

OCR

Sistemas de Reconocimiento Óptico de Caracteres

Joaquim Arlandis Navarro - Investigador responsable del área OCR y Análisis de Documentos
ITI - Instituto Tecnológico de Informática

Introducción: Digitalización+Ventajas de Automatización

La digitalización de la información (textos, imágenes, sonido, etc.) se ha convertido en los últimos años en un punto de creciente interés para la sociedad.

Por lo que respecta a los textos, existen y se generan continuamente grandes cantidades de información escrita, tipográfica o manuscrita en todo tipo de soportes. Especialmente en papel, soporte susceptible de ser digitalizado, para poder gozar de las ventajas que del procesamiento de datos por computador se derivan.

En este contexto, automatizar la introducción de caracteres al sistema evitando la entrada por teclado, implica un importante ahorro de recursos para las empresas, incrementando la productividad al mismo tiempo que se preserva o mejora la calidad de los servicios ofrecidos a los clientes.

Los sistemas de reconocimiento óptico de caracteres (OCR), así como los de reconocimiento de texto en general, tienen como objetivo ayudar en el desarrollo de estas tareas. Se presentan en forma de aplicaciones diversas dirigidas al tratamiento automático de textos, ofreciendo así, claros beneficios a la sociedad actual.

De hecho, la tecnología OCR llega hoy en día, tanto a empresas directamente relacionadas con la digitalización y la gestión documental

con requerimientos de procesamiento de grandes volúmenes de datos, como a la Administración Pública y a gran variedad de Pymes. Esto se debe en parte a la gran versatilidad de campos de aplicación y el coste asequible que presentan estos sistemas.

La Problemática Científico-Técnica

El reconocimiento óptico de caracteres u OCR (*optical character recognition*), consiste en la identificación automatizada de símbolos o caracteres pertenecientes a un determinado alfabeto, a partir de una imagen recogida mediante la lectura óptica de un texto grabado en un apoyo real.

El problema del reconocimiento óptico de caracteres es bien conocido y ha estado abordado de manera intensa por disciplinas científicas, como el Reconocimiento de Formas y Visión Artificial.

Las características de este problema y las diversas vertientes que presenta, hacen que sea un proceso en continua investigación. Desde el reconocimiento de caracteres impresos, de uno o múltiples tipos de letra (considerado en la práctica un problema superado), como la escritura continua, restringida o no (de gran dificultad intrínseca), pasando por el reconocimiento de caracteres manuscritos aislados o la disposición fija o flotante del texto a reconocer, son hoy por hoy, objeto de estudio.

Sin embargo, cabe distinguir el reconocimiento "off-line" versus "on-line", este último más sencillo porque recoge información temporal

del proceso de escritura. La existencia de esta variedad de contextos hace que esta área de conocimiento aplicado siga abierta a múltiples vertientes hoy en día.

En este sentido, se ha de resaltar la dificultad que implica el reconocimiento automático de la escritura humana. Por un lado, hay distintos patrones válidos para un mismo carácter y por otro, las distorsiones en los trazos hacen que la forma de un carácter se distinga radicalmente de cualquier patrón caligráfico, convirtiendo a veces los caracteres en irreconocibles.

Así, nos podríamos preguntar, ¿Es posible que un sistema automático reconozca algún día las prescripciones de un médico?

Acerca de Sistemas OCR

Existen sistemas OCR muy diversos según los tipos de problemas que abordan y las funcionalidades que ofrecen.

Respecto de reconocimiento de caracteres y símbolos propiamente dicho, existe una terminología de empleo extendido en la industria para referirse a cada una de las variantes específicas de sistemas OCR.

Así, ICR (*intelligent character recognition*) se utiliza para designar el reconocimiento de caracteres manuscritos.

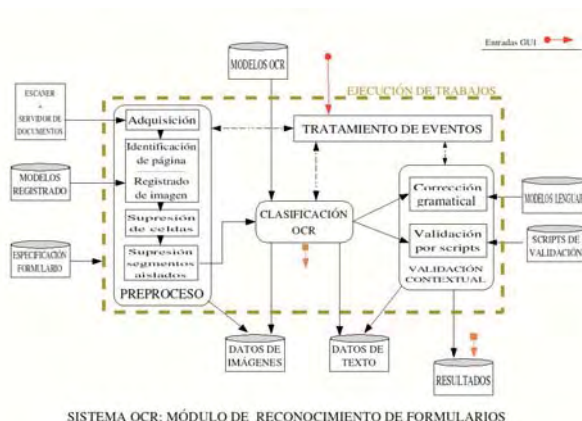
OCV (*optical character verification*) hace referencia a la verificación de contenidos previamente conocidos.

Y OMR (*optical mark recognition*) designa una funcionalidad de reconocimiento de marcas.

Aunque habitualmente se asocia el problema del OCR, a la extracción de caracteres de una imagen proveniente de un papel adquirido mediante un escáner, existen sistemas OCR diseñados para trabajar con otros soportes, dispositivos y entornos de adquisición diferentes.

La captura mediante una cámara, con luz visible o infrarroja, tableta digital, capturas de pantalla, etc, proporcionan imágenes de diferente naturaleza.

Existen también requerimientos relativos a velocidad de adquisición y procesamiento, implantación en entornos dispares como: oficinas, industrias tipográficas o de envasado, entornos exteriores como el urbano o las carreteras en el caso del reconocimiento de matrículas de vehículos.



En el caso de sistemas OCR para documentos, la problemática se enmarca en el campo del análisis de documentos. En este contexto, los sistemas pueden resolver tareas como: identificación de documentos, registro, segmentación del documento en bloques lógicos, texto, gráficos, tablas, títulos y columnas, entre otros.

La detección en el texto, de líneas, palabras y caracteres también puede ser necesaria para poder aplicar finalmente el reconocimiento a nivel de carácter.

Como parte del post-proceso, se contempla la corrección contextual de los errores del clasificador de caracteres junto con una confianza del resultado obtenido.

Finalmente, los resultados obtenidos pueden ser proporcionados de diferentes maneras: simple salida de texto o *xml*, salida formateada para un procesador de textos o utilizada para la recuperación de información, indexación de documentos para bases de datos, u otras finalidades específicas.

Los sistemas de reconocimiento óptico de caracteres se presentan en forma de aplicaciones diversas dirigidas al tratamiento automático de textos, ofreciendo así, claros beneficios a la sociedad actual

OCR en Formularios

En un formulario, el texto de interés se encuentra dentro de alguno de sus campos. Cada campo del formulario, suele estar dividido en un número fijo de casillas que pueden contener caracteres manuscritos o estar en blanco, formando una o más palabras. A partir de una especificación del formato del documento, los campos a reconocer son localizados y los caracteres contenidos en las casillas son segmentados y pasados al clasificador de caracteres manuscritos.

Así, el diseño de un formulario, puede jugar un papel importante en la calidad de los resultados obtenidos. La distribución de los contenidos, el interlineado, el espaciado y tamaño de las casillas, así como la utilización de marcas de guía e identificación de formulario, son

FORMULARIO DE DATOS PERSONALES

NOMBRE: RAQUEL D.N.I.: 74834

PETICIONES QUE SOLICITA POR ORDEN DE
ESCRÍBASE CON LA MAYOR CLARIDAD POSIBLE, DADO QUE CUALQUIER E
QUE SE ANULE LA PETICIÓN O SE OBTENGA DESTINO EN UN

Nº ORDEN	CENTRO O LOCALIDAD	ESPECIALIDAD
	CODIGO	ESP. VERBALES / ITINERANTE
103	11022000137	
104	11022000138	
105	11013000237	
106	11013000238	
107	11008000137	
108	11008000138	

Nº ORDEN	CENTRO O LOCALIDAD	ESPECIALIDAD
	CODIGO	ESP. VERBALES / ITINERANTE
142	29039000338	
143	29011000237	
144	29011000238	
145	29043000237	
146	29043000238	
147	29091000537	

elementos que suelen tenerse en cuenta en el diseño del documento. Además, para evitar que las casillas que rodean el carácter interfieran en su reconocimiento, es habitual la impresión de estos en colores, con características espectrales particulares que los hacen invisibles al escáner óptico ("*tinta ciega*" o "*blind color*"). No obstante, en algunos casos, los formularios pueden contener campos no encasillados destinados a escritura continua.

Un sistema OCR integral para formularios, deberá resolver la extracción del texto de interés de una imagen de formulario. En líneas generales, esto implica abordar las siguientes fases: adquisición de la imagen, reconocimiento automático de caracteres o texto, validación manual de los textos reconocidos con baja confianza y exportación de resultados.

Entre las anteriores, el reconocimiento automático es la parte con mayor carga tecnológica y la que, por este motivo, puede llegar a marcar grandes diferencias entre unos sistemas y otros, en cuanto a la calidad de los resultados.

Dicha parte debería comprender las siguientes etapas:

1. Identificación de página (en formularios multi - página).
2. Registro o corrección de las deformaciones producidas por el dispositivo de entrada durante la captura.
3. Detección y supresión de casillas visibles.
4. Extracción de características y clasificación de los caracteres aislados.
5. Validación/Corrección automática del reconocimiento OCR: empleo de información gramatical o contextual.

Validación/Corrección Contextual

La habilidad que los humanos tenemos para leer texto manuscrito, se apoya en nuestra extraordinaria capacidad de recuperación de errores, gracias a las restricciones léxicas, sintácticas, semánticas, pragmáticas y de lenguaje discursivo que somos capaces de aplicar. Así, una parte muy importante en un sistema OCR, consiste en la verificación o corrección contextual de los errores que pueda haber producido el clasificador de caracteres o texto.

A nivel de un campo de formulario, la capacidad de verificación de que el contenido reconocido en un campo, forme parte del conjunto de palabras o frases válidas para ese campo (lenguaje), tiene una gran repercusión en las prestaciones del sistema. Por ejemplo, detectar que "HARIA" no es un nombre de persona, evita producir un dato erróneo. Si además de verificar, el sistema tiene la capacidad de corregir de modo que sugiera un contenido válido: "MARIA", y además con un cierto valor de confianza, la potencialidad del sistema es aún mucho mayor.

Técnicamente, este tipo de corrección tiene por objeto maximizar la probabilidad de que palabras o frases generadas por el sistema sean correctas, es decir, que pertenezcan al lenguaje definido por su contexto.

Diferentes campos de formulario como: números, apellidos, direcciones, códigos postales, poblaciones, etc. tienen asociados intrínsecamente sus propios modelos de lenguaje. Existen gran cantidad de lenguajes que pueden encontrarse frecuentemente en formularios u otro tipo de documentos que son fácilmente modelizables. Otros lenguajes de carácter más específico, también se pueden modelizar a partir de una muestra de palabras o frases del lenguaje asociado al campo.

Otro tipo de validación contextual es la de campos que llevan asociados restricciones semánticas. Un ejemplo típico, es el de una fecha cuyo rango de valores numéricos posibles está acotado, o un NIF, al que se le puede exigir la restricción consistente en que para una determinada serie numérica sólo se válida una letra.

OCR en el ITI

El ITI dispone de tecnología OCR avanzada adaptable a las necesidades de un amplio rango de sectores empresariales. La experiencia de nuestros investigadores en este campo, es constatable e incluye la implementación de un motor OCR y Técnicas de Corrección Gramatical, empleadas en la digitalización de los datos del Censo Padrón de España del año 2001.

Actualmente, el trabajo de investigación y desarrollo llevado a cabo por los grupos de Visión Artificial, Tecnologías del Lenguaje y el Área de Desarrollos Tecnológicos del ITI, ha permitido el desarrollo de un sistema OCR integral para formularios, adaptable tanto a pymes como a empresas del sector de la digitalización y gestión documental, con requerimientos de procesamiento de grandes volúmenes de datos. La capacidad de particularización de los diversos módulos del sistema en función de los requerimientos específicos de las empresas demandantes, lo convierte en un producto atractivo para muchos entornos de trabajo.

El sistema es adecuado para cualquier tipo de modelo de formulario en el que se quiera reconocer caracteres manuscritos aislados e impresos (incluyendo facturas, formularios administrativos, contables, etc.). Dispone de una interfaz gráfica de usuario completa incluyendo una validación manual de resultados optimizada y permite una exportación de resultados a medida. El software, puede ser adaptado a otros sistemas informáticos a través de su API, pudiendo además integrar los resultados en sistemas ERP.

El sistema OCR desarrollado por el ITI, incorpora técnicas de reconocimiento y corrección gramatical contrastadas, ofreciendo resultados de alta precisión. Esto, junto con una política de innovación continuada del producto, tanto a nivel de investigación como de incorporación de nuevas funcionalidades, nos permite tener claras expectativas de mejora e innovación del producto.

Actualmente, se están incorporando nuevas funcionalidades en el reconocimiento de texto continuo, por ejemplo, para la lectura de cheques, la detección de caracteres anómalos y la especificación automatizada de documentos, entre otras.