Capítulo 3.

Sistemas OCR e ICR para grandes volúmenes de información

En este capítulo se revisarán las distintas soluciones disponibles para la transformacion de grandes volúmenes de imágenes textuales en texto. Esto es para elegir la solución más apta para el problema de las tarjetas de Hu. Se busca una solución confiable, rápida y con buen índice de reconocimiento. Además debe tener alguna forma de procesar grandes cantidades de imágenes en forma continua.

A los sistemas de reconocimiento se han añadido nuevas funcionalidades. Por ejemplo, Se han creado sistemas que capturan texto en formas (cheques, solicitudes). De esta adición nació el término llamado Reconocimiento Inteligente de Caracteres (ICR). Los sistemas de ICR se especializan en reconocer caracteres escritos a mano en formas, y tienen idea de la distribución del texto en dichas formas, así que extraen información ya clasificada de acuerdo a su localización en el documento.

En primera instancia se verán los sistemas OCR e ICR públicos, para posteriormente revisar las soluciones comerciales de los mismos.

3.1 Sistemas Públicos

3.1.1 NIST

El Instituto Nacional de Estándares y Tecnología (NIST) [Garris 1994] ha desarrollado un sistema de reconocimiento basado en formas para escritura a mano. NIST ha hecho el sistema disponible al público sin cargo alguno, la distribución es en forma de CD-ROM. Se distribuye el código fuente, que está escrito en C (K&R, no ANSI), y algunas librerías están escritas en Fortran. Se distribuyen librerías para registrar tipos de formas, aislar campos de las formas, segmentación de campos, clasificación de caracteres y post-procesamiento usando diccionarios. El sistema, que funciona bajo el sistema operativo Unix, puede usarse sin restricciones, pues fue creado con fondos del gobierno de los Estados Unidos de América.

Se instaló una versión disponible por FTP de este software, sin embargo, no se logró hacer funcionar, por no contar con la versión necesaria de Fortran en las máquinas en las que se instaló. La distribución en CD-ROM fue solicitada, pero no ha llegado a nuestras

manos hasta la fecha de la edición de este documento.

3.1.2 Xocr

Esta es una solución *amateur*, diseñada para ser usada en máquinas PC bajo el sistema Linux; sin embargo, fue posible compilarlo y usarlo en Unix 5.5 en una SparcStation 5, y comprobar que tiene un pobre desempeño. Se obtuvo del servidor WWW de Entendimiento de Documentos y Reconocimiento de Caracteres de la Universidad de Maryland (http://documents.cfar.umd.edu/ocr) . El sistema usa segmentación directamente "inundando" los símbolos a reconocer, sin reconocer la orientación de la hoja, o si lo que selecciona es texto o no. Resultó ser un sistema poco estable, además de presentar una interfaz que era difícil de manipular, pues el despliegue de la imagen a reconocer era muy pobre. El único formato aceptado por este OCR es el de archivos Bitmap de Windows (BMP). Las imágenes de las HuCards tienen el formato de Tagged ImageFile Format (TIFF), así que para usar Xocr, se tendría que hacer la conversión de las imágenes. El desempeño del software no lo amerita. El autor, Martin Bauer lo hizo disponible como shareware a un costo de 20 dólares.

3.1.3 OCRChie

OCRChie nació como un proyecto de Ingeniería de software en la universidad de Berkeley. Después se convirtió en el proyecto de Tesis de Kathey Mardsen [1996]. Este software tiene mejores características que xocr. Puede detectar errores de alineamiento en el texto y corregirlos, segmenta basándose en palabras y admite el popular formato de imágenes TIFF. Para instalarlo se procedió a obtener la distribución disponible en la página web de la autora, se revisaron los requerimientos y se advirtió que no se contaba con todas las librerías necesarias, entre ellas estaban las versiones de TCL 7.0 y TK 7.4, además de una librería para manipular imágenes en formato TIFF. Se obtuvieron estas distribuciones, y posteriormente se instaló y compiló el sistema. El sistema compiló perfectamente, pero no fue usable debido a un error de Tcl, debido a la falta de un archivo script del código del proyecto. La autora se deslindó públicamente de dar soporte para este desarrollo, así que fue inutil el intento de instalación.

3.1.4 SOCR

SOCR es un esfuerzo por construír librerías estándar de OCR, no obstante, este proyecto tiene mucho por avanzar. Es un proyecto financiado por el departamento de Ciencias de la Computación de la Universidad de Waikato, Nueva Zelandia. El fin de este proyecto es hacer este OCR disponible gratuitamente. El desarrollador comenzó a trabajar en este OCR desde 1994, pero no fue hasta julio de 1998 cuando consiguió financiamiento para trabajar en él de tiempo

completo. Se obtuvieron versiones de prueba con poca funcionalidad del sitio WWW de SOCR (http://www.socr.org), se trató de instalarlo, obteniendo por aparte las librerías *flex* y *bison*, en las versiones que se solicitaban en el software, no se logró construir el binario por completo. El sistema aún no tiene una interfaz integrada para el usuario y aún no reconoce caracteres, la integración de este software está pendiente.

3.2 Sistemas Comerciales

Se revisaron varios sistemas comerciales, incluyendo sistemas que van dirigidos a usuarios inexpertos, pero que responden en buena forma a necesidades más fuertes de las que se cree que puedan manejar. Se investigó también en el campo de los negocios de ICR, los cuales se encargan de la captura de formas al detalle directamente con el cliente. Además se revisaron las librerías disponibles que las diferentes compañías proveen a los desarrolladores para efectuar OCR en sus aplicaciones. Aquí se presenta un buen resumen de esta investigación.

3.2.1 Aplicaciones Integradas

Los sistemas de OCR comerciales revisados fueron OmniPage Pro 7.0 , TextBridge Pro 8.0, Omniforms 2.0 y El "plugin" Capture de Adobe Acrobat 3.1.

Los primeros dos son reconocedores de textos que se pueden emplear en computadoras personales. Ambos tienen interfaz gráfica de usuario para efectuar la labor de reconocimeinto de imágenes con texto. La plataforma de ejecución en la que se evaluó fue Macintosh puesto que no fue posible encontrar versiones para equipos compatibles con IBM PC..

Al probar el software disponible, en primera instancia quedaron eliminados los sistemas de reconocimiento de formas, OmniForms y Capture, pues no contaban con algún mecanismo para manejar un volumen grande de imágenes, además de estar específicamente hechos para capturar y reproducir formas como inventarios, libros de contabilidad, etc. Posteriormente se hicieron pruebas en los últimos dos, ambos tienen mecanismos de procesamiento por lotes de archivos.

TextBridge puede ser manejado totalmente por el lenguaje script de programación del sistema operativo de las Macintosh: AppleScript. Mediante ésta facilidad el programador puede crear scripts que procesen imágenes en TextBridge, si dichas imágenes son copiadas a un folder determinado. Su adaptabilidad para controlar es magnífica para el procesamiento en lotes, sin embargo, TextBridge de Xerox no puede procesar imágenes TIFF [Aldus Developer's Desk, 1992] con más

de dos colores, lo cual no se cumple para las tarjetas Hu. Por ello tuvo que ser eliminado.

OmniPage, de Caere, tiene opcion para ajustar un folder que éste puede observar, para procesar los archivos que se copien a él. Y aunque no es controlable por AppleScript, el reconocimiento automático es estable. Este fúe el software elegido para el sistema de reconocimiento de este trabajo.

	Plataforma evaluada	Plataformas disponibles	Desempeño OCR	Procesamiento por lotes	Formato TIFF de tarjetas soportado	Precio
OmniPage Pro 7.0	Mac	Mac/PC	100	si (incluido)	si	$475 USD
TextBridge 8.0	Mac	Mac/ PC	90	si (por medio de applescript)	no	$49 USD
OmniForms 2.0	Mac	Mac/ PC	80	no	si	$149 USD
Acrobat Capture Plugin	Mac	Mac/ PC	80	no	si	$ 295 USD

Figura 3.1: Desempeño de software comercial de reconocimiento.

La figura 3.1 muestra las características del software que se tuvo disponible para evaluar. Las versiones para PC compatibles tienen mejores opciones en algunos casos, pero no se tuvieron a la mano.

3.2.2 Librerías de Desarrollo

Diversas compañías de reconocimiento de caracteres, al observar la creciente demanda por parte de los desarrolladores, han convertido sus productos en diferentes librerías de desarrollo, que pueden ser empleadas por los programadores para crear sus propios tipos de aplicaciones a la medida de sus necesidades. Entre éstas se pueden mencionar las libreirías de Compañías como Mitek ($4500 USD), Caere ($5495 USD), NewSoft ($3995 USD), entre otros.

Las librerías de Mitek y Caere pueden ser utilizadas en máquinas PC mediante los lenguajes VisualBasic y VisualC++, de Microsoft, además de poder ser usadas en forma de controles ActiveX.

Ambas reconocen caracteres de máquina y escritos a mano, sólo el de caere tiene opción para reconocer diferentes lenguajes, pues tiene diccionarios para validar documentos en 100 lenguajes diferentes.

La librería de Caere además incluye soporte para reconocer códigos de barras en documentos, opción no disponible para el producto de Mitek (http://www.miteksys.com).

Comparable con la de Caere (http://www.caere.com), la librería ofrecida por NewSoft (http://www.newsoftinc.com), reconoce el mismo tipo de documentos, aunque no tiene soporte para reconocimiento de documentos multilingues. NewSoft hace disponible su "engine" con el nombre de Recore, y su plataforma de uso es también Windows 95.

Para este proyecto se consideró usar estas librerías, sin embargo, por su costo alto se optó por usar las soluciones existentes.

Dircio Palacios Macedo, R. 1998. Reconocimiento y Consulta de Imágenes Textuales en Bibliotecas Digitales. Tesis Licenciatura. Ingeniería en Sistemas Computacionales. Departamento de Ingeniería en Sistemas Computacionales, Escuela de Ingeniería, Universidad de las Américas-Puebla. Diciembre.