OCR: occhi e cervello per il PC

di Mauro Gandini

Parleremo in questo numero dell'OCR — Optical Character Recognition. Fino ad oggi i sistemi disponibili sul mercato non hanno brillato per affidabilità, ma la tecnica software ha fatto ultimamente grandi passi in questo campo ed i risultati iniziano ad essere interessanti. Vedremo alcune problematiche dell'OCR, una breve comparazione tra alcuni prodotti leader di questo mercato e una prova su strada di un prodotto che, specialmente nell'ambiente Apple Macintosh, sta raccogliendo molti consensi: Typist della Caere

Prologo

Appena la tecnologia ha permesso di poter avere uno scanner sulla propria scrivania a fianco di un personal computer, e più tardi, grazie alla Logitech con i suoi ScanMan, anche nel palmo della mano, c'è stato subito chi ha pensato di andare oltre alla semplice acquisizione di immagini, sebbene i problemi da risolvere siano stati enormi, provando ad acquisire anche il testo: il solo fatto di dover leggere materiale scritto con differenti font di varie grandezze ha comportato infatti grandi limitazioni.

La prima soluzione è consistita nell'applicazione di tecniche di intelligenza artificiale: si trattava in pratica di insegnare al computer a leggere. Di base il programma di lettura riconosceva a grandi linee l'andamento delle differenti lettere, ma non era in grado di riconoscere i differenti font con cui queste lettere venivano utilizzate. In pratica una volta acquistato il programma bisognava attivare la sua parte di apprendimento e iniziare le prove. Per tutte le lettere il programma chiedeva la corrispondenza, fino ad interpretare l'intero alfabeto.

Ovviamente questo metodo era alquanto scomodo, anche se le ultime versioni di questi programmi erano già abbastanza furbe da accontentarsi di leggere un paio di pagine per essere poi operativi con il carattere letto (l'operazione andava poi ripetuta con tutti i tipi di caratteri che si pensava dovessero essere letti dal programma durante la normale attività).

Ma quello a cui volevano arrivare i progettisti era ben altra cosa: volevano infatti che il programma potesse, senza alcuna necessità di training, leggere un po' di tutto anche con caratteri differenti nello stesso testo, di diverse grandezze. Alla fine si è arrivati alla elaborazione di algoritmi estremamente complicati in grado di interpretare testo senza una specifica istruzione.

La nuova generazione

I programmi basati su questi algoritmi sono comunemente chiamati «omnifont» proprio per questa loro abilità di leggere differenti tipi di font senza la necessità di attuare uno stage di «insegnamento». Quasi tutti i produttori di programmi per OCR dichiarano che la precisione di lettura arriva ora al 99 %: forse come dato è un po' esagerato, ma senza dubbio non si discosta molto dall'85-90%.

Ovviamente i risultati dipendono da molti fattori, in primo luogo la bontà del documento che si intende leggere, la qualità dello scanner utilizzato e la sua taratura. Esaminiamo come questi tre fattori influiscono con la qualità dei risultati ottenibili.

Documento da leggere

Si può decidere di acquistare un programma di lettura per diverse ragioni: registrare in forma elettronica dati da vecchi documenti, immagazzinare informazioni tratte da libri e giornali, trasferire materiali stampati in forma elettronica per successive elaborazioni, archiviare tutti i materiali possibili su disco.

Nel primo caso probabilmente ci troveremo in una situazione nella quale i documenti sono simili tra loro, ma di

Typist

Produttore:
Caere Corporation - 100 Cooper Court
Los Gatos, CA 95030.
Distributori:
Per Macintosh: Delta - Via Brodolini, 10
- 21046 Malnate (VA) - Tel. 0332/860780.
Per MS-DOS: S.I.A. - Via Brodolini, 30
21046 Malnate (VA) - Tel. 0332/860795.
Prezzo *(IVA esclusa):*
Vers. Macintosh e vers. MS-DOSL. 1.070.000.

scarsa qualità dovuta all'usura del tempo. Nel secondo caso, invece, dovremmo trovarci nella miglior situazione visto che si tratta di materiale stampato. Il terzo caso ci consente di prestare meno attenzione alla qualità dell'interpretazione da parte del programma di scrittura, in quanto poi il materiale va rielaborato.

Infine, il terzo caso è senza dubbio anche quello dove si richiede l'impegno maggiore poiché la qualità dei documenti è molto variegata: si passa dalla lettera stampata con stampante ad aghi, alla pagina stampata, al telefax (quanto di peggio si possa augurare ad un programma di lettura).

Qualità dello scanner

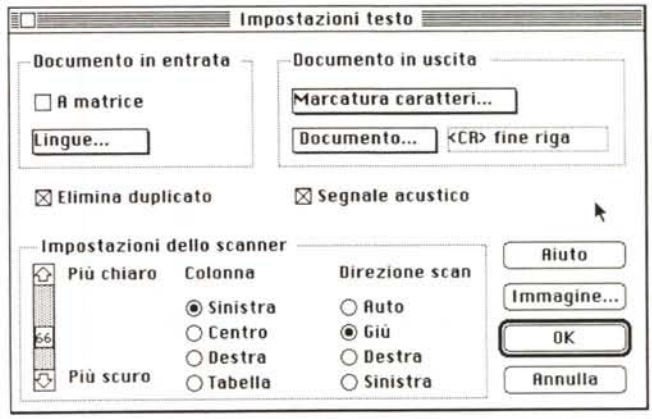
Sul mercato vengono proposti differenti tipi di scanner con differenti risoluzioni (mediamente tra i 200 e i 400 punti per pollice); diverse configurazioni, diversi ambienti operativi (dal mondo MS-DOS all'Apple Macintosh senza tralasciare MS-Windows), diverse dimensioni (dagli scanner a piano fisso in formato A4 agli scanner dello stesso formato a foglio mobile all'ultima generazione di scanner manuali «inventati» dalla Logitech). Utilizzando un vecchio scanner da 200 punti/pollice, i risultati saranno senza dubbio peggiori rispetto ad un moderno scanner da 400 punti per pollice. Inoltre la velocità di interpretazione risulterà notevolmente più alta poiché il carattere in acquisizione risulterà meno ricco di informazioni da interpretare.

Taratura

Anche questo è un fattore molto importante. Molte volte si incolpano gli scanner di scarsa qualità e affidabilità, ma è invece colpa nostra solo perché non si ha la pazienza di cercare la miglior soluzione di luminosità e contrasto per la lettura del documento. Ciò vale indifferentemente sia per la lettura di testo che per l'acquisizione di immagini.

	AccuText	Omnipage	Omnipage 386	Recognize	Recognize	TextPert
Costruttore/Importatore	Xerox/Delta	Caere/Delta	Caere/S.I.A.	DEST/Modo	DEST/Modo	CTA/Thema
Hardware	Macintosh	Macintosh	MS-DOS	Macintosh	MS-DOS	Macintosh
Memoria RAM minima	4 Mb	4 Mb	4 Mb	1 Mb	640 Kb	1 Mb
Errori per pagina:						
Testo in Courier	2	0	0	4	2	15
Testo variato	7	4	3	10	9	10
Testo da MC	3	5	2	18	3	4
File per il salvataggio	5 tra cui: Word Excel	9 tra cui: Word Excel	oltre 30 tra cui: Word RTF	2 ASCII RFT-DCA	2 ASCII RFT-DCA	3 ASCII MacWrite
Scanner supportati	6	11	oltre 20	3	3	oltre 10
Costo (Iva esclusa)	L.1.950.000	L.1.950.000	np	L.1.200.000	L.1.200.000	L.1.900.000

Indirizzi:
Delta - Via Brodolin, 10 - 21046 Malnate (Va) - Tel. 0332/860780
S.I.A. - Via Brodolini, 30 - 21046 Malnate (Va) - Tel. 0332/860795
Modo - Via Masaccio, 11 - 42100 Reggio Emilia - Tel. 0522/512828
Thema - Corso Vitt. Emanuele II, 20 - 12100 Cuneo - Tel. 0171/60983



Questa è la finestra che appare selezionando Typist dal menu mela. Si notino le possibilità di scelta della lingua utilizzata, il controllo della luminosità, direzione di scan e tipo del testo. Selezionando «Immagine...» si attiva il programma per utilizzare Typist come normale scanner.

Qualche trucco

Bisogna anzitutto dire che un testo letto da OCR quasi mai sarà perfetto. Esistono tuttavia due tipi di consigli che vorremmo dare a chi si accinge ad iniziare una attività di lettura con OCR: consigli pre-lettura e consigli di correzione dei testi acquisiti.

Consigli pre-lettura

In pratica si tratta di quanto detto pri-

Caratteri nitidi

Regolare il controllo della luminosità su Più scuro

Regolare il controllo della luminosità su Più chiaro

Ecco tre tipologie di caratteri: il primo caso è il migliore; nel secondo notiamo i caratteri leggermente sfocati a causa del poco contrasto; nel terzo caso i caratteri sono troppo pieni ed è necessario agire sul controllo della luminosità riducendola.

ma sull'utilizzo dello scanner e sulla sua taratura. Prima di tutto ricordiamo di utilizzare lo scanner alla sua massima definizione per ottenere il massimo di informazioni che saranno senza dubbio maggiormente gradite dal vostro programma di OCR. Inoltre ciò comporta anche tempi di interpretazione decisamente più contenuti.

Per quanto riguarda la taratura dei parametri di lettura dello scanner, come per esempio la luminosità (es. lettura documenti su carta colorata), vi consigliamo un approfondito stage di prova con le differenti tipologie di documenti che intendete leggere. Per ogni tipologia di documento preparate quindi un prospetto relativo alla miglior taratura dello scanner per la lettura del documento in questione. Sarete così certi di non dover perdere troppo tempo ogni qualvolta sia necessario leggere un determinato tipo di documenti.

Consigli di correzione

Questi consigli servono per risparmiare tempo in fase di correzione degli errori di lettura. Partiamo proprio dalle differenti tipologie di errori per stabilire le migliori procedure da adottare per la correzione dei testi.

Caratteri di segnalazione — I programmi di lettura utilizzano dei caratteri poco utilizzati per indicare eventuali caratteri irriconoscibili o riconosciuti, ma con un'alta probabilità di errore. Con la funzione di ricerca del vostro programma di scrittura potrete trovarli ed apportare le dovute correzioni.

Spazi multipli — Molte volte i programmi di scrittura non riescono ad interpretare perfettamente gli spazi bianchi tra le varie parole, specialmente se si sta leggendo un testo di tipo giustificato: in questo caso vengono infatti distanziate le parole in maniera regolare, ma non corrispondente alla classica singola spaziatura. Ci si potrebbe quindi trovare con un testo pieno di doppi o tripli spazi. Per eliminarli utilizzeremo la funzione di «cerca e sostituisci» del proprio programma di scrittura.

Si chiederà alla funzione di ricercare due spaziature consecutive e di sostituirle con una singola (utilizzando la possibilità di sostituzione automatica lungo tutto il documento). Per essere sicuri di aver eliminato tutte le doppie spaziature, vi consigliamo di eseguire questa procedura tre volte di seguito: in tal modo sarete sicuri di aver eliminato fino a 8 spaziature consecutive. Con questo sistema si potranno eliminare anche altri caratteri come per esempio tabulatori, ecc. Se il programma di riconoscimento dei caratteri ha inserito nel testo un ritorno a capo ad ogni fine riga incontrata, dovremmo anche in questo caso eseguire un «cerca e sostituisci», ma la procedura sarà leggermente differente. Per prima cosa dovremo essere sicuri che tra ogni paragrafo ci siano due a capo consecutivi (in pratica una linea di spaziatura). Poi andremo a sostituire tutti i doppi a capo con un qualsiasi simbolo non comune (^, #, @, £) oppure con due x (xx). A questo punto elimineremo tutti gli a capo con un cerca e sostituisci: ogni a capo dovrà essere sostituito con uno spazio. Infine si andranno a cercare e sostituire i simboli che avevamo inserito in luogo dei doppi a capo (ovviamente con dei nuovi doppi a capo). L'ultima operazione da svolgere è un bel passaggio con il correttore ortografico. Alcuni programmi di lettura hanno già al proprio interno un correttore ortografico che provvede ad inserire le corrette parole, per esempio avendo interpretato la parola 'femmonile', questa verrà modificata dal correttore in 'femminile'. Alla fine è consigliabile rileggere tutto in quanto esistono una serie di vocaboli che possono sfuggire al correttore come per esempio parole al singolare e plurale (es. «femminile» e «femminili») o altre parole molto simili tra loro (es. «quando» e «quanto»).

Qualche prova

Oltre agli articoli specifici già pubblicati sull'argomento in questa ed in altre rubriche ed apparsi nei precedenti numeri di MCmicrocomputer (ai quali vi rimandiamo per conoscere le caratteristiche dei prodotti già presentati), in questa occasione sono state eseguite anche alcune prove su programmi OCR attualmente disponibili (tra i quali Accu-Text, Omnipage, Recognize, TextPert) e dei quali riportiamo i risultati nella relativa tabella. Per eseguire il test abbiamo provato a far leggere singoli fogli simulando una lettera scritta a macchina (con classico font Courier), testo con differenti font da 8 a 14 punti e la copertina di MCmicrocomputer.

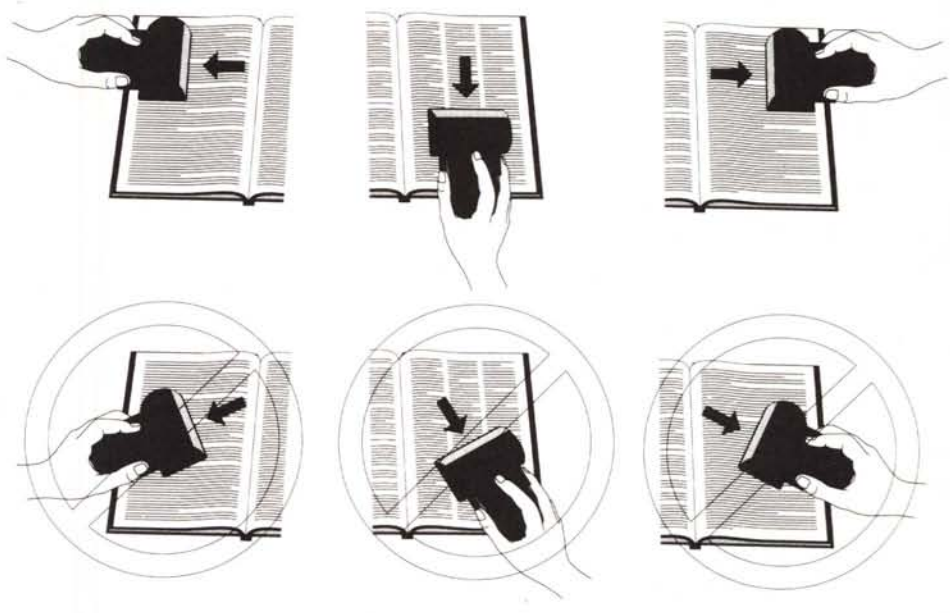
La configurazione utilizzata per i programmi su Mac era un Macintosh II con 8 Mb di RAM, mentre per i programmi MS-DOS abbiamo utilizzato un IBM PS/2 70 con 6 Mb di memoria RAM.

Typist: leggere con la mano

Come già è stato detto in apertura, abbiamo rivolto la nostra attenzione su un prodotto che sta riscuotendo molto successo negli Stati Uniti d'America; un prodotto che non è un semplice programma, ma quasi un sistema integrato.

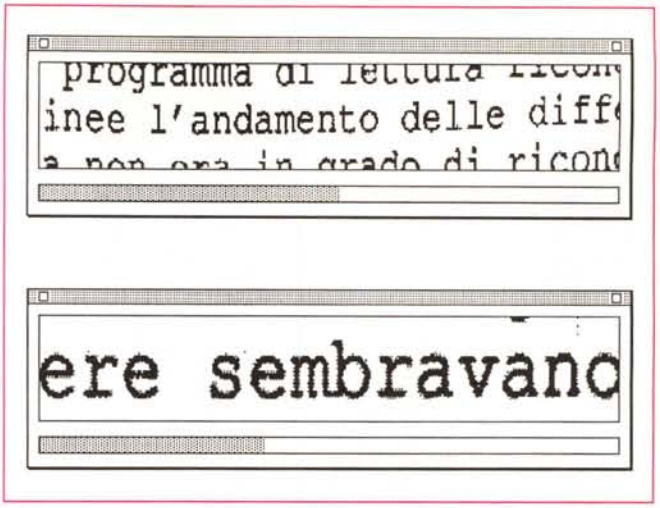
Typist della Caere Corp. è infatti composto da un lettore hardware, in pratica uno scanner manuale a scorrimento sul foglio, di colore nero e di dimensioni non certo ridotte, e da un software di gestione per l'utilizzazione OCR. Ad un costo che comprende non solo il software di lettura, ma anche l'hardware necessario, Typist fornisce prestazioni molto buone.

Typist viene fornito in 3 versioni: per Macintosh (quella da noi utilizzata per la prova), per macchine MS-DOS e per



Angoli di scansione

Queste sono le direzioni di scansione possibili e quelle «vietate».



Una volta finita la lettura inizia l'elaborazione del testo per la sua interpretazione: normalmente il programma agisce a passi successivi. Vediamo infatti che prima l'interpretazione avviene ad un certo grado di definizione e poi ingrandendo le parti non interpretate: per eseguire questa operazione serve moltissima memoria. Il programma visualizza in questo modo l'operazione.

macchine operanti in ambiente Windows 3. Il lettore è identico per tutte le versioni: si tratta, come già detto, di uno scanner manuale di ragguardevoli dimensioni, con una testa di lettura larga 16 cm (12,7 cm di finestra utile) e una comoda impugnatura. Da notare i rulli di gomma di scorrimento che con una distanza tra loro di 9 cm consentono una lettura sufficientemente precisa evitando le possibilità di ondeggiamenti a destra o sinistra.

Nella parte superiore una cupola trasparente consente di vedere cosa sta leggendo lo scanner; lo scanner è poi completato da un grande pulsante per la sua attivazione, una spia per confermare il funzionamento ed un commutatore che consente di regolare la sensibilità alle tonalità di grigio (esiste anche la posizione per leggere immagini al tatto, cioè solo bianchi e neri, da utilizzare quando si usa Typist come lettore di testo).

La versione per Macintosh comprende l'interfaccia per il collegamento alla presa SCSI (l'interfaccia è settata con il commutatore relativo al dispositivo SCSI sul valore di default 6, ma può essere variato a piacere), completa di alimentatore. La versione per PC ha invece una scheda di interfaccia da inserire in uno slot libero. In entrambi i casi il lettore è collegato all'interfaccia mediante un cavo terminato con un connettore di dimensioni sufficientemente comode.

Anche il software risponde alle tre diverse tipologie: per Macintosh, per personal computer MS-DOS e per sistemi in ambiente Windows 3.

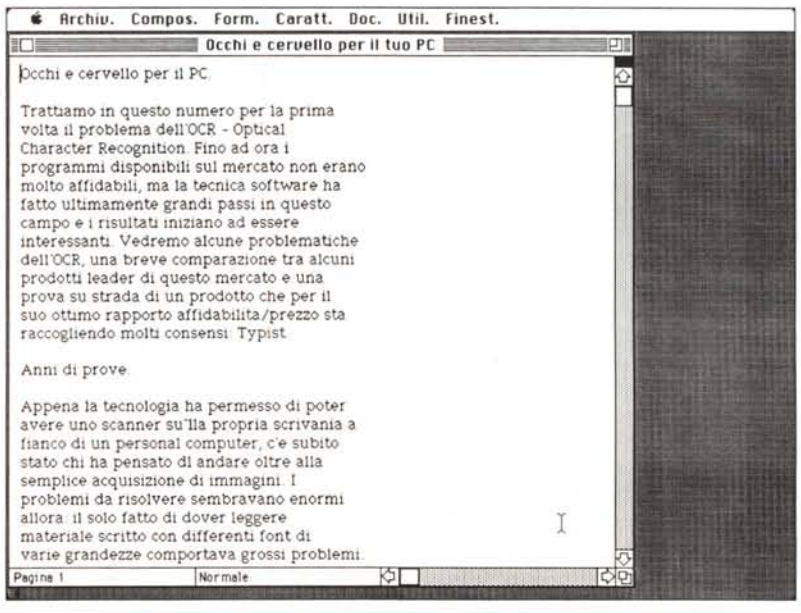
Per installare il software in versione MS-DOS e Windows è sufficiente inserire il solito disco nell'unità del computer e digitare a:TINSTALL. Appariranno alcune maschere con una serie di domande sulla configurazione del sistema, risposto alle quali il software sarà installato secondo i parametri indicati. Se si desidera vengono anche aggiornati i file AUTOEXEC.BAT e WIN.INI: in tal caso il programma viene automaticamente caricato all'accensione del computer (versione per PC) oppure nel momento in cui viene attivato Microsoft Windows 3. Inoltre è possibile ottimizzare la memoria estesa o espansa mediante una utility interna: il funzionamento di questo programma è abbastanza inconsueto, poiché lavora per passi successivi. In pratica avviata l'utility, il computer si blocca e va fatto ripartire, dopo di che si ripete l'avvio dell'utility che si chiama TUNE, ripetendo l'operazione fino a quando non avviene più alcun blocco (3 o 4 passaggi). La configurazione minima del sistema richiesta consiste in un personal computer MS-DOS compatibile con processore 286, 386 o superiori, 640 Kbyte di memoria RAM e 2 Mbyte di memoria espansa o estesa, almeno 2 Mbyte su hard-disk, MS-DOS 3.1 o versioni successive.

Il software per Macintosh consiste in tre programmi: l'accessorio di scrivania che consente la regolazione di alcuni parametri di utilizzo dello scanner, il programma di acquisizione delle immagini e quello di lettura del testo. L'accessorio di scrivania va installato nel System attraverso la classica procedura attraverso il D/A Font Mover, mentre gli altri due programmi sono in pratica degli INIT che devono essere inseriti nella cartella di sistema e che vengono attivati all'accensione del sistema. La configurazione minima richiesta consiste in un Macintosh SE o superiore con 4 Mbyte di memoria RAM (a dire il vero noi siamo riusciti a farlo funzionare anche con 2.5 Mbyte), almeno 2 Mbyte disponibili sul disco rigido, Multifinder e System 6 o versioni successive, un cavo SCSI (stranamente non è fornito in dotazione).

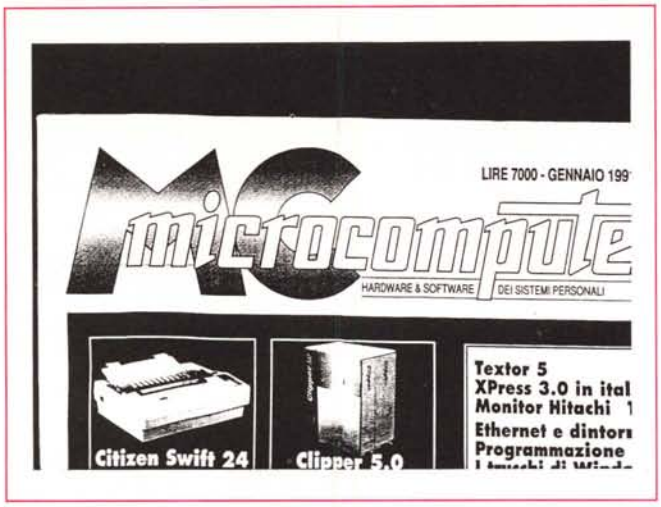
Typist è in grado di leggere qualsiasi carattere non stilizzato da 6 a 72 punti e gruppi di caratteri in 11 lingue differenti. La capacità di scansione è di 258 cm^2 (12,7×20,3 cm), ma se la lettura del testo è eseguita dall'alto in basso e la colonna di testo letta è più stretta dei 12,7 cm della finestra di lettura massima dello scanner, si può raggiungere un'area massima di lettura di 355 cm^2.

Macintosh, attento lettore

Typist è un'applicazione sempre atti-



Ecco il testo interpretato: la prova è stata eseguita stampando la prima parte del testo di questo articolo in Courier 12 punti. Il risultato è molto buono.



Questa è un'immagine acquisita utilizzando Typist come normale scanner.

va in Macintosh per cui necessita di Multifinder, non è assolutamente possibile utilizzarlo con il Finder. La memoria RAM necessaria al solo Typist è di almeno 2 Mbyte: per questa ragione è consigliata una memoria di almeno 4 Mbyte. Il programma di lettura esegue automaticamente un'operazione di «Incolla» del testo appena letto nella applicazione attiva in quel momento: in pratica sarà indispensabile avere aperto un documento di scrittura dove Typist, dopo averlo interpretato, possa inserire il testo.

Una volta attivato dal menu mela l'accessorio della scrivania, appare una finestra di dialogo con alcune richieste di informazione e consenso. Innanzitutto due finestre consentono di indicare alcuni parametri relativi al testo in entrata e al testo in uscita verso il word processor. Le indicazioni sul testo in entrata sono in pratica due: se il testo è stato scritto con stampanti a matrice di punti (9 aghi, non alta definizione) o meno e la lingua in cui è scritto il testo (per la ricerca di caratteri speciali nelle varie lingue, 11 per l'esattezza: danese, francese, inglese, irlandese, italiano, norvegese, olandese, portoghese, spagnolo, svedese e tedesco). Si possono anche tenere attive più lingue contemporaneamente, ma ciò rallenta l'operatività della lettura a causa del maggior numero di controlli necessari.

Il testo in uscita verso il word processor può contenere alcune indicazioni a scelta dell'utente: all'inizio del testo acquisito da scanner può apparire un'indicazione riguardante l'inizio del blocco letto dal computer (es./inizio/); i caratteri non letti possono essere indicati da un simbolo, il programma propone " ~ ", ma possiamo variare il carattere a piacere; stessa cosa dicasi per i caratteri sospetti che vengono di default indicati con ^.

Inoltre, il documento in uscita potrà contenere un ritorno a capo per ogni riga o solo alla fine del paragrafo. Sempre a livello di documento potrà essere richiesta la lettura di un foglio elettronico e la sua uscita sempre su foglio elettronico (ogni 5 spazi Typist inserisce un tabulatore).

Più sotto troviamo altre due richieste: la prima è relativa alla duplicazione del testo, la seconda riguarda l'attivazione o meno di un avviso acustico relativo alla duplicazione. La duplicazione del testo può essere disattivata quando il testo viene acquisito con passaggi multipli (vedremo poi come è possibile ciò) in modo da non ottenere la lettura della stessa riga due volte. L'indicazione acustica è comoda quando si acquisiscono testi di particolare lunghezza: in questi casi il computer può avere la necessità di dedicare alcuni minuti ad esaminare e «leggere» tutto il testo; grazie all'avvisatore acustico sarà possibile dedicarsi ad altre attività sicuri che il computer ci avviserà con un beep al termine dell'elaborazione.

Infine, le impostazioni dello scanner: prima di tutte la luminosità che ci consente di variare le condizioni di ripresa a seconda del tipo di documento che stiamo leggendo, cioè con caratteri (o immagini) chiari o scuri. Segue l'indicazione della posizione della colonna che dobbiamo leggere (valida nel caso di testo su più colonne o tabelle). Infine la direzione della lettura: automatica, verso il basso, verso destra e verso sinistra. In questa maniera sarà possibile riprendere tutte le tipologie di documenti, anche parti di libri, difficilmente accessibili al centro, dove c'è la rilegatura.

Sempre nella finestra di dialogo troviamo anche una opzione che consente di utilizzare lo scanner per le immagini: una volta attivata questa funzione viene richiamato il programmino per la lettura delle immagini e si può iniziare il relativo lavoro di acquisizione. Alla fine l'immagine apparirà a video e potrà essere salvata come immagine TIFF, TIFF compressa o PICT 2 (PCX nel caso di computer MS-DOS). Per meglio regolare la resa alle tonalità di grigio si potrà agire su un piccolo commutatore a 4 posizioni presente sullo scanner stesso.

Ma veniamo alla lettura del testo che avviene in tre fasi: attivazione del lettore e scansione del testo; interpretazione del testo; funzione di trasferimento con una semplice operazione «incolla» del testo letto nel documento di scrittura. Come già detto prima Typist funziona solo sotto Multifinder ed è sempre attivo: basta posizionare lo scanner sul testo, premere il tasto di attivazione, attendere l'accensione della spia verde che indica l'avvenuta attivazione e iniziare la scansione. Finita questa operazione apparirà automaticamente una finestra che visualizzerà le operazioni d'interpretazione e fornirà l'indicazione dell'andamento della stessa. Quando quest'operazione sarà terminata, automaticamente il testo verrà incollato nel documento di scrittura che stiamo preparando, esattamente nel punto in cui abbiamo lasciato il cursore.

A questo punto bisognerà controllare che la lettura sia stata eseguita esattamente adottando le varie procedure di controllo del testo letto consigliate precedentemente.

Con Typist è possibile anche leggere testi più larghi dei 12,7 cm consentiti fisicamente dal lettore: infatti è possibile eseguire letture multiple: il programma penserà ad eliminare le eventuali parti acquisite due volte.

Conclusioni

I risultati sono stati abbastanza buoni, anche se comunque il testo acquisito va senza dubbio rivisto per apportare qualche correzione. Se vengono adottati gli accorgimenti consigliati in questo articolo si riescono ad ottenere risultati migliori. L'importante è perdere un po' di tempo all'inizio per poi non perderne più in seguito.

Senza alcun dubbio il costo consente di acquistare questo dispositivo sapendo già in partenza che le prestazioni sono rapportate al prezzo: al dire il vero bisogna ammettere che la qualità ottenuta nella scansione di immagini è ottima, in alcuni casi migliore di quella di altri scanner di costo molto maggiore.

Per quanto riguarda il riconoscimento dei caratteri (OCR), il discorso non finisce certo qui: vedremo cosa il futuro ci porterà (per esempio la lettura del testo scritto a mano) e soprattutto quale peso potranno avere questi sistemi all'interno del desktop publishing.

MC