Resumen

El Análisis de la Estructura de Documentos (*Document Layout Analysis*), aplicado a documentos manuscritos, tiene como objetivo obtener automáticamente la estructura intrínseca de dichos documentos. Su desarrollo como campo de investigación se extiende desde los sistemas de segmentación de caracteres desarrollados a principios de la década de 1960 hasta los sistemas complejos desarrollados en la actualidad, donde el objetivo es analizar estructuras de alto nivel (líneas de texto, párrafos, tablas, etc.) y la relación que existe entre ellas.

Esta tesis, en primer lugar, define el objetivo del Análisis de la Estructura de Documentos desde una perspectiva probabilística. A continuación, la complejidad del problema se reduce a un conjunto de subproblemas complementarios bien conocidos, de manera que pueda ser gestionado por medio de recursos informáticos modernos. Concretamente se abordan tres de los principales problemas del Análisis de la Estructura de Documentos siguiendo una formulación probabilística. Específicamente se aborda la Detección de Línea Base (*Baseline Detection*), la Segmentación de Regiones (*Region Segmentation*) y la Determinación del Orden de Lectura (*Reading Order Determination*).

Uno de los principales aportes de esta tesis es la formalización de los problemas de Detección de Línea Base y Segmentación de Regiones bajo un marco probabilístico, donde ambos problemas pueden ser abordados por separado o de forma integrada por los modelos propuestos. Este último enfoque ha demostrado ser muy útil para procesar grandes colecciones de documentos con recursos informáticos limitados.

Posteriormente se aborda el subproblema de la Determinación del Orden de Lectura, que es uno de los subproblemas más importantes, aunque subestimados, del Análisis de la Extructura de Documentos, ya que es el nexo que permite convertir los datos extraídos de los sistemas de Reconocimiento Automático de Texto (*Automatic Text Recognition Systems*) en información útil. Por lo tanto, en esta tesis abordamos y formalizamos la Determinación del Orden de Lectura como un problema de clasificación probabilística por pares. Además, se proponen dos diferentes algoritmos de decodificación que reducen la complejidad computacional del problema.

Por otra parte, se utilizan diferentes modelos estadísticos para representar la distribución de probabilidad sobre la estructura de los documentos. Estos modelos, basados en Redes Neuronales Artificiales (desde un simple Perceptrón Multicapa

hasta complejas Redes Convolucionales y Redes de Propuesta de Regiones), se estiman a partir de datos de entrenamiento utilizando algoritmos de aprendizaje automático supervisados.

Finalmente, todas las contribuciones se evalúan experimentalmente, no solo en referencias académicas estándar, sino también en colecciones de miles de imágenes. Se han considerado documentos de texto manuascritos y documentos musicales manuscritos, ya que en conjunto representan la mayoría de los documentos presentes en bibliotecas y archivos. Los resultados muestran que los métodos propuestos son muy precisos y versátiles en una amplia gama de documentos manuscritos.