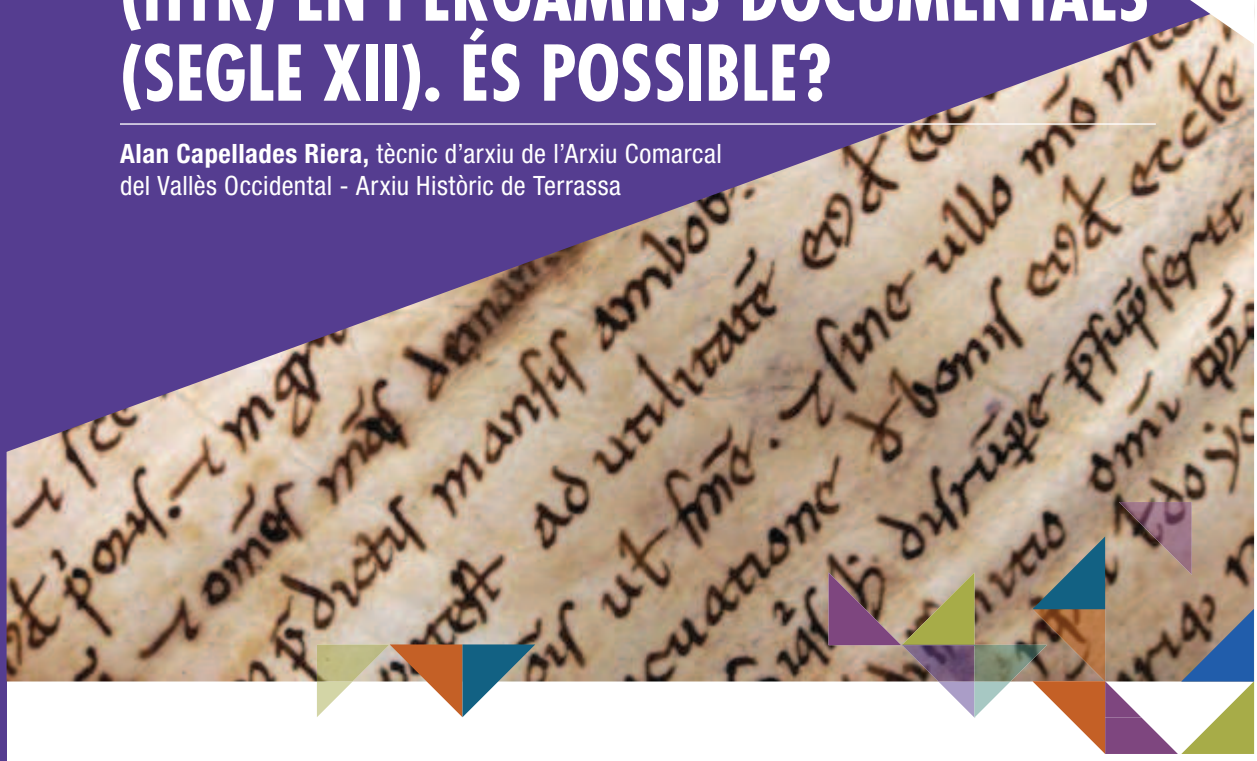
HANDWRITTEN TEXT RECOGNITION (HTR) EN PERGAMINS DOCUMENTALS (SEGLE XII). ÉS POSSIBLE?

Alan Capellades Riera, tècnic d'arxiu de l'Arxiu Comarcal del Vallès Occidental - Arxiu Històric de Terrassa



INTRODUCCIÓ

Els usuaris experts –així és com ens anomenen els tecnòlegs als paleògrafs– tenim davant nostre una oportunitat d'or amb els diferents projectes del Centre de Visió per Computador (CVC) de la Universitat Autònoma de Barcelona (UAB). Volem destacar entre tots ells el Five Centuries of Marriage[1] i el projecte EINES,[2] que, amb la col·laboració del Centre d'Estudis Demogràfics (CED), s'han proposat sistematitzar diversos registres històrics de població.

Davant de propostes tan innovadores, la resposta d'un professional que està en contacte amb documentació de l'Antic Règim no va ser la indiferència, sinó una nova pregunta: seria possible aplicar-ho en una documentació diferent, com ara els pergamins documentals? El present estudi té per objectiu respondre aquesta

pregunta, però no serà amb un monosíl·lab. Ara bé, si el lector espera trobar-se amb resultats reveladors de transcripcions íntegres de pergamins prement un únic botó, està ben equivocat. El reconeixement òptic de caràcters aplicat a l'escriptura manuscrita encara està lluny d'assolir aquesta fita. Calen certes condicions prèvies, com ara digitalitzacions en alta resolució, documents en bon estat de conservació i documents originals escrits per una mateixa mà i íntegrament descrits. Requeriments difícils d'aconseguir, però possibles per a un arxiver que coneix els fons documentals.

L'objectiu d'aquesta investigació ha estat experimentar amb les eines informàtiques de què disposa el CVC per tal d'analitzar-ne els resultats i treure'n conclusions útils per a la recerca històrica o per a la millora de la tecnologia vigent.

1. ENTENDRE (UNA MICA) EL *HANDWRITTEN TEXT RECOGNITION* (HTR)

Abans de començar, volem recomanar-vos la lectura dels tècnics del CVC sobre la visió per computador com a eina per a la interpretació automàtica de fonts documentals. No ens entretindrem a explicar en què consisteix l'OCR ni en els seus mètodes associats, com els processos de millora d'imatge. El nostre punt de partida és l'escriptura manuscrita i saber per què és tan difícil el seu reconeixement automàtic.[3]

L'escriptura manuscrita és un exercici neuronal vinculat a la musculació. El moviment de l'escriptura, a base de repetició, és tan continu i estable que l'escriptura es converteix en una codificació neuronal. Així doncs, l'escriptura d'una sola mà pot ser prou regular com per ser reconeguda per la màquina. Ara bé, en tractar-se d'un exercici neuronal, també pot quedar alterada per malalties neuronals (com l'estrès) i, en bona part, per l'edat. És possible esbrinar amb mètodes quantitatius i científics, com els que proporciona l'HTR, l'edat d'una persona a partir de la seva escriptura.[4] Ara bé, cal que l'escriptura tingui un nombre limitat de variacions i això s'assoleix delimitant el reconeixement a una sola mà, és a dir, centrant-se només en un únic escrivent, com hem fet nosaltres en aquest estudi.

La dificultat que representa el reconeixement d'escriptures manuscrites i les seves particularitats han exigit als enginyers informàtics que siguin creatius. Un

exemple és convertir el document en una imatge 3D mitjançant la geometria diferencial. Els punts més foscos s'interpreten com a elevacions, mentre que l'absència de negre o color s'entén com a depressions; d'aquesta manera, la segmentació automàtica és possible.

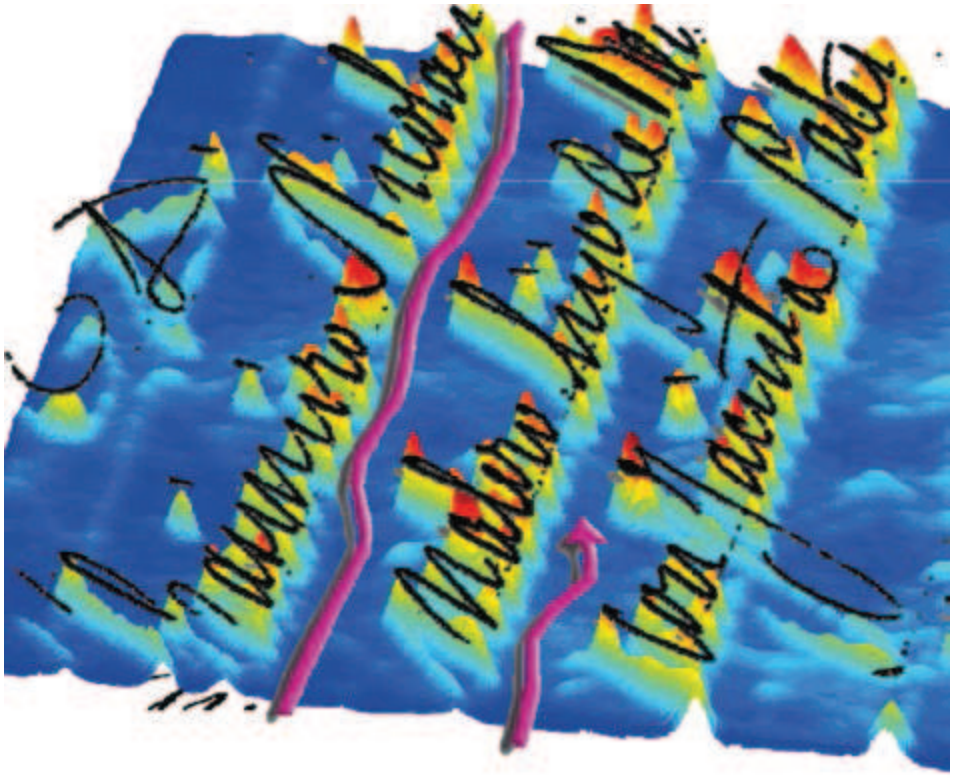


Figura 1[5]

Però aquesta no és l'única solució proposada. Per resoldre una empresa tan ambiciosa cal combinar diversos mètodes. L'aplicació del 3D, per exemple, no resol la superposició de les astes i els fusos. Per aquest motiu, altres especialistes han proposat abordar l'escriptura manuscrita d'una manera diferent: tractar la paraula com un tot, és a dir, identificar la paraula com una seqüència de característiques, com ara l'extracció de *profiles*, *graphs* i histogrames. Tots aquests mètodes són complementaris i tenen com a objectiu traduir la imatge en un llenguatge que la màquina pugui entendre i representar en vector, com l'exemple que mostrem a continuació.

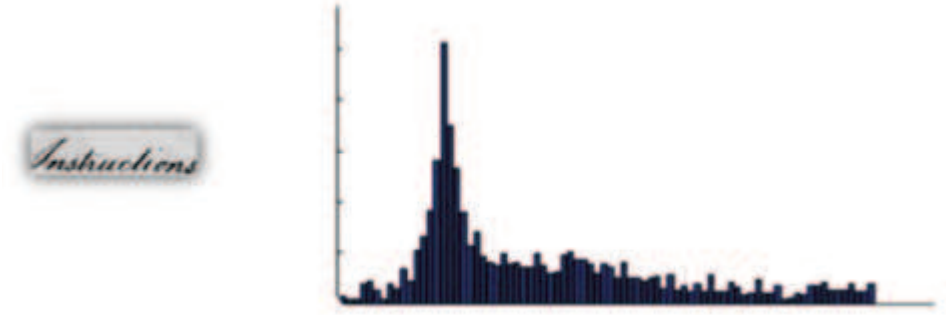


Figura 2[6]

El perfeccionament d'aquests mètodes i la seva combinació han permès dur a terme càlculs de probabilitat, com ara el mètode de reconeixement dels **models ocults de Markov**, el qual afegeix un model de llenguatge que restringeix un seguit de combinacions de lletres: les que no apareixen en un diccionari o en una base de models de paraules.

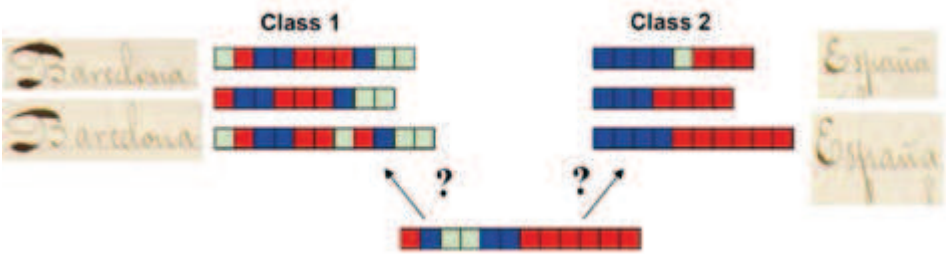


Figura 3[7]

Seguint aquest darrer apunt, hem de destacar que, sigui quin sigui el mètode de reconeixement, el **context** és clau. De fet, es reconeix més per context que per reconeixement directe, sobretot quan es determina el model de llenguatge, és a dir, l'estructura lèxica i de llenguatge, com ara determinar si existeix una alta probabilitat de reconèixer números abans de la paraula *solidus* (un tipus de moneda). Aquest mètode de contextualitzar les paraules es coneix per **n-*grames*** o, dit d'una altra manera, la combinació de lletres.

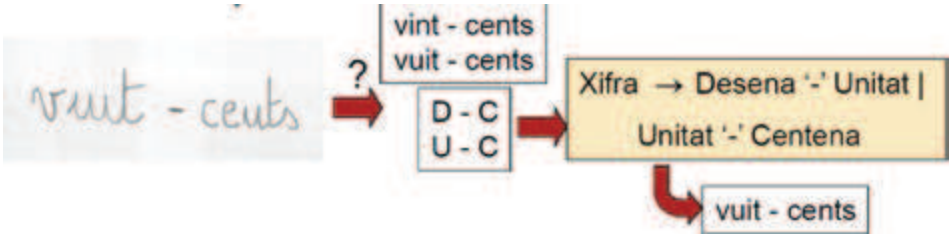


Figura 4[8]

Quan la màquina està davant d'una incertesa, pren una decisió d'acord amb la probabilitat més alta que li oferiex l'*n*-grama.

Un altre mètode per contextualitzar les paraules és mitjançant **diccionaris**, els quals permeten la correcció ortogràfica de les paraules reconegudes, però tenen un problema: què cal fer amb aquelles paraules que no preveu un **diccionari tancat**, com ara noms propis o topònims. Hi ha una alternativa: que el diccionari acceptés paraules noves, és a dir, un **diccionari obert**; en contrapartida, és possible que la correcció ortogràfica no sigui del tot efectiva. Per a l'aplicació adequada d'aquest mètode caldria disposar d'un diccionari adaptat a la documentació que es volgués tractar. Per exemple, en el nostre cas d'estudi seria suficient un tesaurus de paraules extretes de centenars de pergamins transcrits d'un mateix àmbit geogràfic i cronològic, juntament amb repertoris toponímics i antroponímics. Creiem que és viable la creació d'un diccionari específic per a aquesta cronologia en la mesura que s'han editat un gran nombre de pergamins documentals de la segona meitat del segle XII gràcies a la col·lecció de protocols i diplomataris de la Fundació Noguera.[9] El repte en aquest cas seria datificar les edicions documentals per tal de fer un diccionari que permetés, per una banda, afinar i millorar el reconeixement i, per l'altra, fer cerques per conceptes.[10] La problemàtica que generen els topònims està resolta per l'Institut Cartogràfic de Catalunya,[11] i també ho està la problemàtica que podrien generar els antropònims.[12]

2. SELECCIÓ I TRACTAMENT PREVI DE LA DOCUMENTACIÓ

Els documents que proposem com a objecte d'estudi són els pergamins documentals del segle XII. Atès que l'OCR aplicat a documents manuscrits exigeix una certa regularitat, o bé una escriptura cal·ligràfica, hem cregut oportú cen-

trar-nos en els pergamins elaborats per una sola persona. L'escollit ha estat **Pere de Corró**, un dels escrivents més prolífics del segle XII. A part de la gran quantitat de documents que s'han conservat d'ell, aquest escrivent té un altre benefici: bona part de la seva obra ha estat editada en els darrers cinquanta anys.

La selecció i la digitalització dels pergamins ha significat més de la meitat d'aquest estudi, tot i que això no quedi del tot reflectit en el present article. A partir d'uns criteris de selecció que exposarem a continuació, hem arribat a analitzar més de 190 testimonis documentals del nostre escrivent:

- Vam seleccionar **pergamins documentals**, primerament, perquè són els documents que més respecte generen als investigadors i als usuaris d'un arxiu històric. El fet que estiguin escrits en llatí, la utilització d'abreviatures i l'escriptura en si mateixa són factors que dificulten l'accés a un usuari novell.

- **Documents originals** escrits per Pere de Corró o pel seu fill, Pere de Corró. S'han descartat aquelles referències documentals provinents de còpies, trasllats i extractes.

- **Documents editats**. Era una condició que el gruix dels pergamins estiguessin editats per tal de prescindir de la transcripció i centrar-nos en l'aplicació de mètodes i tècniques d'OCR.[13]

- La **reproducció** dels documents no podia comportar una gran inversió de temps i recursos, calia centrar-se en arxius públics, amb els mitjans i el personal suficients.[14]

- L'altre criteri aplicat era el de l'**estat de conservació**. Calia que fossin pergamins el més íntegres possible, sense forats o fragmentacions que en dificultessin el reconeixement de caràcters.

Com ja hem comentat, de gairebé 200 ítems analitzats, n'hem seleccionat 99, 74 dels quals seran la base dels supòsits pràctics. La resta de documents, és a dir, 25 pergamins, són els inèdits amb què es comprovarà l'efectivitat d'un mètode de reconeixement, el *word spotting*. Tots aquests documents seleccionats han estat introduïts en una base de dades. Sense la sistematització d'aquest elevat volum de dades seria molt difícil extreure conclusions.

Finalment, cal apuntar les característiques tècniques de les digitalitzacions: tots els documents han estat digitalitzats amb càmera fotogràfica a 300 DPI en format JPEG, llevat dels pergamins de Sant Llorenç del Munt, que estan a una resolució de 200 DPI.

3. PROCESSOS DIGITALS DE MILLORA D'IMATGES

El pas previ al reconeixement de caràcters és la millora de la imatge, però, abans d'exposar les millores d'imatge, recomanem una lectura ràpida a la bibliografia especialitzada sobre les degradacions que pateixen els pergamins documentals i les seves causes.[15] Així doncs, davant de les alteracions que pot patir un pergamí, quins mètodes informàtics poden aplicar-se per millorar les imatges?[16]

- **Detecció de la pàgina:** eliminació dels marges, sempre que els marges siguin negres o d'un color proper al negre.

- **Correcció de l'orientació**: determinar l'angle d'orientació i aplicar-hi una rotació, en cas que el document estigui pla i no hi hagi curvatura.

- **Correcció de les curvatures**: les curvatures generen problemes greus a l'hora d'aplicar l'OCR, primer, per segmentar cada una de les línies i, després, per identificar la capsa de cada una de les lletres.

- **Transmissió de tintes** (*bleed through o showthrough*): és la degradació més abundant en documents d'època moderna. A diferència del paper, el pergamí no queda afectat per aquesta alteració; tot i això, hem cregut convenient explicar-ho perquè ha estat el procés que s'ha aplicat per a la millora d'imatges de la nostra selecció documental.

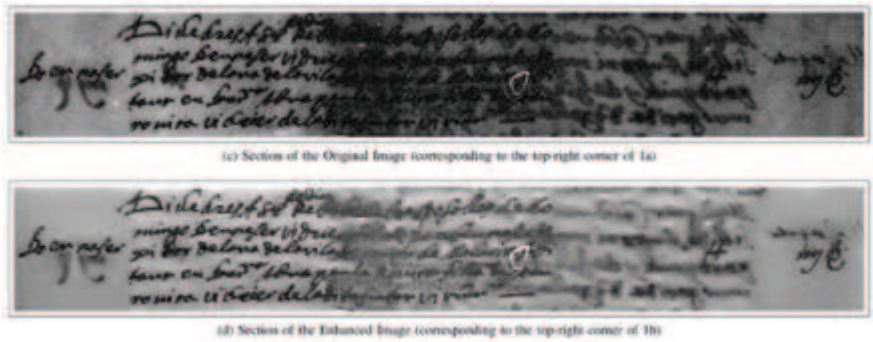


Figura 5[17]

› Millora del **contorn** de l'escriptura mitjançant el procés de **binarització**, a partir del qual una imatge en color o en escala de grisos esdevé una imatge en blanc i negre. Aquest procés s'assoleix gràcies a la manipulació de l'histograma d'una imatge. Recordem que un histograma és la representació gràfica dels píxels, a cada un dels quals se li atribueix un valor numèric d'entre 255 (blanc) i 0 (negre). A partir d'aquesta gràfica, es divideixen els píxels en dos grups: els superiors a 128 passen a 255, blanc; l'altre grup, els píxels inferiors a 128, passen a 0, negre. Aquesta mitjana pot modificar-se amb la finalitat de millorar el contrast entre l'escriptura i el fons. La binarització adaptada a cada imatge dóna uns resultats òptims; ara bé, quan s'aplica de manera massiva, ignorant les particularitats de cada imatge, pot ser contraproduent, com en el cas de les tintes esvaïdes.[18]

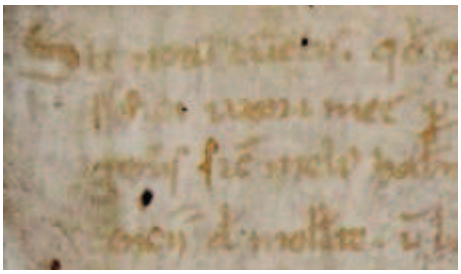


Figura 6.1 Document original afectat per tintes. esvaïdes

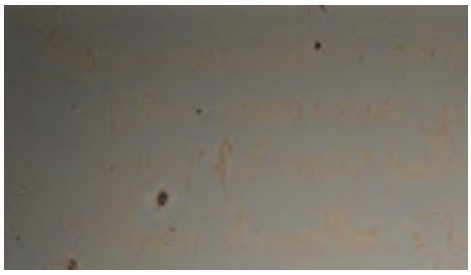


Figura 6.2 Document després del "procés de millora d'imatge". Il·lustra com pot empitjorar la lectura sinó hi ha una adaptació prèvia a les característiques de cada document.

La pràctica habitual és aplicar la binarització en zones concretes de cada imatge per tal de segmentar correctament la imatge i fer ressaltar el contorn de les lletres. Els experts recomanen, però, aplicar la binarització amb un ventall més ampli de valors, com el de l'escala de grisos, amb la possibilitat de reduir un 15% els errors. En el nostre cas s'ha aplicat la binarització per fer desaparèixer o disminuir les taques i eliminar així elements que provocarien soroll durant el reconeixement.

› **Eliminació de taques**, sigui quin en sigui l'origen o naturalesa. Les taques per a un ordinador són una dificultat afegida perquè generen **soroll** en la imatge, és a dir, es poden confondre amb l'escriptura, sobretot quan tenen el mateix color que les tintes. El procés bàsic per minvar el perjudici de les taques és la manipulació del contrast. També es poden aplicar algorismes que eliminin les taques aïllades, similars als algorismes aplicats en la segmentació automàtica del text.

Figura 7.1. Document original

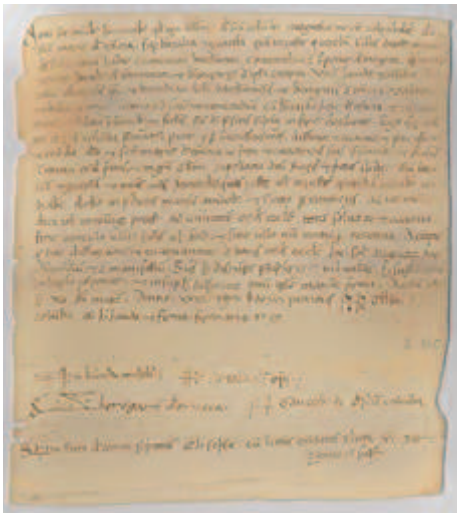


Figura 7.2[19]. Document després del procés de contrast de multiresolució

4. RESULTATS DEL PROCÉS DE MILLORA D'IMATGE

Després d'identificar cada una de les degradacions, ens hem disposat a analitzar els resultats de l'aplicació del procés de contrast de multiresolució, pel qual s'eliminen les taques i una part substancial del soroll que dificulta el reconeixement de caràcters.[20]

Els resultats més notables els trobem en el contorn de les lletres, que queden clarament definides fins i tot quan han estat afectades per taques d'humitat, i hi disminueix també la presència de fongs. Vint-i-set documents han donat uns resultats òptims, és a dir, un 46,55%. En aquest grup la presència de fongs i de taques d'humitat no desapareix del tot, però queden suficientment tènues per no dificultar la lectura. Volem destacar que també ressalten el contorn de l'escriptura quan aquesta pertany a l'hebreu.

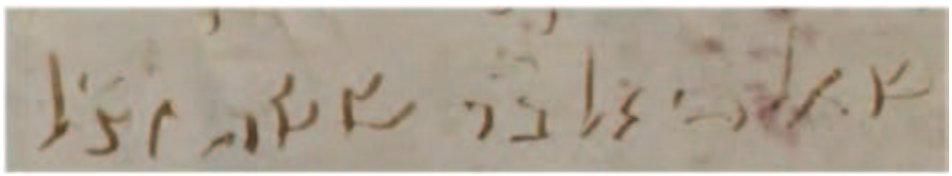


Figura 8.1 Detall de firma hebrea hològrafa sense aplicació de processos de millora.[21]

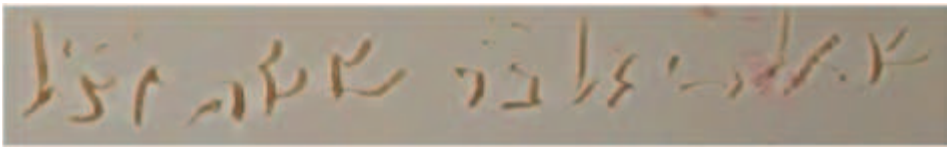


Figura 8.2 Detall de firma després de l'aplicació dels processos de millora.

La resta de documents, el 54%, presenten alteracions que abans de l'aplicació del procés no hi eren i que empitjoren la lectura. A continuació exposem breument els resultats i les seves causes:[22]

- Lectura difícil de fragments on s'han utilitzat tintes diferents de les del gruix del text.
- Les taques d'humitat desapareixen parcialment.
- Lectura difícil en aquells documents que presenten tintes esvaïdes.
- Lectura difícil en el vers del pergamí.

Entenem que el fet que el mètode aplicat sigui específic per combatre la transmissió de tintes és la raó d'uns resultats positius inferiors al 50%. Creiem que altres mètodes de millora de la imatge proporcionarien millors resultats, sobretot aquells creats expressament per a pergamins documentals i tintes esvaïdes.[23]

5. APLICACIÓ DEL *WORD SPOTTING*

El *word spotting* és una altra manera d'aproximar-se a l'OCR. La finalitat d'aquell no és transcriure les paraules pròpiament dites, sinó identificar el conjunt de caràcters a partir d'un model o *query-by-example.* Això vol dir que podem trobar una paraula o una lletra a partir d'una selecció d'un exemple o model. El *word spotting* requereix un procés de processament previ, que implica la segmentació de paraules i lletres, amb la finalitat de delimitar les imatges que s'han de cercar.

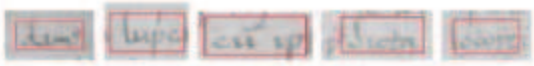
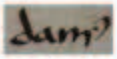
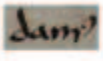
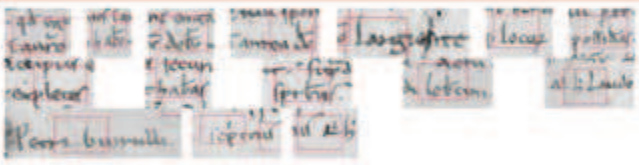
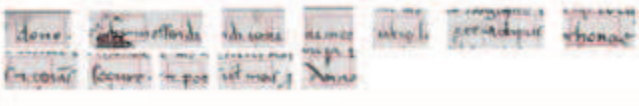
A continuació plantegem el segon supòsit pràctic: la detecció automàtica de verbs dispositius mitjançant word spotting. En el treball original hem dedicat un apartat als verbs dispositius, juntament amb una graella que en facilita l'anàlisi.[24] Creiem que la informació que pot aportar el simple verb dispositiu pot ser sufi-

cient per delimitar les cerques. Per exemple, hi ha verbs dispositius que aporten més informació que altres, com els testaments, on apareix un gran volum de dades referents a persones, llocs, possessions i les relacions que s'estableixen entre elles. Davant d'un volum de pergamins digitalitzats sense cap mena d'identificació, un investigador podria estalviar-se hores de recerca amb l'aplicació del *word spotting*.

Per fer-ho possible hem retallat el verb dispositiu principal de cada document, el qual hem plasmat en una taula. Una vegada tramesos els verbs als investigadors del CVC, aquests han extret les característiques de cada retall mitjançant l'aplicació del ***histogram of oriented gradients*** (HOG).[25]

El pas següent ha estat aplicar un mètode de *word spotting* concret, l'**Exemplar-SVM**,[26] el qual proporciona millors resultats en comparació amb altres mètodes de *word spotting*. Com el seu nom indica, l'Exemplar-SVM cerca a partir d'un exemple, una imatge base o *query*; aquesta és deformada per tal que s'adapti a la variabilitat que exigeix l'escriptura manuscrita. El descriptor generat és en realitat una selecció de les regions que millor defineixen la paraula, però preveient les variacions amb què pot aparèixer.

Els documents on s'ha aplicat la cerca de verbs dispositius són una selecció dels pergamins inèdits facilitats pel Servei d'Arxiu de la Federació Catalana de Monges Benedictines. Com es pot observar a les imatges, ha estat necessari aplicar filtres i tractaments de millora de la imatge per tal de generar el mínim soroll possible (suavització del contrast i conversió de color a escala de grisos). A continuació proporcionem una relació dels resultats de l'aplicació del *word spotting* de tres verbs dispositius (*damus*, *debeo* i *dono*) en deu pergamins documentals, tots ells sense cap mena de tractament previ, ni identificació ni datació.

Document	Exemple	Resultats
AMSPP, MSPP, Col·lecció de perg., núm. 84		
AMSPP, MSPP, Col·lecció de perg., núm. 87bis		
AMSPP, MSPP, Col·lecció de perg., núm. 97		
AMSPP, MSPP, Col·lecció de perg., núm. 15		
AMSPP, MSPP, Col·lecció de perg., núm. 17		
AMSPP, MSPP, Col·lecció de perg., núm. 19		
AMSPP, MSPP, Col·lecció de perg., núm. 32		
AMSPP, MSPP, Col·lecció de perg., núm. 19		
AMSPP, MSPP, Col·lecció de perg., núm. 21		
AMSPP, MSPP, Col·lecció de perg., núm. 25		

Com es pot veure, de 75 deteccions només 11 han donat un resultat positiu, és a dir, un 14,66% d'efectivitat. Certament són uns resultats realment baixos; ara bé, si ho mirem des d'una altra òptica, en tots els documents s'ha trobat el verb dispositiu, és a dir, que el mètode funciona. Per reduir les identificacions fallides només caldria reduir el percentatge de variabilitat de la paraula a l'hora de determinar el descriptor de cada verb dispositiu.

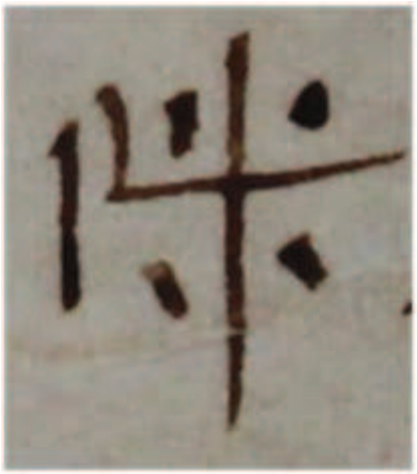


Figura 9[27]. Signum

6. COM ES POT DETERMINAR L'AUTORIA D'UN MANUSCRIT DEL SEGLE XII AMB OCR

La tercera proposta del treball s'ha centrat en allò que poden oferir la tecnologia i els mètodes de reconeixement de caràcters actuals. Per aquest motiu hem valorat diverses línies de treball i la que hem trobat més adient ha estat la identificació de l'autor. Per a una empresa com aquesta cal el reconeixement total de l'escriptura, és el que assegura la identificació adequada d'un escrivent mitjançant l'aprenentatge (ICR) a partir de més de 2.000 exemples de cada caràcter. Amb paraules de Josep Lladós, director del CVC, aquest objectiu requereix uns recursos materials i humans propis d'una tesi doctoral. Per aquest motiu ens hem vist obligats a escollir quina lletra o signe pot ser característic d'un escrivent del segle XII. La resposta és molt senzilla: el *signum*, és a dir, el dibuix característic de la persona que signa un document, sigui com a testimoni o com a escrivent.

A partir del segle X el *signum* de l'escrivent era regular i diferent del de la resta de testimonis i corresponia a una interpretació de l'abreviatura de *subscripsit*, expressat en tres esses entortolligades. Cada escrivent tenia el seu propi *signum* i es poden distingir a simple vista. El cas de Pere de Corró no és una excepció i és difícil de confondre'l amb el de qualsevol altre escrivent. El *signum* del seu fill, Pere de Corró el Jove, també és característic.

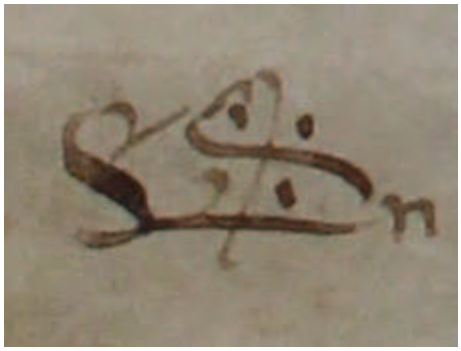


Figura 10.1 Pere de Corró[28]

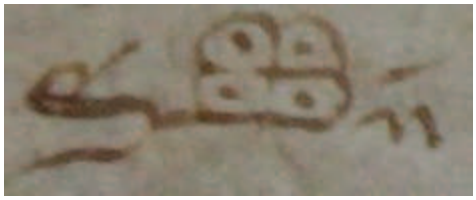


Figura 10.2 Pere de Corró el Jove[29]

Una vegada seleccionat el caràcter identificatiu de l'escrivent, calia retallar el màxim d'exemples possibles i facilitar-los als tècnics del CVC. Recordem que el nostre objectiu és esbrinar si els mètodes actuals d'OCR són capaços de reconèixer i identificar els caràcters d'una mà concreta, per aquest motiu no vam indicar quins eren el *signa* de Pere de Corró i quins eren del Jove. El procés que han aplicat al CVC ha constat de diverses fases:

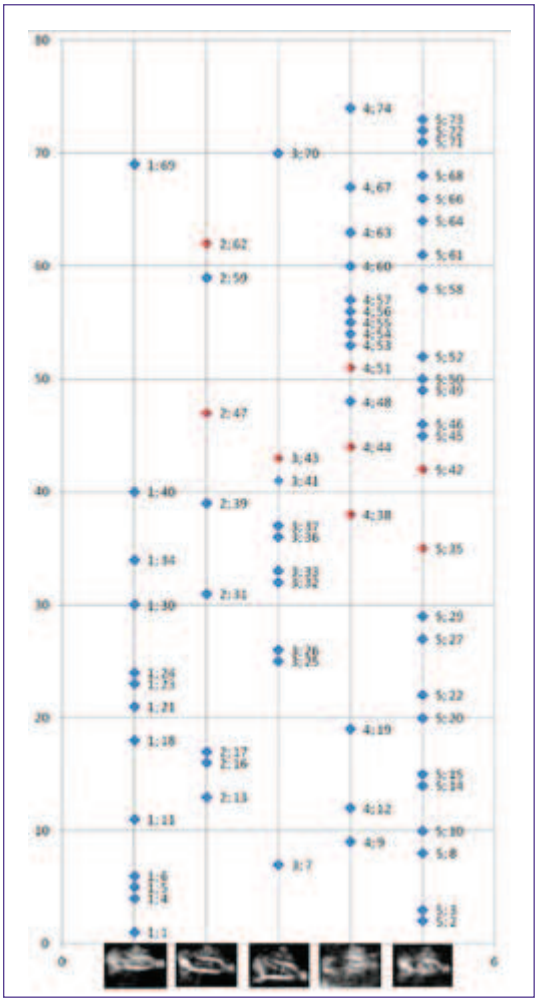
- Binaritzar la imatge per extreure'n el soroll que pot generar el color.
- Extreure les característiques mínimes de cada *signum* mitjançant el procés del *blurred shape model* (BSM), el qual mesura la densitat de píxels de cada graella aplicada a la imatge seleccionada.[30]
- Establir grups de descriptors amb característiques comunes o clusterització.

Figura 11. Procés de manipulació de la imatge per al reconeixement d'autor.

› Comparar cada descriptor amb cada document i atribuir-li un grup de l'1 al 5.

La finalitat d'aquest procés és determinar les mans que poden haver escrit els *signa*. L'algorisme genera i agrupa de manera automàtica cada *signum* elaborant una graella final amb els valors de l'1 al 5 en cada document.

Atès que el procés és complex per al profà en la matèria, com un servidor, hem cregut pertinent elaborar uns gràfics de dispersió a partir dels quals hom pot interpretar els resultats ràpidament. Cada gràfic disposa d'una miniatura de cada descriptor per tal de visualitzar i entendre millor els resultats; corresponen a l'eix horitzontal. L'eix vertical correspon a l'ordre cronològic dels documents.[31]



32

Abans de rebre els resultats, crèiem que observaríem dos grups de *signa*: els *signa* de Pere de Corró i els *signa* de Pere de Corró el Jove, sense comunicar prèviament quins eren quins. Hem marcat en vermell aquells documents signats pel Jove. Els resultats no ens permeten associar un descriptor concret al *signum* del fill, és a dir, que l'experiment ha resultat fallit.

Però volem destacar-ne un detall: cap dels documents on es presenta el *signum* de Pere de Corró no pertany al grup 1, sempre basculen entre els grups 2 i 5.

Els resultats negatius els atribuïm a dos factors: la manca de documents aportats i l'avançat estat de degradació dels documents. Recordem que el *signum* de l'escrivent es troba sempre al marge inferior del document, i aquest el trobem rossegat, tacat i perdut. Seria necessari repetir l'exercici només amb aquells *signa* en un estat òptim de conservació, amb una resolució més alta i que superessin el centenar d'exemples.

8. NO FUNCIONA, PROVEM-HO PER UNA ALTRA BANDA

No hem pogut agrupar adequadament els *signa* dels autors dels pergamins mitjançant l'OCR, però podem obtenir altres resultats.

Saber l'autor material dels pergamins no ens permetria determinar si un document és fals o no; ens calen més dades, dades que ens proporciona la diplomàtica. En l'estudi original hem analitzat àmpliament les característiques diplomàtiques dels documents seleccionats i podem afirmar que els documents són autèntics, és a dir, entenem que contenen totes les característiques formals pròpies de l'època. Ara bé, i si realment l'autor material és un altre, tot i que sembla autèntic a primer cop d'ull? Què voldria dir? L'única resposta que podem proporcionar és la participació de diverses mans en un context específic: l'aprenentatge.[33]

Per desgràcia, no ens podem estendre en aquest apartat tal com ho hem fet al treball final del postgrau. Només apuntem la hipòtesi de l'experiment. Creiem que, per la quantitat de pergamins conservats de Pere de Corró, aquest requeria els serveis d'aprenents. Tenim testimonis d'aprenents d'escrivania en segles posteriors, sobretot després de la irrupció del notariat a Catalunya (segle XIII),

però en el període que ens pertoca, inicis del segle XII, no s'ha demostrat de manera quantitativa la participació activa dels aprenents.

En aquest punt volem identificar en quins documents ha participat un aprenent. A diferència de l'experiment anterior, no podem identificar amb seguretat la intervenció d'un aprenent; només un paleògraf expert podria fer-ho.

Figura 12. Abreviatura *hoc*

A partir del mateix principi que el supòsit dels *signa*, hem seleccionat un caràcter especial que es repetís en el cos del text i en la subscripció de l'escrivent. Ens hem decantat per l'abreviatura d'*hoc* ('això', 'aquest').

El procés a partir del qual s'han comparat les abreviatures d'*hoc* ha estat el següent:

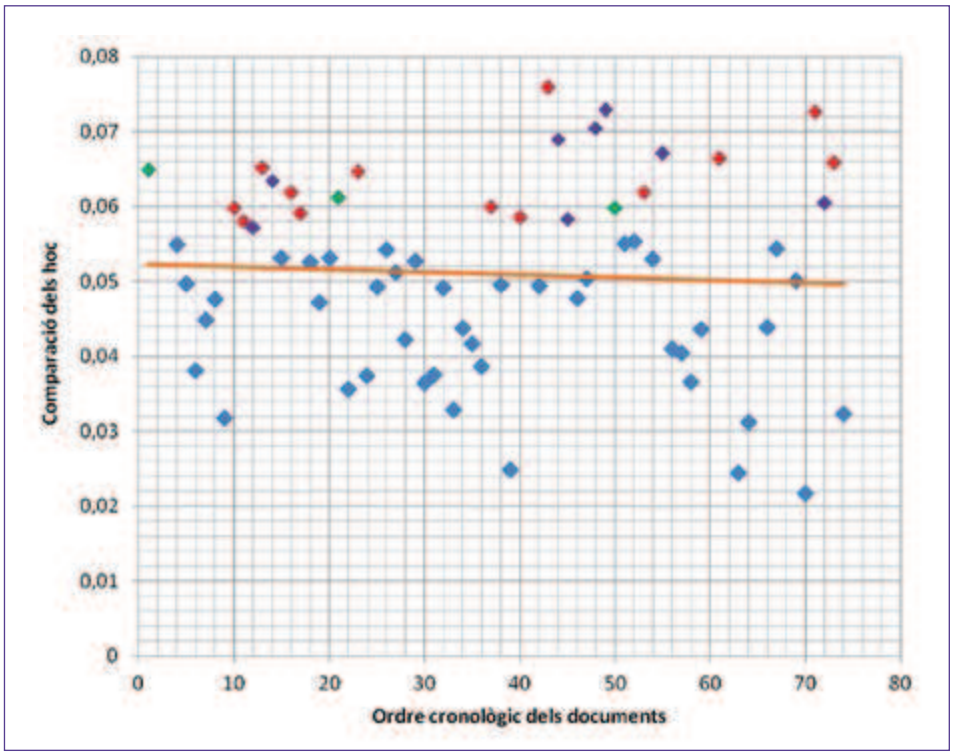
- Retallar cada abreviatura identificant-ne cada una amb el número de referència del document. Per una banda, una carpeta amb els retalls pertanyents als *hoc* del text, és a dir, aquells que possiblement haurien d'estar elaborats pel deixeble. L'altra carpeta conté els retalls dels *hoc* pertanyents a la subscripció de l'escrivent.

- Una vegada trameses les respectives carpetes, aquestes han comparat les lletres a partir del mateix procés del BSM, que mesura la densitat de píxels a cada graella. A partir d'aquí se n'extreu un descriptor, és a dir, un valor numèric que defineix la imatge seleccionada.[34]

- Comparar aquests valors entre ells a partir de la lògica següent: l'*hoc* de l'escrivent té el valor 0; si el resultat de la comparació supera el 0, voldrà dir que es tracta d'imatges diferents. Com més elevat sigui el valor, menys semblant a l'exemple serà. Si la comparació parteix de dues lletres iguals, voldrà dir que aquestes estan fetes per dues mans diferents, o bé en períodes temporals diferents.[35]

Per facilitar la comprensió també hem creat un gràfic a partir del qual podem veure la distància mitjana entre l'*hoc* del text i l'*hoc* de l'escrivent. L'eix horitzontal equival als documents ordenats cronològicament. L'eix vertical indica la distància entre l'*hoc*, és a dir, si tendeixen a 0 voldrà dir que són similars i, per tant, d'una mateixa mà. Si el valor s'allunya de 0, voldrà dir que són possible-

ment de mans diferents. En el gràfic s'observa una línia de tendència que indica la mitjana, que es manté al voltant del 0,05 en tot el gràfic. Seria temerari pensar que tots els documents per damunt de 0,05 han estat elaborats per un deixeble. Tal com ens han apuntat els investigadors del CVC, qualsevol valor inferior a la segona dècima indica que es tracta de lletres molt similars, és a dir, hi ha un alt percentatge de possibilitats que sigui feta per una mateixa persona.



[A la gràfica, el mot *hoc* ha d'anar en cursiva.]

Si partim de valors superiors a 0,06, els resultats són sorprenents. Hem analitzat aquells 24 documents que superen el 0,06 i podem dir que en **més d'un 50% dels documents hi ha participat un aprenent** (13 de 24 documents, marcats en vermell).[36] Només un 12,5% (3 de 24, marcats en verd) era de la mateixa mà, tant el text com la subscripció de l'escrivent. La resta de documents, un 33,3% (8 de 24, marcats en lila), es trobava per damunt del 0,06 a causa de la degradació del document o per una mala digitalització.

Així doncs, la nostra premissa de l'aprenentatge la podem ratificar a partir de les dades quantitatives d'aquest experiment. Caldria la comparació de més lletres per poder obtenir resultats que suportessin aquesta hipòtesi.

CONCLUSIONS

En aquest treball hem demostrat que l'OCR es pot aplicar als pergamins documentals del segle XII. Ara bé, si el resultat que esperem obtenir és la transcripció del text íntegre, hem de dir que actualment no és possible.

El reconeixement de caràcters és una disciplina de l'enginyeria informàtica i disposa de múltiples vies de recerca, una n'és el reconeixement d'escriptura manuscrita que, per la seva irregularitat i variacions, és un repte encara no superat per les eines i els mètodes d'OCR actuals. A més a més, la tecnologia vigent exigeix uns requeriments específics a l'hora d'aplicar l'OCR en escriptures manuscrites: documents originals, escrits per una sola mà, un estat de conservació òptim i una digitalització en alta resolució (300 DPI) i de qualitat (sense sobreexposició ni ombres). Tot i les dificultats apuntades, hem dut a terme diferents experiments, traduïts en quatre supòsits pràctics:

El primer supòsit l'hem dedicat al tractament i la millora de la imatge, pas previ al reconeixement pròpiament dit. Hem aplicat a les imatges el procés de *multiresolution contrast* i n'hem analitzat detalladament els resultats, determinant que l'estat de conservació i les degradacions de cada pergamí són un factor clau a l'hora d'obtenir resultats reeixits.

El segon supòsit, dedicat al reconeixement de paraules a partir d'un model, *word spotting*, l'hem centrat en la detecció dels verbs dispositius, amb uns resultats satisfactoris. En tots els 10 pergamins inèdits, s'ha identificat el verb dispositiu, ara bé, juntament amb un seguit de deteccions fallides que en redueixen substancialment l'efectivitat.

El tercer supòsit pràctic tenia la finalitat d'identificar l'escrivent mitjançant l'extracció i la comparació de descriptors mitjançant el procés del BSM. Els resultats no han acompanyat en la mesura que no hem donat un model de cerca, com en el cas del *word spotting*. Tot i això, hem detectat altres particularitats a tenir en compte, com la generació de clústers segons el període en què s'ha elaborat el document, és a dir, una pauta cronològica.

El quart supòsit, i per a nosaltres el més engrescador, s'ha plantejat a partir d'una premissa: Pere de Corró tenia aprenents. Els resultats quantitatius que ens ha proporcionat l'OCR han confirmat aquesta premissa mitjançant la comparació d'una abreviatura, *hoc*, que apareix en el cos del text i en la subscripció de l'escrivent.

Malgrat que els resultats presentats no han estat del tot satisfactoris, volem recordar que és una primera aproximació, no ens consta cap altre treball similar enlloc més. Amb més recursos, humans i econòmics, podríem afinar les eines tecnològiques. Tenim la convicció que l'aplicació de l'OCR comportarà la creació de mecanismes de recuperació de la informació més àgils i eficaços.

Encara ens queda un llarg camí per a la transcripció íntegra d'un pergamí, però el repte més gran per a qualsevol arxiver és treure-li el màxim profit a les eines que té a l'abast, i no només per fer-nos la vida més fàcil, sinó per fer accessibles els documents.

NOTES

1. Five Centuries of Marriage, <http://dag.cvc.uab.es/infoesposalles/> [consulta: 24 de juliol de 2016].
2. Projecte EINES, <http://xac.gencat.cat/ca/detalls/Noticia/Nova-Noticia-04767> [consulta: 24 de juliol de 2016].
3. Per al desenvolupament d'aquest apartat hem utilitzat principalment els apunts i la presentació de J. Lladós «Reconeixement de text (imprès i manuscrit)», I Postgrau de Gestió i Tractament Digital de Documents Històrics, Barcelona/Terrassa: Escola Superior d'Arxivística i Gestió de Documents (ESAGED), 28 de novembre de 2013; i d'A. Fornés, «Forensics: Writer Identification / Signature Verification», I Postgrau de Gestió i Tractament Digital de Documents Històrics, Barcelona/Terrassa: ESAGED, 5 de desembre de 2013.
4. Plamondon, R.; O'Reilly, C.; Rémi, C.; Duval, T. «The longnormal handwriter: learning, performing, and declining». Frontiers in psychology. <http://journal.frontiersin.org/article/10.3389/fpsyg.2013.00945/full> [consulta: 24 de juliol de 2016].
5. Lladós, J. «Reconeixement…». Op. cit., diapositiva 101.
6. Fornés, A.; Ramos, O. Eines i programari de textos manuscrits. Curs de formació continuada de l'Associació d'Arxivers i Gestors de Documents de Catalunya (AAGDC). Barcelona: 29 de gener de 2015, diapositiva 28.
7. Lladós, J. «Reconeixement…». Op. cit., diapositiva 108.
8. Lladós, J. «Reconeixement…». Op. cit., diapositiva 116.
9. La col·lecció de llibres de privilegis, de diplomataris, o l'Acta Notariorum Cataloniae són només algunes d'aquestes. Vegeu: <http://www.fundacionoguera.com/col-leccions.asp>.

10. Els mètodes d'explotació i recuperació de dades a partir de textos on s'ha aplicat el ROC ha estat estudiat per M. Rossinyol a «Searching by Content in Digital Archives», I Postgrau de Gestió i Tractament Digital de Documents Històrics, Barcelona/Terrassa: ESAGED, 27 de novembre de 2013.

11. Institut Cartogràfic de Catalunya (ICC), Nomenclàtor oficial de toponímia major de Catalunya. Barcelona: Generalitat de Catalunya; Institut d'Estudis Catalans, 2009.

12. Bolòs, J.; Moran, J. Repertori d'antropònims catalans (RAC). Barcelona: Institut d'Estudis Catalans, 1994.

13. Els pergamins editats s'han extret de les publicacions següents:
    - Alturo, J. L'Arxiu antic de Santa Anna de Barcelona de 942 al 1200: aproximació històrico-lingüística, vol. 1. Barcelona: Fundació Noguera, 1985.
    - Puig, P.; Robles, J.; Ruiz, V.; Soler, J.; Capellades, A. Diplomatari de Sant Llorenç del Munt (1101-1230). Barcelona: Fundació Noguera, 2013.

14. Les reproduccions han estat obtingudes de manera gratuïta, en la mesura que un treball de recerca com el present interessa als centres on es custodia aquesta documentació. Agraïm el tracte rebut del personal de cada centre:
    - Pergamins custodiats per la Biblioteca de Catalunya i publicats al portal Memòria Digital de Catalunya. <http://mdc.cbuc.cat/cdm/landingpage/collection/pergamiBC>. [consulta: 24 de juliol de 2016].
    - Pergamins custodiats a l'Arxiu Comarcal del Vallès Occidental - Arxiu Històric de Terrassa (ACVOC-AHT).
    - Pergamins custodiats a l'Arxiu Diocesà de Barcelona (ADB).
    - Pergamins de Sant Llorenç del Munt custodiats a l'Arxiu de la Corona d'Aragó (ACA).
    - Pergamins custodiats pel Servei d'Arxiu de la Federació de Monges Benedictines (SAFMB).

15. Al treball presentat en el postgrau, hem dedicat un apartat sencer a les degradacions que afecten els pergamins documentals. Aquest punt ha estat elaborat a partir de C. Bello, A. Borrell, El patrimonio bibliográfico y documental: claves para su conservación preventiva, Gijón: Trea, 2001. Recomanem l'obra de referència per al tractament integral dels pergamins del professor i mestre P. Puig, Els pergamins documentals, col·lecció Normativa Arxivística, núm. 3, Barcelona: Departament de Cultura de la Generalitat de Catalunya, 1995, p. 27-60.

16. Valveny, E. Op. cit.

17. La descripció del mètode de millora d'imatge per transmissió de tintes el trobareu a: Fornés, A.; Otazu, X.; Lladós, J. «Show-through cancellation and image enhancement by multiresolution contrast processing» [en línia], p. 3. <http://www.cvc.uab.es/~afornes/publi/conferences/2013_ICDAR_AFornes.pdf> [consulta: 24 de juliol de 2016].

18. ADB, carpeta 3A, perg. 137. Fons de Santa Eulàlia del Camp.

19. ACVOC-AHT, 9/1 Esglésies de Sant Pere de Terrassa, pergamins I-140.

20. Fornés, A.; Otazu, X.; Lladós, J. «*Show-through…*». *Op. cit*.

21. ADB, carpeta 6, perg. 18. Fons de Santa Eulàlia del Camp.

22. Els resultats individuals de l'aplicació de les millores han estat relacionats en una taula al treball original.

23. El Centre of Image and Material Analysis in Cultural Heritage (CIMA), adscrit al Computer Vision Lab de la Universitat Tècnica de Viena (Àustria), té un projecte específic de millora de la imatge per facilitar la transcripció de documents en un estat de conservació molt dolent. <http://www.caa.tuwien.ac.at/cvl/project/cima/> [consulta: 24 de juliol de 2016].

24. En aquesta secció hem apuntat que a vegades el verb dispositiu no concorda amb la naturalesa concreta, com ara les donacions, que poden tractar-se de compravendes o arrendaments.

25. Rodríguez, J. A.; Perronnin, F. «Local gradient histogram features for word spotting in unconstrained handwritten documents». *The 11th International Conference on Frontiers in Handwriting Recognition*. Mont-real, Quebec, Canadà: Universitat de Concordia, 19-21 d'agost de 2008. <http://www.iapr-tc11.org/archive/icfhr2008/Proceedings/papers/cr1015.pdf> [consulta: 24 de juliol de 2016].

26. Almazán, J.; Gordo, A.; Fornés, A.; Valveny, E. «Segmentation-free Word Spotting With Exemplar SVMs». *Patter Recognition*, vol. 47, 18 de juny de 2014, pp. 3967-3978. <http://www.cvc.uab.es/people/afornes/publi/journals/2014_PR_Almazan.pdf> [consulta: 24 de juliol de 2016]

27. Detall del *signum* d'un testimoni fet per l'escrivent del document (ADB, carpeta 2A, perg. 5, Fons de Santa Anna).

30. Escalera, S.; Fornés, A.; Pujol, O.; Radeva, P.; Sánchez, G.; Lladós, J.«Blurred Shape Model for Binary and Grey-level Symbol Recognition». *Pattern recognition letters*, vol. 30, núm. 15. Elsevier, 19 de març de 2009, pp. 1424-1433.

31. Només facilitem un dels quatre gràfics. Vam elaborar-ne una per cada resolució de descriptor. Hem representat les dades d'aquesta manera per un motiu concret. Com ja hem apuntat anteriorment, l'escriptura és un exercici físic que relaciona la musculació amb el sistema neuronal. Amb el pas del temps un escrivent modifica la seva escriptura a mesura que el seu cos canvia. Era necessari per aquest motiu que l'eix vertical indiqués l'ordre cronològic dels documents, així podríem veure si els clústers es deuen a l'edat de l'escrivent.

32. Comparació de *signa*: descriptors de 32 columnes i 32 files (nh32-nv32).

33. Per a l'elaboració d'aquest apartat hem consultat: Pagarolas, L. «El notariat i cultura: els registres notarials». *Actes del I Congrés d'Història del Notariat Català. Barcelona 11, 12 i 13 de novembre de 1993*. Josep M. Sans i Travé (coord.). Barcelona: Fundació Noguera, 1994, pp. 333-350.

34. Escalera S., *et al. Op. cit*. <http://www.cvc.uab.es/people/afornes/publi/journals/2009_ PRL.pdf> [consulta_ 24 de juliol de 2016].

35. Hem de confessar que la interpretació de les graelles de resultats ha estat difícil. Agraïm a Alícia Fornés la paciència i el seu esperit didàctic.

36. Aquesta afirmació se sustenta a partir de l'anàlisi paleogràfica només d'aquells documents que superen el valor 0,055. Estem segurs que paleògrafs amb més experiència podrien tenir opinions diferents, fins i tot contràries a les conclusions a què hem arribat.

RESUM

Aquest és el treball final de la primera edició del postgrau de Gestió i Tractament Digital de Documents Històrics, organitzat per l'ESAGED. La recerca es basa en l'aplicació de diferents mètodes de reconeixement de text manuscrit (HTR per la sigla en anglès) en pergamins documentals d'un escrivent concret, Pere de Corró (mitjan segle XII). A partir de la selecció de documents originals, en alta resolució i en un estat de conservació òptim, s'han dut a terme diversos supòsits pràctics: determinar si els processos de millora de la imatge fan més fàcil la lectura dels pergamins, distingir l'autor a partir del *signum* sense processament previ, analitzar els resultats del reconeixement de paraules clau (*word spotting*) mitjançant la detecció dels verbs dispositius i esclarir l'existència d'aprenents a partir de les dades quantitatives que ens proporciona el reconeixement òptic de caràcters (ROC; OCR per a la sigla anglesa).

RESUMEN

Trabajo final de la I edición del Postgrado de Gestión y Tratamiento Digital de Documentos Históricos, organizado por la ESAGED. La investigación se basa en la aplicación de diferentes métodos de reconocimiento de texto manuscrito (HTR por la sigla en inglés) en pergaminos documentales de un escribiente concreto, Pere de Corró (mediados del siglo XII). A partir de la selección de documentos originales, en alta resolución y en un estado de conservación óptimo, se han llevado a cabo varios supuestos prácticos: determinar si los procesos de mejora de la imagen facilitan la lectura de los pergaminos; distinguir el autor a partir del *signum* sin aprendizaje previo; analizar los resultados del reconocimiento de palabras clave (*word spotting*) mediante la detección de los verbos dispositivos; esclarecer la existencia de aprendices a partir de los datos cuantificativos que nos proporciona el ROC.

RÉSUMÉ

Travail final de la première édition du troisième cycle de gestion et de traitement numérique des documents historiques organisé par l'ESAGED. La recherche est basée sur l'application de différentes méthodes de *Handwritten Text Recognition* (HTR) sur des parchemins documentaires d'un écrivain donné, Pere de Corró (milieu du XIIe siècle). À partir de la sélection de documents originaux, en haute résolution et dans un état de conservation optimale, plusieurs cas pratiques ont été menés : déterminer si les procédés d'amélioration de l'image facilitent la lecture des parchemins ; reconnaître l'auteur à partir du *signum* sans processus préalable ; analyser les résultats du *word spotting* via la reconnaissance des verbes de dispositifs ; expliquer l'existence d'apprentis à partir des données quantitatives fournies par l'OCR.

ABSTRACT

Final project of the first postgraduate course on Digital Management and Treatment of Historical Documents, organise by ESAGED. The research is based on the application of various handwritten text recognition (HTR) methods to documentary parchments from a specific scribe, Pere de Corró (mid-12th century). Using a selection of original documents, in high resolution and an optimum state of preservation, different practical experiments have been carried out: testing whether processes to improve the image make it easier to read the parchments: determining the author based on the signum without prior processing; analysing word spotting results, through the detection of dispositive verbs; clarifying the existence of apprentices based on quantifying date provided by OCR.