OCR, Ordinadors que "llegeixen"

Lluís Soler i Carrascosa

Actualment, el reconeixement òptic de caràcters ha esdevingut una potent eina professional en microinformàtica. A continuació es detalla, breument, què és el que hi ha darrere de la tècnica de l'OCR (Reconeixement Òptic de Caràcters). Avantatges i desavantatges de la tècnica, casos en què és possible aplicar-la, limitacions actuals, són alguns dels aspectes tractats aquí.

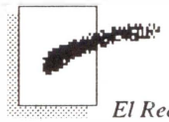
La lectura òptica amb reconeixement de caràcters és un mitjà ràpid i molt automatitzat d'entrar, en la memòria d'un ordinador, documents ja dactilografiats o impresos sense necessitat de tornar a escriure'ls manualment des del teclat. Els textos que s'obtenen per aquest mètode estan en format de fitxers de TdT (Tractament de Textos), cosa que en fa molt flexible la manipulació posterior.

Des de sempre, s'ha pres l'ull humà com a model per a la tecnologia de l'OCR. Quan es llegeix un text, l'ull humà en digitalitza una línia; la retina envia la imatge digitalitzada al cervell, i aquest descodifica la informació i n'obté el missatge. La tècnica de l'OCR ha progressat contínuament des de fa aproximadament un segle. C. R. Carey, de Boston (Massachusetts), va desenvolupar la primera retina-escànner el 1870. El 1890, P. Nipkow, polonès, va fabricar un disc d'escombratge que ha estat utilitzat per les modernes càmeres de televisió. Companyies com IBM i *Bell Laboratories* també han contribuït notablement al desenvolupament d'aquesta tecnologia a partir dels anys 50. Sheppard i Rabinow, cap allà als anys 60, van participar en el desenvolupament d'importants eines relacionades amb l'OCR dedicades a organismes tals com el govern, bancs i editorials. Posteriorment, en alguns moments, la tecnologia OCR va caure en desgràcia i alguns la tenien per una tècnica massa exòtica i futurista. Actualment, aquesta tecnologia torna a tenir el seu lloc al món de la microinformàtica gràcies als esforços d'algunes empreses privades especialitzades en aquest camp.

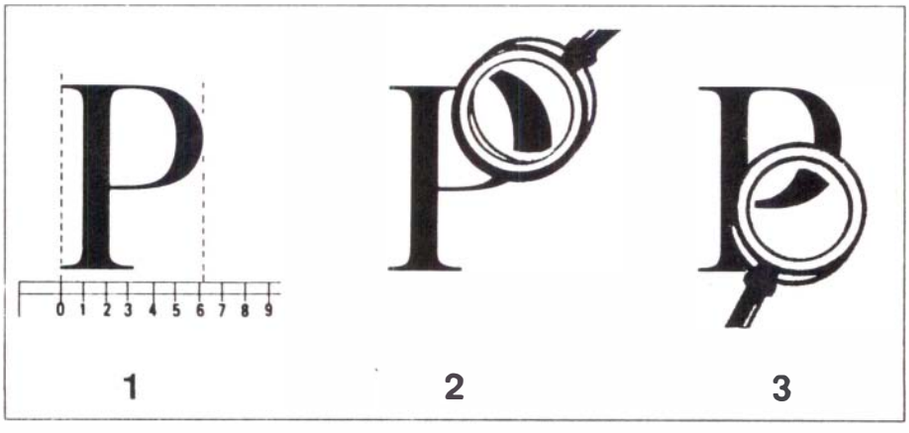
Components d'un sistema de lectura òptica

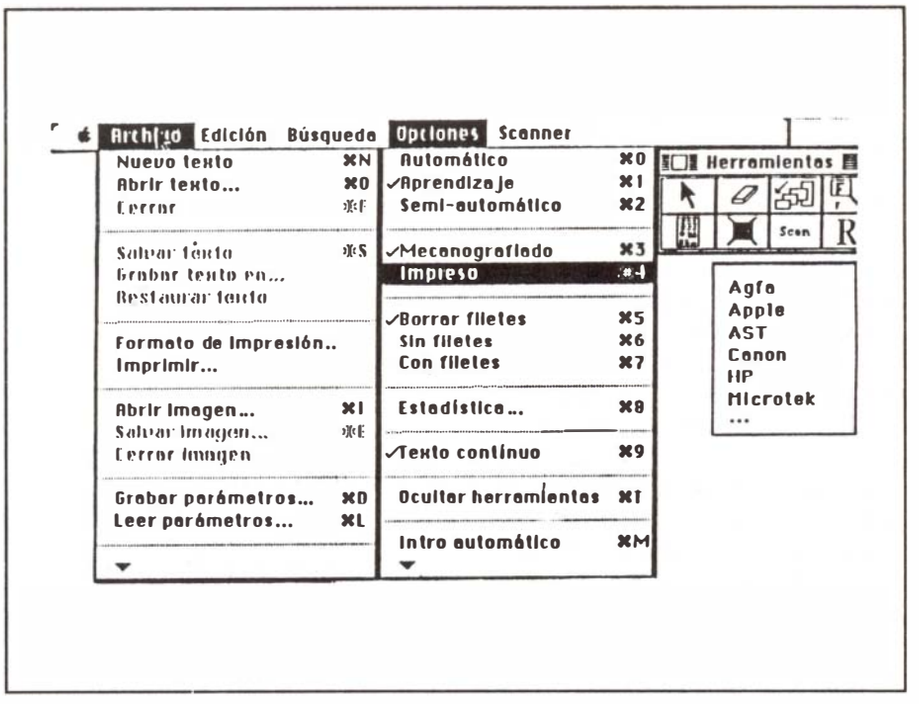
Les instal·lacions de reconeixement òptic de caràcters es componen dels elements següents: un microordinador que processa la informació (sovint un IBM-PC o un APPLE-MACINTOSH), eventualment amb una pantalla d'alta resolució; un lector òptic (escànner o cambra) connectat al microordinador, per a la digitalització dels textos; i un programa de captació d'imatges i de reconeixement de caràcters.

Sovint, aquest equip de digitalització de textos compta amb una impressora



El ReadStar fa servir una anàlisi topològica i un algoritme per reconèixer caràcters en dues fases. Estudia els caràcters del text d'un a un i els analitza comparant els models matemàtics de les formes dels caràcters específics. En aquest cas, el caràcter a identificar és una 'P' (1). La forma que ha detectat és comuna a tres lletres: 'B', 'P' i 'R' (2). Aquesta altra forma, en canvi, ja és exclusiva de la 'P': es tracta doncs d'una 'P' (3).





Software per a OCR en MACINTOSH. La pantalla presenta el menú principal de les diferents funcions per passar del caràcter imprès en el paper al disc sense que s'hagi de teclejar.

làser, que permet que els textos introduïts a l'ordinador puguin tenir una sortida en paper en el format que l'usuari necessiti.

El microordinador

La missió de l'ordinador és processar la informació enviada per l'escànner i pel programa de reconeixement òptic de caràcters a la major velocitat possible. Tenint en compte la sofisticació dels programes d'intel.ligència artificial, la massa d'informació a tractar enviada per l'escànner, la quantitat de models i de tests d'identificació necessaris per al reconeixement dels caràcters, l'ordinador haurà d'estar equipat amb una memòria RAM direccionable com més poderosa millor (640 Kb sobre PC-AT), un processador ràpid (AT-286 o AT-386) i una velocitat de procés elevada (de 8 a 20 MHz). Es pot dir que la majoria de sistemes d'OCR utilitzen un entorn IBM-PC o APPLE-MACINTOSH per funcionar.

El lector òptic

L'escànner és, de fet, una càmera que analitza línia per línia un document i enregistra una successió de punts blancs (0) o negres (1). Aquesta anàlisi es denomina digitalització de la pàgina.

L'escànner es defineix per diverses característiques:

El **tipus**: segons la manera com es posa el document per fer-ne la lectura. Pot ser de plat o de corró. El de plat o estil fotocopiadora permet de llegir llibres o revistes sense necessitat d'arrencar les pàgines de l'obra. En el corró, el document entra per uns corrons que roden davant de la finestra de l'escànner.

L'**alimentació**: manual (full per full) o automàtica, amb un carregador d'entre cinc i cinquanta fulls.

La **resolució** (o poder separador): és el nombre de punts que és capaç d'analitzar l'escànner per longitud de línia. S'expressa normalment en punts per polzada (dpi). Les resolucions corrents que es poden trobar entre els escànners del mercat són de 200, 300 i 400 dpi (actualment en comencen a aparèixer amb resolucions de 600 i 800 dpi, però que encara no són de gaire utilitat en aquest camp). Com més gran sigui la resolució, millor serà la qualitat de la imatge obtinguda del document.

■ L'escànner és una càmera que analitza línia per línia un document i enregistra una successió de punts blancs o negres.

La **qualitat òptica, mecanica i electrònica**: aquests paràmetres són importants de cara a la fidelitat de la imatge donada per l'escànner. Un bon objectiu, sense distorsió, l'absència de vibracions i de paràsits, la constància de paràmetres fixats, són importants per a la qualitat de la imatge obtinguda.

La **velocitat d'escombratge i de transmissió**: és el temps necessari perquè l'escànner pugui captar la imatge que s'ha d'analitzar i en transmeti les dades a l'ordinador. La transmissió de les dades pot ser directa (de l'escànner a la memòria de l'ordinador) o indirecta, mitjançant una memòria tampó incorporada a l'escànner o a la placa d'interfície ordinador-escànner.

El programa de reconeixement de caràcters

El programa de reconeixement de caràcters permet digitalitzar la imatge del document i extreure fitxers de text després d'haver efectuat una anàlisi intel·ligent de la imatge. Després d'haver captat la imatge mitjançant l'escànner,

en primera aproximació, es pot dir que existeixen tres tipus d'algoritmes per al reconeixement òptic dels caràcters.

En primer lloc, hi ha el sistema que funciona a base de superposar models de lletres, que el sistema ja coneix, sobre el nou caràcter aïllat que es vol identificar. Es calcula un factor de correlació per cada model comparat i, si aquest factor supera un valor estipulat de tolerància, s'interpreta el nou caràcter com a reconegut. El segon algoritme fa servir criteris simples per classificar els diferents símbols a reconèixer; i el tercer mètode es basa en una anàlisi d'estructura i un estudi morfològic dels elements gràfics. Aquests dos últims mètodes es poden classificar com a mètodes intel·ligents de reconeixement. Aquests consisteixen a estudiar la forma segons un procés quantitatiu o qualitatiu. El reconeixement no s'obté per comparació, sinó mitjançant un estudi analític o lògic, en el sentit matemàtic del concepte.

La via lògica es basa en la descomposició d'un caràcter en aquelles formes bàsiques de què es pot considerar que està format (línies rectes, corbes, obliqües, interseccions...). El reconeixement s'obté, per tant, per comparació amb una biblioteca de caràcters descompostos. Segons que les formes bàsiques del nou model resultin ser unes o altres, aquest caràcter s'associarà al seu corresponent més pròxim de la biblioteca. Per contra, la via analítica assigna a cada model de la biblioteca una funció contínua. Al nou model a identificar, li és assignada una funció contínua que, per correlació amb les de la biblioteca, determina quin dels caràcters d'aquesta és el més semblant al caràcter en qüestió. Aquests dos mètodes intel·ligents tenen l'inconvenient que fan un reconeixement amb certes ambigüitats. La incertesa es resol, teòricament, pel coneixement lèxic i semàntic que té el programa, és a dir, pel coneixement del lloc que cada lletra té dintre d'una paraula, i d'aquesta dintre d'una frase. El programa, per tant, hauria "d'entendre" la informació que va generant. Val a dir, però, que la realitat demostra que la informàtica encara no ha arribat tan lluny en aquest aspecte.

El text generat després del reconeixement òptic té, per tant, un reduït tant per cent de caràcters no identificats (normalment representats per un caràcter arbitrari que sol ser un asterisc). Posteriorment, els caràcters no reconeguts poden ser substituïts pel caràcter real, després d'una lectura informatitzada prèvia del document, és a dir, se'n fa la correcció mitjançant correctors ortogràfics, amb diccionaris de paraules com més complets millor (utilització del verificador ortogràfic o *spelling checker*).

■ L'entrada de textos amb OCR és de cinc a quinze vegades més ràpida que amb el mètode manual.

Aquest mètode de recerca ràpida i validació presenta grans avantatges quant a rapidesa i fiabilitat de reconeixement. Alguns exemples d'aplicacions dedicades al món de l'OCR són els següents: en un entorn APPLE-MACINTOSH, hi tenim MacinTEXT, ReadStar II i III Plus, TextScan, Read-It i Publish Pac. En un entorn IBM-PC, alguns exemples són AutoRead, ReadStar II i III Plus, TextRight i MultiReader (en la Taula I s'expressen els resultats comparatius per algunes aplicacions funcionant en un entorn APPLE-MACINTOSH).

POSSIBILITATS DE LA LECTURA ÒPTICA

La lectura òptica de documents permet una entrada de textos amb rendiments de reconeixement molt elevats (de vegades és del 99 % i més), amb velocitats considerables (fins a 100.000 caràcters per hora) amb un microprocessador AT-386 i un escànner d'oficina.

Els sistemes de reconeixement òptic, desgraciadament, no poden llegir qualsevol mena de text. Els límits poden ser de diferents tipus i depenen, al mateix temps, dels microordinadors, dels escànners i dels documents a llegir.

Límits que presenta l'ordinador

Per ser operatiu, un sistema de lectura ha de ser ràpid. La gran quantitat de dades que han de tractar-se exigeix que els ordinadors siguin molt ràpids i tinguin una memòria RAM important.

Límits que presenta l'escànner

La qualitat òptica i la fiabilitat de l'escànner són una primera limitació. Per llegir caràcters de mida molt petita o molt pròxims és necessari tenir un bon poder separador (resolució). D'aquesta resolució dependrà la possibilitat de llegir o no certs documents. Depenent també del format de l'original que s'ha de digitalitzar, si difereix molt d'un format DIN A4, l'estàndard en molts escànners, hi pot haver dificultats de lectura si no es descompon el document en parts.

Taula I

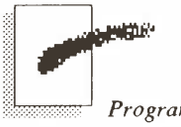
RESULTATS	Paraules per minut	%Reconeixement caràcters	%Reconeixement paraules
MacinText	325	99.53	96.91
Publish Pac	203	99.07	93.82
Read-It	63	98.30	88.65
ReadStar II+	150	96.90	78.98
TextScan	150	96.91	79.54

Límit que presenten els documents a llegir

Els conflictes per al reconeixement són nombrosos i la capacitat per resoldre'ls depèn de l'eficàcia del sistema. L'anàlisi de formes, necessària per al reconeixement de caràcters, implica que la imatge del caràcter ha de ser:

- Aïllada: els caràcters han d'estar separats (no junts). El sistema ha de poder separar caràcters conflictius (com ara lletres itàliques, parelles de caràcters com ara 'fi', 'ff', 'rt' amb tendència a unir-se...).
- Repetitiva: un bon sistema ha de ser capaç de tolerar variacions de tintatge del text, que s'han de reconèixer. En cas contrari, l'exactitud del sistema baixa notablement.
- De qualitat: la qualitat del document que s'ha de llegir influeix en el rendiment de la lectura. Caràcters tallats, en forma dispersa, taques de tinta, menes de paper de poca qualitat, són obstacles seriosos per al reconeixement òptic.

La multiplicitat de tipus de caràcters i també la mescla de tipus fan més dificultosa la identificació de caràcters. És freqüent que certs originals que s'han de digitalitzar, com ara moltes obres impreses, facin ús d'una gran varietat de tipus de lletres, cosa que dificulta el procés de lectura. A més a més, normalment es demana als sistemes de digitalització que siguin capaços de marcar correctament, en el text digitalitzat, els diversos -i en ocasions molt nombrosos- canvis de tipus de lletra que hi ha en l'obra original. Aquest és un procés de gran complexitat en un sistema d'OCR, ja que certs caràcters de diferents tipus de lletra són molt semblants i poden originar confusions notables (de fet existeixen més de deu mil tipus de lletra diferents, encara que normalment tan sols se'n fa servir una desena part i, en una obra normal, se'n solen fer servir entre cinc i quinze).

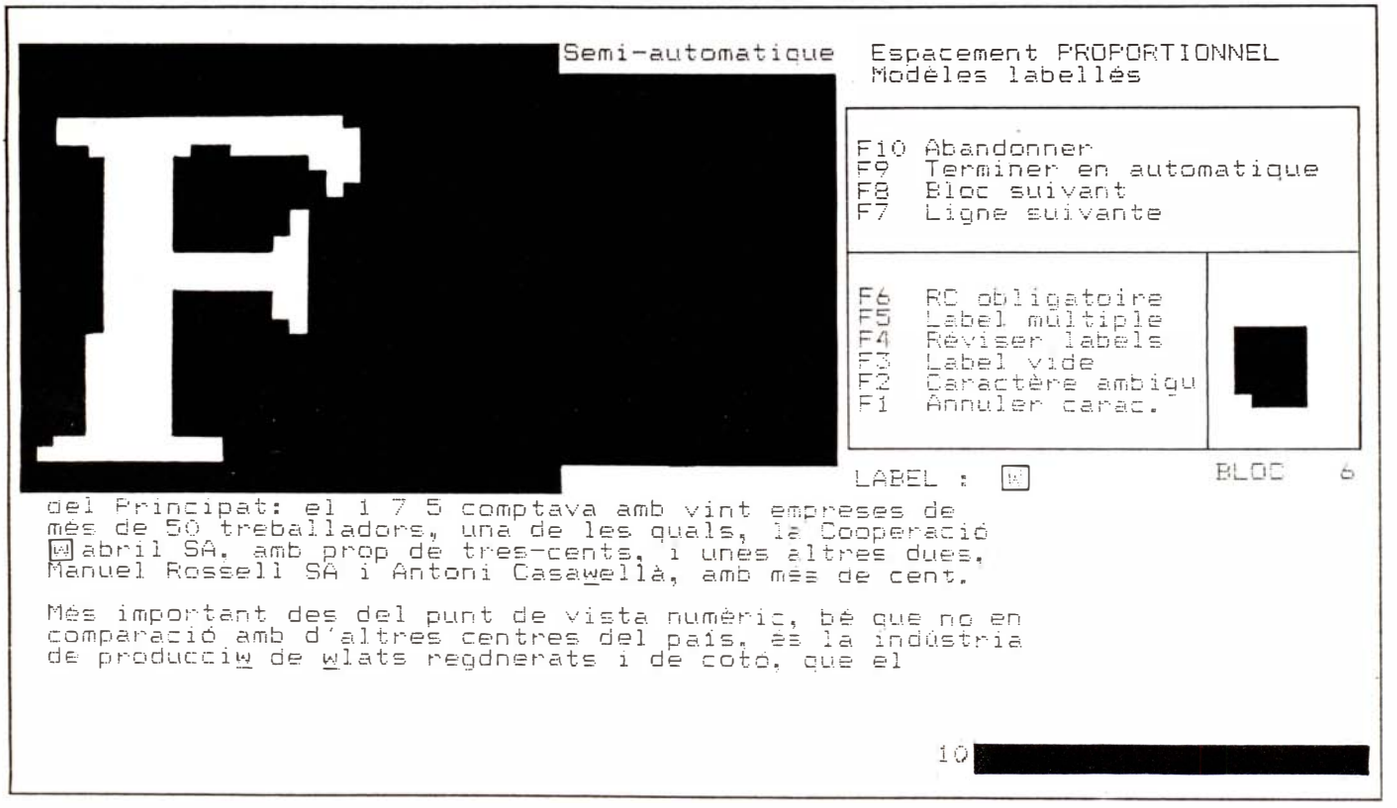


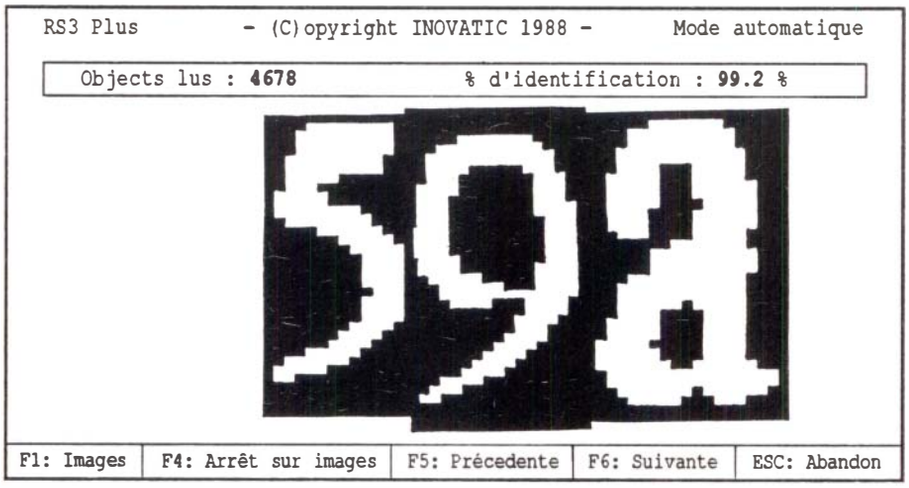
Programa ReadStar III Ver.6.01. Procés d'aprenentatge del programa. En aquesta figura s'observa un moment del procés d'aprenentatge del programa. Com es veu en el text de la part inferior, el programa ja és capaç de reconèixer gran part de l'alfabet, tot i que no reconeix encara certs caràcters de manera correcta (aquests caràcters estan simbolitzats amb una "w"). El programa va fent preguntes sobre tots aquests caràcters que no ha identificat per tal d'aprendre-se'ls per a posteriors lectures (en la figura, el programa no ha pogut identificar la "F" del mot "Fabril" i demana que se li especifiqui quin codi ASCII correspon a aquesta imatge).

També l'organització de la pàgina aporta dificultats, com ara les següents:

- Presència d'elements pertorbadors (logos, imatges...).
- Presència de taules i columnes amb separadors o sense.
- Subratllats.
- Línies de text inclinades.
- Línies irregularment espaiades o de longitud variable.

Aquests diversos aspectes han de ser resolts automàticament pel sistema (recerca de columnes, d'espais, d'interlineats) o de manera interactiva (sistema de finestres, correcció d'errors de lectura, etc.), amb l'objectiu de no fer el sistema lent o ineficaç.





Programa ReadStar III+ Ver.6.01. Procés de lectura automàtica d'un document. En aquesta figura s'observa com el programa de lectura òptica de caràcters digitalitza automàticament un document. Durant el procés, ha digitalitzat 4.678 caràcters i n'ha reconegut bé un 99.2 %, és a dir, 4.641 caràcters han estat identificats correctament mentre que solament 37 no ho han estat. Les imatges que surten a la pantalla representen aquells caràcters que, pel motiu que sigui, no han estat identificats correctament.

AVANTATGES

Són de tipus econòmic i pràctic.

Pel que fa als primers, hi ha el fet que l'entrada manual d'una pàgina dactilografiada que contingui a la ratlla de mil cinc-cents caràcters requereix aproximadament uns quinze minuts de feina amb el cost corresponent que això representa.

L'entrada del text de manera "semiautomàtica", mitjançant un sistema d'OCR, necessita de 45 segons a 3 minuts (és, doncs, de cinc a quinze vegades més ràpid). El cost del treball és, evidentment, de cinc a quinze vegades menor. Així, el guany de temps resulta considerable.

L'entrada de caràcters en modalitat text economitza igualment la memòria d'emmagatzematge. Un document en modalitat text ocupa molta menys memòria de disc magnètic o òptic que un text emmagatzemat en modalitat imatge. Un text d'uns mil cinc-cents caràcters ocupa de 2 a 3 Kbytes emmagatzemat en format ASCII, quan ha estat digitalitzat via OCR, i entre 125 Kbytes i 1 Mbyte en modalitat imatge, quan senzillament s'ha "fotografiat" el document.

Els avantatges pràctics rauen en això: un text digitalitzat en modalitat text és directament utilitzable en programes de TdT, de recerca de paraules i altres, mentre que un text entrat en modalitat imatge (per exemple, per ser arxivat) pot ser solament visualitzat (amb la condició d'utilitzar una pantalla de molt alta resolució) o ser reimprès per una impressora gràfica (impressora làser).

APLICACIONS

La utilitat fonamental d'aquesta tècnica radica en la possiblitat de recuperar textos impresos per reutilitzar-los tal com estan o modificar-los.

Les aplicacions, per tant, poden ser molt nombroses:

- Utilització de textos en TdT o en PAO (Publicació Assistida per Ordinador). Aquí, els potencials usuaris de la tecnologia OCR serien impressors, editors o periodistes. En general, empreses consumidores de paper o de text dactilografiat -com ara bancs o companyies d'assegurances- són, també, susceptibles d'utilitzar aquesta tècnica.
- Construcció de bases de dades. Per a les grans empreses, l'existència de gran nombre de documents dactilografiats no beneficia de cap manera la seva estructura, i prefereixen tenir els documents en base magnètica o en qualsevol altre format susceptible de ser llegit electrònicament. Per a elles, l'OCR pot ser un element a considerar.
- Entrada de textos per a utilització en fotocomposició.
- Entrada de textos per ser utilitzats en un sistema de TAO (Traducció Assistida per Ordinador).
- Recerca lingüística en els textos.
- Recerca documental per «paraules clau».

FUTUR DE L'OCR

En vista del que s'ha comentat fins ara, sembla clar que aquesta nova tecnologia s'anirà instal·lant progressivament en les empreses. La lectura òptica amb reconeixement de caràcters pot portar importants guanys als seus usuaris, encara que certs detractors tinguin notables reserves sobre la seva aplicabilitat actual.

Hi ha interessants rumors que parlen de la pròxima aparició de novetats importants tant de *hardware* com de *software*, en l'àmbit de la tecnologia OCR (empreses com Xerox, DataCopy, Kurzweil, Palentir, entre altres, treballen intensament per millorar aquesta tècnica).

Avui en dia, el reconeixement òptic de caràcters és ja una eina consolidada en el món de la microinformàtica, tot i que encara necessita la indispensable participació de l'operador. L'arribada de sistemes totalment automatitzats serà el pròxim repte, esperat amb impaciència. ■

Lluís Soler i Carrascosa
és Enginyer químic i director tècnic de Comunicació i Càlcul SA

BIBLIOGRAFIA

Carmona, P.: *Del Papel al Disco sin Teclear.* Delibros. Revista profesional del Libro, abril 1989, núm. 11, 61.

CD-ROM 1989-1990 Yearbook. Compiled by Salley Oberlin and Joyce Cox. MICROSOFT PRESS, Washington, 1989.

Dorange, C.: *La Reconnaissance Optique de Caractères: un instrument de choix à la page.* Décision Informatique, maig 1988, núm. 184, 38.