

OCR (Optical Character Recognition)

Este artículo pretende dar una visión general de algunos métodos, desarrollados en el campo del reconocimiento de formas, aplicados al problema concreto del reconocimiento óptico de caracteres (OCR, Optical Character Recognition), así como una descripción de tecnologías ya disponibles en el grupo de investigación del ITI y algunos ejemplos de posibles aplicaciones de las mismas.

Descripción de la tecnología de OCR

La tecnología OCR (Optical Character Recognition) engloba a un conjunto de técnicas que complementándose entre sí, se emplean para distinguir de forma automática entre los diferentes caracteres alfanuméricos existentes. En realidad no se reconocen exactamente los caracteres de un determinado alfabeto, sino que es posible distinguir entre cualquier conjunto de ideogramas. Sin embargo, se debe tener en cuenta que la precisión que se obtiene en la práctica al intentar distinguir entre un conjunto de símbolos no es del 100%. Por lo tanto, es fácil deducir que cuanto más numeroso es el conjunto de símbolos entre los que se debe decidir, mayor es la probabilidad de que se produzca un fallo de clasificación.

En todo sistema de reconocimiento óptico de caracteres (OCR) se distinguen al menos estas 4 etapas:

- Adecuación de la imagen (preproceso).
- Selección de la zona de interés (segmentación).
- Representación digital de la imagen (extracción de características).
- Distinción del carácter contenido en la imagen (reconocimiento).

Y para cada una de las cuatro etapas es posible aplicar multitud de técnicas ya existentes o desarrollar alguna específica en función de las condiciones en las que se presentan los datos de entrada, que en el caso de OCR se puede traducir por las imágenes de entrada.

A continuación se describen algunos de los métodos que se aplican dentro del grupo de visión del ITI para resolver cada una de las etapas.

1. Preproceso

Normalmente, las técnicas de OCR son útiles para digitalizar textos de algún libro (caracteres impresos) o formularios rellenos manualmente (caracteres manuscritos). Tanto en un caso como en el otro el desglose de los caracteres individuales es mucho más sencillo que en el caso de texto manuscrito continuo, para el que es necesario la aplicación de técnicas de preproceso y segmentación más complejas que en el caso de OCR.

En esta fase de preproceso (o adecuación de la imagen) el objetivo que se persigue es eliminar de la imagen de cualquier tipo de ruido o imperfección que no pertenezca al carácter, así como normalizar el

tamaño del mismo. Además, para el caso de OCR, la normalización de la imagen también puede implicar un binarizado de la misma.

Para la eliminación del ruido que puede aparecer en una imagen digital, bien provocado por manchas reales o grafías imperfectas, o bien por defectos técnicos en la adquisición o binarizado de la imagen, se utilizan diversos algoritmos:

- Etiquetado: para la división de la imagen en regiones de componentes conectadas.
- Erosión / expansión: para la eliminación de pequeños grupos de píxeles.
- Umbralizado de histograma: para eliminar/seleccionar los objetos más brillantes o más oscuros de la imagen.

2. Segmentación

Como ya se ha comentado anteriormente, la segmentación del texto manuscrito es un caso más complejo que el tratado en OCR, donde los caracteres, bien se encuentran claramente separados en la imagen original (formularios con campos perfectamente delimitados) o bien es posible separarlos de manera relativamente fácil, ya que su escritura es regular y presenta características aprovechables para este fin.

En el primer caso nos encontramos en las condiciones más favorables, puesto que la segmentación de los caracteres viene dada por la demarcación de los límites de los campos en los que se espera que se rellene el formulario. Esta información la conocemos a priori y

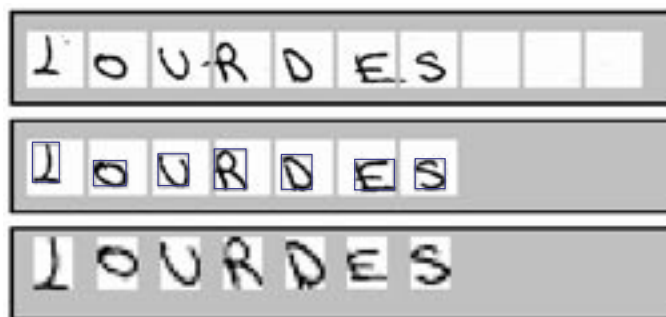


Figura 1: Ejemplo de segmentación y normalización. De arriba hacia abajo: imagen original dividida por las marcas de un campo de un formulario, eliminación de ruido y detección de la caja mínima de inclusión, y normalización del tamaño de los caracteres.

es una de las formas más fiables de realizar la segmentación con éxito. En la figura 1 se muestra el proceso de segmentación y normalización de un campo de texto extraído de un formulario manuscrito.

Sin embargo, para el caso de texto continuo se requiere la explotación de alguna característica del mismo, como puede ser la longitud de los caracteres (en el caso más sencillo), los valles de separación entre letras o números distintos, las proyecciones del texto sobre líneas imaginarias y el posterior análisis de los perfiles obtenidos, etc. La utilización de éstos métodos está supeditada a las características concretas del texto que se desea segmentar, por lo tanto el uso de los mismos o de otros distintos se decide tras un análisis de los datos con los que se debe trabajar.

3. Extracción de características

Una vez realizada la segmentación, se dispone de una imagen normalizada en la que se encuentra la información susceptible de ser “reconocida”. La información así representada, una matriz bidimensional de valores binarios, niveles de gris o color RGB, no codifica de forma óptima las características más discriminativas del objeto al que representa.

Desde el punto de vista del reconocimiento de formas, la matriz bidimensional se ve como un vector de tantas dimensiones como componentes tiene la matriz. La dimensión de estos vectores (el número de componentes) es normalmente elevado, lo que supone un gran coste computacional a la hora de procesar el mismo. Y no solo eso, y más importante aún, es que está comprobado que al intentar clasificar (“reconocer”) vectores de este tamaño aparece un efecto, llamado maldición de la dimensionalidad, que provoca que los resultados, independientemente del método de clasificación utilizado, sean malos. Por ello se han desarrollado multitud de técnicas, denominadas “técnicas de selección y extracción de características”, mediante las cuales es posible obtener una representación del objeto a reconocer más eficiente.

Eficiencia, en este caso, significa que con una representación más compacta se consigue un poder discriminativo igual o superior al que se tenía con la representación original. Esto no es solo importante por el ahorro de espacio en el almacenamiento de las muestras, sino que durante el proceso de reconocimiento reduce los costes computacionales, debido a la reducción en el volumen de información procesado.

En el campo de investigación del reconocimiento de formas se tiene experiencia en el uso de algunos métodos de extracción de características basados en transformaciones del espacio de representación de las muestras. Ejemplos de estos métodos son:

- PCA (Principal Component Analysis);
- LDA (Linear Discriminant Analysis);
- ICA (Independent Component Analysis);
- NDA (Non-linear Discriminant Analysis).

En el caso del grupo de investigación del ITI, uno de los más usados es el método PCA. El objetivo de esta técnica es definir una transformación lineal desde el espacio de representación original a un nuevo espacio en el que las distintas clases de las muestras quedan mejor separadas. Además, esta transformación permite reducir la dimensión del nuevo espacio sin perjudicar sensiblemente la capacidad discriminativa de la nueva representación.

4. Reconocimiento

Objetivo final del OCR: clasificar una imagen entre un conjunto de símbolos posibles.

En esta última etapa es donde el reconocimiento (o clasificación) de los objetos, en nuestro caso imágenes de caracteres, se realiza. El problema que en esta etapa se plantea consiste en desarrollar algún método que sea capaz distinguir la clase a la que pertenece un objeto entre un conjunto limitado de clases posibles. Este planteamiento general del reconocimiento de formas, se traduce, en el caso de la aplicación de OCR, en asignar un carácter hipótesis a una imagen de entrada.

Método: comparación con patrones de referencia (knn).

Actualmente existe una gran variedad de métodos de clasificación que han ido surgiendo durante el desarrollo del campo de investigación del reconocimiento de formas. La variedad es tan amplia que se define una taxonomía entre los distintos métodos de clasificación en función de algunas características de los mismos (paramétricos/no-paramétricos, supervisados/no-supervisados, etc.). Dentro de toda esta amplia gama de clasificadores, no todos son adecuados para cualquier problema, sino que algunos presentan ciertas ventajas sobre el resto en función de las características de los datos con los que se debe tratar.

En el caso de aplicaciones OCR, existe un método estadístico, no paramétrico y supervisado, al que se conoce con el nombre de los “k vecinos más próximos” (knn, k-nearest-neighbours), ampliamente usado. Este método es muy popular debido a su sencillez y a cierto número de propiedades estadísticas bien conocidas que le proporcionan un buen comportamiento para afrontar diversos tipos de problemas de clasificación, siendo uno de ellos el de OCR.

Básicamente el método funciona de la siguiente forma: dado un conjunto de objetos prototipo de los que ya se conoce su clase (es decir, dado un conjunto de caracteres de muestra) y dado un nuevo objeto cuya clase no conocemos (imagen de un carácter a reconocer) se busca entre el conjunto de prototipos los “k” más parecidos al nuevo objeto. A este se le asigna la clase más numerosa entre los “k” objetos prototipo seleccionados.

Fase de entrenamiento y fase de test.

Conociendo el funcionamiento básico del método de clasificación de los “k vecinos más próximos” es obvio que para poder empezar a trabajar con este método es necesario reunir un conjunto de datos etiquetados, es decir, un conjunto de muestras prototipo con las clases a las que pertenecen.

En OCR, esta recolección implica disponer de una base de datos de imágenes de los tipos de caracteres que posteriormente se esperen reconocer. A este conjunto de datos se le denomina conjunto de entrenamiento. Sin embargo, la fase de entrenamiento no solo consiste en la recopilación de estos datos, sino que, típicamente, los datos originales que se dedican al entrenamiento deben ser preprocesados adecuadamente para obtener representaciones compactas y coherentes. En el ejemplo de OCR, esto quiere decir que las imágenes deben ser segmentadas (eliminación de ruido y selección de la caja mínima de inclusión), normalizadas y transformadas (extracción de características) para obtener los

vectores de baja dimensionalidad que finalmente se almacenan como conjunto de entrenamiento.

Con este conjunto de entrenamiento ya construido, el clasificador "knn" ya puede ser utilizado para reconocer la clase de una nueva muestra. Esta es la fase de test y lógicamente, también aquí es necesario aplicar todo el preproceso descrito anteriormente a cada una de las nuevas muestras. Por lo tanto, aquí se ve la necesidad de disponer de métodos rápidos de realizar estas tareas de preproceso, puesto que la velocidad de reconocimiento dependerá, en parte, de ellos. En la práctica se tiene que este preproceso es posible realizarlo muy rápidamente, aunque justo a continuación aparece la parte del proceso de reconocimiento que normalmente más carga computacional conlleva, la clasificación.

Técnicas de búsqueda rápida de vecinos

Se ha visto que el método de clasificación "knn" requiere la construcción de un conjunto de prototipos. El tamaño, entre otras cosas, de este conjunto influye en la precisión del clasificador. Debido a la naturaleza estadística del método de clasificación, cuantos más prototipos contiene este conjunto mayor exactitud se consigue aunque al mismo tiempo mayor complejidad se introduce para realizar las búsquedas, aumentando el coste computacional.

En tareas de OCR es frecuente utilizar conjuntos de referencia de más de 200.000 muestras. Con estos tamaños surge la necesidad de diseñar estructuras de datos adecuadas para realizar las búsquedas de forma optimizada, pues una búsqueda exhaustiva requeriría demasiado tiempo y se degradarían las prestaciones del sistema de OCR.

En esta línea se han desarrollado algunos algoritmos de búsqueda rápida de vecinos y sus correspondientes estructuras de datos (voronoy polygons, k-d-trees, r-trees, etc.), que intentan paliar el problema del coste de realizar búsquedas en grandes conjuntos de datos multidimensionales.

Para diversas tareas que combinan técnicas de reconocimiento de formas y visión por computador, los k-d-trees son una buena opción para implementar los algoritmos de búsqueda. De hecho, en el ITI se han empleado algoritmos de búsqueda aproximada sobre k-d-trees con diversas tareas (reconocimiento de caras, matrículas y caracteres) obteniendo resultados competitivos, tanto en velocidad como en precisión.

Tecnologías desarrolladas

1. Reconocimiento de formularios manuscritos

Dentro de un proyecto de colaboración con la empresa ODEC, se desarrolló un sistema de OCR que fue integrado en el software que ya utilizaba la empresa para tratar de forma semiautomática formularios que estaban rellenos manualmente, como el mostrado en el ejemplo de la figura 2. Inicialmente los formularios rellenos son digitalizados por medio de un escáner y de forma automática se les aplica el software de OCR para reconocer automáticamente los caracteres que se encuentran en cada campo. Posteriormente un operador humano revisa el resultado del reconocimiento automático y rectifica aquellos casos en los que el OCR falla.

Este procesado automático de los formularios incrementa notablemente la producción que se consigue realizando un proceso completamente manual.

Figura 2: Formulario con caracteres manuscritos no continuos. El sistema OCR procesa los campos marcados por los rectángulos, generando un registro con las cadenas de texto reconocidas para cada campo.

2. Reconocimiento de texto manuscrito

Otra aplicación llevada a cabo mediante el uso de técnicas de OCR consiste en el reconocimiento de cantidades numéricas sobre cheques bancarios. Las cantidades se encuentran escritas manualmente, en cifras y en letra. Esta aplicación tiene una dificultad especial en cuanto a la segmentación de los caracteres. Por otra parte también tiene una característica que permite mejorar los resultados frente al reconocimiento de texto manuscrito no restringido. Puesto que solo se pretende reconocer cantidades numéricas, y la estructura de estas sigue una gramática perfectamente definida y mucho más limitada que la gramática del lenguaje natural, es posible aprovechar esta circunstancia para diseñar un método de reconocimiento restringido a la gramática de las cantidades numéricas que proporciona un resultado de mayor precisión que el se conseguiría con un método generalista.

Aplicaciones futuras

Siguiendo la línea de investigación actual, se presentan diversos proyectos que podrían ser abordados con la tecnología ya desarrollada y que únicamente requerirían del esfuerzo necesario para implementar y adaptar las técnicas a los problemas planteados en situaciones

