los documentos y las búsquedas de información son mucho más efectivas, fáciles y rápidas).
- Beneficios para la ecología (ayuda a evitar la alteración voluntaria del medio ambiente, contribuyendo a depurarlo y mantenerlo sano).
- Ahorro de espacio físico (los documentos digitalizados disminuyen el consumo de espacio y a su vez los costes que esto implica).
- Facilidad de acceso (no es necesario revisar caja tras caja de papeles, todos los documentos se encuentran a un clic de distancia).
- Rapidez (lo que llevaba 10 minutos realizar ahora puede realizarse en tan solo segundos).
- Reducción de costos (una vez digitalizada la información, el envío y la consulta de la misma no tiene costo adicional).
- Facilidad de distribución (puede enviarse la información que se desee a través del correo electrónico o mediante la utilización de las redes. Se elimina la necesidad de distribuir múltiples copias de un mismo documento).
- Conservación de la información (la calidad del documento no se deteriora con el paso del tiempo).
- Seguridad (los documentos almacenados en formato digital son mucho más seguros que en formato duro. Se pueden definir las autorizaciones para acceder a los distintos tipos de documentación).

- Conexión (se pueden compartir datos con varias personas que se encuentran en lugares diferentes en tiempo real. En caso de mudanza de oficina sólo se desplazan unos cuantos discos en lugar de varias cajas de papeles).
- Calidad (agiliza los procesos creando un flujo de trabajo más veloz).

1.2. Actualidad de las soluciones de digitalización.

A nivel internacional existen diferentes aplicaciones que permiten la digitalización de documentos. Cada una con sus características y funciones específicas, las cuales están enfocadas mayormente a las necesidades de empresas e instituciones que trabajan con un gran volumen de información. A continuación se refleja una breve descripción de algunas de estas herramientas tanto en el ámbito nacional como el internacional.

1.2.1. Soluciones nacionales.

En agosto del año 2001, el Consejo de Estado resolvió dictar leyes que establecen las normas y principios que rigen la actividad archivística en el territorio nacional, debido a que esta actividad en nuestro país fue poco valorada durante mucho tiempo. A partir de entonces comenzaron a crearse y extenderse las soluciones elaboradas por equipos de desarrollo de software nacionales, destacándose entre ellos la Empresa Nacional de Software, DESOFT SA[2], la cual incluye dentro de sus productos el ejemplo que se cita a continuación:

Papiro

Papiro es un sistema que permite la conservación, digitalización, gestión y socialización de información documental de los archivos en Cuba. Este software fue implementado por personal de la empresa DESOFT de la provincia Granma. Es un producto informático de uso libre que emplea

[2]DESOFT SA: Empresa Cubana de Desarrollo de Software.

herramientas igualmente libres y permite conservar documentación de valor histórico al evitar su manipulación directa. Está concebido para correr bajo la plataforma Microsoft Windows en las versiones 2000 y XP.

El sistema es utilizado en el Archivo Histórico de la Ciudad de Manzanillo, y si bien fue diseñado para su empleo en repositorios históricos, también puede aplicarse en archivos de gestión, resultando al mismo tiempo un intento pionero - por lo menos en Cuba -, al vincular la gestión de bases de datos con las imágenes de documentos originales digitalizados[2].

Con el estudio realizado de las soluciones de digitalización existentes en nuestro país, se evidencia la ausencia de aplicaciones capaces de suplir las necesidades referentes al tema de la digitalización de documentos, como parte inherente a la incorporación de activos documentales a un sistema de gestión documental digital. La solución de digitalización que propone la empresa DESOFT, resuelve parcialmente los problemas de incorporación de documentos a sistemas digitales de gestión documental, pues no realiza un proceso íntegro de digitalización de documentos, omitiendo la obtención de la información contenida en los documentos digitalizados; además es dependiente del sistema operativo Windows y no posee integración con gestores de contenidos empresariales.

1.2.2. Soluciones en el ámbito internacional.

El desarrollo del software se ha convertido en un poderoso instrumento económico utilizado por instituciones y compañías, las cuales aparejadas al avance tecnológico han creado aplicaciones capaces de brindar servicios con gran calidad. A continuación se hace una breve descripción de algunos ejemplos en el ámbito internacional, donde se exponen algunas de las soluciones de digitalización que más se emplean hoy en día.

Xsane

Xsane es una aplicación liberada bajo licencia GNU/GPL que no posee una interfaz amigable, pero es configurable y está basada en GTK para SANE. Xsane se puede ejecutar como un programa independiente o mediante el programa de manipulación de imágenes GIMP, ya que puede funcionar como plugin del mismo. En modo independiente, puede guardar una imagen a un archivo en una variedad de formatos de imagen, servir como interfaz a un programa de fax, o enviar una imagen a una impresora[3][4].

Simple Scan

Simple Scan se encuentra incluida en la distribución de Ubuntu (10.4) Lucid Lynx. Esta aplicación es libre y está basada en SANE pero utiliza una interfaz gráfica integrada en el escritorio GNOME. Permite aplicar post-proceso simple como recortado o rotación, previsualización de la imagen a digitalizar, pero deja la tarea de aplicar un proceso más complejo a aplicaciones más avanzadas como Gimp e Inkscape. Permite además digitalizar múltiples documentos y agregarlos como una serie imágenes JPG o exportarlos a un documento PDF de igual forma. Facilita grabar en formatos estándares como PNG, PDF, PS y JPEG, y se puede invocar desde otras aplicaciones como Evolution y OpenOffice. Todos los escáners soportados actualmente por SANE funcionan sin problemas con esta aplicación[5][6][7].

Kooka

Kooka es una aplicación con licencia GPL, que ayuda a manejar los más importantes parámetros de escaneo, encuentra el formato correcto de imagen para salvar y manejar las imágenes escaneadas. Además ofrece soporte para diferentes módulos OCR. Es la aplicación de exploración de elección para el proyecto KDE y por lo tanto es parte oficial del paquete de gráficos del mismo. Su instalación es bastante simple, funciona además en GNOME[8][9].

Kofax Capture

Kofax Capture es una aplicación propietaria que ofrece herramientas tales como la indexación automática, identificación de forma automática, precisa de zona de OCR e ICR. Ofrece perfecta integración con gestores de contenido como Documentum, IBM Content Manager, Alfresco, entre otros más. Kofax Ascent Capture forma una única plataforma para el escaneo de documentos e imágenes, que acelera los procesos de negocio, da la posibilidad de interactuar con otras aplicaciones y transferir la información capturada hacia estas e incluye la clasificación automática de documentos, extracción de datos y validación. Kofax es el proveedor líder mundial de soluciones de captura de información[10][11].

Abbyy FineReader

Abbyy FineReader es para muchos, el mejor reconocedor de texto del mercado. Es considerado una aplicación propietaria sencilla, potente y segura con la característica de solo poderse ejecutar en sistemas Windows. Es muy fácil de usar; dispone de una interfaz intuitiva y de diseño sencillo, con un editor de textos integrado, soporte para guardar imágenes, utilidad de búsqueda. Compatible con todos los escáneres. Posee un motor OCR de calidad insuperable y corrector ortográfico; soporte para 186 idiomas. Soporta los formatos (DOC, DOCX, PDF, RTF, TXT, XLS, XSLX, HTML, CSV). El programa hace gala de un alto nivel de precisión en el reconocimiento de caracteres y retención de formato, los documentos digitalizados pueden ser: editados, enviados por e-mail y guardados en formato (doc, pdf, xml, html, etc.). Es una herramienta muy potente que detecta elementos de maquetación, imágenes, tablas e incluso el idioma en el que está escrito el documento. Abbyy FineReader es considerado uno de los mejores de su tipo por sus funcionalidades, facilidad de uso y consumo de recursos[12][13].

Además de las herramientas anteriormente citadas, existen algunas, que a pesar de no poseer novedades de gran impacto con respecto a las expuestas no se deben dejar de mencionar. Tal es el caso de las aplicaciones QuickScan Pro de Captiva[14], Documalis Free Scanner[15] y ADD scan[16].

Estos sistemas no pueden ser tomados como referencias para dar una propuesta de solución de software en aras de la digitalización de documentos, pues aunque contienen herramientas y facilidades de alto nivel para humanizar el trabajo, presentan la limitación de ser en su mayoría privativas y las que no muestran esta limitante, no tienen la capacidad de integración con entornos de gestión documental.

1.3. El procesamiento digital de imágenes como parte de los sistemas de digitalización.

El campo del procesamiento digital de imágenes (PDI) se refiere a procesar imágenes del mundo real de manera digital por medio del uso de computadoras aplicando un conjunto de técnicas, con el objetivo de mejorar la calidad o facilitar la búsqueda de información. El interés del procesamiento digital de imágenes se centra en dos áreas fundamentalmente:

- Mejoramiento de la información pictórica para la interpretación humana.
- El procesamiento de datos de la imagen para su almacenamiento, transmisión y representación para la percepción autónoma de máquinas.

1.3.1. Tipos de procesamientos computarizados.

El procesamiento de las imágenes digitales se divide para su estudio en tres niveles fundamentales[17]:

Procesos de Nivel Bajo: Se refiere a las acciones asociadas en lo fundamental al mejoramiento de las imágenes digitales; tales como, reducción de ruido, realce de contraste así como al realce de las características de la imagen, donde las entradas/salidas siempre son imágenes.

Procesos de Nivel Medio: En este grupo se encuentran las técnicas de segmentación (regiones, objetos), descripción de objetos y clasificación donde la entrada es una imagen y la salida son atributos de objetos (bordes, contornos, identidades de objetos individuales, etc).

Procesos de Nivel Alto: A este grupo pertenecen aquellas técnicas que buscan "darle sentido" al conjunto de objetos encontrados; es decir, análisis de la imagen, llevando a cabo funciones cognitivas normalmente asociadas a la visión.

En el marco del presente trabajo se estará haciendo uso de un grupo de técnicas asociadas al procesamiento digital de imágenes, que forman parte en lo fundamental de los proceso de nivel medio y alto, aunque también se trabaja sobre el nivel bajo, recayendo el mayor peso en los niveles antes mencionados. Los mismos son ampliamente tratados en la bibliografía especializada por lo que en los epígrafes venideros se harán cortas referencias a los mismos dejando al lector que amplíe los detalles en las citas correspondientes.

1.3.2. Mejoramiento de imágenes.

El desarrollo alcanzado por las microcomputadoras ha posibilitado un incremento del uso de las imágenes en diversos ámbitos (dígase la medicina, biología, astronomía, historia, geología, criminalística y fotografía propiamente dicha por solo citar algunos de una larga lista)[18], dada la variedad y cantidad de información que las mismas poseen.

La mejora de la imagen es una de las zonas más interesantes y visualmente atractiva del procesamiento de imágenes, constituyendo un proceso casi obligatorio que precede a la inmensa mayoría de las acciones de extracción, análisis y despliegue de información visual. Es importante señalar que el propio proceso de mejora no aumenta la información inherente contenida en los datos sino simplemente ayuda a eliminar las características no deseadas en las imágenes, dándole énfasis a ciertas características especificadas de las mismas. El principal objetivo de este tipo de técnica es procesar una imagen dada, de forma que la imagen resultante sea más apropiada que la imagen original para aplicaciones específicas.

El presente trabajo se enfoca fundamentalmente en las técnicas de mejora mediante procesamiento puntual, debido a que estas operan directamente sobre los píxeles, realizando operaciones de memoria

cero[19][20], ya que no tienen en cuenta información local para ello. En el marco de la propuesta de digitalización se propone emplear una técnica de mejora de brillo y contraste. Con el empleo de la misma se podrá acentuar la intensidad relativa de los elementos de la imagen.

1.3.3. Transformaciones espaciales de las imágenes.

Toda imagen en proceso de digitalización, transita por ciertas fases de vida, ligado a cada etapa existen pasos que se toman, por lo cual la misma se pre-procesa hasta llegar a su destino final. Los documentos cuando se digitalizan tienden a presentar incongruencias con respecto a su versión en formato duro. Dígase que la misma puede presentar inclinaciones en el dorso o inclinaciones en su totalidad. Todo esto conlleva al trabajo de realizar procesos y/o transformadas para detectar estas incongruencias, tratar las mismas de forma segura y con la mayor eficiencia posible, para que dicha imagen no pierda fiabilidad y originalidad.

Entre las principales transformadas se encuentra las que se encargan de identificar patrones de formas en las imágenes. Dichos patrones se emplean para realizar cambios en las imágenes a conveniencia según el criterio que se siga. Para el caso que compete a este trabajo de diploma, se trata sobre tomar la imagen y procesarla, de tal manera que si la imagen proviene con errores de escaneo, dígase que la imagen presenta cierto ángulo de inclinación, estas transformadas nos permiten corregir la misma empleando ciertos artificios matemáticos.

Transformada de Hough

La transformada de Hough es un algoritmo empleado para el reconocimiento de patrones en imágenes enfocándose en lo fundamental hacia el reconocimiento de formas dentro de una imagen (líneas, círculos, etc). El objetivo central de este algoritmo es encontrar puntos alineados que puedan existir en la imagen, es decir, puntos en la imagen que satisfagan la ecuación de la recta, para distintos valores de ρ y θ. Su modo de operación es principalmente estadístico, y consiste en conocer si un punto es parte de una

línea. Para ello se aplica una operación dentro de cierto rango, que permite hallar las posibles líneas de las que puede ser parte el punto. Esta acción se aplica a todas los puntos en la imagen, determinándose al final cuales líneas fueron las que más puntos posibles tuvieron y esas son las líneas en la imagen.

Transformada de Radon

La transformada de Radon constituye otra de las alternativas para representar una imagen en un espacio de parámetros de manera similar a como lo hace la transformada de Hough. De modo particular, esta transformada calcula las integrales de línea de múltiples fuentes a lo largo de trayectorias paralelas o "vigas" en una dirección determinada. Las vigas se encuentran separadas a una unidad de píxel y para representar una imagen se toman múltiples proyecciones paralelas de la imagen desde diferentes ángulos de rotación de la fuente tomando como referencia el centro de la imagen(ver Figura 1.1).

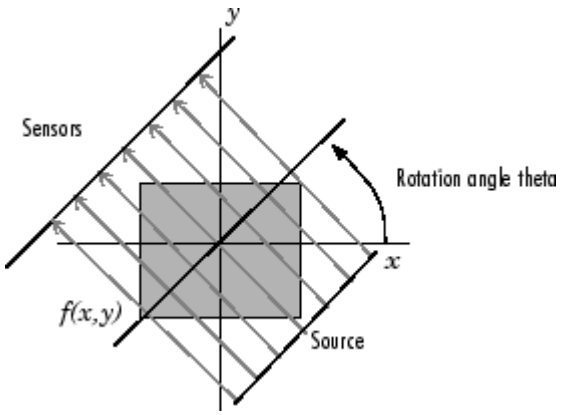

Figura 1.1: Geometría de la Transformada de Radon.

La aplicación de la transformada de Radon en una imagen $f(x, y)$ para un conjunto de ángulos se puede considerar como la computación de la proyección de la imagen a lo largo de los ángulos dados. La proyección resultante es la suma de las intensidades de los píxeles en cada dirección, es decir, una integral de línea. El resultado de lo anterior es una nueva imagen de $R(\rho, \theta)$.

Importante resulta destacar que para poder aplicar el algoritmo de la transformada de Radon a una imagen esta debe ser sometida previamente a un proceso de conversión a escala de grises; es decir, para que el algoritmo pueda procesar esta imagen requiere de una imagen escalada en grises y no en colores. Por su parte se puede hoy encontrar diferentes implementaciones de la transformada de Radon; la primera de ellas se basa en construir una función de los parámetros θ y ρ para cada píxel calculado mientras que la segunda variante puede expresar la imagen original como una función de los píxeles de la misma (i.e. de la forma clásica). Esta última es la más fácil de implementar y por sus características muchos autores la consideran como una derivación de la transformada de Hough.

Una vez culminado el análisis de los algoritmos de las transformadas de Hough y Radon nos encontramos en condiciones de concluir que el algoritmo más factible a utilizar en el contexto del presente trabajo es la transformada de Radon. Esta decisión está sustentada en el hecho de que ambos algoritmos pueden ser empleados para alcanzar el mismo objetivo sin embargo, Radon resuelve el problema de una manera más directa y sencilla quedando por su parte la transformada de Hough reservada para al reconocimiento de formas y patrones en imágenes.

1.4. El reconocimiento de patrones dentro del proceso de digitalización.

"El reconocimiento de patrones es una disciplina científica cuya meta es la clasificación de objetos en diversas categorías o clases. Dependiendo de la aplicación, esos objetos pueden ser imágenes, señales de transmisión o cualquier tipo de mediciones que necesiten ser clasificadas"[21].

El reconocimiento de patrones tiene una larga historia, pero no es hasta la década de 1960 en la cual realza su verdadero objetivo y aplicación. Con la llegada de las nuevas tecnologías de manejo de información, el reconocimiento de patrones afronta la realidad y es la tecnología que hasta hoy en día perdura como predilecta.

Esta tecnología se ha desempeñado con gran impacto en las siguientes temáticas: visión artificial, diagnóstico asistido por ordenador, reconocimiento por voz, minería de datos, descubrimiento de conocimiento, y en especial también se ha dedicado al reconocimiento óptico de caracteres, donde ha presentado mayor implicación en la automatización del proceso y el tratamiento de imágenes. Es en este tema final donde se centra este epígrafe realizádose una explicación del proceso.

1.4.1. Segmentación de imágenes.

Desde el punto de vista clásico la segmentación de imágenes se define como la partición de una imagen en regiones[3] constituyentes no solapadas, las cuales son homogéneas con respecto a alguna característica como intensidad o textura[22].

La segmentación de imágenes se emplea en muchas aplicaciones, incluyendo la detección de objetos, la codificación basada en objetos, el seguimiento de objetos, la recuperación de la imagen, etc y para llevarla a cabo existe un sin número de métodos y técnicas, las cuales pueden ser clasificadas en técnicas basadas en regiones y técnicas basadas en el conocimiento[23]. Las técnicas de segmentación basadas en regiones de modo particular, se centran en la agrupación de píxeles para convertirlos en regiones con propiedades uniformes mientras que las basadas en el conocimiento hacen hincapié en la detección de cambios significativos en el nivel de gris cerca de los límites de los objetos. Es de suma importancia conocer a fondo cada técnica dado que de esto depende su elección para la aplicación específica.

Las ventajas asociadas a la segmentación basada en regiones radican en que mediante el resultado de la segmentación se pueden obtener regiones coherentes, vinculando los bordes; sin embargo, posee el inconveniente que las decisiones sobre los miembros de una región son a menudo más difíciles que los que van sobre el borde de detección.

[3]Región: una región, en una imagen, es un grupo de píxeles conectados que tienen propiedades similares.

Los mecanismos de segmentación asociados a estos métodos pueden clasificarse ya sea como segmentación supervisada o segmentación no supervisada. En la literatura especializada aparecen varios métodos de segmentación basados en regiones y que a su vez tienen asociado un mecanismo de segmentación no supervisada; entre ellos se destacan:

- Región Estadística Semisupervisada de Refininamiento (SSRR) desarrollado por Nock y Nielsen[24].
- Descubrimiento no supervisado de Objetos en Video (DISCOV) desarrollado por Liu y Chen[25].
- Método de las Líneas Divisorias o Cuencas[26][27][28].
- Árbol de Particiones Binario (BPT)[29][30].

La segmentación basada en el conocimiento por su parte, puede simplificar el análisis por minimizar drásticamente la cantidad de píxeles de una imagen a ser procesada, preservando las estructuras adecuadas del objeto. Sin embargo, su inconveniente radica en que el ruido puede dar lugar a un borde erróneo. En la literatura también aparecen varios ejemplos de este tipo de método, entre ellos:

- Detector de bordes Canny[31].
- The Live Wire On the Fly (LWOF) propuesto por Falco[32].
- Snake o esquema de contornos activos presentado por primera vez por Kass[33].
- Método de segmentación multiescala[34].
- Extensión del Gradiente del Vector de Flujo (E-GVF), que no es más que una mejora de los contornos activos[35].
- Aquellos basados en morfología matemática[36][37].
- Método basado en el seguimiento de contornos, que puede ser aplicado a cualquier tipo de imágen[38][39].

Este último es el propuesto a ser empleado como parte del futuro sistema de digitalización, debido a que es un método de segmentación robusto conveniente para cualquier tipo de imagen no específica. Las ventajas de la adopción de este método son su simplicidad en cálculos y la reducción en el tiempo de búsqueda de los puntos de umbralización. Esta propuesta se sustenta en el análisis computacional que realizó la Universidad de California en Berkeley, de los métodos LWOF, E-GVF, Snake, Watershed y el propuesto. Los resultados de la simulación demostraron que el método seleccionado es superior a los métodos convencionales, ya que es más robusto y adecuado para diversas aplicaciones de imagen que el resto de los métodos antes mencionados.

Aunque el método del que se ha venido hablando hasta el momento, es el ideal para darle solución al problema de la segmentación de imágenes, este no pudo ser implementado por razones del tiempo con el cual se dispone para realizar una tesis de pregrado y la novedad del mismo. Por tal motivo se hizo la implementación del método clásico de Otsu[40], el mismo es muy reconocido, utilizado y probado a nivel internacional. A pesar de ser menos eficaz que el método propuesto, este permite exponer claramente el objetivo de nuestra investigación.

1.4.2. Redes neuronales artificiales (RNA).

Mediante el estudio y comparación realizada sobre las Redes Neuronales y la técnica Matching Template[41] para el reconocimiento de caracteres. Se evidencia, que lo más factible a utilizar en el marco del presente trabajo es el uso de una red neuronal. Los resultados obtenidos[41] en una demostración realizada evidencian que con una red neuronal se puede lograr un reconocimiento preciso y eficaz, requiriendo menos memoria y con una mejora considerable luego de haber sido entrenada dicha red. Por otra parte el software Matlab, herramienta que fue seleccionada para desarrollar el prototipo de la solución propuesta, permite manejar redes neuronales de manera sencilla, proporcionado herramientas fáciles de usar en este sentido, como la toolbox de redes neuronales.

En el caso que nos ocupa, las redes neuronales artificiales[42], son una muy buena alternativa para la etapa de reconocimiento de caracteres, teniendo en cuenta lo potente que resultan a la hora de lidiar con problemas que involucran reconocimiento de patrones.

Luego de un estudio realizado sobre las diferentes tipologías de RNA, se seleccionó una red Perceptrón Multicapas también conocida como MLP(MultiLayer Perceptron) para realizar el reconocimiento de caracteres en la solución propuesta. Fundamentado principalmente en la habilidad que tienen estas redes para la solución de problemas de clasificación de patrones[43] pues son capaces de aprender cualquier función o relación continua entre un grupo de variables, posibilitándole actuar como aproximador universal de funciones. Propiedad que convierte a este tipo de red en una herramienta de propósito general, flexible y no lineal; mostrando un rendimiento superior respecto a los modelos estadísticos clásicos[44].

Esta red utiliza aprendizaje supervisado, por lo que se le presentó los patrones de entrada y las salidas esperadas para cada uno de ellos. Aprovechando su topología no recurrente (en la que todas las señales van desde la capa de entrada hacia la salida, es decir, todas sus conexiones son hacia delante) fue seleccionado el algoritmo de entrenamiento de retropropagación del error o backpropagation. Debido a que es el método de entrenamiento más utilizado en RNA con conexión hacia delante[45] y tal es el caso de la red neuronal seleccionada MLP. La función de activación o transferencia de la RNA propuesta es la función sigmoide[4], pues cumple con la condición de ser derivable, requisito del algoritmo de entrenamiento utilizado.

1.4.3. Reconocimiento óptico de caracteres (OCR).

El reconocimiento óptico de caracteres, generalmente abreviado a OCR, consiste en la traducción mecánica o electrónica de imágenes escaneadas, texto escrito a mano, mecanografiada o impresa en texto

[4]Función sigmoide: función de transferencia o activación que describe una progresión temporal desde niveles bajos al inicio, hasta acercarse a un climax transcurrido en un cierto tiempo. Su gráfica tiene forma de “S”.

a máquina codificada. Los primeros trabajos sobre el reconocimiento óptico de caracteres se remontan al año 1950, cuando David Shepard y Louis Tordella comenzaron la investigación del procedimiento para la automatización de datos de la entonces Agencia de Seguridad de las Fuerzas Armadas (AFSA) de los Estados Unidos, la misma que dos años después se convertiría en la Agencia de Seguridad Nacional (NSA)[46].

El reconocimiento óptico de caracteres es ampliamente utilizado para convertir documentos en archivos electrónicos, informatizar sistemas de registro, o publicar el texto en una página web. El OCR permite además editar el texto, buscar una palabra o frase, almacenar de manera más compacta, mostrar en pantalla o imprimir una copia sin hacer uso de los artefactos de escaneo, aplicar técnicas como la traducción automática de texto a voz y minería de texto. El OCR brinda también un amplio campo de investigación en esferas como el reconocimiento de patrones, la inteligencia artificial y la visión por computador.

1.4.4. OCR y la gestión documental.

La gestión documental en combinación con la tecnología OCR ofrece múltiples ventajas, las que se traducen en: el control de todo el ciclo de vida de un documento, el aumento de la rapidez en sus transacciones, contracción de los tiempos en la localización y recuperación de los documentos, la disminución del espacio físico de almacenamiento de la información y cierto incremento en la seguridad de la información.

Gracias al surgimiento del OCR, la gestión documental hoy en día, se encuentra en un nivel superior. En épocas de antaño, la gestión de documentos se encontraba limitada por la cantidad de documentos a tramitar en formato duro. La poca o bien difícil localización de los archivos por metadatos era prácticamente un caos. Esta tecnología ha permitido gestionar con mucha más calidad los documentos y realizar mejores flujos de trabajos con los mismos.

El OCR se distingue por ser un proceso viable y confiable en el proceso de la gestión de documentos. Toda empresa o institución, que gestione grandes cantidades de volúmenes de documentos, debe hacer uso de esta tecnología de punta, en aras de ofrecer un servicio sencillo, simple y transparente.

1.4.5. Extracción de resumen de texto.

La generación automática de resúmenes juega un papel de suma importancia. El propósito de los resúmenes es facilitar el procesamiento de información optimizando el tiempo de lectura necesario para localizar la información requerida. La descripción compacta del contenido relevante de un documento, permite el incremento de la eficiencia en el procesamiento, clasificación y recuperación del material textual.

En la actualidad son muchas las aplicaciones comerciales disponibles relacionadas con la generación de resúmenes, lo que da una idea, tanto del interés, como de la necesidad existente de este tipo de herramientas. Estas poseen un propósito general y por tanto son aplicables a cualquier clase de documento. En Cuba de modo particular, los sistemas de extracción automática de resúmenes aún constituyen proyectos en estado de gestación. Las investigaciones en este campo no son abundantes y las necesidades actuales se cubren con sistemas privativos, que no siempre de ajustan a las necesidades reales.

A pesar de las limitaciones antes mencionadas, es importante destacar el papel de la extracción de resúmenes de textos como técnica asociada a los procesos que se llevan a cabo, hasta la obtención de la información contenida en los documentos digitalizados. Mediante el empleo de la misma se podrá mejorar el proceso de recuperación de información desde el ECM Alfresco. Posibilitando que la solución de software propuesta disponga de un proceso íntegro de digitalización de documentos.

Para el desarrollo del prototipo, se seleccionó como herramienta para aplicar la extracción de resumen de texto, la librería **libots0 v0.5.0**, pues es una tecnología libre, de fácil uso. La misma está disponible

en varias distribuciones GNU/Linux, supliendo las necesidades que se presentan hasta el momento para realizar la extracción de resumen de texto.

1.5. Tecnologías, herramientas y metodología utilizada.

Lenguaje unificado de modelado (UML)

Es uno de los lenguajes más reconocidos y utilizados en la actualidad para la modelación de software. Es un lenguaje gráfico para visualizar, especificar, construir y documentar un sistema. Brinda un estándar para representar un "plano" del sistema, agregando aspectos conceptuales tales como: procesos de negocios, funciones del sistema, expresiones de lenguajes de programación, esquemas de bases de datos y componentes de software reutilizables.

Se puede aplicar de varias formas para sobrellevar una metodología de desarrollo, aunque no define cual usar o que proceso emplear. UML tiene diversos tipos de diagramas, los cuales expresan disímiles aspectos de las entidades personificadas:

- Diagramas de clases para personalizar la estructura estática de las clases en el sistema.
- Diagramas de objetos para simbolizar la estructura estática de los objetos en el negocio.
- Diagramas de casos de uso para representar los procesos del negocio.
- Diagramas de actividad para modelar el comportamiento de los casos de uso, objetos u operaciones.
- Diagramas de secuencia para escenificar el paso de mensajes entre objetos.
- Diagramas de colaboración para modelar interacciones entre objetos.
- Diagramas de estado para simbolizar el comportamiento de los objetos en el sistema.

- Diagramas de componentes para modelar componentes.
- Diagramas de implementación para personificar la distribución del sistema.

Lenguaje M o MATLAB

El software de cálculo científico MATLAB se presenta indisolublemente ligado a su lenguaje de trabajo comúnmente conocido como lenguaje M o lenguaje Matlab. Es un lenguaje de computación técnica de alto nivel para el trabajo esencial con matrices, sentencias para control de flujo, creación de funciones y estructuras de datos, funciones de entrada/salida, análisis de datos y cálculo numérico. MATLAB proporciona de forma adicional una serie de funciones para documentar y compartir su trabajo además de poder integrar su código con otros lenguajes como C++, Java y .Net.

Ventajas del lenguaje MATLAB:

- Permite un rápido desarrollo y ejecución de aplicaciones.
- Desarrollado esencialmente para el trabajo con matrices lo cual es fundamental en el universo del cálculo científico.
- El lenguaje posibilita la programación orientada a objetos (POO), el control de flujo, la depuración e implementa diversas estructuras de datos.

Se escoge el lenguaje M como lenguaje de programación (durante el prototipado) en primer lugar porque es el que brinda Matlab; no obstante, se aclara que no es razón obligada sino que formó parte de las variables de decisión durante el proceso de selección de la herramienta para el desarrollo del prototipo que acompaña a este trabajo. En adición a esto, no es de despreciar el interés que despierta su empleo desde la perspectiva de las facilidades que brinda, como ser un lenguaje fácil de aprender, entendible y legible a la hora de interpretarlo por solo citar algunas de una lista un tanto más extensa.

Librería Libots0 v0.5.0

Es de código abierto bajo licencia GPL. Se puede emplear en varias distribuciones de GNU/Linux, como: Ubuntu, Fedora, Gentoo, Debian, Mandrake y RedHat. La misma está disponible en los repositorios de estas distribuciones. Su funcionamiento es simple y de fácil uso. Esta lee un texto y decide que oraciones son importantes y cuales no, luego crea un corto resumen donde se destaca la idea principal en el texto (la salida puede ser HTML o texto plano), trabaja con codificación UTF-8. Resume textos en inglés, alemán, español, ruso, hebreo, esperanto y otros.

Se escoge esta librería para la extracción de resumen de texto, pues es una tecnología libre, secilla y de fácil uso. Brinda soporte para realizar los resúmenes en diversos idiomas, es la herramienta más empleada en cuanto al tema en la extracción de resumen en los sistenas GNU/Linux.

Visual Paradigm

Es una herramienta profesional, es decir, un software de modelado que utiliza UML como lenguaje de modelado. Soporta el ciclo completo de vida del software (análisis y diseño orientados a objetos, implementación, pruebas y despliegue); permitiendo en cada una de sus etapas generar los diagramas necesarios sin ningún tipo de problema. Está orientado para posibilitar tanto ingeniería directa como inversa, pues posee varios lenguajes de programación que aprueban la generación de código.

Esta herramienta CASE[5] soporta la importación y exportación de varias versiones de XML. Facilita la modelación de diversos tipos de diagramas, transformando códigos de estos modelos, concibiendo de esta manera, los códigos fuentes de los diagramas.

[5]Constituyen un cúmulo de programas y ayudas, que suministran apoyo a los ingenieros de software y desarrolladores, durante el período de desarrollo.

Algunas características de Visual Paradigm:

- Posee generación de código para Java y la exportación de todos los diagramas a formato HTML y JPG.
- Ostenta de un medio de creación de diagramas para UML 2.0.
- Posibilita la integración a los principales IDE.
- Cuenta con un diseño enmarcado en casos de uso y dirigido al negocio.
- Contiene facilidades para representar especificaciones de casos de uso del sistema.

Visual Paradigm se empleará como herramienta para el modelado de diagramas debido a que es multiplataforma. Posee una distribución automática de diagramas, contando con una reorganización de las figuras y conectores de los diagramas UML. Permite además exportar los diagramas a imágenes y páginas HTML.

Software MATLAB

MATLAB es un entorno de computación y desarrollo de aplicaciones totalmente integrado; orientado para llevar a cabo proyectos en donde se encuentren implicados elevados cálculos matemáticos y la visualización gráfica de los mismos. MATLAB integra el análisis numérico, cálculo matricial, cálculo simbólico, trabajo con cadenas y visualización gráfica en un ambiente amigable, donde los problemas y sus soluciones son expresados del mismo modo en que se escribirían tradicionalmente, sin necesidad de hacer uso de la programación tradicional.

MATLAB dispone también en la actualidad de una amplia variedad de programas especializados de apoyo, denominados Toolboxes, que extienden significativamente el número de funciones incorporadas en el programa principal. Estos Toolboxes cubren en la actualidad prácticamente casi todas las áreas

principales de la ingeniería y la simulación, destacando entre ellos los de Procesamiento de Imágenes y Redes Neuronales por ser los más significativos en el marco de este trabajo. Matlab es un entorno de cálculo técnico que se ha convertido en estándar de la industria, con capacidades no superadas hasta el momento en computación numérica y visualización.

Ventajas de MATLAB:

- Cálculos intensivos desde el punto de vista numérico.
- Gráficos y visualización avanzada.
- Lenguaje de alto nivel basado en vectores, arrays y matrices.
- Colección muy útil de funciones de aplicación.

Se escoge Matlab como herramienta de desarrollo toda vez que es líder mundial para realizar implementaciones en las temáticas antes mencionadas y que compete al presente trabajo de diploma. Es importante destacar que el papel esencial de esta herramienta dentro del trabajo que se describe es el de "prototipar"; es decir, construir un prototipo funcional de la solución (no desde el punto de vista de la ingeniería del software en particular sino de la ingeniería en general)(ver Anexo 1) que muestre la posibilidad y factibilidad de desarrollar la solución de software que se propone para mejorar los procesos de recuperación de información desde el ECM Alfresco.

Xtreme Programming (XP)

XP es una metodología ágil centrada en potenciar las relaciones interpersonales como clave para el éxito en el desarrollo de software, ya que promueve el trabajo en equipo, se preocupa por el aprendizaje de los desarrolladores y propicia un buen clima de trabajo. Se basa en realimentación continua entre el cliente y el equipo de desarrollo, comunicación fluida entre todos los participantes, simplicidad en las soluciones implementadas y coraje para enfrentar los cambios.

Esta metodología se define como especialmente adecuada para proyectos con requisitos imprecisos y muy cambiantes, y donde existe un alto riesgo técnico. Para la especificación de las funcionalidades del sistema se utilizan tarjetas de papel (historias de usuarios) en las cuales el cliente describe brevemente las características que el sistema debe poseer, sean requisitos funcionales o no funcionales. El tratamiento de las historias de usuario es muy dinámico y flexible. Cada historia de usuario es lo suficientemente comprensible y delimitada para que los programadores puedan implementarla a lo sumo en varias semanas.

SCRUM

SCRUM, más que una metodología de desarrollo de software, es una forma de autogestión de los equipos de trabajo de forma general. Un grupo de integrantes del equipo de trabajo decide cómo hacer sus tareas y cuánto van a tardar en ello. Ayuda a que trabajen todos juntos, en la misma dirección, con un objetivo claro. Permite además seguir de forma clara el avance de las tareas a realizar, de forma que los jefes puedan ver día a día como progresa el trabajo. Sin embargo, no es una metodología de desarrollo, puesto que no indica qué se debe hacer para realizar el código. Debería, por tanto, complementarse con alguna otra metodología de desarrollo.

SXP

SXP es una metodología compuesta por la metodología XP y SCRUM que es un método adaptativo de gestión de proyectos que se basa en los principios ágiles. SXP ofrece una estrategia tecnológica, a partir de la introducción de procedimientos ágiles que permitan actualizar los procesos de software para el mejoramiento de la actividad productiva fomentando el desarrollo de la creatividad, aumentando el nivel de preocupación y responsabilidad de los miembros del equipo, ayudando al líder del proyecto a tener un mejor control del mismo.

SXP consta de cuatro fases principales:

1. ***Planificación-Definición:*** se establece la visión, se fijan las expectativas y se realiza el

aseguramiento del financiamiento del proyecto.

2. ***Desarrollo:*** se realiza la implementación del sistema hasta que esté listo para ser entregado.
3. ***Entrega:*** se entrega el software y la documentación correspondiente al mismo.
4. ***Mantenimiento:*** se realiza el soporte para el cliente.

SXP está especialmente indicada para proyectos de pequeños equipos de trabajo, rápido cambio de requisitos o requisitos imprecisos, muy cambiantes, donde existe un alto riesgo técnico y se orienta a una entrega rápida de resultados y una alta flexibilidad. Ayuda a que trabajen todos juntos, en la misma dirección, con un objetivo claro, permitiendo además seguir de forma clara el avance de las tareas a realizar, de forma que los jefes pueden ver día a día como progresa el trabajo.

Se escoge SXP para el desarrollo del presente trabajo de diploma, pues esta metodología tiene como premisa la no duplicación de esfuerzos, así como la integración del cliente en el equipo de desarrollo, lo que garantiza que no haya necesidad de documentaciones extensas, solo se documenta lo necesario para una futura reutilización. Lo cual da al traste con nuestra problemática científica a solucionar.

1.6. Indicaciones generales para el desarrollo.

A partir de un análisis realizado sobre las principales herramientas y tecnologías existentes en la actualidad, se seleccionaron las siguientes como las apropiadas a utilizar en la implementación de la propuesta de solución.

Lenguaje Java

Java es un lenguaje de programación orientado a objetos. Fue desarrollado por Sun Microsystems a principio de los años 90 con los siguientes criterios de diseño: independiente de la máquina, seguro para trabajar en red y potente para sustituir código nativo.

Este lenguaje de programación utiliza una sintaxis muy familiar a la de C y C++, aunque se debe aclarar que no es ni la evolución, ni la mejora de C++. Actualmente está bajo la licencia GNU/GPL, de acuerdo con las especificaciones del Java Community Process.

Características de Java:

- ***Orientado a objetos:*** se acopla a un diseño moderno de lenguaje de programación y los elementos de los datos son tipos primitivos (enteros, caracteres, etc.) u objetos. Los mismos adoptan los valores de datos con las funciones que operan sobre ellos.
- ***Seguro:*** facilita la validación de los programas evitando que violen las reglas semánticas del lenguaje. Por ejemplo, no permite que las referencias a un arreglo superen los límites de éste, eliminando de esta forma un tipo común de error y de brecha en la integridad. Además cuenta con características de seguridad que evitan la infección por virus o trampas de programación.
- ***Uniforme y posee una arquitectura neutra:*** la precisión de los formatos para los tipos de datos primitivos se especifican por completo, asegurando que los programas funcionen por igual en cualquier locación. Además está diseñado para un ambiente de red, donde estén presentes sistemas con diferentes arquitecturas.
- ***Portable:*** no posee especificaciones que dependan de un tipo de procesador en particular, permitiendo crear programas que sin ningún cambio puede ser utilizado en cualquier plataforma.
- ***Robusto:*** no admite el manejo directo de memoria ya que no se utilizan apuntadores sino arreglos y cadenas, que libera al usuario de preocuparse por la administración de la memoria.
- ***Dinámico:*** es adaptable a cualquier cambio que ocurra en los componentes fundamentales del mismo.
- ***Multihilos:*** es capaz de atender varias tareas al mismo tiempo.

Ventajas de Java:

- Permite crear programas modulares y códigos reutilizables.
- Es una fuente abierta, evitando la lucha con los impuestos sobre patente cada año.
- Es más fácil de usar, de escribir, compilar, depurar y aprender que otros lenguajes de programación.

La recomendación del lenguaje de programación java, viene dada porque el sistema al cual se le incluirá este módulo está desarrollado completamente en este lenguaje. Además existen otros insentivos para arribar a esta sugerencia como: ser multiplataforma, orientado a objetos, seguro, robusto y de fácil de uso.

Librería OpenCV

Librería libre Open Source, desarrollada por Intel. Esta se encuentra publicada bajo licencia BSD, la cual permite que sea usada libremente para propósitos comerciales y de investigación con las condiciones expresadas en la misma. Es una librería de C/C++ y Python para ponerle visión a la computadora, la misma cuenta con varios ejemplos en C y Python. Su instalación es muy sencilla, fácil de utilizar y altamente eficiente, se ha utilizado en infinidad de aplicaciones.

Esta proporciona un alto nivel de funciones para el procesado de imágenes; le permite a los programadores crear aplicaciones poderosas en el dominio de la visión digital; ofrece muchos tipos de datos de alto nivel como juegos, árboles, gráficos, matrices, etc. Además es multiplataforma, existiendo versiones para GNU/Linux, Mac OS X y Windows.

Se recomienda el uso de la librería openCV en el desarrollo del módulo, pues la misma es una de las librerías más eficaces a la hora del trabajo con imágenes. Es multiplataforma, proporciona un grupo de

funciones de alto nivel para el procesado de imágenes basadas en algoritmos robustos. Actualmente es una de las más empleadas para implementar en esta rama, con una amplia documentación disponible.

NetBeans IDE

Es un entorno de desarrollo, hecho principalmente para el lenguaje de programación Java. NetBeans IDE es un producto libre y gratuito sin restricciones de uso. Es un IDE multilenguaje completo y modular, con soporte para JavaSE, JavaME y JavaEE. Posee una gran variedad de módulos de terceros conocidos como plugins. Desarrollo intuitivo de drag-and-drop. Es una herramienta para programadores pensada para escribir, compilar, depurar y ejecutar programas. Opera sobre sistemas Unix, OpenSolaris, Mac OS y Microsoft Windows.Soportando varios lenguajes como: Java, C/C++, Python, PHP, JavaScript.

Características de NetBeans:

- Instalación y actualización más simple.
- Diseñador visual de formularios para Swing GUI.
- Profiling integrado.
- Características visuales para el desarrollo web.
- Creador gráfico de juegos para celulares.

Se recomienda el IDE de desarrollo Netbeans para la implementación del módulo, por ser un entorno de desarrollo fácil de usar, robusto y factible a la hora de realizar la escritura y compilación del código. Además es multiplataforma y la aplicación donde se incluirá este módulo esta desarrollado completamente con este IDE, de esta forma se asegura una completa integración.

1.7. Conclusiones del capítulo.

En el capítulo recién concluido se analizó el proceso de digitalización de documentos. Se estudiaron las principales soluciones de software que contemplan este proceso en la actualidad, tanto en el contexto nacional como el internacional; permitiendo con ello identificar la necesidad de diseñar un prototipo de software, para realizar el post-proceso del documento escaneado, así como extraer el contenido del mismo, facilitando con ello las búsquedas de estos, una vez que se encuentren en el ECM Alfresco. De manera adicional, se precisaron los basamentos matemáticos a considerar dentro del propio proceso. Por otro lado se seleccionaron las herramientas, algoritmos, tecnologías y la metodología necesaria para el correcto desarrollo de la propuesta de solución.

Capítulo 2

Planificación y Diseño del Módulo de Digitalización

En el presente capítulo se comienza el desarrollo de la solución propuesta guiado por la metodología ágil SXP. Se especifican los principales artefactos generados en las primeras fases; se definen las actividades de las cuales se obtienen los documentos relacionados con la concepción inicial del sistema, los requisitos, el diseño y las tareas a realizar.

2.1. Concepción del sistema.

Al comienzo de todo proyecto y para su correcta planificación se realizan un conjunto de encuentros entre el cliente y el equipo de desarrollo. De las mismas se obtiene la visión general del producto a implementar. Dando paso con ello a la selección de los diferentes roles que intervendrán en el desarrollo del software y las tecnologías a utilizar.

2.1.1. Descripción de la propuesta de solución.

Con el objetivo de mejorar el proceso de digitalización de documentos, se propone el diseño de un módulo que incorpore el proceso íntegro de digitalización.

El proceso inicia cuando un usuario desea digitalizar uno o varios documentos. Una vez escaneado el documento deseado, se procede a realizar el mejoramiento del documento digitalizado. Este incluye dos etapas, la primera se encarga de eliminar las incongruencias en cuanto a la orientación, la segunda y última etapa del mejoramiento consiste en aplicar ajustes de brillo y contraste.

Concluido este post-proceso se procede al reconocimiento óptico de caracteres, con el objetivo de extraer toda la información disponible en la imagen en forma de texto. Por último se realiza la extracción del resumen del texto reconocido en la imagen. Obteniendo como salida el resumen de la información contenida en el documento escaneado al inicio del proceso.

Para el diseño del mismo, se deben tener en cuenta una serie de funcionalidades las cuales se listan a continuación:

- Orientar la imagen previamente escaneada de forma automática.
- Aplicar brillo deseado a la imagen.
- Aplicar el contraste deseado a la imagen.
- Realizar el proceso de reconocimiento óptico de caracteres.
- Extraer resúmenes de texto de la salida del OCR.

2.1.2. Planificación del proyecto por roles.

Rol	Responsabilidad	Nombre
Gerente	Encargado de tomar las decisiones finales, acerca de estándares y convenios a seguir durante el proyecto.	Frank Rosales Muñoz
Cliente	Participa de forma activa en el levantamiento de los requisitos del software.	Msc. Alexeis Companionis Guerra
Programadores	Definir las tareas de ingeniería y producir el código fuente del sistema.	Frank Rosales Muñoz y Caridad Odil Reyes Duconger
Analista	Confecciona la plantilla concepción del sistema, las historias de usuario y los diseños de caso de pruebas funcionales. Labora en conjunto al cliente en la realización de estas tareas.	Caridad Odil Reyes Duconger
Diseñador	Responsable del diseño del sistema, los prototipos de interfaz y la realización del diseño de las metáforas.	Frank Rosales Muñoz
Encargados de Prueba	Ejecutar las pruebas al producto y divulgar los resultados de las mismas al equipo.	Frank Rosales Muñoz y Caridad Odil Reyes Duconger

Cuadro 2.1: Planificación del Proyecto por Roles.

2.2. Modelo de dominio.

Un sistema por pequeño que sea, generalmente es complicado, por ello se establece una técnica para la especificación de los requisitos más importantes del sistema, que va a dar soporte al negocio, “el

modelo del negocio". En dependencia de la situación o escenario se determina si es necesario un modelo completo del negocio o de lo contrario se procede a definir el modelo conceptual o modelo de dominio.

El mismo representa las clases conceptuales u objetos del mundo real en un dominio de interés, siendo su función principal ayudar a entender el problema a tratar. Este se crea con el objetivo de ayudar a comprender los conceptos que utilizan los usuarios, los conceptos con los que trabajan y con los que deberá trabajar nuestra solución.

Seguidamente se muestra el modelo de dominio de la propuesta de solución:

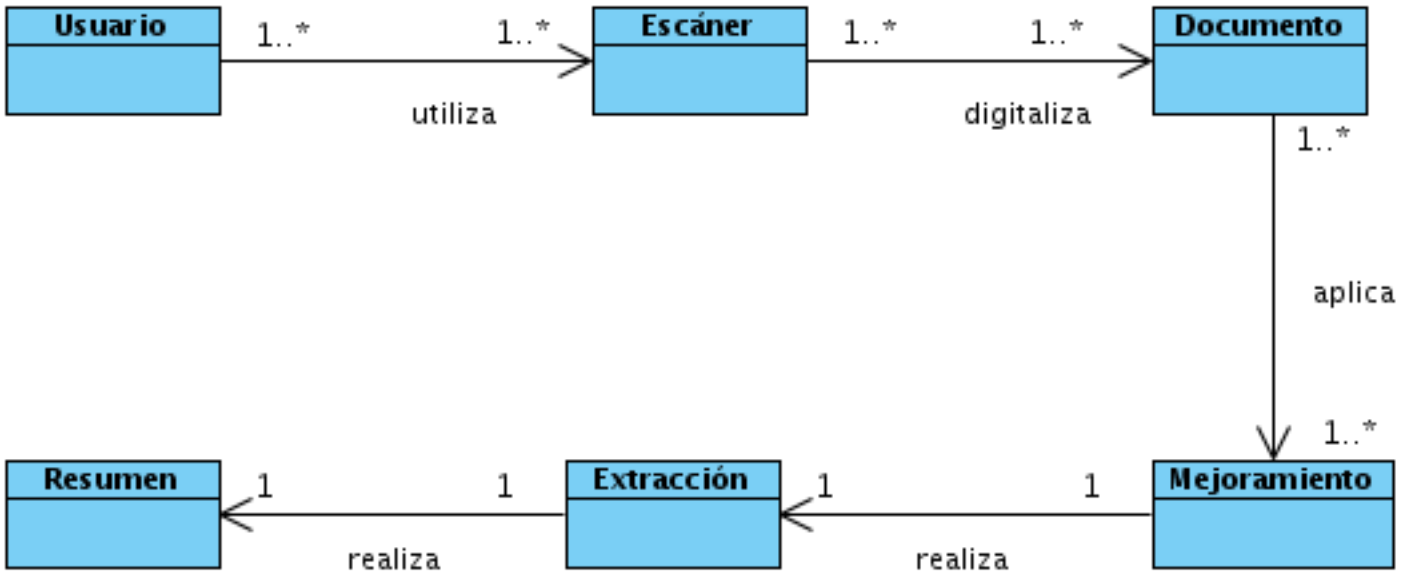

Figura 2.1: Modelo de Dominio de la Propuesta de Solución.

Conceptos del modelo de dominio:

Usuario:	Persona interesada en la funcionalidad del sistema para acceder al servicio de digitalización de un documento, el cual le permitirá aplicar mejoramiento al documento ya digitalizado, además de la obtención de la información contenida en el mismo para luego ser resumida si así lo desea.
Escáner:	Periférico externo conectado a una computadora, que posibilita la conversión a formato digital de cualquier documento impreso o escrito en forma de imagen.
Documento:	Escrito que contiene información susceptible de ser visualizada o compartida a través de una computadora. Es el testimonio de una actividad humana fijada en cualquier tipo de soporte (papel, cintas, discos magnéticos, películas, fotografías, etc.) a través de un lenguaje natural o convencional.
Mejoramiento:	Método que puede ser aplicado a una imagen previamente escaneada para mejorar su calidad y/o facilitar la búsqueda de información. En el presente trabajo se realiza un mejoramiento referido a la acentuación de brillo, contraste y se orienta la imagen.
Extracción:	Proceso que facilita la recuperación de la información contenida dentro del documento previamente escaneado.
Resumen:	Proceso que facilita el procesamiento de información, optimizando el tiempo de lectura necesario para localizar la información requerida. En la propuesta este se realizará a la salida del proceso anterior.

2.3. Captura de requisitos.

La captura de requisitos es un paso de vital importancia en el flujo de desarrollo de cualquier sistema informático, pues permite la fácil obtención de ciertas pautas que son decisivas para el futuro perfeccionamiento del mismo. Con la vista conceptual del sistema ya elaborada, el próximo paso es la

generación de la Lista de Reserva del Producto (LRP), en la cual se plasman los requerimientos a tener en cuenta por el equipo de desarrollo en la elaboración de las historias de usuarios y la implementación.

2.3.1. Lista de reserva del producto (LRP).

La plantilla lista de reserva del producto es elaborada en conjunto entre el cliente y el analista. Es el primer artefacto que se genera en la etapa de captura de requisitos, siendo esta una lista priorizada de todas las tareas a realizar. Está conformada por los requerimientos técnicos y del negocio, funciones, errores a reparar, defectos, mejoras y actualizaciones tecnológicas requeridas.

Prioridad	**Item**	**Descripción**	**Estimación (semanas)**	**Estimado por:**
Muy Alta				
	1	Orientar Imagen	2	Analista y Desarrollador
	2	Aplicar Reconocimiento Óptico de Caracteres	4	Analista y Desarrollador
	3	Aplicar Extracción de Resumen de Texto	2	Analista y Desarrollador
Alta				
	1	Aplicar Brillo a Imagen	2	Analista y Desarrollador
	2	Aplicar Contraste a Imagen	2	Analista y Desarrollador
Media				
Baja				
RNF	**Software**			
	1	La PC requiere tener instalado el software Matlab (vR2009b o superior) para poder ejecutar el prototipo de la solución propuesta.		

Continúa en la próxima página

	2	La PC precisa la instalación de la librería Libots0 v0.5.0 en sistemas GNU/Linux para poder ejecutar la opción de extracción de resúmenes de texto.
	Portabilidad	
	3	Se podrá ejecutar el prototipo en Windows y Linux siempre que esté instalado el software Matlab (vR2009b o superior).
	Restricciones en el diseño y la implementación	
	4	Lenguaje de programación M o Matlab.
	5	Librería Libots0 v0.5.0.
	Hardware	
	6	La PC necesita como mínimo de RAM 1.0 GB.
	7	La PC requiere como mínimo 40.0 GB de disco duro.
	Legal	
	8	El prototipo y toda la documentación generada pertenecen al grupo de proyecto Gestión Documental y Archivística de la Universidad de las Ciencias Informáticas.

Cuadro 2.3: Lista de Reserva del Producto.

2.3.2. Historias de usuario y prototipos de interfaz de usuario.

"Las historias de usuario son la técnica utilizada para especificar los requisitos del software. Se trata de tarjetas de papel en las cuales el cliente describe brevemente las características que el sistema debe poseer, sean requisitos funcionales o no funcionales. El tratamiento de las historias de usuario es muy dinámico y flexible. Cada historia de usuario es lo suficientemente comprensible y delimitada para que los programadores puedan implementarla en unas semanas"[47].

A continuación se exponen las historias de usuarios correspondientes a la propuesta de solución:

Historia de Usuario	
Número: HU-1	**Nombre Historia de Usuario:** Orientar Imagen.
Modificación de Historia de Usuario Número: 0	
Usuario: Frank Rosales Muñoz.	**Iteración Asignada:** 1
Prioridad en Negocio: Alta	**Puntos Estimados:** 2
Riesgo en Desarrollo: Alto	**Puntos Reales:** 2
Descripción: Esta sección garantiza la orientación automática de la imagen previamente escaneada en caso de poseer alguna inclinación. Posibilitando de manera adicional rotar 90°en ambos sentidos, la creación de un espejo tanto vertical como horizontalmente y corregir el ángulo de inclinación manualmente(siempre que se encuentre entre -45°a 45°) para la orientación de la misma, si así se desea.	
Observaciones:	
Prototipo de interfaz: (Ver Anexo 2)	

Cuadro 2.4: Historia de Usuario Orientar Imagen.

Historia de Usuario	
Número: HU-2	**Nombre Historia de Usuario:** Aplicar Brillo a Imagen.
Modificación de Historia de Usuario Número: 0	
Usuario: Caridad Odil Reyes Duconger.	**Iteración Asignada:** 1
Prioridad en Negocio: Alta	**Puntos Estimados:** 2
Riesgo en Desarrollo: Medio	**Puntos Reales:** 2
Descripción: En esta sección el usuario le aplicará el brillo que desee a la imagen previamente escaneada. De manera adicional se le posibilita corregir la transparencia de la imagen y el restablecimiento de la misma a su estado inicial.	
Continúa en la próxima página	

Observaciones:	
Prototipo de interfaz: (Ver Anexo 3)	

Cuadro 2.5: Historia de Usuario Aplicar Brillo a Imagen.

	Historia de Usuario
Número: HU-3	**Nombre Historia de Usuario:** Aplicar Contraste a Imagen.
Modificación de Historia de Usuario Número: 0	
Usuario: Caridad Odil Reyes Duconger.	**Iteración Asignada:** 1
Prioridad en Negocio: Media]	**Puntos Estimados:** 2
Riesgo en Desarrollo: Medio	**Puntos Reales:** 2
Descripción: En esta sección el usuario le aplicará el contraste que desee a la imagen previamente escaneada. De manera adicional se le posibilita corregir la transparencia de la imagen y el restablecimiento de la misma a su estado inicial.	
Observaciones:	
Prototipo de interfaz: (Ver Anexo 4)	

Cuadro 2.6: Historia de Usuario Aplicar Contraste a Imagen.

	Historia de Usuario
Número: HU-4	**Nombre Historia de Usuario:** Aplicar Reconocimiento Óptico de Caracteres.
Modificación de Historia de Usuario Número: 0	
Usuario: Frank Rosales Muñoz. Caridad Odil Reyes Duconger.	**Iteración Asignada:** 2
Prioridad en Negocio: Alta	**Puntos Estimados:** 4
Continúa en la próxima página	

Riesgo en Desarrollo: Alto	**Puntos Reales:** 4
Descripción: En esta sección garantiza la realización del proceso de OCR a la imagen previamente escaneada.	
Observaciones: Para realizar esta HU la imagen debe haber sido sujeta previamente a los procesos de las HU-1, HU-2 y HU-3.	
Prototipo de interfaz: (Ver Anexo 5)	

Cuadro 2.7: Historia de Usuario Aplicar Reconocimiento Óptico de Caracteres.

Historia de Usuario	
Número: HU-5	**Nombre Historia de Usuario:** Aplicar Extracción de Resumen de Texto.
Modificación de Historia de Usuario Número: 0	
Usuario: Frank Rosales Muñoz.	**Iteración Asignada:** 2
Prioridad en Negocio: Alta	**Puntos Estimados:** 3
Riesgo en Desarrollo: Alto	**Puntos Reales:** 3
Descripción: En esta sección se realizará un proceso de extracción de resumen de texto de la salida del proceso de OCR.	
Observaciones: Para realizar esta HU la imagen debe haber sido sujeta previamente al proceso de la HU-4.	
Prototipo de interfaz:	

Cuadro 2.8: Historia de Usuario Aplicar Extracción de Resumen de Texto.

2.4. Lista de riesgos.

En todo proceso de desarrollo de software se debe tener en cuenta los aspectos negativos que puedan retrasar el proceso de elaboración de un sistema. Para ello se elabora la lista de riesgos, en la cual se plasman las posibles afectaciones que puedan atentar en contra del cumplimiento del cronograma y la estrategia trazada para mitigar las mismas (Ver Anexo 6).

2.5. Diseño con metáforas.

En la metodología SXP no se destaca la definición temprana de una arquitectura estable para el sistema; si no se asume de forma evolutiva y los posibles inconvenientes que se generarían por no contar con ella explícitamente en el comienzo del proyecto se solventan con la existencia de una metáfora. La misma es una historia compartida que describe cómo debería funcionar el sistema. Su objetivo es proporcionar a todo el equipo una misma visión del fin del sistema y de su arquitectura general.

La metáfora definida para el sistema a desarrollar es: *el diseño de un módulo para la digitalización de documentos posibilitará, tras su implementación, recuperar y manejar la información contenida en los documentos escaneados de una forma más eficaz.*

Basados en la metáfora definida, se diseña una solución de baja complejidad, funcional y de fácil implementación la cual incluye el diagrama de componentes. Este diagrama muestra las dependencias lógicas entre componentes software, sean éstos componentes fuentes, binarios o ejecutables. Además permite modelar sistemas de software de cualquier tamaño y complejidad. A continuación se muestra el diagrama de componentes correspondiente a la propuesta de solución:

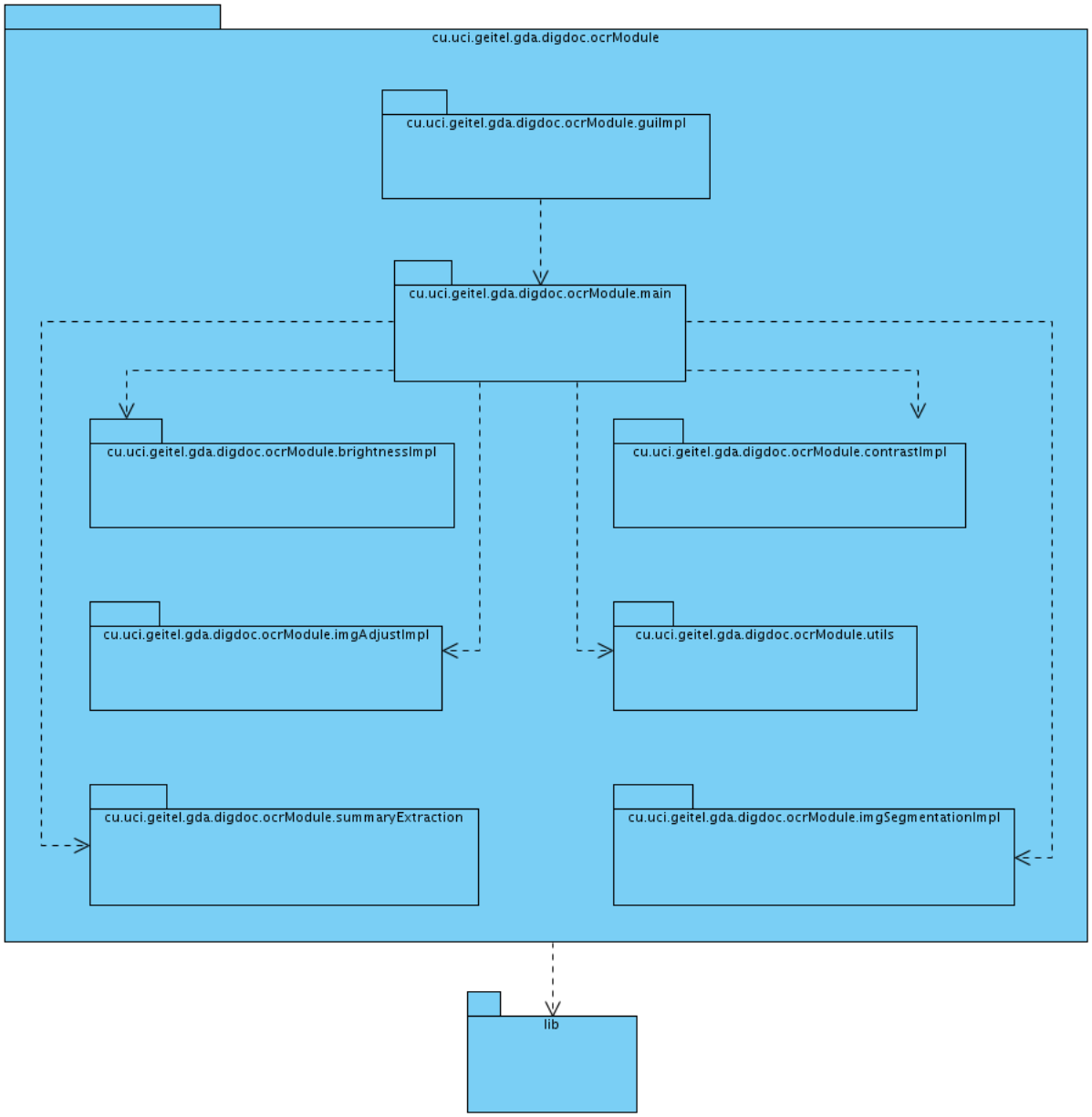

Figura 2.2: Diagrama de Componentes General.

Descripción de los paquetes que componen el diagrama de componentes:

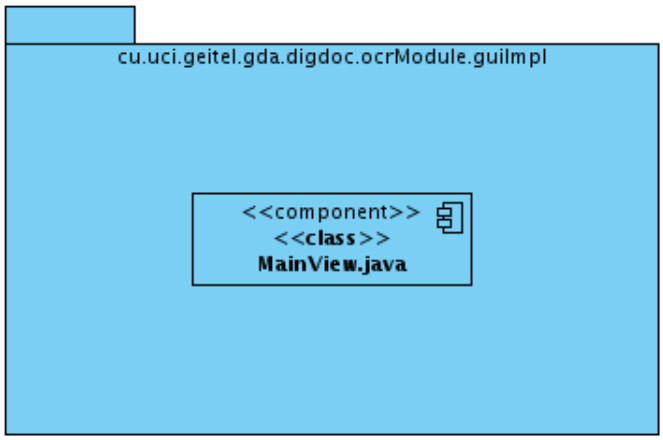

Figura 2.3: Paquete guiImpl.

Contiene el componente asociado a la interfaz del sistema, almacenando la vista principal del módulo a implementar.

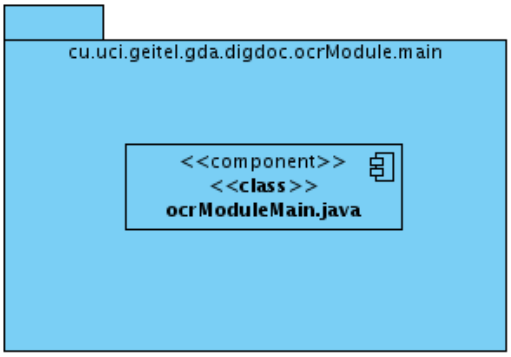

Figura 2.4: Paquete main.

Contiene el componente asociado a la inicialización de las variables y objetos del sistema, almacenando la clase principal.

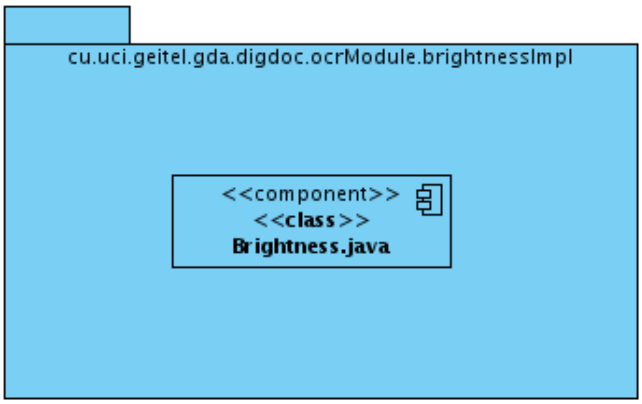

Figura 2.5: Paquete brightnessImpl.

Contiene el componente asociado al algoritmo de mejora de imágenes, específicamente, la aplicación del brillo. Almacenando la clase encargada de implementar esta técnica.

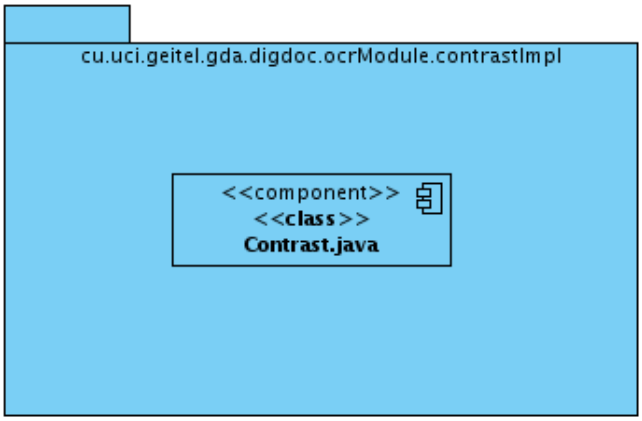

Figura 2.6: Paquete contrastImpl.

Contiene el componente asociado al algoritmo de mejora de imágenes, específicamente, la aplicación del contraste. Almacenando la clase encargada de implementar esta técnica.

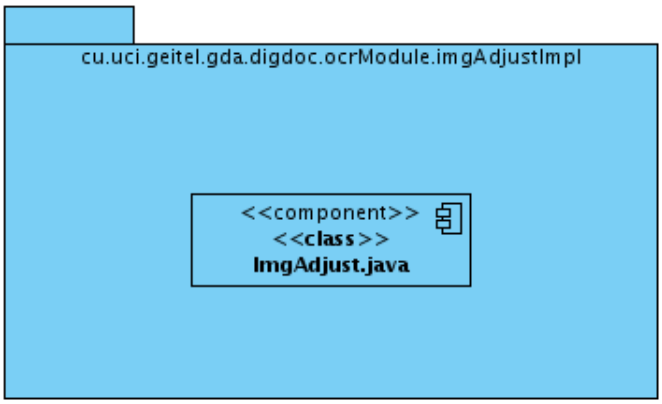

Figura 2.7: Paquete imgAdjustImpl.

Contiene el componente asociado al algoritmo de transformación espacial de la imagen, específicamente, la aplicación de la transformada de Radon. Almanenando la clase encargada de implementar esta transformada.

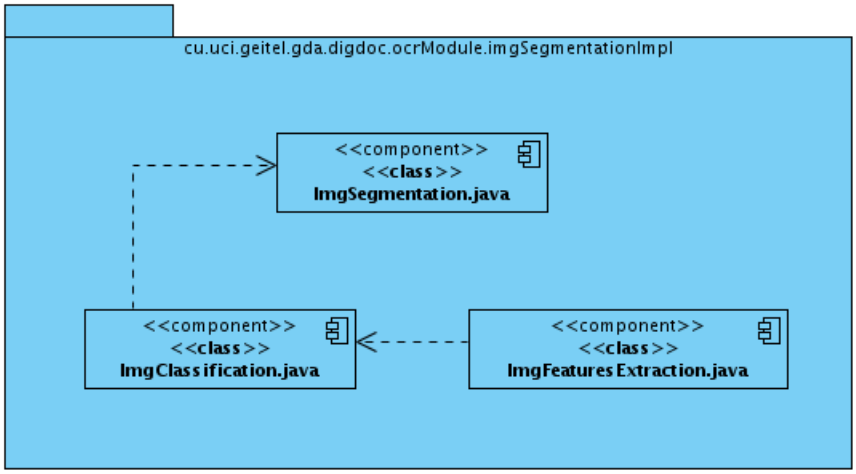

Figura 2.8: Paquete imgSegmentationImpl.

Agrupa los componentes asociados a la aplicación del algoritmo de OCR a la imagen. La clase ImgSegmentation.java, se encarga de realizar la segmentación, luego la ImgClassification.java clasifica la imagen y por último la clase ImgFeaturesExtraction.java extrae las características de la imagen.

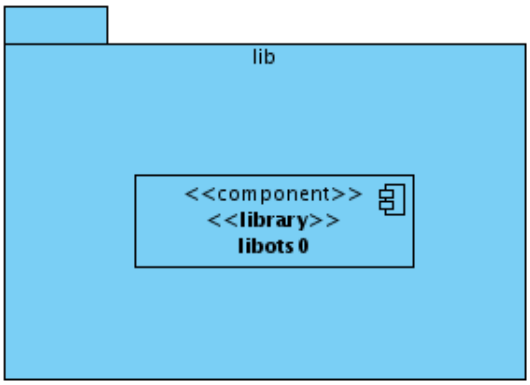

Figura 2.9: Paquete lib.

Contiene el componente asociado a las librerías. Contiene la librería libots0, la cual se emplea para extracción de resumen de texto.

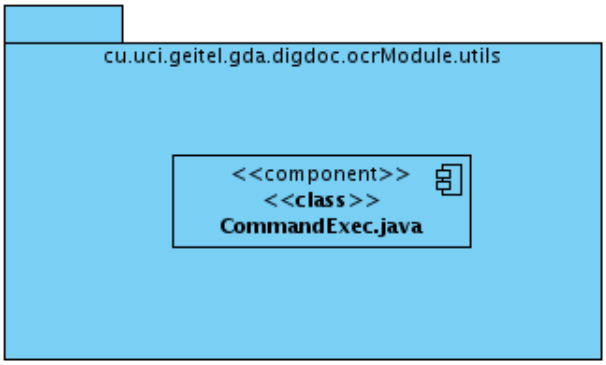

Figura 2.10: Paquete utils.

Contiene el componente asociado a las clases auxiliares a emplear. Contiene la clase CommandExec.java, la cual se emplea para hacer uso de la librería de extracción de resumen de texto.

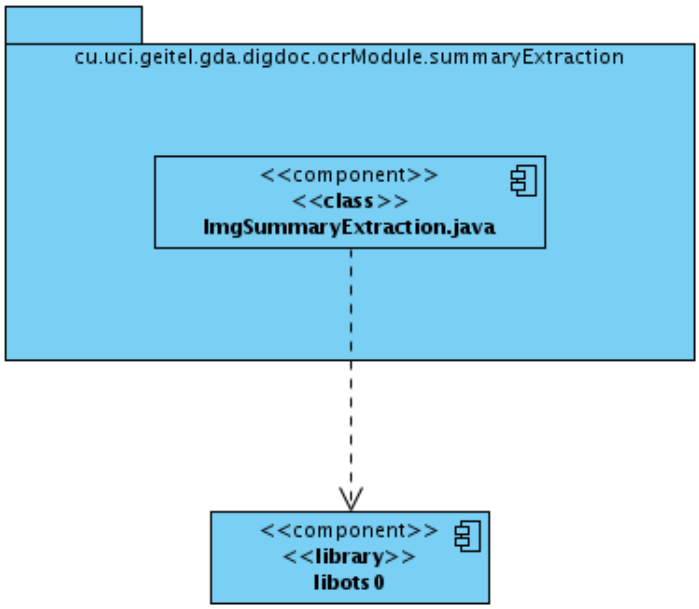

Figura 2.11: Paquete summaryExtraction.

Contiene el componente asociado a la técnica de extracción de resumen. Contiene la clase ImgSummaryExtraction.java, la cual se emplea para hacer la extracción de resumen de texto, esta depende del paquete lib, específicamente con la libots0.

2.6. Tareas de ingeniería.

La plantilla tareas de ingeniería es el primer artefacto creado en la fase de desarrollo. Facilita la definición de cada una de las actividades que estarán asociadas a las historias de usuario y que permitirán su implementación. Permite conocer a quien está asignada la misma y el tiempo que se necesita para su implementación, facilitando con ello su estimación.

Tarea de Ingenería	
Número Tarea: 1	**Número Historia de Usuario:** HU-1
Nombre Tarea: Orientar Imagen.	
Tipo Tarea: Desarrollo.	**Puntos Estimados:** 2
Fecha Inicio: 17/2/2011	**Fecha Fin:** 3/3/2011
Programador Responsable: Frank Rosales Muñoz.	
Descripción: Esta tarea de ingeniería permite la implementación de la orientación de la imagen previamente escaneada en caso de poseer alguna inclinación.	

Cuadro 2.9: Tarea de Ingeniería Orientar Imagen.

Tarea de Ingenería	
Número Tarea: 2	**Número Historia de Usuario:** HU-2
Nombre Tarea: Aplicar Brillo a Imagen.	
Tipo Tarea: Desarrollo	**Puntos Estimados:** 2
Fecha Inicio: 4/3/2011	**Fecha Fin:** 18/3/2011
Programador Responsable: Caridad Odil Reyes Duconger.	
Descripción: Esta tarea de ingeniería permite la implementación de la funcionalidad, en la cual el usuario le aplicará el brillo que desee a la imagen previamente escaneada.	

Cuadro 2.10: Tarea de Ingeniería Aplicar Brillo a Imagen.

Tarea de Ingenería	
Número Tarea: 3	**Número Historia de Usuario:** HU-3
Nombre Tarea: Aplicar Contraste a Imagen.	
Tipo Tarea: Desarrollo	**Puntos Estimados:** 2
Fecha Inicio: 20/03/2011	**Fecha Fin:** 4/4/2011
Continúa en la próxima página	

Programador Responsable: Caridad Odil Reyes Duconger.
Descripción: Esta tarea de ingeniería permite la implementación de la funcionalidad, en la cual el usuario le aplicará el contraste que desee a la imagen previamente escaneada.

Cuadro 2.11: Tarea de Ingeniería Aplicar Contraste a Imagen.

	Tarea de Ingenería
Número Tarea: 4	**Número Historia de Usuario:** HU-4
Nombre Tarea: Segmentar Imagen.	
Tipo Tarea: Desarrollo	**Puntos Estimados:** 2
Fecha Inicio: 5/4/2011	**Fecha Fin:** 19/4/2011
Programador Responsable: Frank Rosales Muñoz.	
Descripción: Esta tarea de ingeniería permite encontrar la unidad lógica más pequeña a reconocer, el caracter, de la imagen previamente escaneada.	

Cuadro 2.12: Tarea de Ingeniería Segmentar Imagen.

	Tarea de Ingenería
Número Tarea: 5	**Número Historia de Usuario:** HU-4
Nombre Tarea: Extraer Características de la Imagen.	
Tipo Tarea: Desarrollo	**Puntos Estimados:** 1
Fecha Inicio: 20/4/2011	**Fecha Fin:** 27/4/2011
Programador Responsable: Frank Rosales Muñoz.	
Descripción: Esta tarea de ingeniería permite a partir de cada uno de los caracteres obtenidos asociar las características que los van a diferenciar.	

Cuadro 2.13: Tarea de Ingeniería Extraer Características de la Imagen.

Tarea de Ingenería	
Número Tarea: 6	**Número Historia de Usuario:** HU-4
Nombre Tarea: Clasificar Imagen.	
Tipo Tarea: Desarrollo	**Puntos Estimados:** 1
Fecha Inicio: 28/4/2011	**Fecha Fin:** 5/5/2011
Programador Responsable: Caridad Odil Reyes Duconger	
Descripción: Esta tarea de ingeniería permite el reconocimiento de los caracteres.	

Cuadro 2.14: Tarea de Ingeniería Clasificar Imagen.

Tarea de Ingenería	
Número Tarea: 7	**Número Historia de Usuario:** HU-5
Nombre Tarea: Aplicar Extracción de Resumen de Texto.	
Tipo Tarea: Desarrollo	**Puntos Estimados:** 2
Fecha Inicio: 6/5/2011	**Fecha Fin:** 20/5/2011
Programador Responsable: Frank Rosales Muñoz.	
Descripción: Esta tarea de ingeniería permite realizar el proceso de extracción de resumen de texto de la salida del proceso de OCR.	

Cuadro 2.15: Tarea de Ingeniería Aplicar Extracción de resumen de texto.

2.7. Plan de liberación.

Planificar y cronometrar las actividades a efectuar en un proyecto es una vía utilizada para estimar el tiempo de duración del mismo. Por tal razón en SXP se elabora el plan de liberación o release, en el cual el cliente decide qué historias de usuario deben ser incluidas en un lanzamiento. La correcta

elaboración de este artefacto facilita la ubicación de las historias de usuario más significativas y permite la planificación del proceso de desarrollo de software en iteraciones.

Release	**Descripción de la iteración**	**Orden de las HU a implementar**	**Duración total**
1	Al concluir esta iteración el usuario será capaz de realizar un mejoramiento a la imagen escaneada previamente.	HU-1, HU-2 y HU-3	17/2/2011 - 4/4/2011
2	Al concluir esta iteración el usuario será capaz de realizar la extracción de caracteres y resumen a la imagen escaneada previamente.	HU-4 y HU-5	5/4/2011 - 20/5/2011

Cuadro 2.16: Plan de Liberación.

2.8. Conclusiones del capítulo.

En el capítulo se detalló la propuesta de solución como aspecto inicial para lograr un mejor entendimiento entre el cliente y los desarrolladores. Se realizó la planificación del proyecto por roles, para distribuir las tareas y de esta forma garantizar la realización de un trabajo organizado. Además se identificaron las necesidades del cliente, plasmándolas mediante requerimientos, los cuales se describieron a través de las HU. Por último, tomando en cuenta la prioridad para el negocio de las HU se creó el plan de iteraciones, estimando el tiempo de desarrollo en semanas para cada una de ellas.

Capítulo 3

Validación de la Propuesta de Solución

En el ciclo de desarrollo del software la etapa de pruebas constituye un aspecto fundamental, pues permite verificar y revelar la calidad que posee el producto desarrollado mediante la revisión final de las especificaciones del diseño y de la codificación. Por tal motivo este capítulo tiene como objetivo principal elaborar y aplicar los casos de prueba de aceptación de cada historia de usuario.

3.1. Casos de prueba.

Las pruebas de aceptación son definidas por el cliente y preparadas por el equipo de desarrollo, aunque la ejecución y aprobación final corresponden al cliente. La utilización de estas, proporcionan grandes ventajas, permitiendo a los programadores principalmente estimar la calidad de su trabajo y garantizar la entrega de un producto de mayor calidad, que responderá siempre a las necesidades del cliente. Tienen como objetivo además, validar que el sistema cumpla con el funcionamiento esperado y permitir al cliente determinar su aceptación. Con este propósito se realizaron un conjunto de pruebas de aceptación para cada una de las historias de usuario definidas en el marco de este trabajo, y de las cuales a continuación se presenta una descripción.

3.1.1. Caso de prueba para la historia de usuario HU-1.

Esta sección cubre el conjunto de pruebas realizadas a la historia de usuario: **Orientar Imagen**. Con ella se pretende comprobar que es posible orientar la imagen previamente escaneada.

Caso de Prueba de Aceptación	
Código Caso de Prueba: HU-1-1	**Nombre Historia de Usuario:** Orientar Imagen.
Nombre de la persona que realiza la prueba: Frank Rosales Muñoz.	
Descripción de la Prueba: Se comprueban las opciones: autorrección, rotar 90°en ambos sentidos, crear espejo vertical y/o horizontal y corregir ángulo de inclinación manualmente (siempre que se encuentre entre -45°a 45°) para la orientación de la imagen.	
Condiciones de Ejecución: La imagen debe estar previamente escaneada o guardada en algún directorio de la PC para que luego pueda ser cargada por el prototipo.	
Entrada/Pasos de ejecución: Para las opciones autorrección, rotar 90°y creación de espejo vertical y/o horizontal, hacer click en la opción **Autocorrección** o **Rotar +90 /- 90** o **Espejo H/EspejoV** respectivamente y luego pulsar el botón **Aplicar** para realizar la operación. Para la opción corregir ángulo introducir el valor numérico (siempre que se encuentre entre -45° a 45°) o desplazar la barra de desplazamiento, luego pulsar el botón **Aplicar** para realizar la operación.	
Resultado Esperado: La imagen queda orientada.	
Evaluación de la Prueba: Satisfactoria.	

Cuadro 3.1: Caso de Prueba de Aceptación HU-1.

3.1.2. Caso de prueba para la historia de usuario HU-2.

Esta sección cubre el conjunto de pruebas realizadas a la historia de usuario: **Aplicar Brillo a Imagen**. Para esta historia de usuario se comprueba la opción de aplicar el brillo deseado a la imagen previamente escaneada.

Caso de Prueba de Aceptación	
Código Caso de Prueba: HU-2-1	**Nombre Historia de Usuario:** Aplicar Brillo a Imagen.
Nombre de la persona que realiza la prueba: Caridad Odil Reyes Duconger.	
Descripción de la Prueba: Se procede a verificar que sea posible aplicar el brillo deseado a la imagen, corregir la transparencia de la misma y el restablecimiento a su estado inicial.	
Condiciones de Ejecución: La imagen debe estar previamente escaneada o guardada en algún directorio de la PC para que luego pueda ser cargada por el prototipo.	
Entrada/Pasos de ejecución: Seleccionar la opción **Brillo**, luego para establecer los límites inferior y superior, desplazar las barras a gusto, pulsar el botón **Aplicar** para realizar esta operación. Para corregir el valor de gamma en la imagen desplazar la barra correspondiente a gusto, pulsar el botón **Aplicar** para realizar esta operación. Para restablecer la imagen a su estado original oprimir el botón **Rest. Orig.**	
Resultado Esperado: La imagen resultante contiene el brillo deseado.	
Evaluación de la Prueba: Satisfactoria.	

Cuadro 3.2: Caso de Prueba de Aceptación HU-2.

3.1.3. Caso de prueba para la historia de usuario HU-3.

Esta sección cubre el conjunto de pruebas realizadas a la historia de usuario: **Aplicar Contraste a Imagen**. Mediante esta se verifica la posibilidad de poder aplicarle el contraste deseado a la imagen previamente escaneada.

Caso de Prueba de Aceptación	
Código Caso de Prueba: HU-3-1	**Nombre Historia de Usuario:** Aplicar Contraste a Imagen.
Nombre de la persona que realiza la prueba: Caridad Odil Reyes Duconger.	
Continúa en la próxima página	

Descripción de la Prueba: Se procede a verificar que sea posible aplicar el contraste deseado a la imagen, corregir la transparencia de la misma y el restablecimiento a su estado inicial.
Condiciones de Ejecución: La imagen debe estar previamente escaneada o guardada en algún directorio de la PC para que luego pueda ser cargada por el prototipo.
Entrada/Pasos de ejecución: Seleccionar la opción **Contraste**, luego para establecer los límites inferior y superior, desplazar las barras a gusto, pulsar el botón **Aplicar** para realizar esta operación. Para corregir el valor de gamma en la imagen desplazar la barra correspondiente a gusto, pulsar el botón **Aplicar** para realizar esta operación. Para restablecer la imagen a su estado original oprimir el botón **Rest. Orig.**
Resultado Esperado: La imagen resultante contiene el contraste deseado.
Evaluación de la Prueba: Satisfactoria.

Cuadro 3.3: Caso de Prueba de Aceptación HU-3.

3.1.4. Caso de prueba para la historia de usuario HU-4.

Esta sección cubre el conjunto de pruebas realizadas a la historia de usuario: **Aplicar Reconocimiento Óptico de Caracteres**. Mediante esta se verifica la posibilidad de poderle realizar el proceso de OCR a la imagen previamente escaneada.

	Caso de Prueba de Aceptación
Código Caso de Prueba: HU-4-1	**Nombre Historia de Usuario:** Aplicar Reconocimiento Óptico de Caracteres.
Nombre de la persona que realiza la prueba: Frank Rosales Muñoz.	
Descripción de la Prueba: Se procede a realizar el proceso de OCR a la imagen.	
Continúa en la próxima página	

Condiciones de Ejecución: La imagen debe estar previamente escaneada o guardada en algún directorio de la PC para que luego pueda ser cargada por el prototipo. Además dicha imagen debe haber sido sujeta previamente de las acciones HU-1, HU-2 y HU-3.
Entrada/Pasos de ejecución: Click en el botón **Ref** del panel **Creación de la colección** para cargar una colección, luego dar click en el botón **Out** del mismo panel para indicar la salida del alfabeto de dicha colección, click en el botón **Aplicar** para efectuar esta operación. Luego click en el botón **Cargar** del panel **Entrenamiento de la R.N.** para cargar el alfabeto, con el cual se entrenará la red neuronal, click en el botón **Aplicar** para efectuar esta operación. Finalmente click en el botón **Ref** del panel de **Reconocimiento** para cargar la imagen a reconocer, luego click en el botón **Out** del mismo panel para dar la salida del reconocimiento de caracteres, click en el botón **Aplicar** para efectuar esta operación.
Resultado Esperado: Se realizan con éxito el proceso de OCR a la imagen.
Evaluación de la Prueba: Satisfactoria.

Cuadro 3.4: Caso de Prueba de Aceptación HU-4.

3.1.5. Caso de prueba para la historia de usuario HU-5.

Esta sección cubre el conjunto de pruebas realizadas a la historia de usuario: **Aplicar Extracción de Resumen de Texto**. Con esta prueba se pretende comprobar que es posible realizar el proceso de extracción de resumen de texto de la salida del proceso de OCR.

Caso de Prueba de Aceptación	
Código Caso de Prueba: HU-5-1	**Nombre Historia de Usuario:** Aplicar Extracción de Resumen de Texto.
Nombre de la persona que realiza la prueba: Caridad Odil Reyes Duconger.	
Continúa en la próxima página	

Descripción de la Prueba: Se procede realizar un proceso de extracción de resumen de texto de la salida del proceso de OCR de la imagen.
Condiciones de Ejecución: La imagen debe estar previamente escaneada o guardada en algún directorio de la PC para que luego pueda ser cargada por el prototipo. Dicha imagen debe haber sido sujeta previamente del proceso HU-4. El sistema operativo debe ser plataforma GNU/Linux, para que el prototipo pueda hacer uso de esta funcionalidad.
Entrada/Pasos de ejecución: Abrir una terminal y ejecutar el siguiente comando: **ots [-r <porciento de compresión>] <ruta texto entrada>-o <ruta texto salida>** **Nota aclaratoria:** El radio de compresión por defecto es 20 %, si se desea especificar otro indicar su valor numéricamente entre 1 y 100.
Resultado Esperado: Se obtiene el resumen del texto.
Evaluación de la Prueba: Satisfactoria.

Cuadro 3.5: Caso de Prueba de Aceptación HU-5.

3.2. Conclusiones del capítulo.

En el capítulo recién concluido se elaboraron y aplicaron los casos de pruebas de aceptación a cada historia de usuario para dar validez y veracidad a la propuesta de solución. Mediante lo anteriormente citado, se arriba a la obtención de un prototipo funcional, con todos los algoritmos probados, evidenciándose la presencia de las cualidades necesarias para la implementación del futuro módulo que se integrará con el software DigDoc[6].

[6]DigDoc: Herramienta de digitalización que permite la digitalización, procesamiento y almacenamiento de documentos.

Conclusiones

De manera general en el presente trabajo se diseñó un módulo para la digitalización de documentos que incorpora técnicas de procesamiento digital de imágenes, clasificación y extracción de resúmenes posibilitándose así la extracción y representación de manera compacta de la información contenida en los documentos digitalizados.

De manera adicional y partiendo de un carácter más específico pueden destacarse los aspectos siguientes:

- Se caracterizaron las soluciones de digitalización de documentos tanto nacionales como internacionales existentes en la actualidad, evidenciándose la necesidad de diseñar una solución a la medida para estos fines.
- Fueron seleccionadas las herramientas y tecnologías más apropiadas para el futuro desarrollo de la propuesta de solución presentada.
- Se desarrolló, asociado a la propuesta presentada, un detallado proceso de ingeniería hasta la etapa de análisis y diseño.
- Se validó el eficaz funcionamiento de la propuesta a través de la implementación de un prototipo real de la misma en ambiente Matlab.

Recomendaciones

Las ineludibles necesidades a las que se enfrenta nuestro país en el orden organizativo y de perfeccionamiento de los sistemas de trabajo y control imponen un ritmo acelerado a la introducción y desarrollo de los gestores de contenidos empresariales. En este anhelo el trabajo presenta una primera aproximación a un estado deseado donde convergen soberanía, seguridad, fiabilidad y facilidad de uso en un todo único; quedando como recomendaciones fundamentales las siguientes:

1. Implementar, de acuerdo con las especificaciones realizadas, la propuesta de solución contenida en el informe.

2. Sustituir los métodos de segmentación incorporados en el prototipo por aquellos de umbrales determinados automáticamente por el contenido de la imagen.

3. Incorporar el uso de los diccionarios en el proceso de reconocimiento óptico de caracteres como vía para elevar la eficacia de este proceso.

Glosario de Términos

A

Alfresco: Gestor de Contenido Empresarial de código abierto.

API: Application Programming Interface, Interfaz de Programación de Aplicaciones. Conjunto de funciones y procedimientos(o métodos si se refiere a programación orientada a objetos) que ofrece cierta librería para ser utilizado por otro software como una capa de abstracción.

E

ECM: Enterprise Content Management, Gestor de Contenidos Empresariales. Identifica a los sistemas informáticos que manejan la captura, almacenamiento, seguridad, control de versiones, recuperación, distribución, conservación y destrucción de documentos y contenido a nivel empresarial.

G

GIMP: Images Manipulation Program. Programa de manipulación de imágenes del proyecto GNU. Trabaja con capas y tiene multitud de plugins. Tiene versiones tanto para sistemas libres, como GNU\Linux, como para sistemas propietarios.

GNOME: GNU Network Object Model Environment. Entorno de trabajo totalmente libre desde su comienzo, de muy sencillo uso y configuración, es además uno de los más completos entornos

de desarrollo, basando ésta en las librerías GTK.

GTK: Es un grupo importante de bibliotecas multiplataforma para desarrollar interfaces gráficas de usuario. Fue creado para desarrollar el programa manipulador de imágenes GIMP, sin embargo actualmente es muy usada por muchos otros programas en los sistemas GNU\GPL.

K

KDE: K Desktop Environment. Es una de las GUI más importantes para sistemas UNIX.

L

Librería: Término informático para referirse a las bibliotecas de vínculos dinámicos.

M

Metadatos: Datos que describen el contexto, el contenido y la estructura de los documentos de archivo y su gestión a lo largo del tiempo.

Microsoft Windows: Es un sistema operativo desarrollado por la compañía norteamericana Microsoft. Se trata de un conjunto de programas que permiten administrar los recursos de una computadora y gestionar el hardware desde los niveles más básicos.

O

OCR: Optical Character Recognition, Reconocimiento Óptico de Caracteres. Facultad que tienen ciertas computadoras para reconocer y procesar caracteres escritos. Se precisa para ello un periférico de entrada de datos con capacidades ópticas y un software específico que permite convertir la foto digital del texto en texto editable con un procesador de texto.

OpenCV: Open Source Computer Vision Library, Librería de Visión por Computador de Código Abierto. Librería Open Source para el tratramiento de imágenes.

P

PDF: Portable Document Format, Formato de Documento Portátil. Es un formato de almacenamiento de documentos.

Plugin: Pequeño programa que añade funcionalidades a otro programa, habitualmente de mayor tamaño que no posee. Un programa puede tener uno o más Plug In.

Prototipo: Es un modelo (representación, demostración o simulación) fácilmente ampliable y modificable de un sistema planificado, probablemente incluyendo su interfaz y su funcionalidad de entradas y salidas.

S

SANE: Scanner Access Now Easy (SANE). Es un estándar en sistemas UNIX's para la adquisición de imágenes. Proporciona una Interfaz de Programación de Aplicaciones (API) que proporciona acceso estandarizado a cualquier dispositivo de escaneo, dígase escaner de sobremesa, escaner de mano, cámaras y videocámaras.

Referencias Bibliográficas

[1] H. G. Riveros and L. Rosas, *El método científico aplicado a las ciencias experimentales.* 2000.

[2] O. G. Delio G. and F. B. Víctor, "Papiro.un sistema de conservación, digitalización, gestión y socialización de información documental para los archivos en cuba.," *Fórum de Ciencia y Técnica, Cuba*, 2006.

[3] "Aplicaciones debian." http://nux.ula.ve/tutoriales/sub_debian.html#xsane.

[4] "Ubuntu. details of package xsane in lucid." http://packages.ubuntu.com/es/lucid/amd64/xsane.

[5] "Simple scan: Digitalizar documentos facilmente en lucid lynx. ubuntumexico.com." http://ubuntumexico.com/2010/03/simple-scan/.

[6] "Simple scan: nueva aplicación de escaneo para ubuntu 10.04." http://www.genbeta.com/imagen-digital/simple-scan-nueva-aplicacion-de-escaneo-para-ubuntu-1004.

[7] "Ubuntu lucid lynx simplificará el uso de scanners. fayerwayer." http://www.fayerwayer.com/2009/12/ubuntu-lucid-lynx-simplificara-el-uso-de-scanners/.

[8] "Kooka, escanea en ubuntu." http://www.softhoy.com/kooka-escanea-en-ubuntu.html.

[9] "Capítulo 19. kooka. el programa de escaneo." http://softwarelibre.unsa.edu.ar/slw/HTML/suse/ch19.html.

[10] "Global solutions catalog." http://h20147.www2.hp.com/hpgsc/solView.aspx?solID=1564.

[11] "Kofax capture." http://siticleon.com/index.php?option=com_content&view=article&id=51&Itemid=37.

[12] "Abbyy finereader. descargar." http://abbyy-finereader.softonic.com/.

[13] "Abby finereader v9.0.724 corporate edition multilenguaje." http://www.dregus.com/f261/abby-finereader-v9-0-724-corporate-edition-multilenguaje-59611/.

[14] "Emc captiva isis quickscan pro-desktop scanning." http://spain.emc.com/products/detail/software2/quickscan-pro.htm.

[15] "Escaner gratuito documalis documalis free scanner por scanpoint software - reporte y descarga." http://www.freedownloadmanager.org/es/downloads/Documalis_Explorador_Libre_47382_p/.

[16] "Escaneo de documentos-digitalización de documentos-software de escaneo-document imaging-scanning software." http://www.cmsrl.com.ar/index.html.

[17] "Curso de procesamiento digital de imágenes.." http://turing.iimas.unam.mx/~elena/PDI-Mast/Tema_5_C.pdf.

[18] R. de la Rosa Flores, "Procesamiento de imágenes digitales." http://www.itpuebla.edu.mx/Eventos/MemoriasyResSemanaInformatica2007/29-ProcesamientoImagenesDigitales.pdf, octubre 2007.

[19] J. A., *Fundamentals of Digital Image Processing.* Prentice Hall, USA, 1989.

[20] R. Molina, "Introducción al procesamiento y análisis de imágenes digitales.," *Dpto de Ciencias y Computación e I.A , Universidad de Granada*, pp. 179–211, 1998.

[21] S. Theodoridis and K. Koutroumbas, *Pattern Recognition.* Elsevier Inc. UK, 2009.

[22] E. Coto, *Métodos de Segmentación de Imágenes Médicas.* IBM Journal of Research and Development, 2003.

[23] Y. B. Chen and O. T. C. Chen, "Image segmentation method using thresholds automatically determined from picture contents," *EURASIP Journal on Image and Video Processing*, pp. 2–14, 2009.

[24] R. Nock and F. Nielsen, "Semi-supervised statistical region refinement for color image segmentation," *Pattern Recognition*, vol. 38, no. 6, pp. 835–846, 2005.

[25] D. Liu and T. Chen, "Discov: a framework for discovering objects in video," *IEEE Transactions on Multimedia*, vol. 10, no. 2, pp. 200–208, 2008.

[26] L. Vincent and P. Soille, "Watersheds in digital spaces: an efficient algorithm based on immersion simulations," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 13, no. 6, pp. 583–598, 1998.

[27] S. Chien, Y. Huang, and L. Chen, "Predictive watershed: a fast watershed algorithm for video segmentation," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 13, no. 5, pp. 453–461, 2003.

[28] C. J. Kuo, S. F. Odeh, and M. C. Huang, "Image segmentation with improved watershed algorithm and its fpga implementation," *Proceedings of the IEEE International Symposium on Circuits and Systems ISCAS '01*, vol. 2, pp. 753–756, May 2001.

[29] P. Salembier and L. Garrido, "Binary partition tree as an efficient representation for image processing, segmentation, and information.," *Proceedings of the IEEE International Symposium on Circuits and Systems ISCAS '01*, vol. 2, pp. 753–756, 2001.

[30] J. C. W. H. Lu and M. Ghanbari, "Binary partition tree for semantic object extraction and image segmentation," *EEE Transactions on Circuits and Systems for Video Technology*, vol. 17, no. 3, pp. 378–383, 2007.

[31] J. Canny, "Computational approach to edge detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 8, no. 6, pp. 679–698, 1986.

[32] A. X. Falcao, J. K. Udupa, and F. K. Miyazawa, "An ultra-fast user-steered image segmentation paradigm: live wire on the fly," *IEEE Transactions on Medical Imaging*, vol. 19, no. 1, pp. 55–62, 2000.

[33] M. Kass, A. Witkin, and D. Terzopoulos, "Snakes: active contour models," *International Journal of Computer Vision*, vol. 1, no. 4, pp. 321–331, 1988.

[34] S. X. Yu, "Segmentation using multiscale cues," *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition CVPR '04*, vol. 1, pp. 247–254, June 2004.

[35] C. H. Chuang and W. N. Lie, "A downstream algorithm based on extended gradient vector flow field for object segmentation.," *IEEE Transactions on Image Processing*, vol. 13, no. 10, pp. 1379–1392, 2004.

[36] B. G. Kim and D. J. Park, "Novel noncontrast-based edge descriptor for image segmentation.," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 16, no. 9, pp. 1086–1095, 2006.

[37] H. Gao, W. C. Siu, and C. H. Hou, "Improved techniques for automatic image segmentation.," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 11, no. 12, pp. 1273–1280, 2001.

[38] I. Pitas, "Digital image processing schemes and application," 2000.

[39] Y. B. Chen and T. C. Chen, "Semi-automatic image segmentation using dynamic direction prediction.," *Speech and Signal Processing ICASSP '02 Proceedings of the IEEE International Conference on Acoustics*, vol. 4, pp. 3369–3372, may 2002.

[40] C. Platero, "Apuntes de visión artificial," *ELAI Universidad Politécnica de Madrid*, 2006.

[41] K. Koche, "Comparison of neural network and template matching technique for identification of characters in license plate," *Proceedings of the Int. Conf. on Information Science and Applications ICISA*, pp. 2–5, 2010.

[42] S. M. Rodríguez, "Clasificación automatizada de imágenes para un sistema de filtrado por contenidos de internet," Master's thesis, Instituto Superior Politécnico José Antonio Echeverría,Facultad de Ingeniería Eléctrica Dpto. de Telecomunicaciones y Telemática, 2009.

[43] "Redes neuronales." http://webdelprofesor.ula.ve/economia/gcolmen/postgrado1.html.

[44] J. M. Moreno, *Redes Neuronales Artificiales aplicadas al Análisis de Datos*. PhD thesis, Facultad de Psicología. Universitat de les Illes les Balears, 2002.

[45] "Entrenamiento de redes neuronales basado en algoritmos evolutivos." http://www.fi.uba.ar/laboratorios/lsi/bertona-tesisingenieriainformatica.pdf.

[46] O. L. Juan Pablo, "Reconocimiento Óptico de caracteres," 2009.

[47] P. L. J. H. Canós and M. C. Penadés, "Metodologías ágiles en el desarrollo de software," *DSIC-Universidad Politécnica de Valencia*.

Bibliografía

- ABBYY FineReader - Descargar. [cited 27 January 2011]. Available from world wide web: http://abbyy-finereader.softonic.com/.
- Anil K. Jain. Fundamentals of Digital Image Processing. [USA]: Prentice Hall, 1989.
- Aplicaciones Debian. [cited 27 January 2011]. Available from world wide web: http://nux.ula.ve/tutoriales/sub_debian.html#xsane.
- Becerril C. Francisco Java a su alcance. [Mexico]: McGRAW-HIL INTERAMERICANA EDITORES SA de C.V,, 2000.
- Belmonte Fernández Oscar. Introducción al lenguaje de programación Java. Una guía básica. 2005. Available from world wide web: http://www3.uji.es/~belfern/pdidoc//IX26/Documentos/introJava.pdf..
- DESOFT para la Sociedad Cubana.pdf. [cited 27 January 2011]. Available from world wide web: http://www.desoft.cu/Portals/0/DESOFT%20para%20la%20Sociedad%20Cubana.pdf.
- Digitalizar imágenes en Linux. [cited 27 January 2011]. Available from world wide web: http://www.dacostabalboa.com/es/digitalizar-imgenes-en-linux/527.
- Document Capture – Data Capture - Kofax. [cited 27 January 2011]. Available from world wide web: http://www.kofax.com/.

- EKINSA - DIGITALIZACION Y MICROFILMACION. [cited 27 January 2011]. Available from world wide web: http://www.ekinsa.com/divisiones_digitalizacion.php.
- El Lenguaje de Programación MATLAB. [cited 27 January 2011]. Available from world wide web: http://lc.fie.umich.mx/~calderon/Matlab/indice.html.
- EMC Captiva ISIS QuickScan Pro - Desktop Scanning. [cited 27 January 2011]. Available from world wide web: http://spain.emc.com/products/detail/software2/quickscan-pro.htm.
- Escaneo de Documentos - Digitalización de Documentos - Software de Escaneo - Document Imaging - Scanning Software. [cited 27 January 2011]. Available from world wide web: http://www.cmsrl.com.ar/index.html.
- Escaner Gratuito Documalis (Documalis Free Scanner) por Scanpoint Software - reporte y descarga. [cited 27 January 2011]. Available from world wide web: http://www.freedownloadmanager.org/es/downloads/Documalis_Explorador_Libre_47382_p.
- FichaTecnicaAvilaDoc.pdf. [cited 27 January 2011]. Available from world wide web: http://www.desoft.cu/Portals/0/FichaTecnicaAvilaDoc.pdf.
- Flanagan David . Java en pocas palabras. McGRAW-HILL INTERAMERICANA EDITORES, S.A. de C.V., 1999.
- Gnome Ubuntu para todos by Leonciokof. [cited 27 January 2011]. Available from world wide web: http://ubuntuparatodos.wordpress.com/tag/gnome/.
- González Rafael C. and Woods Richard E.. Digital Image Processing. 2nd [Upper Saddle River, New Jersey]: Tom Robbins, by Prentice-Hall, Inc., 2002.
- Historia del lenguaje Java. [cited 27 January 2011]. Available from world wide web: http://www.cad.com.mx/historia_del_lenguaje_java.htm.

- Introducción a Matlab. [cited 27 January 2011]. Available from world wide web: http://www.fisica.unav.es/~angel/matlab/matlab0.html.
- Introducción a Matlab y Octave. [cited 27 January 2011]. Available from world wide web: http://iimyo.forja.rediris.es/tutorial/intro.html#matlab-es-un-lenguaje-de-programacion.
- Kooka - Scan and OCR Suite for KDE. [cited 27 January 2011]. Available from world wide web: http://kooka.kde.org/.
- L. Ávila Estrada, and V. I. Álvarez Morell. IMPORTANCIA DE LA DIGITALIZACIÓN PARA LA CONSERVACIÓN DE DOCUMENTOS. Available from world wide web: http://innovacion.ciget.lastunas.cu/index.php/innovacion/article/viewFile/55/51.
- Más razones para digitalizar - Digitalizacion documentos - Escaneo - Digidoc. [cited 27 January 2011]. Available from world wide web: http://www.digidoc.es/.
- MathWorks r 2010 a MATLAB - Fotografía Digital / Diseño Gráfico. [cited 27 January 2011]. Available from world wide web: http://www.applesana.es.
- MATLAB. [cited 27 January 2011]. Available from world wide web: http://www.worldlingo.com/ma/enwiki/es/MATLAB#Limitations.
- Molina, R. Introducción al Procesamiento y Análisis de Imágenes Digitales. [Universidad de Granada], 1998.
- NetBeans - El único IDE que necesitas. [cited 27 January 2011]. Available from world wide web: http://www.slideshare.net/felipecerda/netbeans-el-nico-ide-que-necesitas.
- NetBeans IDE - NetBeans Rich-Client Platform Development (RCP). [cited 27 January 2011]. Available from world wide web: http://netbeans.org/features/platform/index.html.

- OCR - Qué es OCR, Reconocimiento Optico de Caracteres - Readsoft. [cited 27 January 2011]. Available from world wide web: http://www.readsoft.es/automatizacion-facturas/ocr.aspx.
- OCR, Software de Reconocimiento Óptico de Caracteres - Yerbabuena Software. [cited 27 January 2011]. Available from world wide web: http://www.yerbabuena.es/ocr.
- OCR y su aplicación en la Gestión Documental. [cited 27 January 2011]. Available from world wide web: http://www.pixelware.com/filesc-ocr-gestion-documental.htm.
- Orozco González Delio G. and Fernández Bertot Víctor. PAPIRO.Un sistema de conservación, digitalización, gestión y socialización de información documental para los archivos en Cuba.
- Pressman Roger. Un enfoque práctico. [Madrid], 2001.
- Procesos De Ingenieria Del Software. [cited 27 January 2011]. Available from world wide web: http://www.slideshare.net.
- Proyectos de Sistemas Informáticos: Historia de OCR. [cited 27 January 2011]. Available from world wide web: http://psi-g5.blogspot.com/2007/08/historia-ocr.html.
- Razones para digitalizar documentos, imágenes y archivos - Digitalización de documentos - Sediex. [cited 27 January 2011]. Available from world wide web: http://www.sediex.com.
- Reconocimiento ótico de caracteres - Enciclopedia Encydia, de reconocimiento ótico de caracteres wikipedia. [cited 27 January 2011]. Available from world wide web: http://es.encydia.com/pt/Reconocimiento_%C3%B3tico_de_caracteres.
- Savit Jeffrey, Wilcox Sean and Jayaraman Bhuvana. Java para la empresa. [Mexico]: McGRAW-HILL INTERAMERICANA EDITORES, S.A. de C.V., 2000.
- Simple Scan: nueva aplicación de escaneo para Ubuntu 10.04. [cited 27 January 2011]. Available from world wide web: http://www.genbeta.com.

- Sistema de control vehicular utilizando reconocimiento óptico de caracteres [cited 27 January 2011]. Available from world wide web: http://www.dspace.espol.edu.ec/bitstream/123456789/1458/1/2973.pdf.
- Software libre y software propietario en ingeniería — Enchufa2. [cited 27 January 2011]. Available from world wide web: http://www.enchufa2.es.
- Sun Microsystems Conozca el nuevo NetBeans. [cited 27 January 2011]. Available from world wide web: http://www.sun.com/emrkt/innercircle/newsletter/spain/0207spain_feature.html.
- Ubuntu 10.04 Lucid Lynx. [cited 27 January 2011]. Available from world wide web: http://foro.el-hacker.com/f61/ubuntu-10-04-lucid-lynx-180058.
- Ubuntu – Details of package xsane in lucid. [cited 27 January 2011]. Available from world wide web: http://packages.ubuntu.com/es/lucid/amd64/xsane.
- Ubuntu Lucid Lynx simplificará el uso de scanners - FayerWayer. [cited 27 January 2011]. Available from world wide web: http://www.fayerwayer.com.
- Universidad Distrital Francisco Jose de Caldas [cited 27 January 2011]. Available from world wide web: http://gemini.udistrital.edu.co/comunidad/estudiantes/ocala/matlabTut/acerca.php.
- Ventajas y desventajas de hacer la entrada de datos contra el OCR. [cited 27 January 2011]. Available from world wide web: http://www.articleset.com.
- Welcome to NetBeans. [cited 27 January 2011]. Available from world wide web: http://netbeans.org/.