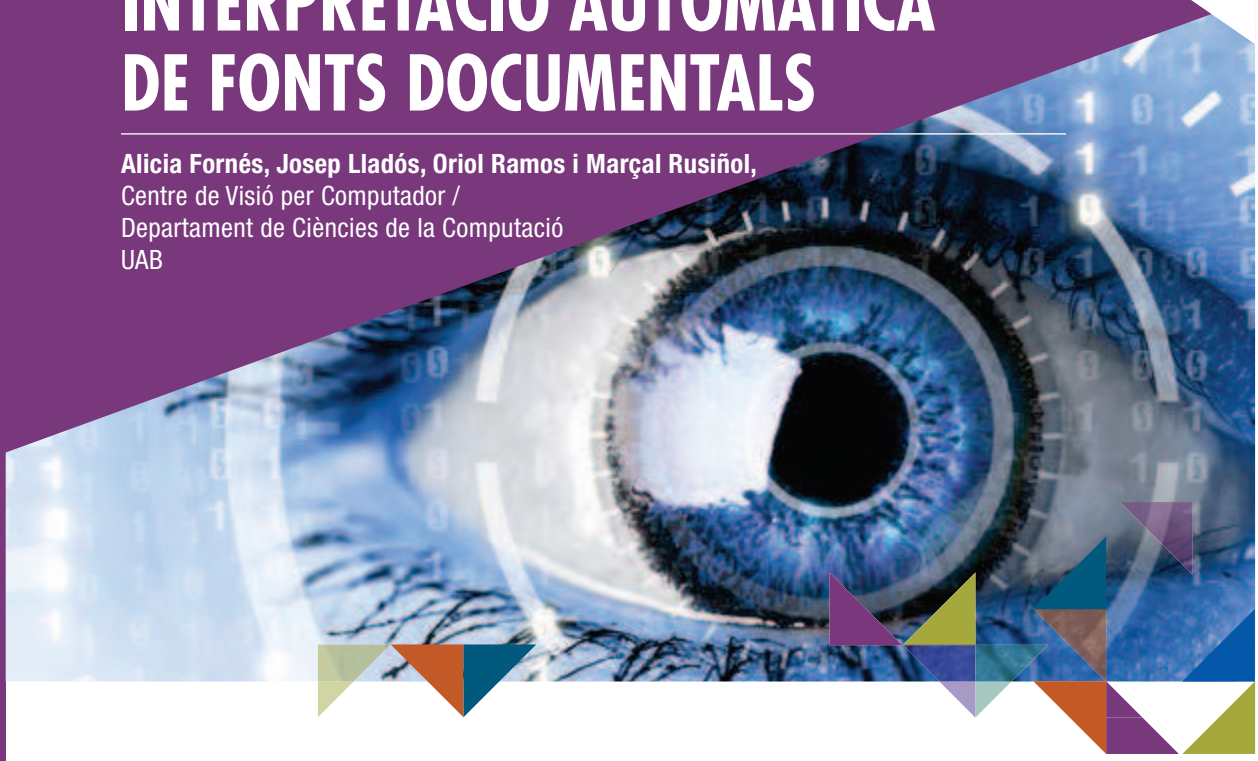
LA VISIÓ PER COMPUTADOR COM A EINA PER A LA INTERPRETACIÓ AUTOMÀTICA DE FONTS DOCUMENTALS

Alicia Fornés, Josep Lladós, Oriol Ramos i Marçal Rusiñol,
Centre de Visió per Computador /
Departament de Ciències de la Computació
UAB



1. INTRODUCCIÓ: PROGRAMARI PERQUÈ ELS ORDINADORS LLEGEIXIN

La visió per computador es pot definir informalment com la disciplina de la informàtica que fa que les màquines hi vegin. La visió humana rep la llum per l'ull, que excita els fotoreceptors (cons i bastons) que formen la retina. Aquests senyals arriben al cervell, que interpreta els estímuls donant significat a allò que veiem. En visió artificial, els ulls són les càmeres, que contenen una «retina electrònica» formada per una matriu de sensors de llum. Aquests sensors converteixen la intensitat de la llum que els arriba en valors numèrics proporcionals a aquesta i generen el que anomenem *imatges digitals*, amb les quals tots estem familiaritzats, atès que diàriament en generem amb els nostres mòbils i càmeres domèstiques. Aquestes imatges digitals són per tant funcions bidimensionals

que retornen matrius de valors anomenats *píxels* (de l'anglès *picture element*). La resolució de les imatges depèn del nombre de píxels que contenen. Fent el símil de nou amb la visió humana, com més píxels té una imatge, amb més definició s'observen els objectes, igual que una persona amb més índex d'agudesa visual, la qual té més cons concentrats al centre de la seva retina (fòvea). Però les imatges adquirides per una càmera, que no són més que matrius de punts, necessiten el seu «cervell», que són els programes d'ordinador que es dissenyen per associar els conjunts de píxels a conceptes, segons quina sigui la seva forma, color, disposició, etc. Si una imatge la dibuixem en tres dimensions com una superfície en què cada punt (píxel) té una alçada equivalent al valor que conté, podem utilitzar eines matemàtiques que processen imatges en termes de geometria diferencial (per exemple, fer més contrastada una imatge correspon a fer més alts els pics de la superfície i més baixes les valls). La figura 1 il·lustra el concepte de representació d'imatges digitals.

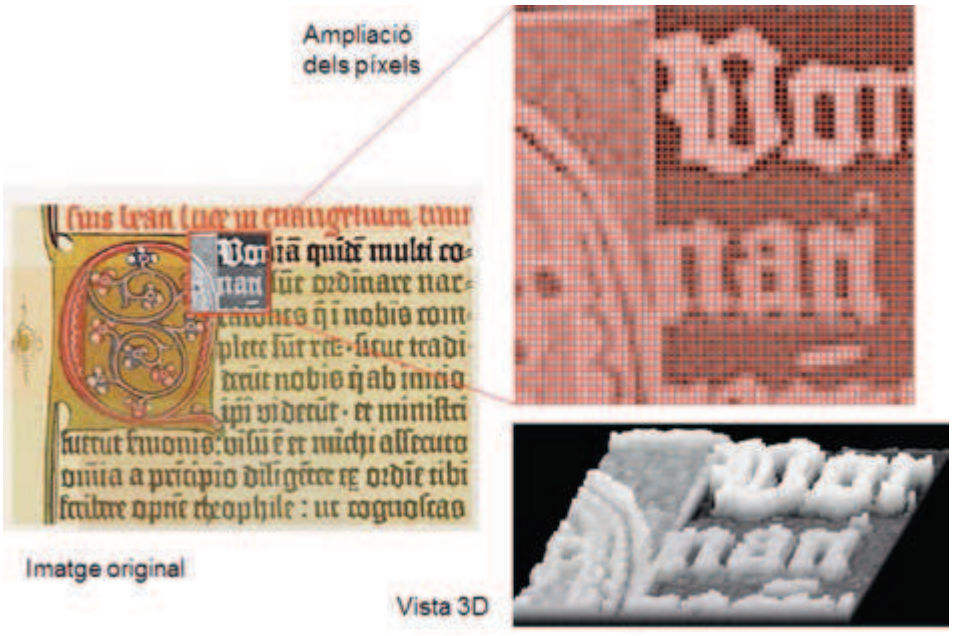


Figura 1. Una imatge digital és una representació matricial en què els valors dels píxels corresponen al color i a la intensitat de la llum en aquell punt.

La **visió per computador** ha esdevingut en els últims anys una tecnologia emergent i ubiqua. L'abaratiment de la tecnologia i l'augment de la capacitat de càlcul dels ordinadors ho han fet possible. Per altra banda, en la nostra vida quotidiana usem sovint dispositius amb càmeres que contenen programes de visió (per

exemple, càmeres nocturnes per vigilar els nadons, càmeres incorporades en consoles de videojocs que detecten el nostre moviment, càmeres que llegeixen matrícules a l'entrada d'aparcaments, o infinitat d'aplicacions als nostres telèfons mòbils). La visió és una tecnologia facilitadora en sectors com l'automoció, l'esport i l'entreteniment, el gran consum i la tecnologia mòbil, la robòtica i la manufactura avançada, la salut o la seguretat, en els quals apareixen constantment nous productes i serveis. Es calcula que el mercat de la visió ha mogut 5.700 milions de dòlars el 2014, i se'n preveu un creixement anual del 42% per arribar al 2020 amb un volum de 46.000 milions de dòlars. Aquests mercats constantment proposen reptes a una comunitat científica també en creixement.

Quan les imatges digitals corresponen a documents fotografiats o escanejats, ens referim a la subàrea denominada **anàlisi i reconeixement d'imatges de documents** (*anàlisi de documents* per referir-nos-hi de manera més abreujada). L'anàlisi de documents aborda el problema de reconèixer de manera automàtica el contingut del document (text imprès, text escrit a mà o elements gràfics). Es pot considerar que l'origen de l'anàlisi de documents sorgeix als anys seixanta, quan apareixen els primers sistemes de reconeixement òptic de caràcters (ROC). Un sistema de ROC integra un model òptic de la forma de les lletres i un model lingüístic sobre les probabilitats que aquestes es combinin segons el llenguatge d'escriptura. Per tant, els programes de ROC reconeixen agrupacions de píxels com a lletres i, en un nivell superior, validen les interpretacions conjuntes per acabar transformant una imatge en un arxiu editable de paraules.

El programari de ROC ha evolucionat molt i té avui en dia bones prestacions, especialment en documents impresos i amb digitalització de qualitat. Les aplicacions d'ofimàtica i els escàners domèstics solen incorporar un programari de ROC que ens permet transcriure automàticament els documents quotidians per a processadors de textos. Comercialment, grans corporacions com Nuance (OmniPage), Abbyy (FineReader) o Google (Tesseract) ofereixen bons sistemes que després altres empreses de serveis adapten a determinats escenaris, com ara el processament postal, la lectura de xecs bancaris o la incorporació de factures a sistemes ERP. Tanmateix, aquests sistemes tenen encara restriccions, i la recerca en anàlisi de documents ha d'avançar per poder oferir solucions a gran escala. Un cas de gran interès són els documents manuscrits, i en particular els documents històrics, que poden estar degradats, escrits en llengües antigues o no estructurats.

Entorn a aquestes fonts documentals, la informàtica i les humanitats convergeixen en l'àmbit de les humanitats digitals, una àrea emergent i interdisciplinària. Els documents custodiats en arxius històrics, administratius o eclesiàstics contenen informació molt valuosa, que recull la memòria històrica de la societat. La digitalització massiva d'aquests fons permet construir arxius digitals d'imatges que sovint són accessibles a través dels portals web de les institucions que els custodien. Aquest accés, però, no sol anar més enllà de navegadors que permeten visualitzar les imatges de manera lineal, pàgina a pàgina. En aquests casos es fan necessaris algorismes de reconeixement de text manuscrit, que sovint estan importats del món del reconeixement de la parla.

Finalment, el repte existent no és la transcripció literal, sinó l'extracció i la interpretació de continguts. Això comporta la reducció del conegut «buit semàntic», és a dir, la distància entre la representació digital dels documents i la interpretació del coneixement contingut. La interpretació de continguts permet identificar les entitats nominals (noms, llocs, dates, valors monetaris, etc.) i, per tant, omplir bases de dades estructurades i oferir grans volums de fonts digitals indexades i cercadors per contingut.

En conclusió, l'anàlisi de documents, en una perspectiva sistèmica (vegeu la figura 2), té per objectiu interpretar els continguts de documents digitalitzats a fi de transferir-los a plataformes o serveis de gestió documental en entorns de productivitat empresarial, cercadors, arxius digitals, etc. Aquest procés es divideix en tres grans passos, que són el processament i la millora de la imatge, el reconeixement i la interpretació. Els continguts documentals poden ser textuals o gràfics, i calen tècniques específiques per abordar cadascun d'aquests continguts. Per als textuals, s'utilitzaran sistemes de ROC si el text és imprès, o d'HTR (de l'anglès *handwritten text recognition*) si és manuscrit. Els elements gràfics (mapes, diagrames, esquemes d'enginyeria, plànols d'arquitectura, partitures musicals, etc.) tenen la seva pròpia idiosincràsia quant a tècniques de reconeixement.

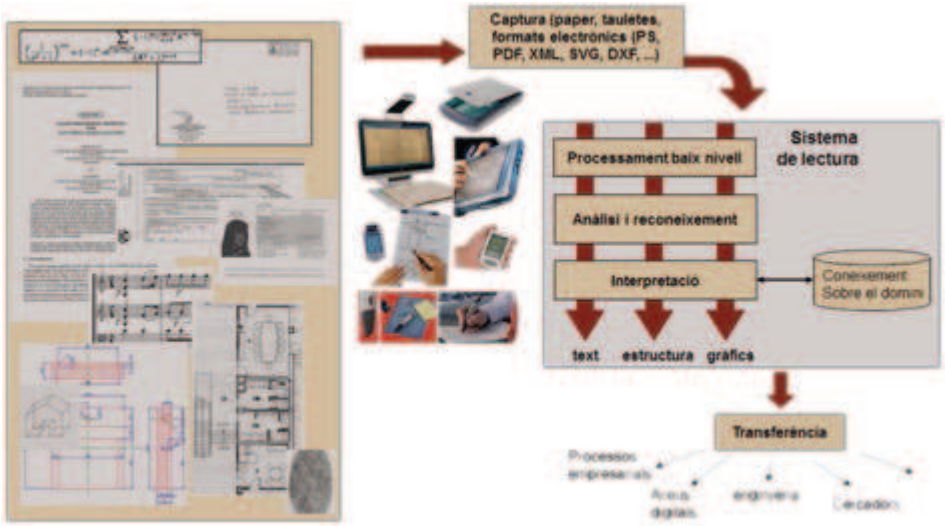


Figura 2. Arquitectura d'un sistema d'anàlisi de documents i tipologies de documents.

2. ANÀLISI I RECONEIXEMENT D'IMATGES DE DOCUMENTS: CONCEPTES BÀSICS

2.1 EL PROCESSAMENT I LA MILLORA DE LA IMATGE

El processament de la imatge comprèn aquelles tècniques que usualment s'apliquen després de la digitalització del document filtrant el valor dels píxels per millorar-ne la qualitat. Aquestes tècniques estan essencialment orientades en dues direccions: d'una banda, millorar la visualització de les imatges amb l'objectiu de fer el document més llegible per a les persones; i, de l'altra, preparar les imatges dels documents per facilitar que els sistemes automàtics n'extreguin posteriorment informació.

En casos de documents antics, o que han estat mal conservats, cal aplicar tècniques generals de millora de la visualització abans d'aplicar altres tècniques de millora per a l'extracció d'informació. Alguns exemples d'aquestes tècniques són les que intenten reduir les taques de tinta, les transparències, la decoloració del paper, la pèrdua d'intensitat de la tinta o la degradació mateixa del paper en forma de forats o esquinçaments (vegeu la figura 3).

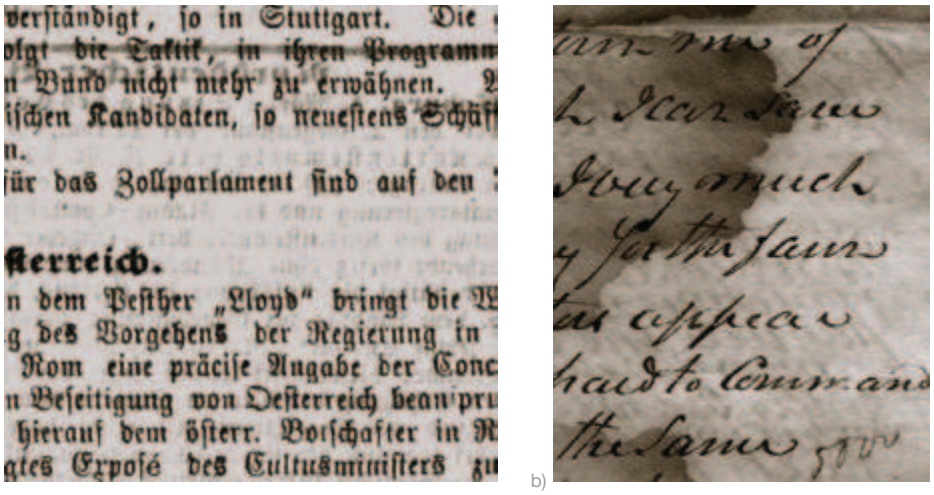


Figura 3. Exemples de degradacions a què estan sotmesos els documents.
a) ***Show-through***: transparència. b) Taques.

Més enllà de la visualització, aquestes tècniques s'apliquen prèviament al reconeixement per millorar el rendiment de l'extracció d'informació. En aquest grup trobem tècniques de reducció del soroll (punts i taques provocats de manera aleatòria per la degradació mateixa del paper, o defectes d'il·luminació en l'escaneig) i d'eliminació del fons de la imatge, i tècniques de correcció geomètrica (vegeu la figura 4). L'objectiu d'aquestes tècniques és reduir la variabilitat de les fonts d'entrada normalitzant-la i facilitar d'aquesta manera el procés posterior d'extracció d'informació i de reconeixement de text.

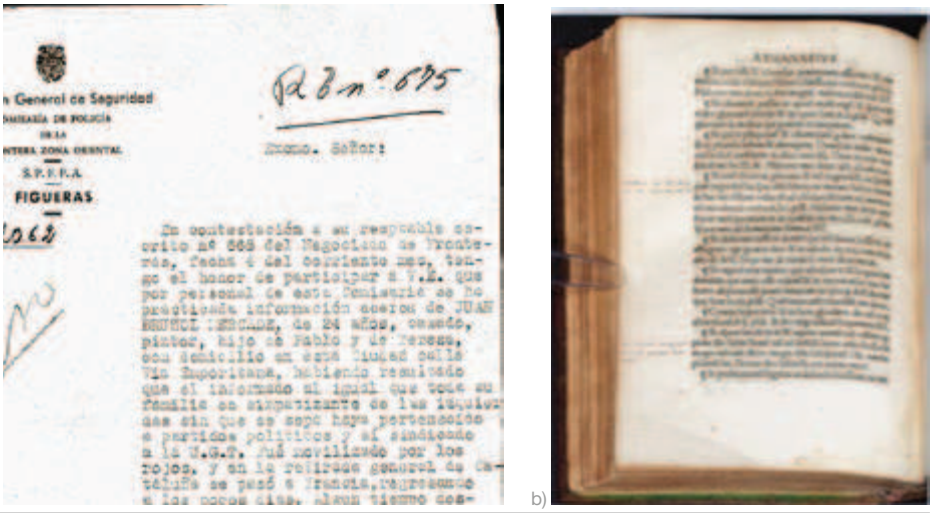
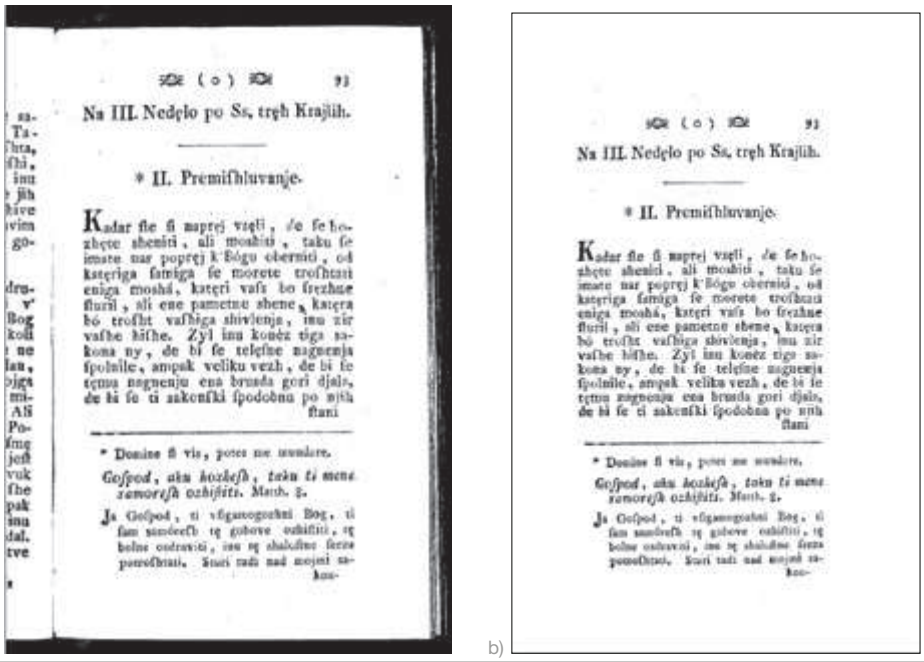


Figura 4. Imatges que cal tractar abans d'extreure'n informació.
a) Imatge amb soroll de l'escàner. b) Imatge d'una pàgina corbada pel llom del llibre.

Una vegada s'ha eliminat el soroll i s'han corregit les deformacions geomètriques, cal identificar la zona central de la pàgina dins la imatge i, a continuació, cadascuna de les regions i el tipus. A la figura 5 es pot veure un exemple del primer procés. A la figura 5a es veu la imatge d'una pàgina d'un llibre digitalitzat on es pot observar l'ombra fosca al voltant del llibre i un tros de text que prové de la pàgina de la dreta. A la figura 5b es veu com hem estat capaços d'eliminar la informació del voltant que no pertany al contingut de la pàgina. Aquest procés de «neteja» del document és necessari per evitar errors de reconeixement i interpretació en les etapes posteriors.



a) b)

Figura 5. Exemple de document en què s'ha detectat i extret la informació de la pàgina.
a) Imatge original. b) Imatge en què s'ha identificat el cos central de la pàgina i s'ha eliminat l'ombra dels contorns i el text de la pàgina precedent.

2.2 L'ANÀLISI DE L'ESTRUCTURA

Un cop hem pal·liat, en la mesura del possible, les degradacions intrínseques dels documents i les hem rectificat i normalitzat per reduir la variabilitat de les fonts d'entrada, cal extreure'n la informació. En l'esquema clàssic aquesta anàlisi de l'estructura es divideix en diferents etapes, que es van aplicant de manera

consecutiva. En la primera etapa s'identifica la informació de l'estructura del document, és a dir, si està organitzat en columnes, si hi ha imatges, figures o taules, etc., o bé si una regió de text és el títol d'un capítol, d'una secció, un peu de pàgina, etc. Aquesta anàlisi es fa en dos nivells: el nivell físic i el nivell lògic.

L'anàlisi física de l'estructura té per objectiu identificar la natura de les regions en què es descompon el document. Pel que fa al reconeixement, no és el mateix tractar amb regions de text que amb imatges, gràfics o informació tabular. En funció del tipus de regió, les tècniques que cal aplicar en cada cas seran diferents. L'estructura física pot servir per identificar la tipologia del document, ja que la distribució d'elements és sovint característica d'una categoria.

L'anàlisi lògica de l'estructura té per objectiu associar les regions físiques a categories i extreure relacions semàntiques entre elles. Per exemple, tot i ser blocs de text, no és el mateix un títol que una capçalera de secció o un paràgraf. En un document estructurat en forma de formulari, una vegada segmentats els diferents blocs, cal saber a quina variable corresponen. El coneixement de l'estructura lògica ajuda al reconeixement, ja que el ROC pot anar a «cercar» certs elements d'informació a determinades posicions de la pàgina i parametritzar-se adequadament. Per exemple, les capçaleres de les seccions, els peus de figures o els números de pàgina poden ajudar a identificar de manera més precisa els continguts dels documents i millorar el procés d'extracció de continguts. En podem veure un exemple a la figura 6, on es mostra una pàgina d'una revista i diverses regions detectades i anotades segons els dos nivells d'anàlisi. En vermell identifiquem les regions de text i en groc les imatges. Al costat de cada regió podem veure una etiqueta amb el tipus de regió de què es tracta: capçaleres, paràgrafs, imatges, peus d'imatges o anotacions (estructura lògica).

Figura 6. Pàgina de revista amb el resultat de fer una anàlisi física i una de lògica de l'estructura del document. En vermell identifiquem les regions de text i en groc les imatges. Al costat de cada regió podem veure una etiqueta amb el tipus de regió: capçaleres, paràgrafs, imatges, peu d'imatges o anotacions.

Una vegada s'ha trobat l'estructura física del document, cal tractar cadascuna de les regions segons el tipus de què es tracti. Així, s'extrauran les imatges del document i s'indexaran, es processaran les taules i se n'extraurà el contingut, i es continuaran processant les regions de text per obtenir la transcripció final.

A continuació descriurem breument alguns dels processos principals que es duen a terme per a cadascun d'aquests tipus de regions.

2.3 EL RECONEIXEMENT ÒPTIC DE CARÀCTERS (ROC)

Com hem comentat a la introducció, els sistemes de ROC són programaris especialitzats que converteixen imatges de documents de text en la seva corresponent versió digital en format editable per un processador de text. De manera genèrica, aquests programes funcionen d'una manera semblant a la descrita en aquest article. Hi ha una primera fase de pretractament en què es corregeixen les degradacions intrínseques dels documents i es normalitza la representació. A continuació hi ha la fase d'anàlisi de l'estructura, en què s'identifiquen les diferents regions del document i es busquen aquelles que són text imprès. Una vegada s'han identificat les regions de text, s'apliquen els algorismes específics de reconeixement de text i que explicarem breument a continuació.

En el cas de text imprès, tots els programaris comercials treballen a partir de regions de text. Primer, s'identifiquen les línies de text; després, les paraules, fins que es retallen les imatges al nivell dels caràcters. És a partir de la forma de cada caràcter que es fa el reconeixement. Aquest reconeixement es fa per comparació de les característiques de la forma amb un conjunt de models apresos per a cada lletra. Les característiques de les lletres utilitzades són diverses i inclouen aspectes geomètrics (alçada, amplada, curvatura, etc.) o topològics (posició relativa entre els components, forats, etc.). Una vegada identificada quina és la lletra amb més probabilitat de cada imatge, és quan es fa servir informació pròpia de l'idioma i diccionaris per reduir els errors de reconeixement. Aquesta informació, que es coneix com a *model de llenguatge*, utilitza les probabilitats de combinació de lletres consecutives (tècnicament es coneixen com a n-*grames*).

Figura 7. Exemples de documents on es pot esperar un nivell de rendiment diferent dels ROC comercials. a) Bon rendiment. b) Rendiment mitjà. c) Rendiment deficient.

A la figura 7 es poden veure exemples de diferents tipus de document amb l'expectativa de rendiment dels ROC comercials. Aquest agrupament l'hem fet sobre la base de la nostra experiència i és el resultat d'haver-los utilitzat en diferents projectes de recerca i transferència. En el primer grup trobem els documents que es poden generar amb qualsevol processador de text d'ús domèstic amb una estructura simple (una columna o dues), fent servir un tipus de lletra estàndard (Arial, Times, Helvetica, etc.), impresos i escanejats amb dispositius comuns de qualitats i resolucions normals (de 200 a 300 ppp). Per a aquests tipus de documents, qualsevol programari de ROC serà capaç de fer una conversió a document digital acceptable (amb pocs errors).

En el segon grup de documents, en trobem alguns amb una estructura més complexa. En aquest cas, variabilitat de documents i complexitat són gairebé sinònims. Si les pàgines contenen taules, tipus de lletres amb mides molt diferents, una estructura poc regular, fonts de lletra poc habituals, etc., pot donar lloc a un rendiment dels ROC molt variable. En el millor dels casos, en què els algorismes d'anàlisi de l'estructura del document funcionen mitjanament bé, podem esperar que el programari sigui capaç d'identificar bé les regions de text i tenir uns resultats acceptables. En les regions on la mida de la font és molt petita o la resolució de la imatge és baixa (72 o 96 ppp), els algorismes de detecció de línies de text, de paraules i d'imatges de caràcters tindran molts problemes per retallar correctament les lletres i reconèixer-les. A més, és molt probable que les taules, i el text que puguin contenir, no es reconeguin correctament.

En l'últim grup de documents trobem aquells en què no hi ha grans regions de text i que tenen força elements gràfics i digitalitzats amb un nivell de qualitat molt baix. Les imatges de fotocòpies de documents o de faxos entrarien en aquest darrer grup. També podríem incloure en aquest grup les imatges de documents fotografiats amb dispositius mòbils en condicions poc controlades (imatges desenfocades amb reflexos, etc.). En aquests casos, els ROC comercials fallen perquè la qualitat de la imatge no és suficient per obtenir resultats mínimament satisfactoris en l'etapa d'anàlisi de l'estructura i el reconeixement de text posterior.

2.4 EL RECONEIXEMENT D'ESCRIPTURA MANUSCRITA

Els programaris comercials de ROC no serveixen per reconèixer el text manuscrit. La principal raó és que aquests programes retallen les imatges al nivell de les lletres i en textos escrits a mà, quan es tracta d'escriptura lligada, és molt difícil retallar-les de manera neta. A més, la gran variabilitat d'estils d'escriptura fa encara més difícil reconèixer què hi ha escrit. En altres paraules, no es pot esperar que una tasca que segons la lletra és pràcticament impossible per a les persones, es pugui automatitzar i generalitzar amb un programa. Tot i això, en certs àmbits d'aplicació molt controlats es pot arribar a obtenir uns bons resultats. L'exemple paradigmàtic és la lectura d'adreces postals (vegeu la figura 8a). Pràcticament tots els serveis postals del món tenen sistemes de lectura automatitzats que processen centenars de cartes i postals per segon. Aquests no són sistemes que fan una lectura completa de l'adreça del destinatari, sinó que busquen paraules concretes, com ara noms de ciutats i de països i codis postals. Es tracta, doncs, d'una tasca molt específica amb un vocabulari molt restringit, que permet desenvolupar sistemes que funcionin bé.

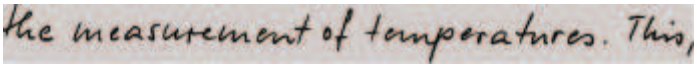
a)

b)

Figura 8. Exemples de documents manuscrits. a) Adreces postals. b) Documents històrics.

Quan es tracta d'altres tipus de documents, com ara la correspondència personal, els dietaris o altres tipus de documents personals, la tasca de reconeixement es complica significativament. En aquest cas, els continguts dels documents poden ser tan amplis que l'ús de vocabulari específic no sigui suficient per assolir quotes de rendiment similars a les dels ROC per a textos impresos.

Per a aquests tipus de documents, el procés de reconeixement és una mica diferent del que es fa servir per reconèixer text imprès. Com hem comentat a l'inici d'aquesta secció, no es pot esperar que un sistema de reconeixement de text manuscrit sigui capaç de retallar les imatges al nivell del caràcter i, per tant, aquests programes només retallen les imatges fins al nivell de les línies de text. Tanmateix, s'han d'aplicar altres tècniques de processament d'imatges per reduir la variabilitat intrínseca a l'escriptura manual (vegeu la figura 9). Un cop s'han aplicat tots aquests processos, s'obté la imatge de sortida. És aquesta imatge la que es dóna al motor de reconeixement de text manuscrit perquè intenti reconèixer-ne el contingut.

a) 

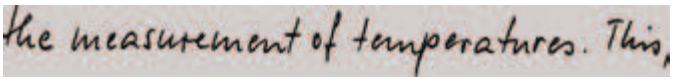
b) 

Figura 9. Resultat d'aplicar tècniques de processament d'imatges per reduir la variabilitat del text manuscrit. a) Imatge d'entrada tal com s'ha retallat del document original. b) Imatge resultant.

Els motors de reconeixement de text manuscrit, segons Plötz i Fink, són força més sofisticats que els de ROC i requereixen moltes més dades per fer-los funcionar. Són sistemes que requereixen una primera fase d'entrenament, és a dir, cal que li donem molts exemples d'imatges de documents, amb les transcripcions corresponents, perquè els sistemes aprenguin les lletres. A més, perquè faci ús de la informació contextual, li hem de donar un conjunt ampli de diccionaris amb les paraules que pensem que es pugui trobar i un model de llenguatge similar als que s'utilitzen per als ROC. Amb aquests dos ingredients (exemples de documents transcrits i diccionaris amb el vocabulari) executem l'algorisme d'aprenentatge.

L'aparició d'ordinadors cada vegada més potents permet que actualment es pugui processar una gran quantitat de dades en unes poques hores, quan fa uns anys requerien de dies a setmanes de càlcul. A més, com més dades, millor rendiment dels sistemes i més capacitat d'adaptació al volum de dades diverses. Això últim fa que aquests tipus de tecnologies es puguin aplicar cada vegada més a un ampli ventall d'aplicacions.

2.5 LA INDEXACIÓ I L'ACCÉS PER CONTINGUT

En alguns casos la transcripció completa dels documents no és necessària, i amb la detecció de certes paraules rellevants (persones, llocs, fets històrics) bastaria per indexar-ne el contingut. Per aquest motiu han sorgit les tècniques de *word spotting* (detecció de paraules clau) amb l'objectiu de localitzar totes les instàncies d'una paraula donada dins una col·lecció. Les tècniques de *word spotting*, segons Almazán et al. (2014) i Rusiñol et al. (2015) es categoritzen en dos grups, depenent de si l'usuari proveeix una imatge de la paraula o bé si hi introdueix una cadena de caràcters mitjançant el teclat (vegeu la figura 10).

En el primer cas, anomenat *query-by-example*, l'usuari selecciona del document una imatge de la paraula que vol cercar (l'exemple), i el sistema retorna totes les imatges de paraules visualment semblants que ha trobat a la col·lecció. L'avantatge és que l'algorisme fa una cerca d'elements semblants visualment; per tant, no necessita fer cap aprenentatge de l'estil d'escriptura. L'inconvenient és que no és capaç de trobar paraules escrites per autors diferents, ja que l'aparença visual de les paraules canvia radicalment a causa de l'estil d'escriptura particular de cadascú.

En el segon cas, anomenat *query-by-string*, l'usuari escriu pel teclat la paraula que vol cercar. El sistema és més flexible en el sentit que l'usuari, en primer lloc, no ha de buscar cap instància de la paraula que vol cercar i, en segon lloc, que no està limitat a l'estil d'escriptura determinat. En contrapartida, per poder cercar dins el text, l'algorisme necessita aprendre l'aparença visual d'aquesta paraula i, per tant, els estils d'escriptura dins aquesta col·lecció. Això implica que l'algorisme necessita dades etiquetades per fer l'aprenentatge.

Figura 10. Il·lustració de les diferències entre les dues famílies de *word spotting*, segons si l'entrada és una cadena de text o una imatge de la paraula. En vermell es mostren les paraules trobades al document.

2.6 EL RECONEIXEMENT D'ELEMENTS GRÀFICS

Certs tipus de documents contenen informació rellevant en forma de gràfic en lloc de text. Per exemple, en mapes, plànols d'arquitectura, diagrames electrònics, partitures musicals, diagrames de flux, documents administratius, etc., la informació rellevant està formada per un conjunt de símbols gràfics i per les relacions espacials entre ells. Per tal de poder interpretar aquests documents i poder-ne extreure informació de manera automàtica, cal saber reconèixer aquests símbols gràfics per poder-los associar a un determinat significat. En un cert sentit, aquesta tasca és més complexa que no pas el reconeixement de text. El reconeixement de text parteix de la base que els caràcters que cal reconèixer poden ser transformats en un senyal lineal (llegint-los d'esquerra a dreta), mentre que aquest paradigma no és vàlid si es vol reconèixer símbols gràfics que no segueixen necessàriament aquesta linealitat.

2.7 LA IDENTIFICACIÓ DE L'ESCRIPTOR I LA DATACIÓ

L'anàlisi forense de documents consisteix a analitzar l'estil d'escriptura. En aquest cas, en lloc de focalitzar-se en el que s'ha escrit, l'interès recau en la manera com s'ha escrit. Així es pot determinar l'autor d'un document (identificació de l'escriptor) o validar-ne l'autenticitat (com la verificació de signatures), o bé datar-lo segons un període històric.

Normalment, les tècniques es basen en l'aprenentatge de les característiques de l'estil d'escriptura i es poden dividir en dues famílies. Primerament, les tècniques es poden basar en la forma de determinades lletres, paraules, símbols o signatures (vegeu els requadres verds i blaus de la figura 11). Segonament, es poden basar en aspectes globals com la curvatura o l'orientació del traç, les anomenades *tècniques independents del text* (vegeu les marques vermelles de la figura 11). Tant en un cas com en l'altre, les tècniques es poden aplicar tant a documents textuals com gràfics (Gordo, Fornés, Valveny, 2013).

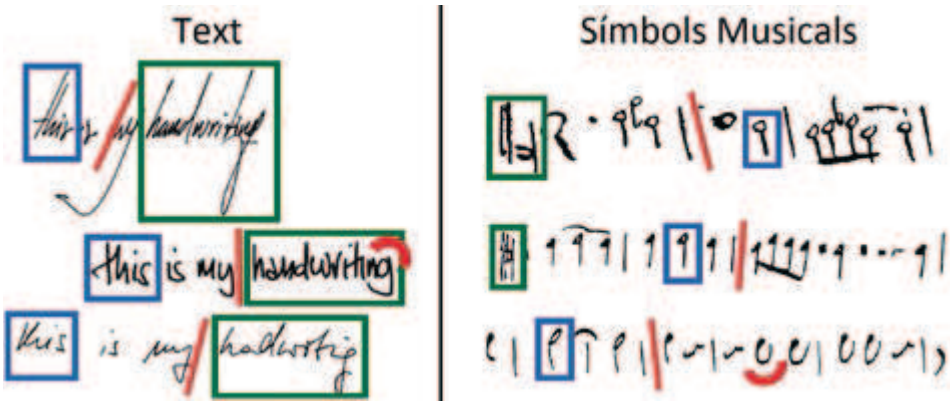


Figura 11. Característiques emprades per a la identificació de l'escriptor. A l'esquerra es mostren tres línies de text escrites per diferents escriptors. A la dreta es mostren els símbols musicals de tres músics diferents. Algunes tècniques es basen en la forma de determinats elements (requadres blaus i verds), mentre que d'altres es fixen en aspectes globals (en vermell).

3. EINES DE TRANSCRIPCIÓ COL·LABORATIVA (*CROWDSOURCING*)

Ateses les dificultats dels algorismes de reconeixement en fonts manuscrites de gran volum, darrerament han emergit amb força i gran eficiència solucions basades en el paradigma de la transcripció massiva (*crowdsourcing*).

Crowdsourcing, de l'anglès *crowd* 'multitud' i *outsourcing* 'externalització', es podria traduir al català com 'col·laboració oberta distribuïda'. El *crowdsourcing* consisteix a dividir una gran tasca complexa en moltes de petites i col·laboratives, distribuïdes entre un grup nombrós de persones o una comunitat mitjançant una convocatòria oberta. Un exemple pioner i reeixit d'aquest paradigma ha estat Wikipedia, una enciclopèdia oberta a Internet en què qualsevol usuari pot afegir articles respectant unes normes de conducta. Wikipedia té actualment un ritme d'entrada de dades de 300 articles diaris.

En l'àmbit del reconeixement d'imatges de documents, el *crowdsourcing* consisteix a implementar plataformes web en què molts usuaris col·laboren en la transcripció (Fornés, Lladós, Mas, Pujades, Cabré, 2014). Per exemple, cada usuari transcriu unes poques pàgines o valida un conjunt de transcripcions automàtiques fetes per ordinador. En el cas de manuscrits històrics, destaca la Family History Library, de l'empresa Family Search, per construir una xarxa històrica de tot el món. Durant anys s'han digitalitzat milions de documents i s'han entrat enllaços creuats de dades genealògiques. Aquest procés el fan manualment milers de voluntaris que utilitzen una plataforma de *crowdsourcing*. D'aquesta manera, el *crowdsourcing* és també un instrument d'innovació social i fa que els consumidors dels actius històrics i culturals passin a ser actors en el procés de producció del coneixement. Darrerament s'està popularitzant el *gamesourcing* (Alavau, Leiva, 2012), que consisteix a fer la transcripció a través d'un joc (per exemple, un joc de Facebook o per a dispositius mòbils Android) i així fer la tasca més atractiva i divertida per als usuaris.

4. CAS D'APLICACIÓ 1: RECONEIXEMENT DE DOCUMENTS DEMOGRÀFICS MANUSCRITS HISTÒRICS

Les fonts de naturalesa demogràfica com els padrons, els registres parroquials o civils de baptismes o naixements, matrimonis i defuncions, etc., ofereixen una gran riquesa d'informació que no solament permeten estudiar el comportament demogràfic, sinó també entendre i explicar millor l'evolució social i econòmica del passat.

Tal com s'ha comentat en la introducció, un cop acabada la seva digitalització per assegurar-ne la preservació, el repte és l'extracció automàtica de continguts per permetre'n així la valorització, l'accés i la interpretació. En

aquest escenari, en lloc d'una transcripció paraula per paraula, l'objectiu és extreure'n i interpretar-ne les entitats nominals (cognoms, llocs, oficis, dates) per crear les bases de dades demogràfiques. Per aquest motiu, les tècniques aplicades es basen en el reconeixement semàntic i en la transcripció col·laborativa emprant el *word spotting*. Per visualitzar un vídeo demostratiu, visiteu: http://dag.cvc.uab.es/infoesposalles/media-gallery/.

Reconeixement semàntic de les esposalles de la catedral de Barcelona

La col·lecció de registres matrimonials de la catedral de Barcelona (anomenats *Llibres d'esposalles*) està formada per 287 volums i conté informació de més de 600.000 matrimonis contrets entre el 1451 i el 1905 en 250 parròquies de la diòcesi de Barcelona. Les esposalles contenen informació dels esposos i els seus pares, els seus oficis i llocs d'origen, la parròquia on es van casar i la taxa que van pagar segons el seu estatus socioeconòmic. La continuïtat d'aquesta font al llarg dels segles permet fer estudis demogràfics, socials, migratoris i genealògics.

En primer lloc, s'ha dissenyat una eina de transcripció col·laborativa (*crowdsourcing*) a través del web (vegeu la figura 12) en la qual més de 50 transcriptors han col·laborat en paral·lel en el buidat dels *Llibres d'esposalles*. L'aplicació a través del web ha permès no solament facilitar la visualització i l'entrada de dades, sinó també que els experts gestionin i monitorin la feina dels transcriptors. Com a resultat, la transcripció ha acabat en menys de dos anys.

Figura 12. Aplicació de crowdsourcing per al buidat de les esposalles de Barcelona.

En segon lloc, s'han dedicat esforços al desenvolupament de tècniques de reconeixement semàntic (Romero, Fornés, Vidal, Sánchez, 2016). L'objectiu és transcriure el document i alhora detectar les entitats nominals i classificar-les en categories semàntiques (cognoms, llocs, oficis, etc.). Com a resultat, es podria crear automàticament una base de dades que només caldria ser validada per un expert, amb la qual cosa s'estalvia un gran volum d'hores de feina.

La metodologia desenvolupada ha consistit en la combinació de tècniques de reconeixement de text manuscrit (models ocults de Markov) amb models de llenguatge basats en categories semàntiques, per treure profit, d'aquesta manera, de l'estructura sintàctica de cada desposori. En concret, el model de llenguatge s'ha après de manera automàtica gràcies a tècniques d'inferència gramatical (per exemple, la *morphic generator grammatical inference*). A tall informatiu, destaquem que els resultats obtinguts sobre un volum de la col·lecció (amb 175 pàgines) han mostrat que prop del 75% de les entitats nominals s'han detectat, transcrit i categoritzat correctament. Per a més informació, consulteu consulteu Romero et al. (2016).

Transcripció interactiva de padrons emprant el *word spotting*

Els padrons són els recomptes municipals nominatius de població per excel·lència. Contenen informació sociodemogràfica dels habitants d'un municipi (per exemple, l'edat, l'estat civil, l'ocupació, l'alfabetització, la relació amb el cap de la llar) i permeten fer estudis sobre la distribució de la població, l'estructura socioocupacional, la fecunditat, el tipus de famílies o les característiques de l'habitatge. Però el vertader potencial és la utilització dels padrons des d'una òptica longitudinal gràcies a la vinculació dels diferents padrons, que permet resseguir la trajectòria vital dels individus transcendint l'espai geogràfic d'un municipi o parròquia. Com a resultat es crearien xarxes socials històriques (semblants a Facebook o LinkedIn) que permetrien estudiar els moviments migratoris o les formacions familiars dinàmicament, en el temps i en l'espai. A més, els padrons gaudeixen d'una estandardització de format en l'àmbit europeu que permetria la integració i la consolidació a escala global dels seus continguts.

Amb l'objectiu d'agilitzar el buidat del gran volum de padrons existents, s'ha dissenyat una plataforma de *crowdsourcing* interactiva que integra la tècnica de localització de paraules claus (*word spotting*); alhora, el sistema es beneficia de la informació redundant en padrons previs (vegeu la figura 13).

Atès que els padrons es duen a terme cada pocs anys, la informació dels individus que viuen a cada llar sol ser bastant estable. Aquesta redundància és precisament la informació que es pot transferir d'un padró al següent. Per tant, quan el transcriptor comença a buidar un nou padró, el sistema reconeix el carrer i el número de la llar i recupera la informació d'un padró anterior ja transcrit. Llavors el sistema aplica el *word spotting (query-by-string)* per cercar cada un dels individus que vivia en aquella llar al nou cens. Els individus localitzats són automàticament incorporats a la base de dades. El transcriptor només ha d'actualitzar els camps corresponents (per exemple, la seva edat o l'ofici) i afegir els individus de nova aparició. Com a resultat, el temps de transcripció es redueix considerablement. Per a més detalls, consulteu Mas et al. (2016).

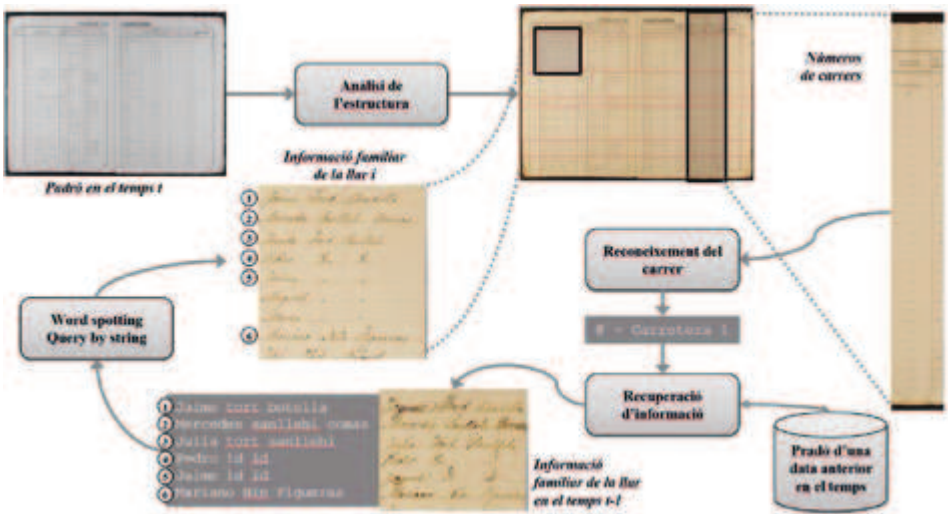


Figura 13. Esquema de la transcripció interactiva de padrons amb la integració de tècniques de *word spotting*.

5. CAS D'APLICACIÓ 2: CLASSIFICACIÓ AUTOMÀTICA DE DOCUMENTS ADMINISTRATIUS EN ENTORNS BANCARIS

Al món empresarial hi ha molts processos en els quals apareix una gran quantitat de documentació. Per tal de tractar aquesta informació, normalment encara es requereix en gran manera una intervenció humana, la qual cosa té associats

uns elevats costos econòmics. És en aquest punt que tractar d'automatitzar certs processos d'interpretació dels documents pot ser molt beneficiós.

Posem per cas els passos que cal seguir a l'hora de contractar un cert producte financer al sector bancari. Normalment, com a clients, abans de poder contractar un producte, se'ns demanarà d'aportar un seguit de documents per tal d'acreditar si l'operació és viable o no des del punt de vista del banc i per veure si se'ns concedeix o no un crèdit, un aval o una hipoteca, i amb quines condicions. Tota aquesta documentació es digitalitza a les oficines i s'envia als serveis centrals per tal que experts l'estudiïn. Per facilitar la tasca dels avaluadors, el primer pas que s'ha de fer és analitzar el conjunt d'imatges rebudes per determinar on comença i on acaba cada document, i de quin tipus es tracta: formularis d'impostos, factures, nòmines, llibres de família, balanços de comptes d'una empresa, etc. Aquest procés, que resulta lent, costós i que no és pas lliure d'errors, es pot automatitzar emprant tècniques de visió per computador i d'anàlisi de documents.

Descripció de documents

Per tal d'automatitzar el procés de classificació, necessitarem un mètode que permeti representar un document de manera numèrica. És el que anomenem *descriptors*. Així doncs, donada una imatge d'entrada, el descriptor extraurà característiques discriminatòries del document per obtenir-ne una representació numèrica. Aquests descriptors hauran de seguir el principi següent per poder ser útils. Si tenim dos documents semblants i en calculem els descriptors, la diferència (distància) entre aquests haurà de ser un valor petit, i contràriament, si calculem els descriptors de dos documents prou diferents, la seva distància haurà de ser elevada.

La classificació d'imatges de documents és un camp de recerca bastant madur i, per tant, han aparegut diversos mètodes a la literatura. Recomanem al lector àvid de més detalls de donar una ullada als articles de revisió de l'estat de la qüestió de Doermann (1998) i Chen i Blostein (2006). En línies generals, els diferents mètodes es poden diferenciar depenent de la representació escollida.

En primer lloc, existeixen mètodes que defineixen les diferents classes de documents en funció de la similitud visual. Els descriptors que es proposen normalment empren estadístiques calculades sobre característiques de baix nivell

(com, per exemple, comptar el nombre de píxels) per tal de codificar quina és l'aparença dels documents. Aquestes tècniques solen ser molt simples, però resulten útils quan es vol agrupar documents que, tot i tenir continguts una mica diferents, tenen una aparença visual prou estable, com pot ser el cas de formularis.

D'altra banda, es poden definir les diferents classes de documents segons la similitud dels seus continguts textuals. Un cop feta una transcripció automàtica dels documents mitjançant alguna eina de ROC, la similitud entre diferents documents es pot expressar com a diferència de les paraules que hi apareixen. Aquests descriptors textuals resulten ideals quan ens enfrontem a documents que «parlen» del mateix tot i que la seva aparença visual pugui ser diferent dins d'una mateixa classe.

En el nostre cas, ens enfrontem, d'una banda, a classes en què la similitud visual és una eina molt potent (com, per exemple, formularis o factures del mateix proveïdor) i, de l'altra, a classes que presenten una similitud textual forta tot i que no segueixin un format i una aparença estàndards (com, per exemple, contractes o informes d'auditoria). Aquest fet motiva que la solució aplicada es basi en la combinació de descriptors de totes dues famílies.

Descriptor visual

Com a descriptor visual n'hem escollit un que es basa en la codificació de la intensitat dels píxels a diversos nivells d'escala. Primer de tot, per tal d'eliminar soroll i detalls petits, les imatges d'entrada es difuminen una mica fent servir un filtre gaussià. Després, seguint la proposta d'Héroux *et al.* (1998), les imatges de documents es parteixen en quatre blocs. En cada un dels blocs es calcula el valor mitjà de la intensitat dels píxels per tal de descriure si una zona és en general més fosca o més clara. Aquest valor es guarda com una de les característiques del vector descriptor. A continuació, cada un dels blocs es torna a dividir en quatre zones i es continua repetint el mateix procés. Podem veure un exemple dels primers nivells de la piràmide a la figura 14, on es pot apreciar com cada vegada tenim una visió més detallada del document que s'ha de tractar. Aquest descriptor, tot i ser extremadament simple, ha demostrat la seva eficiència envers altres propostes més complexes.

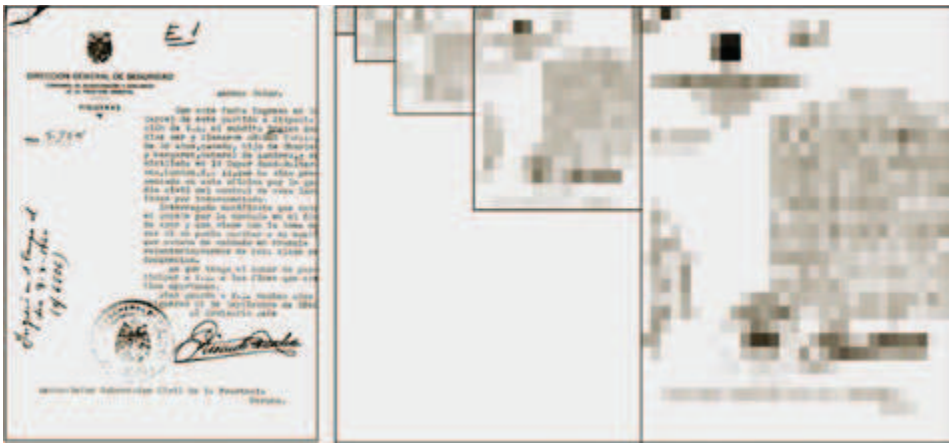


Figura 14: Exemple de visualització del descriptor visual.

Descriptors de contingut textual: dels «sacs de paraules» a l'anàlisi semàntica

Pel que fa a la descripció textual, generalment el que es fa és representar els documents com un conjunt de paraules juntament amb les seves freqüències d'aparició, que és el que es coneix com a representacions de «sacs de paraules» (*bag-of-words model*). El problema, però, és que dos documents poden parlar del mateix tot i fer servir paraules diferents. Per tal de fer front a aquesta limitació, hem escollit una tècnica que representa els continguts textuals a partir de la semàntica latent que hi ha als documents; és a dir, es defineixen els documents segons quins temes tracta i en quina mesura. Aquests temes es poden aprendre de manera automàtica fent servir una tècnica anomenada *latent semantic analysis* com podeu veure a Deerwester et al. (1990).

Classificador estadístic i rebuig

Un cop tenim representats els documents tant amb descriptors visuals com amb descriptors textuals, podem entrenar un classificador per tal que aprengui a classificar els documents entrants mirant-se la seva aparença i els seus continguts.

En una fase de test, emprant un conjunt de més de 40.000 documents, vam obtenir els resultats següents:

Descriptors	Només visual	Només textual	Combinació
Percentatge d'encert	85,54%	94,78%	95,49%

Podem veure que els descriptors textuals permeten obtenir un rendiment molt més elevat que no pas els descriptors visuals; però, si decidim combinar-los, s'aconsegueix un petit guany extra.

Finalment, cal tenir en compte que el classificador, a part de donar com a resposta a quina classe pertany un document entrant, també retorna una certa probabilitat de certesa, que es pot fer servir per tal de rebutjar respostes automàtiques si no n'estem del tot segurs. En aquest sentit, s'arriba a aconseguir un 100% d'encert si es rebutgen un 35% de documents entrants, que es reconduieixen al tractament manual.

Trobarem una descripció molt més detallada d'aquest sistema en l'article publicat a la revista IJDAR per Rusiñol et al. (2014).

6. PERSPECTIVES

En els darrers anys s'ha avançat força en la visió per computador i, en particular, en l'anàlisi de documents. L'avenç de les noves tecnologies, tant de programari com de maquinari, hi han ajudat. Des de la recerca en l'àmbit de les enginyeries s'ha transferit coneixement que s'ha integrat en productes o serveis que avui en dia estan integrats en entorns de productivitat empresarial, en arxius o biblioteques, o en plataformes de gran consum. Un exemple recent és la incorporació del ROC a l'aplicació de traducció de Google per a mòbils, que permet a l'usuari fer una fotografia d'un text i obtenir-ne la traducció en pocs segons. Tanmateix, existeixen encara diversos reptes científics i tecnològics de cara a disposar de serveis de lectura universal. L'arxiver del futur hauria de poder treballar en escriptoris que integrin el físic i el digital. Són les anomenades *interfícies tangibles*. Podem imaginar-nos una taula que integra una càmera aèria i un projector, de manera que el document físic que l'usuari té sobre la taula és digitalitzat per la càmera, reconegut per l'ordinador i, mitjançant el projector, augmentat virtualment amb informació complementària. En aquest tipus d'entorns, la interacció amb l'usuari és molt rellevant si s'integren mecanismes de reconeixement de gestos, anotacions manuscrites sobre tauleta, etc. Actualment hi ha entorns en aquesta direcció com el model Sprout d'HP.

El pas a l'escala és també un repte important. El que fa anys era el tractament de documents en estacions individuals, avui en dia ha evolucionat cap a grans volums documentals (l'era del *big data*). Els ciutadans del futur hauran de poder fer cerques per Internet darrere les quals hi haurà milions de documents provinents d'arxius d'arreu del món. Els algorismes han de ser prou robustos per no baixar el rendiment en aquesta nova dimensió. Lligada a aquest accés universal, s'obre la necessitat dels sistemes intel·ligents, és a dir, no solament de reconeixement i transcripció literal, sinó d'interpretació dels continguts. Només així els sistemes podran relacionar termes en diferents tipologies de documents i en diferents llengües. Aquest problema va més enllà de la visió per computador i requereix altres experteses d'intel·ligència artificial com a sistemes de modelatge i gestió del coneixement, representacions amb ontologies, etc.

Finalment, la ubiqüitat que ens dóna el fet de tenir un mòbil connectat en tot moment fa que els arxius surtin dels seus límits físics. En qualsevol moment i en qualsevol lloc tenim accés a la informació. Les noves tecnologies canviaran, per tant, els modes d'accés a les fonts documentals, democratitzant-les però redefinint els models de servei. Probablement l'accés *in situ* serà només per a historiadors o amb finalitats museístiques pel valor del mateix document. Les eines de gestió documental seran les d'ús diari dels arxivers, que seran els encarregats d'administrar els continguts i els serveis digitals que se'n derivin.

BIBLIOGRAFIA

- J. Almazán, A. Gordo, A. Fornés, E. Valveny. «Word Spotting and Recognition with Embedded Attributes». *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(12), pp. 2552-2566, 2014.
- V. Alabau, L. Leiva. «Transcribing Handwritten Text Images With a Word Soup Game». *Proceedings of Extended Abstracts on Human Factors in Computing Systems*, pp. 2273-2278, 2012.
- N. Chen, D. A. Blostein. «A Survey of Document Image Classification: Problem Statement, Classifier Architecture and Performance Evaluation». *International Journal on Document Analysis and Recognition*, 10(1), pp. 1-16, 2006.
- D. Doermann. «The Indexing and Retrieval of Document Images: A Survey». *Computer Vision Image Understanding*, 70(3), pp. 287-298, 1998.
- S. Deerwester, S. Dumais, G. Furnas, T. Landauer, R. Harshman. «Indexing by Latent Semantic Analysis». *Journal of the American Society for Information Science,* 41(6), pp. 391-407, 1990.
- A. Fornés, J. Lladós, J. Mas, J. M. Pujades, A. Cabré. «A Bimodal Crowdsourcing Platform for Demographic Historical Manuscripts». *Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage*, pp. 103-108, 2014.
- D. Fernández, S. Marinai, J. Lladós, A. Fornés. «Contextual Word Spotting in Historical Manuscripts Using Markov Logic Networks». *Proceedings of the 2nd International Workshop on Historical Document Imaging and Processing*, pp. 36-43, 2013.
- A. Gordo, A. Fornés, E. Valveny. «Writer Identification in Handwritten Musical Scores with Bags of Notes». *Pattern Recognition*, 46(5), pp. 1337-1345, 2013.
- A. Graves, M. Liwicki, S. Fernández, R. Bertolami, H. Bunke, J. Schmidhuber. «A Novel Connectionist System for Unconstrained Handwriting Recognition». *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(5), pp. 855-868, 2009.
- P. Héroux, S. Diana, A. Ribert, E. Trupin. «Classification Method Study for Automatic Form Class Identification». *Proceedings of the Fourteenth International Conference on Pattern Recognition*, pp. 926-928, 1998.
- J. Mas, A. Fornés, J. Lladós. «An Interactive Transcription System of Census Records Using Word-Spotting Based Information Transfer». *Proceedings of the International Workshop on Document Analysis Systems* (DAS), 2016.
- T. Plötz, G. A. Fink. «Markov Models for Offline Handwriting Recognition: A Survey». *International Journal on Document Analysis and Recognition*, 12(4), pp. 269-298, 2009.
- V. Romero, A. Fornés, E. Vidal, J. A. Sánchez. «Using the MGGI Methodology for Category-based Language Modeling in Handwritten Marriage Licenses Books». *Proceedings of the International Conference on Frontiers in Handwriting Recognition*, 2016.
- M. Rusiñol, D. Aldavert, R. Toledo, J. Lladós. «Efficient Segmentation-free Keyword Spotting in Historical Document Collections». *Pattern Recognition*, 48(2), pp. 545-555, 2015.
- M. Rusiñol, V. Frinken, D. Karatzas, A. D. Bagdanov, J. Lladós. «Multimodal Page Classification in Administrative Document Image Streams». *International Journal on Document Analysis and Recognition*, 17(4), pp. 331-341, 2014.

RESUM

La visió per computador és la disciplina de la informàtica que s'encarrega de dissenyar algorismes que interpreten les imatges digitals. Quan les imatges corresponen a documents digitalitzats, estem en la subdisciplina de l'anàlisi i el reconeixement d'imatges de documents. En aquest article fem un repàs de la situació actual d'aquesta tecnologia i de les seves possibilitats d'aplicació en la resolució tant dels problemes més tradicionals de lectura òptica (ROC) com dels que estan associats a altres tipologies de documents com ara els manuscrits, especialment històrics, i els gràfics. En primer lloc, fem un repàs de l'estat de la qüestió de la tecnologia. A continuació, descrivim dos casos pràctics arran de projectes duts a terme al Centre de Visió per Computador de la UAB i de rellevància en l'àmbit arxivístic: l'anàlisi massiva de documents administratius i de documents demogràfics manuscrits històrics.

RESUMEN

La visión por computador es la disciplina informática que se encarga de diseñar algoritmos que interpretan las imágenes digitales. Cuando las imágenes corresponden a documentos digitalizados, nos encontramos en la subdisciplina del análisis y reconocimiento de imágenes de documentos. En este artículo hacemos un repaso de la situación actual de esta tecnología y de sus posibilidades de aplicación en la resolución tanto de los problemas más tradicionales de lectura óptica (ROC), como de los que están asociados a otras tipologías de documentos tales como los manuscritos, especialmente los históricos, y los gráficos. En primer lugar, hacemos un repaso del estado de la cuestión de la tecnología. A continuación describimos dos casos prácticos a raíz de proyectos llevados a cabo en el Centro de Visión por Computador de la UAB y de relevancia en el ámbito archivístico: el análisis masivo de documentos administrativos y de documentos demográficos históricos manuscritos.

RÉSUMÉ

La vision par ordinateur désigne la discipline informatique qui permet de concevoir des algorithmes pour interpréter les images numériques. Lorsque les images correspondent à des documents numérisés, il s'agit de la sous-discipline de l'analyse et de la reconnaissance d'images de documents. Cet article est l'occasion de faire un état des lieux de la situation actuelle de cette technologie et de ses possibilités d'application, aussi bien en ce qui concerne la résolution des problèmes les plus courants en matière de reconnaissance optique de caractères (OCR) que de ceux associés à d'autres types de documents, dont, notamment, les manuscrits historiques et les graphiques. Tout d'abord, nous ferons le point sur les progrès technologiques. Nous décrirons ensuite deux cas pratiques de projets mis en œuvre dans le Centre de vision par ordinateur de l'Université autonome de Barcelone (UAB) et ayant une incidence dans le domaine de l'archivistique : l'analyse de masse des documents administratifs et des documents démographiques manuscrits historiques.

ABSTRACT

Computer vision is a discipline of Computer Science that designs algorithms that interpret digital images. When these images involve digitalized documents, we enter the subdiscipline of document image analysis and recognition. This article reviews the current state of this technology and its possible application to the more traditional problem-solving issues in optical character recognition (OCR) as well as those associated with other types of documents, such as manuscripts, particularly of a historical nature, and graphics. Firstly, we review the state of the art of the technology. We then describe two case studies of projects carried out at the UAB's Computer Vision Centre that are of relevance to archival science: the massive analysis of administrative documents and of historical handwritten demographic documents.