Resum

L'Anàlisi de l'Estructura de Documents (*Document Layout Analysis*), aplicada a documents manuscrits, pretén automatitzar l'obtenció de l'estructura intrínseca d'un document. El seu desenvolupament com a camp d'investigació comprén des dels sistemes de segmentació de caràcters creats al principi dels anys 60 fins als complexos sistemes de hui dia que busquen analitzar estructures d'alt nivell (línies de text, paràgrafs, taules, etc) i les relacions entre elles.

Aquesta tesi busca, primer de tot, definir el propòsit de l'anàlisi de l'estructura de documents des d'una perspectiva probabilística. Llavors, una vegada reduïda la complexitat del problema, es processa utilitzant recursos computacionals moderns, per a dividir-ho en un conjunt de subproblemes complementaris més coneguts. Concretament, tres dels principals subproblemes de l'Anàlisi de l'Estructura de Documents s'adrecen seguint una formulació probabilística: Detecció de la Línia Base (*Baseline Detection*), Segmentació de Regions (*Region Segmentation*) i Determinació de l'Ordre de Lectura (*Reading Order Determination*).

Una de les principals contribucions d'aquesta tesi és la formalització dels problemes de la Detecció de les Línies Base i dels de Segmentació de Regions en un entorn probabilístic, sent els dos problemes tractats per separat o integrats en conjunt pels models proposats. Aquesta última aproximació ha demostrat ser de molta utilitat per a la gestió de grans col·leccions de documents amb uns recursos computacionals limitats.

Posteriorment s'ha adreçat el subproblema de la Determinació de l'Ordre de Lectura, sent un dels subproblemes més importants de l'Anàlisi d'Estructures de Documents, encara així subestimat, perquè és el nexe que permet transformar en informació d'utilitat l'extracció de dades dels sistemes de reconeixement automàtic de text. És per això que el fet de determinar l'ordre de lectura s'adreça i formalitza com un problema d'ordenació probabilística per parells. A més, es proposen dos algoritmes descodificadors diferents que reducix la complexitat computacional del problema.

Per altra banda s'utilitzen diferents models estadístics per representar la distribució probabilística sobre l'estructura dels documents. Aquests models, basats en xarxes neuronals artificials (des d'un simple perceptron multicapa fins a complexes xarxes convolucionals i de propostes de regió), s'estimen a partir de dades

d'entrenament mitjançant algoritmes d'aprenentatge automàtic supervisats.

Finalment, totes les contribucions s'avaluen experimentalment, no només en referents acadèmics estàndard, sinó també en col·leccions de milers d'imatges. S'han considerat documents de text manuscrit i documents musicals manuscrits, ja que representen la majoria de documents presents a biblioteques i arxius. Els resultats mostren que els mètodes proposats són molt precisos i versàtils en una àmplia gamma de documents manuscrits.