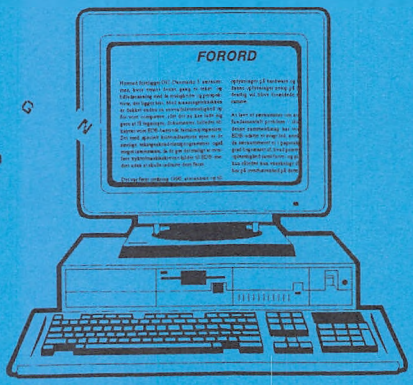
SCANNING

- teknik og erfaringer



DIS-Danmark
Databehandling I Slægtsforskning

INDHOLDSFORTEGNELSE

Redaktion til dette særnummer:

Bent Pilgaard, Randersvej 29, 8800 Viborg. Tlf. 86 67 55 12
Jens Verner Nielsen, Bagergyden 1, Høruphav, 6470 Sydals. Tlf. 74 41 55 80
Svend-Erik Christiansen, Hvedebjergvej 24, 8220 Brabrand. Tlf. 86 25 22 52

Layout:

Hanne Marie Rud, Egebjergtoften 122, 2750 Ballerup. Tlf. 44 66 17 04
Svend-Erik Christiansen, Hvedebjergvej 24, 8220 Brabrand. Tlf. 86 25 22 52

Oplag: 1350
Trykt: "Huset" Århus september 1994.

ISBN 87-90044-01-0

FORORD

Hermed foreligger DIS-Danmarks 3. særnummer, hvor emnet denne gang er tekst- og billedscanning med de muligheder og perspektiver, der ligger heri. Med scanningsteknikken er dukket endnu en anvendelsesmulighed op for vore computere, idet det nu kan lade sig gøre at få tegninger, dokumenter, billeder tilknyttet vore EDB-baserede familieoptegnelser. Set med specielt kildeindtasteres øjne er de særlige tekstgenkendelsesprogrammer også meget interessante, da de gør det muligt at overføre trykte/maskinskrevne kilder til EDB-mediet uden at skulle indtaste dem først.

Det var først omkring 1990, at scannere og tilhørende programmer reelt blev tilgængelig for private PC-brugere, men da det hele stadig er relativt nyt, er der endnu kun skrevet meget lidt herom. Der sker hele tiden nye landvindinger og forbedringer, og det skrevne skal ikke være ret gammelt, før det indeholder forældede oplysninger.

I opbygning er dette særnummer om scanning lidt specielt, idet vi har valgt at lave et to-delt nummer, hvor første halvdel er en bred orientering om emnet, mens anden halvdel er en række brugeres erfaringer med scanning i forskellige sammenhænge. I den brede orientering om emnet har vi på flere punkter været nødt til at støtte os til, hvad der er skrevet om emnet med de aktualitetsproblemer, som hermed følger. Vi har valgt at nedprioritere prisoplysninger på hardware og software, da sådanne oplysninger netop på dette felt formodentlig vil blive forældede næsten med det samme.

At lave et særnummer om scanning, giver et fundamentalt problem - illustrationerne. I denne sammenhæng har vores almindelige EDB-udstyr et svagt led, nemlig printerne, og da særnummeret er i papirudgave, er vi i høj grad begrænset af, hvad printerne formår. Højopløselighed samt farve- og gråtonescanninger kan således kun vanskeligt illustreres. Vi håber på overbærenhed på dette felt.

Dette særnummer har i høj grad været inspireret af DIS-medlem Erik Helmer Nielsens meget informative dobbeltartikel i Slægt & Data 1993 nr. 2 og 3 om scanning.

En særlig tak skal også rettes til DIS-medlem John Thomsen for en række korrigerende oplysninger til særnummerets orienterende del.

Nærværende særnummer er udsendt gratis til alle medlemmer af foreningen DIS-Danmark sammen med medlemsbladet Slægt & Data i september 1994, men det vil i begrænset omfang være mulig at købe for ikke-medlemmer.

Bent Pilgaard
Jens Verner Nielsen
Svend-Erik Christiansen

1. INDLEDNING

Da den personlige computer (PC'en) blev lanceret omkring 1981 var der næppe ret mange, der havde forestillet sig, hvilken revolutionerende udvikling der hermed blev sat igang. PC'en har på de relativt få år fået en enorm udbredelse, og anvendelsesmulighederne har vist sig at være utallige. Ja, det er faktisk kun fantasien, der sætter begrænsninger, og det er hele er gået så stærkt i alle mulige retninger, at vi almindelige brugere har meget svært ved at følge med. Ikke blot med hensyn informationer og viden, men også med hensyn til EDB-udstyr. Det udstyr som i dag er det nyeste og bedste, er i morgen blevet overhalet af den tekniske udvikling og ikke tidssvarende mere.

I dag, kun 13 år efter PC-ens fremkomst, er vi på vej ind i en såkaldt multimedie-verden, hvor computeren bliver et redskab, der kan benyttes i masser af medie-relationer. Lyd og levende billeder er ved at vinde indpas, og telekommunikation er ved at indgå som en naturlig ting ved moderne PC-er. De CD-drev, som kan erhverves til vore computere er endnu kun beregnet til læsning, men vil formodentlig om få år kunne fås til både læsning og skrivning, og almindelige CD-ere med musik og video vil kunne køre på vore personlige computere.

PC'en startede som en lidt avanceret skrivemaskine med tekstbehandlingsprogram og lidt programmeringsmuligheder, men snart blev der udviklet software (programmer) til stort set alt muligt bl.a. til slægtsforskning. Hardwaren (EDB-udstyret) blev løbende videreudviklet og forbedret, sort-hvid-skærmen blev til farveskærm, standard-disketten blev ændret fra 8" via 5 1/4 „ til 3 1/2", og CD-en ser ud til at blive det næste trin, printerne udviklede sig fra 9-nåls matrixprintere til laserprintere o.s.v.

På skærmen er vi efterhånden gået over i grafikkens verden, og har nu mulighed for at arbejde med tegninger og billeder, og kan derved supplere den skrevne tekst på en ny måde. Flere og flere privatpersoner anskaffer sig scanner-udstyr, så de selv kan overføre billeder fra papir og til computeren. Slægtsbøger og slægtsnotater kan nu udvides med f.eks. indscannede slægtsvåben og familiebilleder, og det hele har mulighed for at indgå som en del af de slægtshistoriske optegnelser, som man har liggende, f.eks. i sit slægtsforskningsprogram.

Udstyret til scanning kan fås i mange forskellige prisklasser, udformninger og kvaliteter, og de tilhørende programmer giver en række vidt forskellige muligheder. Der er dog tale om to hovedtyper af programmer til scannere, nemlig billedbehandlingsprogrammer og genkendelsesprogrammer (også kaldet OCR-programmer).

I dette særnummer er det vores hensigt at give en nærmere gennemgang af teknikken omkring scanning, og fortælle om de anvendelsesmuligheder det giver; både med hensyn til almindelig billedscanning og med hensyn til tekstgenkendelse. For at det hele ikke skal blive rent teori, har vi i den sidste halvdel af dette særnummer fået en række brugere til at fortælle om deres erfaringer og problemer i forbindelse med billed- og tekstscanning.

2. DEN TEKNISKE SIDE

For at kunne overføre et dokument (billede) fra papirform til EDB-form er det nødvendig at få det omsat til digitale værdier, som computeren kan arbejde med. Til dette formål er det nødvendigt at have et apparat, der nøje kigger hele dokumentet igennnem, og oversætter hvert lille plet til et signal, der afhænger af farven det pågældende sted. Efter det engelske ord „scan", der betyder „undersøge nøje, se nøje på", kaldes et sådant apparat for en scanner.

Det centrale i scanneren er en særlig fotosensor, der kaldes for en CCD (Charge-Coupled Device), og den kan opfattes som en multi-foto-

celle, der typisk består af et par tusinde små fotoelektriske elementer. Antallet og størrelsen af dem har betydning for billedopløseligheden. Ved lyspåvirkning (d.v.s. bombardering af lyspartikler - fotoner) udsender CCD-en en elektrisk spænding (elektroner frigives), som i størrelse er afhænger af lysmængden, der rammer den.

Scanneren indeholder desuden en lyskilde, som ved scanningen benyttes til at belyse det aktuelle dokument. Det reflekterende lys bliver via en linse og små spejle sendt til CCD-en. Hele teknikken hviler på, at farver reflekterer lys forskelligt. Fra et mørkt område på dokumentet vil scannerens belysning kun resultere i en lille lysreflektering til CCD-enheden, som så kun udsender en lav spænding, hvorimod et lyst område vil betyde en stor reflektering og udsendelse af en tilsvarende høj spænding. Det betyder altså, at den udsendte spændings størrelse vil afhænge af, hvor vi er i farveskalaen. Det er langt hen af vejen den samme teknik, som kopimaskiner og fax-maskiner benytter.

Det viser sig ofte, at farverne ikke bliver gengivet helt korrekt, og det er derfor nødvendigt med justering (også kaldet kalibrering) af scanneren. Det foregår almindeligvis ved, at man lade den indstille ud fra et standard farvekort eller en såkaldt gråkile, hvor en række nøje valgte gråtoner er angivet.

Selve scanningen foregår normalt ved at dokumentet med en jævn hastighed og linievis fra den ene ende til den anden bliver belyst/aflæst/digitaliseret. Nogle typer af scannere (flatbed- scannere m.fl.) har indbygget denne aflæsning hen over dokumentet, mens andre scannere (håndscannere) kræver, at man selv manuelt foretager processen.

Er der tale om indscanning af farvebilleder kræver nogle farvescannere, at der scannes tre gange (3-pass scanning); en gang for hver af grundfarvene rød, grøn og blå. Denne såkaldte RGB-repræsentation kender vi også fra fjernsynsmodtagere (i forbindelse med trykning på papir bruges CMYK-farverne, der er blå, rød, gul og sort). Ved hjælp af et særligt filter for hver af disse bliver farverne aflæst ved scanningen. Nogle af de mere avancerede scannere er dog i stand til at foretage aflæsningen af de tre farver i en arbejdsgang (1-pass scanning). Alt andet lige opnås den mest nøjagtige gengivelse med 1-pass-scannere, idet resultatet her ikke først skal sammensættes ud fra tre mekaniske aflæsninger, som let kan være en lille anelse unøjagtige.

Via en såkaldt ADC (Analog Digital Converter) bliver de forskellige spændinger fra de fotoelektriske elementer konverteret til nogle digitale værdier, og dermed bliver dokumentet omsat til noget, som computeren kan forstå og arbejde videre med.

Billedopløsning

Ved den linievise aflæsning af dokumenter opfattes hver linie som en række punkter, der hver især får tilknyttet en digital værdi. Det sammensatte billede kender vi også fra aviser og ugeblade, hvor punkttætheden dog efterhånden er blevet så stor, at man skal bruge lup for at se, at billederne er sammensat af prikker.

Tætheden af linierne og punkterne kan normalt varieres med scanneren og det tilhørende program, og jo mere finkornet scanningen bliver, des mere nøjagtig bliver den i forhold til originaldokumentet. Punkttætheden måles i DPI (Dots Per Inch), d.v.s. antal punkter pr. tomme (2,54 cm). De fleste scannere og tilhørende programmer kan variere læsetætheden fra 25 x 25 til 400 x 400 DPI, men meget avancerede scannere kan arbejde med en betydelig større punkttæthed, f.eks. 2000 x 2000 DPI.

Billedopløseligheden kan imidlertid ikke betragtes alene. Hvert punkt er på computeren

repræsenteret ved en digital værdi, og i takt med den stigende punkttæthed sker der en kraftig vækst i størrelsen af det indscannede målt i antal digitale værdier.

Eksempelvis vil et dokument på 10 x 10 cm (3,9" x 3,9"), som er scannet med 256 gråtoner (8 bits registreringer) ved 100 DPI fylde ca. 150 KB, men øges scannekvaliteten til 300 DPI, så vil det pludselig fylde ca. 1,4 MB. Ikke blot fylder det indscannede betydelig mere, men det bliver også lidt tungere og langsommere at arbejde med på PC'en. Med den nuværende teknik kan der ikke ændres på de store pladskrav, men udviklingen giver os større og hurtigere computere, så problemer efterhånden reduceres.

Som en lille advarsel skal det lige nævnes, at angivelserne af en scanners billedopløsning (punkttæthed - DPI) kan gøres på to måder. Det kan rent fysisk være den punkttæthed, hvormed et billede kan indscannes, men det kan også være punkttætheden efter at de aflæste punkter har været igennem nogle beregninger, hvor der konstrueres nye punkter, som er middelværdier af de omkringliggende (interpolering). D.v.s. at punkttætheden øges ved at „udglatte" de aflæste farver.

For billedfiler gælder naturligvis også som for tekst og datafiler, at de ved hjælp af såkaldte pakkeprogrammer (ZIP, ARJ o.s.v.) kan komprimeres til lagring. Afhængig af program og billedets kompleksitet er det muligt at reducere den lagerplads et billede optager til et sted mellem 1/10 og 1/100.

Endelig bør den valgte punkttæthed ses i sammenhæng med skærmens opløselighed og naturligvis også printerens opløselighed, hvis det indscannede efterfølgende skal skrives ud. De fleste inkjet- og laserprintere har vel i dag henholdsvis 360 x 360 DPI og 600 x 600 DPI som bedste opløsning ved udskrifter, men det må ikke glemmes, at der normalt er tale om totone-udskrifter (sort og hvid), hvor det er nødvendigt at indlægge raster (se senere herom).

Filformater

Der er desværre ingen standard, når det gælder billedfilernes opbygning og format, d.v.s. den måde hvorpå de aflæste oplysninger om farverne i de scannede punkter er organiseret i filen. Indscannede billeder kan således ikke uden videre læses af alle systemer. Af kendte filformater kan nævnes BMP, PCX, TGA, EPS, GIF, TIF, CLP, IMG og JPG. Billedfilens format kan aflæses af filtypenavnet (extension).

Billedet er gengivet med henholdsvis 75 DPI, 150 DPI og 300 DPI. At billedet er sammensat af punkter ses lettest til venstre ved 75 DPI. Gengivelsen her har kvalitetsmæssig slagside.

Tre filformater skal her fremhæves, og det er TIFF (TIF), BMP og PCX. Meget udbredt på markedet er TIFF-formatet (Target Image File Format), der er udviklet af firmaerne Aldus og Microsoft, og som bruges både blandt gråtone-scannere og blandt fuldfarve-scannere; det er nærmest blevet et standard-format. De to andre formater (BMP og PCX) er måske især kendt blandt Windows-brugere, idet det medfølgende Paintbrush-program opererer med de to formater. Brugere af slægtsforskningsprogrammet Brothers Keeper, har måske også stødt på formatet PCX, idet billeder kan tilkobles i dette format.

Computerens videre behandling af det digitaliserede dokument afhænger nu af det behandlende program; om der er tale om billedbehandling eller tekstgenkendelse.

3. BILLEDSCANNING

Med billedbehandlingsprogrammer bliver det indscannede dokument kun at opfatte som et billede, der kan manipuleres med på forskellige måder (se herom senere). Det er naturligvis også mulig at indscanne tekster som billeder med disse programmer, men der er stadigvæk kun tale om et billede, og det er ikke muligt at arbejde videre med det i tekstbehandling/databaser eller søge på ord eller lignende.

Selv om indscannede billeder/dokumenter optager meget lagerplads, så fylder billederne rent fysisk ikke ret meget i forhold til de tilsvarende papirudgaver. For mange virksomheder, institutioner, arkiver o.l. er mængden af papirbaserede oplysninger, som skal gemmes, ved at være så stor, at det udgør et pladsmæssigt problem. Med dette udgangspunkt vil der utvivlsomt komme flere og flere tilfælde fremover, hvor papirdokumenterne vil blive kasseret til fordel for en indscannet udgave af dem. For arkivernes vedkommende er det imidlertid ikke helt så ligetil, for man kan ikke sådan bare opgive originalkilden til fordel for en scannet kopi, men en indscanning kunne blive en vigtig brik i kopispredningen.

På længere sigt vil der utvivlsomt blive et marked for bl.a. indscannede bøger (nok især opslagsværker). Til dette formål rummer disketter med 1,44 MB alt for få data, og det hele er meget afhængig af masseudbredelsen af medier med betydelig større kapacitet f.eks. CD-ROM. Ikke alene med hensyn til det fysiske omfang, men også prismæssigt vil CD-ROM-udgaver let kunne konkurrere med den trykte bog. Fra 1992 til 1993 fem-dobledes salget af CD-drev til computere, og det ser ud til, at CD-ROM nu vil slå igennem for alvor.

Billedmanipulation

Billedbehandlingsprogrammer giver en række muligheder for at viderebearbejde det indscannede. Det kan være i form af beskæringer, størrelsesændringer, farvejusteringer o.s.v., men det er også muligt at foretage egentlige ændringer af billedet, som f.eks. at retouchere noget bort eller at føje noget til.

Nogle programmer gør det muligt at forbedre skarpheden i grå-tone-billeder. Det foregår ved, at der i grænseområdet mellem to forskellige gråtoner sker en „rensning" af farverne, så overgangen mellem de to gråtoner bliver mere brat. Andre programmer opererer med den såkaldte gammakorrigering, hvor der ved indlæsningen eller senere er muligt at øge kontrasterne i området med næsten ens gråtoner.

Indscanninger af dokumenter kan foregå på tre trin. Enten som to-tone-scanning, gråtone-scanning eller som farve-scanning.

To-tone-scanning

Den mest simple type scannere er de såkaldte bi-level-scannere, der som navnet antyder, kun kan skelne mellem to farver, nemlig sort og

hvid. Alle farver vil blive omsat til digitale koder for sort eller hvid. Da der således kun kan skelnes mellem to farver, er en sådan scanning kun brugbar til simple logoer, tegninger, tekst (som billede) o.l. På trods af navnet indeholder sort/hvide fotografier egentlig også en række gråtoner, og de vil miste detaljer ved indscanningen, og to-tone-systemet er derfor ikke særlig velegnet til disse; det kan sammenlignes lidt med en gammeldags fotokopi af et sort/hvid familiebillede.

Farvemæssigt har hvert indscannede punkt kun to muligheder (sort, hvid), og det kan digitalt repræsenteres med 1 bit, som netop har to muligheder (1 eller 0).

Gråtone-scanning

For at kunne få et rimeligt resultat med sort/hvide billeder skal man over i gråtone-scanning, hvor der mellem de to yderfarver sort og hvid også kan skelnes mellem en række gråtoner. Det drejer sig typisk om 16, 64 eller 256 forskellige gråtoner. Jo flere gråtoner, der kan skelnes imellem, des mere nøjagtigt kan farverne gengives på et sort/hvid fotografi.

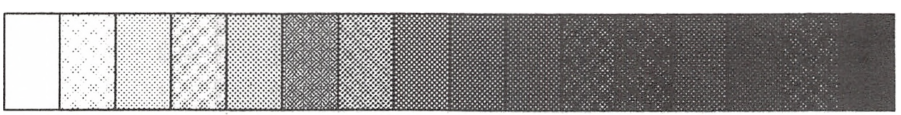
I betragtning af det menneskelige øjes begrænsninger ville det være nok med blot 64 nuancer, men da scannere typisk ikke har en jævn opdeling af gråtoneskalaen anbefales en større opdeling, og 256 gråtoner er næsten blevet standard i dag. Skulle gengivelsen af sort/hvid fotografiet være helt perfekt, så er end ikke 256 gråtoner tilstrækkelig for et sådant billede kan faktisk have et næsten uendeligt antal af forskellige nuancer mellem farverne sort og hvid.

Da hvert indscannede billedpunkt kan antage 256 (=2x2x2x2x2x2x2x2) forskellige gråtoner, skal der lagermæssigt hele 8 bit (=Byte) til at repræsentere hvert punkt. Alt andet lige vil et gråtone-billede (256) optage otte gange mere lagerplads end det samme billede i en to-tone udgave. En gråtone-scanner kan naturlivis også bruges til at lave de mere primitive to-tone-scanninger.

Farve-scanning

Nogle farve-scannere kan kun registrere gengive 4096 farver, og selv om det lyder af meget, så giver det i nogle sammenhænge en utilstrækkelig gengivelse. De 4096 farver fortæller, at DDA-enheden kun kan adskille 16 forskellige udgaver af hver af de tre grundfarver rød, grøn og blå. Farveovergangene vil ikke blive så jævne, som på det originale forlæg. Det er naturligvis også muligt at scanne sort/hvid fotografier med denne farveopløsning, men det kan også kun blive med 16 grå-toner, og det giver som tidligere nævnt en noget ringere billedgengivelse med tab af detaljer.

Scannere med 16,7 millioner farver (d.v.s. 256 nuancer af hver af farverne rød, grøn og blå = 256 x 256 x 256 = 16,7 mill.) er i dag de mest



Eksempel på gråtoneskala med 16 forskellige grånuancer. Bemærk, at de grå nuancer udelukkende er fremkommet ved forskellige tætheder af sort (og hvid), og at "punkterne" også kan have forskellige størrelser. Det ses tydeligst ved 2. og 4. farve fra venstre. Muligvis kan det af skalaen her ses, at nogle af gråtonerne giver et forstyrrende monster. Denne gråtoneskala er udskrevet ved hjælp af tegneprogrammet Paintbrush, der medfølger Windows operativsystem.

almindelige indenfor farvescannere. En scanning af et farvefotografi vil med dette antal farvenuancer kunne give et resultat, der er meget tæt op af forlægget og et resultat, som er betydelig bedre end ved de nævnte 4096-farvescannere.

I forhold til 256-gråtone-scanning vil billedfilen ved farvescanning med 16,7 millioner farver kun fylde 3 gange så meget, idet hver scannet punkt farvemæssigt kan beskrives med 3 Byte; en Byte (8 bit med 256 muligheder) for hver af de tre grundfarver.

Af hensyn til hastighed, den kommende billedfils størrelse og hensigten med det indscannede er det en god idé, at der bliver scannet med de indstillinger, som passer bedst i hvert enkelt tilfælde. At foretage en fuldfarve-scanning med 1200 DPI af en sort/hvid stregtegning vil vist være at skyde lidt over målet.

Princippet med at scanneren aflæser et billede som en række punkter kan give et lidt specielt problem, hvis det, som skal scannes, er et rasterbillede, d.v.s. et billede, som i forvejen består af punkter. Her kan det være svært at få scanneropløsningen til at passe med rasteret, så man populært sagt ikke kommer til at læse mellem punkterne.

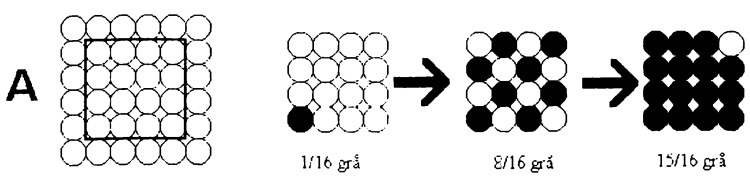
Det indscannede vist på skærmen

Det siger sig selv, at man ikke får det fulde udbytte og de fulde redigeringsmuligheder med det indscannede billede, hvis man ikke har en computerskærm, der kan gengive billedet med mindst samme kvalitet, som scanneren og det tilhørende program kan optage det. En farveskærm med et almindeligt VGA-skærmkort kan eksempelvis kun gengive 256 farver i en opløsning på kun 320*200 punkter (dots), og det er ikke nok, hvis der er tale om en fuldfarve-scanning med stor opløselighed.

Udskrivning af billeder

I mange sammenhænge er det indscannede billede bare en mellemstation, inden det igen skal skrives ud på papir. De almindelige sort/hvid printere (matrix-, inkjet-, laser-printere) kan ikke umiddelbart udskrive et acceptabelt billede med gråtoner. Det er nødvendigt, at indlægge raster i billedet først, d.v.s. at det skal omformes til et billede af sorte prikker, hvor det er muligt at simulere gråtonerne ved enten at operere med forskellige størrelser af prikkerne eller ved at ændre afstanden mellem disse. Jo tættere prikkerne står, des mørkere bliver den gråtone, som vi opfatter med øjet. Metoden nedsætter dog billedets opløselighed (kvalitet), idet f.eks. 16 gråtoner kræver 4*4 punkter, og det vil reducere et 300 DPI-billede til kun 75 DPI. I fagsproget hedder denne metode til at skabe gråtoner via varierende punkttætheder for „dithering".

Ved udskrivninger viser det sig imidlertid, at det „visuelt" er bedst, når punkterne (rasteret) ikke er placeret direkte vandret/lodret. En drejning på f.eks. 15 grader giver et bedre resultat. Ved farvetryk bliver dette yderligere kompli-



Udprintning af et 400 DPI-billede (A) med 16 gråtoner på en almindelig sort hvid laserprinter bruger 4x4 dots til at lave hver af de 16 gråtoner, d.v.s. at det udprintede billede faktisk kun svarer til 100 DPI.

ceret, idet raster for hver farve skal lægges med forskellige drejninger, så der ikke opstår visuelle mønstre i gengivelsen.

4. TEKSTGENKENDELSE

Med billedbehandlingsprogrammer var det naturligvis også muligt at arbejde med indscannede tekster, men de blev kun opfattet som billeder. I de senere år er der imidlertid blevet udviklet særlige tekstgenkendelsesprogrammer, som efter forskellige systemer er i stand til at genkende bogstaverne i en indscannet trykt/maskinskrevet tekst. Den genkendte tekst kan efterfølgende manipuleres med, som enhver anden tekst i et tekstbehandlingsprogram. På det danske marked findes kun en lille håndfuld forskellige programmer af varierende kvalitet, og prismæssigt er de endnu lidt for dyre for os privatpersoner. Disse tekstgenkendelsesprogrammer betegnes ofte OCR-programmer, hvor OCR er en forkortelse for det engelske Optical Character Recognition. Tekstgenkendelse sker normalt på to-tone-niveau (sort/hvid).

Der eksisterer i dag to forskellige grundprincipper, hvorpå OCR-programmerne kan genkende tegn i en indscannede tekst. Enten ved en såkaldt mønstergenkendelse (på engelsk kaldet „matrix matching"/"pattern matching") eller den mere avancerede såkaldte topologiske genkendelse (på engelsk kaldet „feature extraction").

Matrix matching

Som navnet antyder, drejer det sig her om at undersøge om de indscannede bogstaver matcher med nogle i forvejen definerede. Forud for scanningen har programmet fået defineret hvert bogstav i en bestemt font (bogstavsæt) og størrelse, og de gemmes som bitmaps (matricer) i et bibliotek (database). Her er hvert indlæst bogstav beskrevet ved en række punkter. Dette system kan kun håndtere et begrænset antal fonte.

Efter indscanningen af en tekstside finder programmet kort fortalt frem til linierne med tekst ved at lokalisere, hvor tryksværten er koncentreret. I de fundne linier søger fra venstre mod højre efter bogstavmellemrum, så bogstaverne kan afgrænses. Hvert formodede bogstav bliver sammenlignet med bitmaps-biblioteket, og hvis der er en bitmap som passer, så er der sket en genkendelse, og (normalt) afleveres en den tilhørende ASCII- eller ANSI-værdi til resultatfilen.

Denne mønstergenkendelse gælder som nævnt kun den aktuelle font og en konkret bogstavstørrelse, og da der findes mange forskellige fonte og størrelser, er det nærmest nødvendigt med en database for hver font og størrelse. Det er meget besværligt, ikke mindst fordi der indenfor hver font faktisk også er store variationer.

Ikke sjældent vil der blive behov for at oprette bitmaps for et nyt bogstavsæt, og programmet må igennem en indlæsningsfase. Det foregår normalt ved, at man indscanner en hel side med den ukendte font, for efterfølgende at gennemgå linierne bogstav for bogstav og definere hvert bogstav for programmet, der så gemmer bitmaps af dem. Ofte vil det ikke være nok med en enkelt indlæsning af hvert bogstav, da den valgte prototype til bitmapkartoteket kan vise sig at have en lille og næsten usynlig uregelmæssighed. Ujævnheder i papiret og slidte/uskarpe bogstavtyper øger problemet med at finde det rigtige „forbillede" til programmet.

Hvis den indscannede tekst indeholder forskellige fonte og bogstavstørrelser, hvilket ofte er tilfældet, skaber dette naturligvis også problemer, akkurat som når der bruges kursivering i teksten. Et andet genkendelsesproblem i denne sammenhæng gælder tekster, som ganske vist er skrevet med en bestemt font, men med brug af proportional skrift, d.v.s. hvor bogstaverne

får en bredde, som er afhængig deres udseende. Dermed er der ikke mere tale om, at bogstaverne er at finde med faste intervaller på linierne. Et klassisk problem er her sammenblandingen af „rn" med bogstavet „m".

De bedste læseresultater nås ved klare og tydelige tekster, og tekster skrevet med lidt brugte farvebånd eller matrixprintere er ikke sagen, men det kan godt lade sig gøre. En justering af scannerens belysning af forlægget kan dog afhjælpe nogle af disse problemer.

Feature extraction

Et mere avanceret og nyere system til bogstavgenkendelse er den såkaldte feature extraction, hvor der ikke alene er direkte mulighed for samtidig at genkende tekst skrevet med forskellige fonte, men også med forskellige størrelser; derfor kaldes systemet også for omnifont-genkendelse. I denne sammenhæng regnes udgaver med fed, kursiv eller understreget skrift for at være forskellige; der er stadigvæk kun tale om et begrænset antal fonte, som kan være indbygget.

Genkendelsen af bogstaverne foregår ved feature extraction ikke ved at sammenligne med punkter i bogstaverne, men i stedet ved at sammenligne med karakteristiske træk ved hvert bogstav. I en database gemmes informationer om bogstavernes topologi, d.v.s. deres opbygning og sammensætning med hensyn til linier, vinkler, kurver og cirkler. Ved at sammenligne med disse oplysninger bliver bogstaverne genkendt.

Denne måde at beskrive og opfatte bogstaverne på er ganske vist mere kompliceret, men systemet er mere smidigt i brug. Bogstaverne behøver her ikke at ligne forlægget i detaljer, men kan nøjes med at ligne i struktur for at blive genkendt.

Generelt er feature extraction i øvrigt bedre til at håndtere bl.a. flerspaltede tekster og til at gå uden om grafik/billeder i teksten, end det var tilfældet ved matrix matching.

Feature extraction og ICR

Til professionelt brug findes nogle meget avancerede OCR-programmer, som imidlertid ikke kan køre på en almindelig PC; de kræver en anden og større processor. Basis er også feature extraction, men der udover er yderligere tilkoblet særlige læsekontrollerende programmer. Det kan være en indbygget stavekontrol og en „intelligent" læsekontrol, som støtter sig til regler for bestemte bogstavsammensætningers hyppighed og eventuelle umulighed. Dette er med til at gøre bogstav- og ordgenkendelsen endnu mere sikker. Et bogstav bliver således ikke alene bestemt ud fra det observerede udseende, men også ud fra hvilke bogstaver, der står omkring det.

Med denne udvidelse af tekstgenkendelsen bliver programmerne bedre til at adskille f.eks. „rn" og „m", „c" og „e", „l" og „i", „5" og „S", „O" og „0". Nogle skriftfonte er mere drilske end andre, og i den font, som vi har brugt i dette særnummer (Times), er der eksempelvis ikke stor forskel på bogstaverne „i" og „r", idet sidstnævntes øverste streg til højre ikke er andet en meget tyndt forbundet prik, og som ellers svarer til prikken over i'et. Denne type af programmer giver således betydeligt færre fejl og stiller ikke de samme strenge krav til trykke-kvaliteten af den indscannede tekst.

Disse avancerede OCR-programmer henvender sig imidlertid ikke til os private brugere, idet de prismæssigt (endnu) er oppe i 100.000-kroners-klassen. På grund af den tilknyttede „intelligens" kaldes disse ofte for ICR-programmer (Intelligent Character Recognition).

ICR-programmernes tekstgenkendelsesmetode svarer faktisk meget nøje til den måde, hvorpå vi manuelt tyder en svært læselig, håndskrevet tekst, idet vi forsøger at identificere et „ulæselig" bogstav ud fra de omkringstående bogstaver samt sammenhængen, og vi forsøger desuden at finde bogstavet skrevet andre steder, hvor der ikke var problemer med læsningen.

OCR/ICR-programmerne opsamler de genkendte bogstaver (normalt i ASCII-/ANSI-format) i en tekstfil, hvor man så efterfølgende er nødt til at foretage en korrekturlæsning, idet fejllæsninger er svære at undgå, specielt hvis originalen ikke er helt perfekt.

OCR-behandlingen betyder samtidig en betydelig pladsbesparelse, idet den resulterende tekstfil kun fylder en brøkdel af, hvad teksten vil fylde som billede. En A4-side med tekst i ASCII-format fylder kun ca. 4 KB, og afhængig af den valgte billedkvalitet vil det ikke være helt urealistisk at tale om en besparelse i størrelsesorden 1:100.

Selv om der siden 1960-erne har været eksperimenteret med OCR-teknikken, er det først i de senere år, at det teknisk er blevet muligt at hamle op med den manuelle tekstindtastning. Dette gælder både m.h.t. indscanningshastigheder, fejlprocenter og ikke mindst omkostningerne. Det var først omkring 1990 at scannere og tilhørende programmer kom ned i priser, som gjorde det bredt tilgængeligt.

Fejlprocent

Selv om tekstgenkendelsen er blevet forbedret meget, så er det vanskeligt at undgå fejllæsninger. Efterhånden kan opnås fejlprocenter på under 1 %, når forlægget ellers er af en ordentlig kvalitet, og dette kan vist sagtens konkurrere med de fleste manuelle indtastninger. Man slipper ikke for en efterbehandling med korrekturlæsning, og her kan der vise sig både globale fejl (f.eks. „%" i stedet for „Æ") og mere eller mindre uforklarlige enkeltfejl. Relativt ofte er det de tre danske vokaler æ, ø og å, som giver problemer.

Læsehastighed

Ved manuel indskrivning af A4-sider (3000 karakterer pr. side) skal man nok regne med, at der maksimalt kan indtastes 4 sider pr. time. I takt med de hurtigere computere er indscanningshastigheden også vokset, således at det efterhånden er muligt at indscanne ca. 30 sådanne A4 sider i timen, og det er vel at mærke med ovennævnte lille fejlprocent.

5. HARDWARE

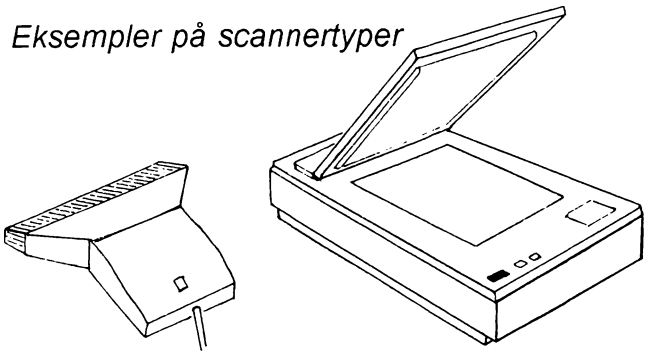
Scannere til billed- og tekst-scanning findes på markedet i forskellige udformninger, men normalt opdeles de i to hovedtyper. Nemlig håndscannere og bordscannere, hvor sidstnævnte fås i lidt forskellige udgaver.

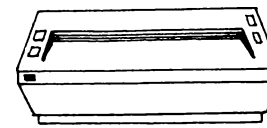
Håndscannere

Mest udbredt blandt private brugere er håndscanner-typen, og den er prismæssigt også den billigste, idet den kan fås til priser fra omkring 1000 kroner og opefter incl. en eller anden form for software til billedscanning. De billigste er de simple sort/hvid-scannere, mens det for de lidt dyreres vedkommende er tale om farvescannere med software; de koster fra ca. 2500 kroner.

Håndscannere, der minder lidt om en lille malerrulle, føres ved håndkraft langsomt henover det, som ønskes indscannes. Håndscannere har imidlertid det problem, at de er begrænset i scannebredde til 6-12 cm. Har man et dokument/billede, som er breddere end dette, må der indscannes over flere omgange. Det er selvsagt svært at føre en håndscanner med en tilpas jævn hastighed uden rystelser, og det kan nogle gange være svært at opnå et pænt sammenstykket indscannet billede.

Eksempler på scannertyper





Sheetfeed-scanner (ovenfor)

Flatbed-scanner (i midten)

Håndscanner (til venstre)

Flatbed-scannere

Til mere professionelt brug benyttes ofte de såkaldte flatbed-scannere, som i funktion kan sammenlignes med en kopimaskine, hvor det, som ønskes scannet, lægges på en glasplade og dækkes med et låg. Scanneren foretager selv den jævne aflæsning af dokumentet. Denne type scanner er fysisk mere skånsom overfor det indscannede og klarer A4-format uden at skulle sammenstykke resultatet som ved håndscannerne. Prismæssigt kan flatbedscannere fås fra omkring 6-8000 kroner. Akkurat som vi kender det fra kopimaskiner, fås der flatbedscannere, som har ADF (automatisk dokument føder), d.v.s. at man ikke skal stå ved scanneren og lægge dokumenter ind enkeltvis, hvis der er tale om en scanning af en hel stak.

Andre scannertyper

Håndscannere og flatbedscannere er de mest almindelige i Danmark, men der findes også andre typer, som f.eks. sheetfeed-scannere, tromlescannere og overhead-scannere. I sheetfeed-scannere føres dokumentet ind via en sprække, og scanneren trækker det selv igennem scanneafsnittet med en jævn hastighed; her er det ikke muligt at scanne direkte fra bøger og andet, der er tykkere end papir. Overhead-scanneren aflæser dokumentet på afstand, og i nogle udgaver er de i stand til at scanne tredimensionelt.

6. SOFTWARE

Foruden scanneren er det selvfølgelig nødvendigt med noget tilhørende software. Almindeligvis følger mere eller mindre avancerede billedbehandlings-programmer med, når man køber en scanner. Enkelte scannere leveres også med et OCR-program. Vil man senere anskaffe sig et nyt billedbehandlingsprogram eller et OCR-program til tekstgenkendelse, så skal man være opmærksom på, at alt software og hardware ikke kan arbejde sammen. En række firmaer har forsøgt at afhjælpe noget af problemet ved at lave et såkaldt TWAIN-modul, som skulle standardisere brugerfladerne og gøre det muligt at bruge scannerne i forskellige operativsystemer, men det er endnu kun et skridt på vejen.

Blandt de mest udbredte programmer til billedbehandling kan nævnes bl.a. Aldus PhotoStyler, Adobe Photoshop og Micrografx Picture Publisher.

Når det drejer sig om tekstgenkendelsesprogrammer er det først og fremmest Omnipage, Recognita man støder på, men der findes også Perceive samt det netop lancerede TextBridge. Programmerne fås til priser fra ca. 1500 kroner og opefter; Omnipage og Recognita koster eksempelvis hver omkring 8000 kroner.

Mulighederne med de enkelte programmer til billedbehandling og tekstgenkendelse er svære at læse ud af en brochure eller ud fra en programomtale, hvorfor det må stærkt anbefales også at se programmerne i funktion. En demonstration eller om muligt en afprøvning, som ikke er styret af sælgeren, giver en meget bedre fornemmelse af et programs muligheder.

7. GENEALOGISKE PERSPEKTIVER

Den fremtidige udvikling er det naturligvis svært at gisne om, men vi må nok regne med, at slægtsforskningsprogrammerne vil udvide mulighederne for personregistreringer i takt med PC-ens udvikling. PC-en ser ud til at blive et medie, hvor håndtering af levende billeder og lyd vil blive en naturlig side af dens muligheder. Det vil derfor ikke være urealistisk at gætte på, at vores fremtidige programmer til slægtsforskning udover den almindelige tekstlige beskrivelse af vore slægtninge vil kunne rumme f.eks. en scannet kopi af oldefars dagbog, en lydoptagelse, hvor oldemor fortæller om sin barndom og en videooptagelse fra deres guldbryllup.

Håndteringen og bearbejdningen af alle vore slægtshistoriske oplysninger vil kunne foregå på PC-en, modsat i dag, hvor vi må have en eller anden form for manuelt parallelkartotek med de oplysninger, som ikke er tekstorienterede (lydoptagelser, fotografier, centrale dokumenter). Dette ekstrakartotek vil så blive reduceret til et arkiv, der blot indeholder de originale dokumenter og ting, som er blevet os overdraget.

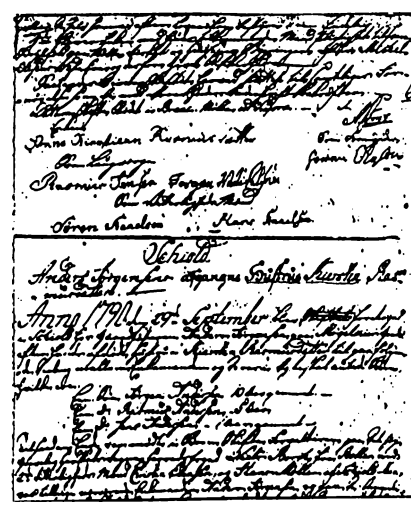
Scannerens muligheder

Scanneren ser ud til at blive et centralt redskab, hvis vore fotografier og dokumenter skal overføres til EDB som billeder. Med tilknyttede tekstgenkendelsesprogrammer vil scannerne også kunne indlæse maskinskrevet tekst. Tidligere trykte tekster uden komplette registre (f.eks. slægtsbøger og kildeafskrifter) vil kunne overføres, uden at man er nødsaget til at skulle indtaste dem, og i EDB-udgaven vil det være let at foretage søgninger f.eks. efter slægtsnavne. Drømmen om at programmerne også skulle kunne genkende håndskrevet tekst, lurer i baggrunden, og tænk om det engang blev muligt at tekstgenkende håndskrevne kilder. Dette ville i rumme nogle spændende perspektiver for os slægtsforskere.

Scanning af håndskrevet tekst

Genkendelse af håndskrift eksperimenteres der en del med, men det er jo helt klart mere vanskeligt at genkende bogstaverne i individuelle håndskrifter, hvor bogstavstørrelser, linier, bogstavudformninger og sidestrukturering ikke ligger i faste rammer, og hvor overstregninger, tilføjelser og sløsede ordafslutninger er hyppige. Et andet central problem er at bogstaverne er sammenhængende, d.v.s. at det maskinelt er svært at afgrænse, hvornår et eventuelt bogstav begynder og slutter i en tekst.

Ex. på kompliceret håndskrevet tekst



En af de veje, som man har forsøgt sig med, når det drejer sig om håndskrevne tekster, er at focusere på den måde skriveredskabet (pennen) er ført hen over papiret med retning og varierende hastighed taget i betragtning. Efter en optræningsfase vil computeren indlære bogstavernes karakteriske træk med centrale stregers begyndelsespunkter.

Når vi som mennesker foretager en tydning af en gotisk tekst, har vi en fleksibilitet og en logik/erfaring at trække på, som er meget svær (umulig ?) at omsætte i regler, som en computer ville kunne håndtere og arbejde efter.

Hvor kvalificerede læsere af gotisk skrift må give op m.h.t. en tydning må en computergenkendelse også anses for at være urealistisk. Da vi mennesker jævnligt tyder en håndskreven tekst forkert, vil det heller ikke være realistisk at tro, at den samme tekst OCR-behandlet ville blive fejlfri. Til hjælp for OCR-programmet ville det være naturligt at tilknytte en ordbog, som kunne støtte bogstavgenkendelsen, men der ville ikke blive tale om en helt almindelig ordbog, da man tidligere ikke havde nogen fastlagt retskrivning.

Hvor langt det rent teknisk er muligt at nå m.h.t. OCR-genkendelse af håndskreven tekst, skal man nok være lidt varsom med at udtale sig om, og vi skal ikke glemme, at landvindingerne i høj grad er styret af, hvor der er størst „behov" og økonomi i videreudvikling.

Der ser ud til at være mere fremtid og penge i at arbejde med stemmegenkendelse, og det er ikke usandsynligt, at vi i stedet for at tekstgenkende håndskreven tekst med scannere lettere vil kunne få teksten ind på PC-en ved højtlæsning for den.

8. SAMMENFATNING

Går man med planer om at anskaffe sig en scanner, bør man nøje overveje, hvad man nu og måske på lidt længere vil kunne tænke sig at bruge scanneren til. Kan man vente lidt, så er det ikke utænkeligt, at de vil blive endnu billigere eller give flere muligheder til samme pris.

Skal der arbejdes med billedscanning, så skal man bl.a. være opmærksom på scannerens muligheder m.h.t. farver (lige fra to-tonescanninger og til fuldfarvescanninger med 16,7 mill. farver), hvilken min./max. opløselighed (DPI) kan scanneren arbejde med, hastighed, RAM-krav og hvilke billedstørrelser kan den arbejde med.

Det er ikke alle scannere og programmer, der kan „tale" sammen, og nogle programmer er meget restriktive m.h.t. mulige filformater, hvilket kan gøre den videre brug af resultatet af en indscanning ekstra besværlig.

OCR-programmerne er ikke så udbredte blandt private brugere, og da programmerne ofte er af amerikansk oprindelse, er der ikke sjældent problemer med netop æ, ø og å, ligesom en eventuel tilknyttet stavekontrol måske ikke bygger på det danske sprog.

Scanner-udstyret stiller generelt store krav til computerens hurtighed og lagerplads, og her er det ikke sikkert, at ens nuværende PC er stor nok, eller at den simpelthen er for langsom. Og så kan indførslen af scanning blive dyrere end umiddelbart forudset.

Håndscanning under WINDOWS

af Kjeld Nymand

Det var i lang tid et stort ønske hos mig at kunne håndtere billeder, tegninger, kortudsnit, gengivelser af gamle håndskrevne dokumenter og andet med relation til min familiehistorie, og at kunne sætte dem ind præcist dèr, hvor de hører hjemme med henblik på gengivelse og præsentation for andre interesserede i slægten. For er det ikke netop dèt at kunne formidle det slægtshistoriske stof og gøre det tilgængeligt og forståeligt for en lidt større kreds, som bliver desto vigtigere, jo længere tid man arbejder med det? - Og så gør det vel egentlig heller ikke noget, at beretningen måske antager en spændende, fortællende eller ligefrem underholdende form i kontrast til de trivielle, skematiske opremsninger af forlængst henfarne aner i lange baner?

En metode, som mange af os har brugt i årevis, kræver fotokopimaskine, saks og masser af klister, og ulemperne er bl.a., at fotokopier taber ret meget for hver generation, som kopieres. Det er besværligt at ændre størrelse af det kopierede og at placere det tilstrækkelig præcist, samt på et senere tidspunkt at redigere eller indføje nye elementer. Nej, det skal ind i kassen, så man frit kan manipulere med tekst og billeder og kort og skifter og kirkebogsudskrifter og tipoldemor Alvildas berømte brødopskrift i hendes egen gotiske håndskrift! Scanning fremstod mere og mere som en spændende mulighed for mig.

Hvad er en scanner?

Uden at gå ret meget ind på det tekniske - her vil jeg henvise til de to grundige artikler af Erik Helmer Jensen, som stod i SLÆGT & DATA nr. 1 og 2, 1993 - så kan en scanner betegnes som et elektronisk kamera, som affotograferer et billede - det være sig en kortskitse, et fotografi eller et stykke tekst på en måde, som ligner den måde, fotokopimaskinen arbejder på. Scanneren oplyser emnet, idet den passerer hen over dette. Nogle fotoceller aflæser samtidig de enkelte punkter i billedet og overfører gradvist hele billedet til PC'en.

Herefter kan man arbejde videre med billedet i forskellige programmer. Man kan trace det, hvorved programmet finder omridsene i et billede og omdanner dem til vektorer - matematiske formler - som herefter kan forstørres uden kvalitetstab, og billedet fylder også mindre på harddisken, når det er tracet. Man kan også sende billedet igennem et OCR program - OCR står for Optical Character Recognition, eller optisk tegngenkendelse. OCR-programmet læser teksten i det indscannede billede og omdanner den til en elektronisk tekstfil, som herefter frit kan formateres og indsættes i de programmer, man har installeret.

Scan, scan ikke...

Svaret på mit ønske lignede mere og mere en håndscanner i en god kvalitet og til en lav pris, og det er i løbet af det sidste års tid blevet muligt. Jeg har med iver læst alt, hvad jeg kunne få fat på om scanning her og hist, bl.a. var der for ca. et år siden en artikel i det engelske blad „Computing in Genealogy", som konkluderede, at HÅNDSCANNEREN ikke kunne bruges til noget seriøst.

Det vil jeg i det følgende påtage mig at tilbagevise, for der er på det allerseneste sket en væsentlig udvikling både på software-og hardwaresiden. Udslaggivende for min beslutning om at købe en håndscanner var et prisfald til under 2000 kr. inklusive de nødvendige programmer. Til den pris syntes jeg, det var værd at gå i gang, vel vidende at det ikke var nok at købe en scanner, men at jeg også måtte udvide RAM fra 4 til 8Mb for at kunne vende tegningerne i luften. Endvidere måtte jeg anskaffe

en ny, stor og hurtig harddisk - den gamle var ret så fuld på trods af Doublespace - og under et besøg på en stor EDB-udstilling, hvor jeg fik rig lejlighed til selv at scanne forskellige medbragte emner og lege Spørge-Jørgen, forsynede jeg mig med en ordentlig stak brochurer og prislister om håndscannere.

Ikke noget for fummelfingrede...

Mit valg faldt på en PRiMAX GreyMobile motordrevet håndscanner, som scanner i sort/hvid og rastermønster, eller i 16-256 gråtoner og med en opløsning som kan justeres fra 50 til 400 DPI. En farvescanner forekom mig for dyr en løsning, da hovedparten af mine gamle billeder er i sort/hvid og farvebilleder optager meget plads på harddisken. Det er en ny type håndscanner, som tager noget af det vanskeligste ved processen væk: Selve dét, med fri hånd at styre retning og hastighed, som variérer fra 1 til 10 cm i sekundet under scanningen kræver stor præcision og påpasselighed. Men i denne pakkeløsning startes scanneren med et klik med musen og den optimale hastighed kontrolleres under scanningen af programmet. Sammen med scanneren leveredes en bakke med skridsikre puder, hvorpå man kan placere emnet, der kan have en størrelse på indtil A4. I hver side af bakken sidder en skinne, der passer op i en styrerille i scanneren, som sikrer at retningen er præcis.

Scanneren kører altså af sig selv på en stor gummidrivrulle og to små støtteruller. Ved scanning af en A4 side, må man flytte scanneren over i næste spor manuelt, da den kun har en scannebredde på 105 mm, men sammensætningen af de to scanninger foregår helt automatisk, usynligt og ret hurtigt. Dersom emnet er større end en A4 side, kan man sætte siderne sammen med kommandoen Stitch efter at man har udvalgt et par fikspunkter i tegningerne (det er her at det begynder at kræve muskler i form af RAM).

Installering

Installeringen af scanneren forløb uden problemer. Efter jeg havde åbnet motorhjelmen, blev det medfølgende kort stukket ned i en passende 16 bit revne - der var lige en enkelt tilbage imellem CD-ROM interface, lydkort, internt modem, Windows acceleratorkort og hvad der ellers sådan hen ad vejen er forekommet mig nødvendigt i en veltrimmet slægtsgranskercomputer. Den medfølgende vejledning til valg eller ændring af IRQ og DMA-kanal er det smarteste, jeg har set til løsning af et problem, der ellers nok kan få flertallet til at tilkalde højere magter (søn eller svigersøn?).

Hvem er (MARC) TWAIN?

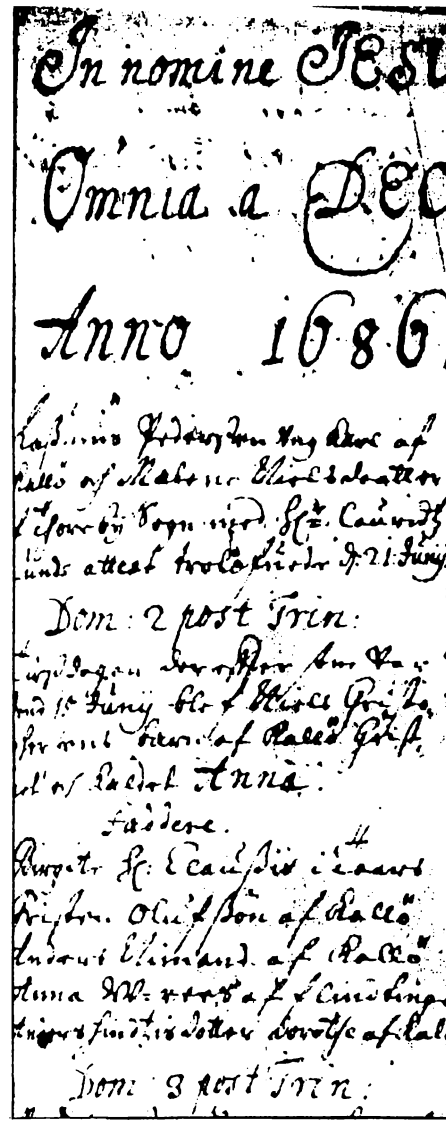
De medfølgende programmer var nemme at installere - og overraskende gode - alle naturligvis tilpasset Windows brugerfladen. Det vigtigste af programmerne er TWAIN, som ikke er en forkortelse, men et engelsk ord, der betyder at tvinde eller spinde sammen. Vittige hoveder har dog ment at ordet står for: Tool Without An Interesting Name! TWAIN-driveren er et mellemled, der gør det muligt at scanne direkte indefra alle TWAIN-støttede programmer under WINDOWS. Man bør ubetinget stille det krav til en moderne scanner, at den understøtter TWAIN-standarden. Jeg blev glædeligt overrasket, da jeg opdagede, at næsten alle mine programmer ved installationen fik tilføjet en ekstra lille box under menuen FILER med SCAN TEXT, herunder også det lille nye danske slægtsforskerprogram WINFAM. Dette er godt nyt i forhold til tidligere, hvor programmer enten skulle understøtte den aktuelle scanner direkte, eller man var nødt til at scanne i et særligt scannerprogram, gemme billederne og derefter importere billedfilerne i sine programmer. Det gør det nemt at indscanne tekst i noterne til de enkelte personer i databasen efter nogen øvelse, og især efter at man har lært programmet en del udtryk og tegn, som det ikke på forhånd kendte. Man kan altså

lægge billeder eller tekst ind i sine dokumenter uden overhovedet at forlade det program, man arbejder i! TWAIN-standarden er kun et par år gammel, men alle grafik- og DtP-programmer af betydning understøtter i dag TWAIN-drivere, og det gør jo blot livet nemmere for os brugere.

Scanning af tekst

OCR-programmet, som medfølger hedder INTERPRETER og leveres med ordbøger til de nordiske sprog, foruden hovedsprogene. Æ-Ø-Å har ikke voldt problemer i dette OCR-program, og det fungerer særdeles nemt og med få fejl. Man kan automatisere processen efterhånden, som man får foretaget de mest hensigtsmæssige valg og fintrimmet sin opsætning. Herefter begrænses brugerens arbejde til nogle få klik med musen, så er scanneren klar til en ny side. Det er med sikkerhed betydeligt hurtigere end selv en hurtig indtaster kan gøre det, men kvaliteten af originalen er afgørende for resultatet. Jeg bruger selv MICROSOFT ACCESS som mit vigtigste databaseprogram, og her kan man også scanne direkte ind i filerne - enten enkelte felter, eller flere felter samtidig - og der kommer ikke uorden i det - der er ikke noget, der skrider. Takket være funktionen „sideanalyse" går det hurtigt og præcist. Det er på denne måde blevet overkommeligt at få maskinskrevne folketællinger og andre kilder over i databaser med de fordele, det indebærer for søgning og rapportudskrift.

Det er også muligt i OCR-programmet at arbejde med billedfiler, som er indscannet i andre programmer, f.ex. breve modtaget som telefax i WINFAX. Hvis man har PC og modem er det jo en lynhurtig, behagelig og portobesparende ting at ordne sin korrespondance pr. telefax og efterhånden er der rigtig mange af os, der bruger det.



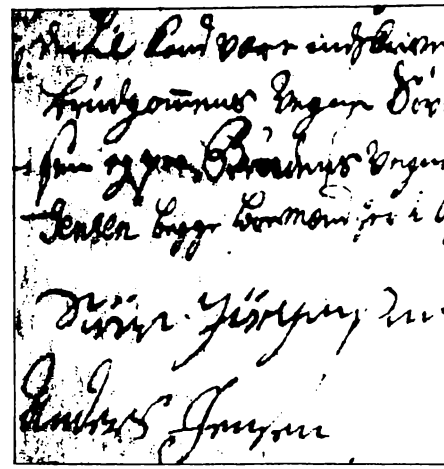
Eksempel på indscanning fra kirkebog. Indscannet med stor opløselighed, men det fremgår ikke klart af gengivelsen her.

Billedbehandling

Det medfølgende billedredigeringsprogram FINISHING TOUCH, har de vigtigste af de faciliteter, som store og dyre programmer fra f.ex. COREL og MICROGRAFX har, og kan håndtere de fleste filformater. Det vil sige overvældende mange muligheder for at videre-behandle indscannede fotografier. Man kan beskære og forstørre så meget, man lyster - og det kan nok være nødvendigt, da næsten alle gamle billeder er optaget med fast optik, eller slet ingen. De næsten skriger efter at blive beskåret og behandlet i „mørkekammeret". Man kan få detaljer til at træde frem i billederne, som man ikke lagde mærke til på de solafblegede, bruntonede dagslyskopier. De virkemidler, som står til rådighed i det moderne elektroniske „mørkekammer" -billedredigeringsprogrammet - overgår faktisk på mange måder dem, man har i det optisk/kemiske mørkekammer.

Scanning fra lysbord

En slægtsforsker fra USA, Alan Robb, videregav for kort tid siden i GENSOFT-konferencen under FidoNet (Volapyk for alle, som ikke har modem - men alligevel...) et tip om, hvorledes han ved at bagbelyse et s/h negativ kunne scanne det, invertere det, dvs. ændre det fra negativ til positiv i redigeringsprogrammet, og på denne billige og nemme måde få et brugbart aftryk. Det måtte jeg naturligvis prøve, og jeg lavede et interemistisk lysbord, som dem man sorterer dias på. Efter få forsøg med forskellige indstillinger, lykkedes det fint, og metoden gives hermed videre, da den næppe står i nogen manual. Der er senere et par slægtsforskere i samme konference, der har foreslået at anvende alufolie eller et topbelagt spejl af samme type, som sidder i de fleste microfiche læseapparater, til at scanne negativer på. Det virker, men ikke særlig godt - lysbordet er der bedste. Her er virkelig en god og billig måde at indscanne de kassevis af negativer, som vi vist alle har liggende i skuffer og skabe, og som vi måske af økonomiske årsager har udskudt kopieringen af „til bedre tider".



Et eksempel på en indscanning fra en mikroficheskærm (256 gråtoner med 400 DPI). Det indscannede er inverteret og siden printet ud. Gengivelsen her kan kun i ringe grad illustrere scanne-kvaliteten.

Scanning fra microfichereader

Det er omsider - efter mange forsøg - lykkedes mig at scanne direkte fra matskærmen på mit microfiche/microfilm læseapparat med et rimelig godt resultat, og jeg giver gerne tippet videre -dog uden nogen form for garanti for, at det vil lykkes med andre opsætninger. Jeg indstiller læseapparatet på højeste lysstyrke, scanneren på 256 gråtoner/400 DPI/laveste gammakorrektion. Derefter viderebehandles billedet i Finishing Touch, til det træder tydeligt frem, hvorefter jeg tracer det. Det kan være nødvendigt at „rense" baggrunden, men man har altså her en metode til at kopiere sine aners sig naturer fra de gamle kirkebøger, når de fx har optrådt som forløftningsmænd, eller fra skifteprotokoller, som de har underskrevet. Det vil jo være meget afhængigt af de pågældende

arkivalier, men her på Guldborgland, hvor de fleste af mine aner stammer fra, har man været gode til at samle autografer - og mange folk har kunnet skrive deres navne helt tilbage i 1600-årene, selv om de var bønder. Håndskriften var jo - før „formskriften" holdt sit indtog i Danmark - en meget personlig og individuel ting - en art miniaturekropssprog, som også for en ikke grafologikyndig kan afsløre ting om et menneskes karakter og fysik, ganske som et fotografi kan det. Jeg „støvsuger" ivrigt alt, hvad jeg kan finde med min håndscanner!

Hvad fylder det?

Alt afhænger af scanningsmetode, opløsning, filformat osv, men et fotografi på ca 10 X 15 cm, indscannet i 256 gråtoner/300 DPI fylder typisk 1 - 3 Mb som ukomprimeret TIFF-fil på harddisken, så det bliver hurtigt småt med pladsen. Ved at bruge GIF eller JPEG-formatet bliver størrelsen noget mindre.

Tapestreamer kan måske være en billig løsning på lagermangel, men „diskettegymnastik" må vist siges at være både for dyrt og for langsomt. CD-ROM-diske, der kan rumme op til 660 - 680 Mb, kan endnu kun laves på meget dyre, professionelle anlæg. De fleste må derfor nok beregne en ekstra udgift til en stor og hurtig harddisk, samt ekstra RAM oveni prisen på scanneren.

Hvad koster det?

Her er udviklingen igen gunstig i disse tider, for prisen på harddiske går i èn retning: NED, NED, NED! Man kan nu købe en hurtig kvalitetsharddisk i størrelsen 400 - 900 Mb for under 5 kr. pr. Mb, og det kan såmænd godt være, at denne oplysning allerede er gammeldags, men man bør se sig godt for - priserne svinger enormt for samme vare. Et tapestreamer backup-drev koster under 1.000 kr. (kan kobles til det ekstra Floppy-kabel), og backup-tape koster 100 kr. pr. stk og kan rumme 250 Mb komprimerede data. Det er 40 øre/Mb - så er vi dernede, hvor det er prismæssigt attraktivt.

Den motor-håndscanner, som er nævnt her, LECTOR GREY MOBILE, inclusive alle tre programmer, scannebakke og manualer har kostet 1.540 kr. Den er købt hos et specialfirma i Tyskland, da det er/var det billigste sted at købe scannere, og det er blevet meget nemt at handle privat pr. postordre med lande i EU.

Priser og afgifter er lavere i udlandet, og med fx VISA/DAN-kort er det hurtigt og nemt at overføre penge og bestille varer pr. FAX, og det koster intet vekselgebyr.

Kjeld Nymand
Rosengaarden 48
4990 Sakskøbing

Indscannet udsnit af emigrantfoto fra Minnesota ca. 1880

OCR-læsning af folketællinger

af Erik Helmer Nielsen

I begyndelsen af 1993 blev jeg opmærksom på, at mange af vore lokalhistoriske arkiver (LHA) har afskrevet en masse folketællinger på skrivemaskine. Da jeg er velkendt med anvendelsen af scannere og OCR-programmer, fik jeg den tanke, at det måtte være muligt at læse disse afskrifter med et OCR-program og derefter lægge dem ind i en database i SAKI eller KIP format. Derved kunne de opbevares af Dansk Data Arkiv (DDA) og blive tilgængelige for en videre kreds.

Det var dog først i efteråret 1993, at jeg ved DIS's mellemkomst fik kontakt med et LHA som var interesseret i et samarbejde med henblik på at afprøve ideen. Det var Søllerød Byhistoriske Arkiv, der tidligere havde afskrevet 4 folketællinger. Jeg blev hurtigt enig med arkivaren om, at jeg skulle afprøve OCR-teknikken på deres afskrifter - mod at disketterne bagefter blev overdraget til DDA.

Folketællingerne omfattede et ret stort sogn på ca. 2000 personer. De fyldte ca. 60 maskinskrevne sider pr. folketælling. Hver årgang var pænt indbundet, og selv om de var i fotokopi, præsenterede de sig pænt; de så ud til at være let læselige.

Scanneren

De krav, der stilles til en OCR-scanner, er ret beskedne. Den skal blot kunne scanne i sort/hvidt med en opløsning på 300 dot pr. inch. Der findes tre hovedtyper af scannere, nemlig håndscannere, 'Sheet-fed'-scannere og 'Flat-bed'-scannere.

Da afskrifterne i dette tilfælde var i A4 format, og scanneren af praktiske grunde skal kunne læse en hel A4 side ad gangen, var brugen af håndscanner udelukket. Afskrifterne var indbundet, og det gjorde det også umuligt at bruge de såkaldte 'Sheet-feed'-scannere. Valget måtte derfor falde på en 'Flat-bed'-scanner til A4 format.

Der er et stort udvalg på markedet af scannere af denne type, og de vil alle kunne tilfredstille de beskedne krav. Det er dog vigtigt at sikre sig, at det OCR-program man bruger eller vælger kan styre scanneren. Da jeg tilfældigvis havde adgang til en HP ScanJet, blev den brugt i forsøget.

OCR-programmet

OCR-programmet skal bruges på en almindelig PC. Der findes forskellige billige OCR-programmer i prisklassen omkr. 500 kr., og der er dyrere professionelle programmer, hvis pris nærmer sig de 10.000 kr. Jeg har prøvet nogle af de billige programmer som f.eks shareware-programmet OCRSHARE. De er efter min mening ikke gode nok til det krævende arbejde, som der her er tale om.

Der findes to professionelle programmer, som er almindelige på det danske marked. Det er programmerne 'OmniPage Pro 2.1' og 'Recognita Plus 2.0'. Det er Windows-programmer med brugervenlige skærmbilleder og en god ydeevne. De kræver begge en PC med kraftig processor og rigeligt hukommelse. Jeg brugte en 486DX2-50 med 8 Mbyte RAM.

Problemet med OCR-læsning af folketællingerne er selvfølgelig i første omgang at kunne læse bogstaverne korrekt. Det viser sig, at begge programmer klarer dette fint med ingen eller kun ganske få læsefejl pr. side. OmniPage Pro har dog visse vanskeligheder med de særlige danske bogstaver æ, ø og å.

Det største problem ligger i, at folketællingerne er opdelt i felter eller kolonner. OCR-program-

met skal kunne analysere og gengive disse felter på en sådan måde, at de senere kan læses ind i de tilsvarende felter i en database. Det viste sig, at OmniPage Pro slet ikke kunne klare opgaven at adskille kolonnerne i det materiale der forelå; resultatet var nærmest kaotisk. Recognita Plus var langt det bedste program i denne henseende, selv om heller ikke dette program kunne løse problemet helt perfekt. Forsøget blev herefter gennemført med Recognita Plus ver. 2.0.

Problemer med kolonnerne

Folketællingerne var afskrevet på A4 sider i alm. tværformat. Den ringe bredde af siden betød, at hver kolonne var ret smal, og at afstanden mellem kolonnerne var meget lille; mange gange var kolonnerne endda skrevet ind i hinanden. Det medførte, at der var to problemer i forbindelse med kolonnerne.

Det første problem var OCR-programmets analyse og gengivelse af kolonnerne. Det skete, at kolonnerne næsten rørte hinanden eller endda løb ind i hinanden. Så kunne OCR-programmet ikke skelne den ene kolonne fra den anden, og to nabokolonner blev slået sammen til een. Så blev det nødvendigt bagefter at skille kolonnerne ad manuelt. Et ret tidskrævende arbejde.

Det andet problem skyldtes afskriverens problem med at få lange navne og betegnelser til at være inden for de smalle kolonner. Det havde afskriveren løst ved at dele lange navne og betegnelser over to linier. Men det kan jo bare ikke fungere i en database, hvor alle oplysninger om een person skal stå i een record, d.v.s. på een linie. Resultatet var igen en omstændelig manuel tilretning.

Et helt tredie problem stammer fra den oprindelige kilde, hvor der ofte var anvendt gentagelsestegn af forskellig art - eller klammer til angive en flok børn eller lignende. Disse tegn var omhyggeligt gengivet i afskriften. Ifølge SAKI skal sådanne former for gentagelsestegn erstattes med den egentlige betegnelse, for at oplysningerne om hver person kan blive så meningsfyldte som muligt. Heri ligger der også et stort redigeringsarbejde, som dog ikke direkte har noget med OCR-læsningen at gøre.

OCR-læsningen

Det tog ca. 20 sec. at scanne en side og derefter ca. 10 sec. at læse den. Det tager også lidt tid at vende bogen og skifte til en ny side, men alt i alt kan man nogenlunde mageligt læse 1 side pr. minut.

Jeg læste ca. 10 sider ad gangen, og lod derefter Recognita automatisk opsøge de bogstaver i teksten, som der havde været problemer med at læse rigtigt. Det er herved let og hurtigt at rette evt. fejl. Dette giver selvfølgelig ikke nogen garanti for, at alle læsefejl bliver fundet, men det viser sig i praksis, at det kun er ganske få, der bliver tilbage til korrekturlæseren.

Redigering i WordPerfect

Fra Recognita læses teksten over i en WordPerfect fil. Herved bliver alle adskillelser mellem kolonnerne gengivet ved tabuleringer i WordPerfect (WP). OCR-programmet laver en individuel tabulering for hver side, og den bliver derfor normalt forskellig fra side til side. Det er imidlertid nemt i WP at ændre alle tabuleringer i et dokument til at være ens, således at kolonnestrukturen er den samme i hele dokumentet.

Sideformatet i WP sættes op, så sidebredden bliver meget stor (ca. 20") for at få plads til alle oplysningerne på een linie med felter, som er tilstrækkeligt store til også at rumme lange navne m.v.. Jeg havde her stor hjælp af at have WP i den sidste version 6.0. Den kan nemlig arbejde med en grafisk skærm med frit valgt opløsning.

På grund af de ovenfor nævnte problemer med kolonnerne og med gentagelsestegnene følger der nu et meget omfattende og tidskrævende redigeringsarbejde i WP. Man kan få nogen hjælp ved at lave forskellige hjælpemakroer i WP, men den samlede redigeringstid bliver alligevel ganske lang. Der var stor forskel på de forskellige oplæg; nogle var rimelig lette at redigere, andre meget besværlige. De mest tidskrævende tog typisk 10 - 15 min. pr. side.

Paradox databasefil

Det endelige mål var at få kildeindtastningen placeret i en Paradox datafil af samme struktur som SAKI/KIP datafilerne.

For at få data'ene ind i Paradox lod jeg først WP lave en almindelig ASCII tekstfil. I denne fil er alle oplysningerne stillet pænt op i de kolonner, som vi så møjsommeligt har rettet til i WP. De enkelte kolonner bliver nu adskilt med mellemrum (space).

Paradox kan ikke selv læse denne type tekstfil, som undertiden kaldes en SDF-fil. Der findes imidlertid et hjælpeprogram til Paradox, som hedder FLIMPORT (FixedLenghtImport), som klarer opgaven let og hurtigt.

Vi har nu en databasestruktur i Paradox, som begynder at ligne SAKI/KIP strukturen. Der skal dog laves en del tilretninger, før opgaven er endeligt løst. Der skal således tilføjes felter for stednavne og recordnumre. Derudover skal der i henhold til SAKI være felter med oplysning om Sogn, Kommune, Amt, Kilde m.m. En del af dette tilretningsarbejde kan laves med nogle af Paradox's små PAL script, og det tager alt i alt ikke særlig lang tid - næppe mere end en time for hver folketælling.

Kildeindtastningsprogrammet KIP

Søllerød Byhistoriske Arkiv havde KIP programmet, men ikke selve Paradox. De ville derfor gerne kunne læse og redigere de OCR-scannede indtastninger i KIP på samme måde som en original KIP indtastning.

Jeg fik et eksemplar af KIP til låns for at prøve, om det kunne lade sig gøre, men det kunne det ikke umiddelbart, da KIP er beskyttet med Password, og det betyder, at alle datafilerne er kryptograferede. I forbindelse med små forskelle mellem SAKI-beskrivelsen og KIP betød det, at jeg ikke uden videre kunne få mine Paradoxfiler ind i KIP systemet, så de kunne læses der. Efter mange forgæves forsøg måtte jeg klage min nød til KIP's forfatter Elsebeth Paikin, og vi fik så til sidst i fællesskab løst problemet.

Konklusion

Forsøget viser, at det er muligt med godt resultat at læse afskrifter af folketællinger med et OCR-program.

Problemerne ligger ikke så meget i selve OCR-læsningen, men mere i den efterfølgende redigering, som viste sig at være nødvendig for at få et tilfredstillende resultat.

Den tid, det er nødvendigt at bruge til redigering, afhænger meget af oplæggets udformning. I tilfælde af uheldige oplæg er den så stor, at det er tvivlsomt, om OCR-scanning kan konkurrere tidsmæssigt med en simpel afskrift i f.eks. KIP. Det må derfor anbefales at være kritisk overfor oplægget på dette punkt.

Erik Helmer Nielsen
Hedager 51
2670 Greve

Billedscanning i mit erhverv

af Jørgen Brandt

Tegninger og fotografier i sort/hvid eller farve er en særdeles spændende dimension, når slægtsforskningen bliver præsenteret i sin færdige form - ikke bare det grafiske udtryk, men også den stemning og information, der ligger i gengivelse at et originalt dokument eller af et fotografi med bedstemor på sin Ellehammer-motorcykel.

Hvis man vil have billeder ind på PC'en alene med det formål at at have billedet på skærmen, så er løsningen „kun" en scanner. Kvalitetskravet vender vi tilbage til. Skal billedet ud på papir er det straks en mere omfattende sag. I sidste ende er det en personlig afgørelse, om et printet fotografi er godt nok.

Jeg hedder Jørgen Brandt og arbejder professionelt med fremstilling af tryksager. Mit værktøj er en Macintosh Centris 610 (i øjeblikket) og en flatbed scanner, der hedder Silverscanner II. En flatbed scanner betyder bare, at den arbejder med originalen liggende på en plan glasplade - modsat højkvalitetsscannere, hvor originalen er spændt op på en tromle.

Virksomheden, jeg arbejder på, indkøbte den første Mac i 1989 sammen med en scanner og en 300 dpi laserprinter. Frem til i dag har jeg haft tre Mac'er og tre scannere, tre printere og en fotosætter. Fotosætteren er i princippet en laserprinter med en høj opløsning og anderledes dyre materialer.

Når jeg skriver dette er det for at fortælle, at vi hele tiden har købt det billigst mulige maskineri med det formål at kunne håndtere billeder i en brugbar kvalitet. Derfor har vi hele tiden måttet købe lidt bedre maskiner, og min know-how er derfor bygget op fra bunden.

Input

Lad mig slå fast, at der skelnes mellem tre typer scanninger, streg, gråtone eller farve (line, grayscale eller colour). Hvis du kun skal have dit billede op på skærmen, skal du sætte dig ind i, hvor stor din skærmopløsning er. Ikke hvor mange linier eller skærmfrekvensen. Opløsningen på min Mitsubitshi farve-skærm er 72 dpi (dots pr inch) dvs. 72 lysprikker pr tomme - og hver lysprik i hver sin farve. Derfor kan jeg nøjes med at scanne et foto ind i 72 dpi, når det blot skal stå på skærmen. Da skærmen ikke kan vise mere end 72 dpi, er det logisk nok. Det gælder både for streg, for gråtone og for farvescanning.

Output

Når man skal trykke et S/H foto (gråtone), bruger trykkeren kun een farve, nemlig sort. Et originalfoto består kun af gråtoner, og før trykkeren kan trykke det, skal der lægges raster i fotoet. Hvis du kigger på fotos i en avis - helt tæt på, vil du se, at billedet består af bittesmå prikker. Størrelsen af prikkerne afgør, om det er den lyse eller den mørke del af fotoet du kigger på. I fagsproget taler vi om højlys og skygge.

En printer er også nødt til at lægge raster i et foto. Der er kun sort pulver i printeren til at generere det materiale, du printer. Printeren har ydermere den begrænsning, at den kan lave striber, og med en opløsning på f.ex. 300 dpi er det kun et meget begrænset antal gråtoner, den kan printe. En 600 dpi printer har også et begrænset antal gråtoner, men dog mange flere end printeren på 300 dpi. Men stadig ikke så godt, at det kan kaldes en brugbar kopi af et foto til mit brug. Min fotosætter printer med en opløsning på 2540 dpi, og det giver et ny-

deligt resultat. Et print kørt ud i en opløsning på 1200 vil være tilstrækkeligt. Nu skal du holde ørerne stive.

Når du har kigget på fotoet i avisen og sammenligner det med et S/H foto i en fin brochure, kan du se, at der er forskel på finheden af fotoet. Det er papiret, der her sætter grænser for for, hvor fint fotoet kan gengives. Denne finhed i fotoet - som ikke har en døjt med hardware eller software at gøre - måles i antallet af prikker pr. cm. Et typisk avisfoto er delt op i ca. 34 prikker pr. cm.- både lodret og vandret. Prikkerne er store eller små, men afstanden fra centrum til centrum af hver prik er den samme. Hvis du kigger på en tryksag af høj kvalitet, vil det typisk være 60 prikker pr. cm. - prøv at tænke dig! 60 prikker fordelt på 1 cm. 3600 prikker pr. cm2, og hver prik har forskellig størrelse. Og ikke nok med det, de har også forskellig facon. - altsammen for at kunne gengive et pænt skarpt foto.

Fotos

Når alle detaljer om prikkerne skal med, er det nødvendigt at scanne fotos ind i en høj opløsning. Når alle informationer skal være i PC'en, fylder de en del. Et foto på 10 x 15 cm fylder således 903 K, og en hel A4 side som gråtonebillede fylder 3,7 MB. De nævnte tal angiver minimumskravene, hvis de skal printes ud i en tålelig kvalitet. Med en tålelig kvalitet mener jeg en kvalitet, der ligger mellem et avisfoto og en pæn trykt brochure. I fagsproget svarer det til en trykt opløsning på 40 liniers raster. Jeg har arbejdet med fotos, der fylder 12 MB - bare for en almindelig S/H. Men så blev det også en lækker kvalitet. Der findes komprimeringssystemer, der redu-cerer filens størrelse, men det kommer jeg ikke ind på her.

Opløsningen på et foto skal være 200 dpi, når du skal lave et foto i 40 liniers raster, og når den rigtige størrelse er defineret. Det betyder, at når fotoet ligger i scanneren, og opløsningen er sat til 200 dpi, skal du også indstille størrelsen af fotoet til den størrelse, du vil bruge som slutresultat. Hvis f.eks. du har et foto i størrelsen 9 x 12 cm, og du vil have det printet ud i størrelsen 10 x 15 cm, skal scanneren indstilles til ca 110 %. Det er den færdige størrelse, der skal have en opløsning på 200 dpi. Hvis du senere ønsker at redu-cerer størrelsen af fotoet i dit dokument, er der ingen problemer; du scalerer bare billedet ned - hvis ellers det program, der håndterer din slægtsforskning, både kan placere et foto og derefter ændre størrelsen. Hvis du derimod vil have billedet gjort større, vil fotoet blive ringere jo større, du vil have det.

Tegninger

Opløsningen på en tegning, en stregscanning, er lidt mere variabel. En tegning, der køres ud på en 300 dpi laserprinter, stiller ikke så store krav til indscanningen. En scannet opløsning på 300 dpi i den endelige størrelse vil i reglen være tilstrækkelig. Selv scanner jeg ind i 800 dpi, men hvis kvalitetskravet ikke er stort, klarer jeg det med 400 dpi. Det er min egen interesse, at filerne er så små som muligt. Derved sparer jeg tid, når jeg skal håndtere f.eks. et blad på 12 sider, der samlet fylder 50 MB.

Fil formater

Mit foretrukne format til gråtone- eller stregscanning er „TIFF". Det betyder „Taged Image File Format", og det er et format, der giver mig tilstrækkelige tekniske muligheder, når det er placeret i et dokument. Der findes en lang række andre muligheder, men dette er et filformat, der umiddelbart lader sig konvertere til DOS verdenen.

Under særlige omstændigheder lagrer jeg mine billeder i „EPS"-formatet, men det format er tidsmæssigt tungere at arbejde med, og det er vanskeligt at konverterer mellem Mac og DOS. Jeg behandler mine fotos i et billedbehandlingsprogram, der hedder Adobe Photoshop.

Det giver mig mulighed for at redigere i mine fotos - og også manipulere dem - inden de bliver placeret i et dokument. Et scannet foto skal altid gøres lysere eller mørkere, eller måske skal det drejes lidt. Jeg kan også gøre et foto elektronisk skarpt - alt sammen for at det færdige resultat bliver af optimal kvalitet.

Scannere i prisklassen 5.000 - 30.000 kr scanner i 8 bit. Det betyder, at der er 8 bit til rådighed pr. pixel i et foto. En pixel er det område på fotoet, der svarer til 1 dpi. En 8 bit scanner kan opfatte 256 forskellige gråtoner. Når scanneren „aftaster" et foto, bestemmer den sig til en af de 256 gråtoner for hver pixel. Det er en tilstrækkelig kvalitet til de fleste S/H fotos i normal kvalitet. Mere professionelle scan-nere arbejder 10 bit og kan derfor opfatte betydelig flere gråtoner. Det giver en bedre kvalitet under alle omstændigheder og allermest, når billedet er meget kontrastløst. Et kontrast-løst foto redigeres op til fuld kontrast i et billedbehandlingsprogram, og her har man brug for en masse gråtoner. Når jeg taler om dækning i foto (rasteromfang), går det fra 5 % i det lyseste sted til 90 % dækning på det mørkeste punkt i fotoet. Når jeg scanner i 256 gråtoner til at dække en skala fra 5 til 90 %, får jeg ca tre gråtoner for hver procent i fotoet - men kun i teorien. Inden fotoet er færdigbearbejdet i billedebehandlingen, er der gået et par bit tabt og inden det når printeren, har det måske nået at tabe mere?

Forvirret? Det ikke så mærkeligt. Disse linier dækker over en del af en årelang uddannelse. Jeg nået dertil, at jeg kan leve af det - men jeg har fået mange gråtoner i håret.

Udstyr og penge

Mit udstyr er til professionelt brug. Med scanner, Mac, printer og fotosætter løber prisen op i 250.000 kr - bare for at få et brugbart foto. Det udstyr er der næppe nogen, der vil købe for at få et foto ind i slægtsforskningen. Mit råd er derfor at foretage en affotografering - evt. hos en fotograf, eller at foretage en fotokopiering. De forretninger, der virkelig har udstyr til fotokopiering, har også specialmaskiner til at håndtere fotos i bedre kvalitet end en 600 dpi laserprinter. Du kan også vælge at købe en scanner og så få resultatet kørt ud på et satsbureau. Du skal først aftale med bureauet om alle de tekniske detaljer. Nogle satssystemer er meget lukkede. Derefter vil du kunne få kørt dit foto ud på fotosatspapir til en pris omkring 70 - 90 kr. pr. A4 side.

Jeg håber at have løftet lidt af sløret omkring scanning til PC'ere. Selv om jeg arbejder på Mac, er principperne stort set ens i billedbehandlingen. God fornøjelse med slægtsforskningen. Artiklen er blevet til med støtte af Torben Larche, der handler med hardware/software i Viborg.

Jørgen Brandt
Fredensborg Offsettrykkeri A/S.

Håndscanner til notebook

af Jørgen Rasmussen

Efterfølgende er en beskrivelse af det udstyr og de programmer, jeg har anvendt i forbindelse med scanner i godt og vel et års tid på en notebook. Det er således en artikel skrevet set med en brugers øjne.

Udstyr

Computer: Compac Contura 3/25 notebook m/ mus (LCD-skærm i gråtoner). Ekstraskærm: Northmann MV-4D (farveskærm).
Scanner: Håndscanner Matador 105 (105 mm scanningsbredde, 64 gråtoner, 100-400 dpi).
Scannerkort: Mustec Prinscan-II I/F Box med separat strømforsyning.
Programmer: Scankit Utility version 2.2 A (Windows). Scankit Gray Utility (DOS). Perceive Personal (OCR program for Windows). Paint Pro Shop (billedredigeringsprogram for Windows, shareware). Graphic Workshop (billedredigeringsprogram for windows, shareware). Winfam (dansk produceret slægtsforskningsprogram for windows).

Anskaffelse

Jeg har dyrket slægtsforskning ret intensivt i ca. 3 år. I hele perioden har jeg anvendt EDB til dataopbevaring og - behandling. På et tidspunkt satte jeg mig for at lave en A5-folder med billeder om familiens rødder baseret på min oldemors livsskildring og de af mig fundne oplysninger om familien. Teksten kunne jeg skrive på min computer. Men billederne blev fotokopieret til en størrelse, der passede til teksten. Billederne blev ikke rigtig tydelige, specielt ikke de ældste billeder fra forrige århundrede. Når kopimaskinen var god til billeder - var den ikke særlig god til tekst. Jeg havde på et tidspunkt set en scanner i et af datatidsskrifterne, altså gik jeg i gang med at finde ud af, hvad en scanner var, hvad den egentlig kunne bruges til og hvad den kostede. Efter studier af diverse datatidsskrifter, besøg i computerforretninger traf jeg endelig mit valg og investerede i en scanner. Valget var ikke særligt svært, da der dengang ikke fandtes andre modeller til brug for en notebook.

Når det blev ovennævnte scanner skyldes det, at det ikke var muligt at indbygge et scannerkort i en notebook. Men kortet sidder i stedet for i en boks, som sluttes til printerporten. Boksen forsynes med strøm fra en medfølgende separat strømforsyning. Alt dette gør det selvfølgeligt mere omstændeligt at anvende scanneren og lidt dyrere, idet den kostede ca. 2.400 kr. mod normalt ca. 1.600 kr. Til gengæld kan jeg medtage scanneren, hvor jeg vil. Det er også årsagen til, at jeg i sin tid anskaffede en notebook.

Jeg havde en del problemer med det medfølgende program, Scankit Utility, idet der var fejl i programmet. Jeg var på nippet til at returnere scanner og program til leverandøren i Viborg. Først da jeg endelig fik version 2.2 A fra leverandøren, kom programmet til at fungere. Siden har det fungeret upåklageligt. Desværre er billedredigeringsmulighederne begrænset i programmet - derfor har jeg for kort tid siden anskaffet de øvrige billedredigeringsprogrammer, men har endnu ikke nogen egentlig erfaring med dem.

Anvendelse

Indledningsvis anvendte jeg scanneren til at indscanne en masse familiebilleder til ovennævnte folder. Det er dog ikke umiddelbart så let at indscanne billeder. For det første skal man have en rolig hånd og føre scanneren i en jævn og konstant bevægelse hen over billedet, ellers kommer der striber eller der mangler dele af billedet. For det andet findes der jo billeder, der er større end scannerens scanningsbredde på 105 mm. Sidstnævnte fænomen medfører,

at man skal til at scanne et billede af flere gange og bagefter flette det sammen til eet billede. Det er spændende, men kræver øvelse for at få et godt resultat. Princippet ved fletning af to billedscanninger er egentlig enkel, idet de to dele af billedet enten scannes ind eller hentes ind fra lageret (harddisken) fra en tidligere scanning. Herefter føres de to billeder sammen på skærmen, til sammenfletningen passer. Endnu nemmere er det, hvis man kan finde et lille markant fælles punkt på begge billeder; så kan man få programmet til at gøre det automatisk. Ved scanningen skal man blot huske, at der skal være overlapning mellem billederne.

For at få et godt resultat bør man anvende en bog eller en lineal til at støtte og trække scanneren langs med. Ellers risikerer man, at billedet bliver trukket skævt. Det samme gælder i øvrigt også for tekst. Scanneren har på undersiden et „løbehjul" og et eller flere støttehjul. Man skal sikre sig, at „løbehjulet" bevæger sig jævnt og konstant henover papiret, idet det virker som en tæller til brug for opbygningen af billedet i computeren.

Jeg tog udstyret med på Landsarkivet for Fyn, hvor jeg fik tilladelse til at anvende scanneren. Jeg har bl.a. anvendt scanneren til at indscanne et skifte i en skifteprotokol, hvorefter jeg hjemme i ro og mag kunne sidde og oversætte det fra gotisk til læsbart dansk. Det er som regel ikke altid muligt at få alt med ved indscanningen fra en protokol eller en bog, fordi nogle protokoller er skrevet helt ind til midten eller pga. krumningen fra indbindingen, så scanneren ikke kan komme ind og få det hele med. Men det er dog væsentligt hurtigere kun at skulle oversætte en række ord langs midten af skifteprotokollen end at skulle oversætte et helt dokument. Jeg scanner normalt siden fra midten og ud mod kanten for at få det bedste resultat. Kun hvor teksten står i spalter, scanner jeg i tekstens længderetning. Ved kanten af siden er det nødvendigt at understøtte scannerens løbehjul for ikke at ødelægge teksten. Man finder også hurtigt ud af at indstille mørkhedsgraden på den indscannede tekst - f.eks. ved lys tekst skal der være en mørkere indstilling. Det har sågar været mig muligt at gøre teksten mere læsbar. I andre tilfælde har jeg oplevet den modsatte reaktion pga. snavs på papirets overflade. Herudover er det muligt at forstørre eller formindske den indscannede tekst.

OCR-programmet, som betyder Optical Character Reading, kan med fordel anvendes til indscanning af maskinskrevet tekst, hvor man ønsker at kunne redigere i teksten. Antallet af fejl i den indscannede tekst er omvendt proportional med indlæsningens omhyggelighed - jo større omhyggelighed, jo færre fejl. Det er ikke hensigtsmæssigt at anvende OCR til håndskrevet tekst. Ganske vist vil programmet kunne læse og omsætte nogle af ordene til redigerbar tekst, men resultatet står ikke mål med indsatsen. Det program, jeg har, kan i øvrigt stilles til forskellige sprog -herunder dansk.

Det er muligt vha. et billedredigeringsprogram at uddrage en del af et billede eller en tekst til indsættelse i et dokument eller f.eks. i et slægtsprogram, som det danskproducerede Winfam fra JamoDat i Ølstykke. Jeg har anvendt både Brothers Keeper og især Personal Ancestral File (PAF), men er gået helt over til Winfam. Programmet er vist nok det eneste slægtsprogram, der kan indlæse billeder til anvendelse i programmet og til udskrifter efter eget valg. Som eksempel herpå kan jeg nævne, at jeg havde fået et billede af mine oldeforældre. Til brug for oplysningerne i slægtsprogrammet ønskede jeg at få et billede af hver af mine oldeforældre. Ved hjælp af PSP-programmet hentede jeg billedet ind af mine oldeforældre, indrammede (klippede ud) den del af billedet, som jeg ønskede skulle til min oldemors oplysninger, og via klippebordet i windows blev

billedet lagt over i oplysningerne i slægtsprogrammet. Proceduren er i øvrigt tydelig beskrevet i en ny udgave af Winfam.

Jeg har anvendt scanneren til at scanne billeder ind til en sang med billeder til min moders 80 års fødselsdag til at illustrere hendes lange livsforløb. Tilsvarende har jeg lavet til en sølvbryllupssang. Herudover har jeg scannet billeder til andre til brug i deres computere.

Forventninger

Har scanneren så levet op til de forventninger jeg havde? Hertil må jeg generelt svare ja, når jeg samtidig indbefatter de øvrige programmer, som jeg har anført under udstyr samt slægtsprogrammet Winfam. Det er muligt, at jeg ikke har forstået at anvende Scankit Utility fuldt ud, men suppleret med billedredigeringsprogrammer får man en bedre udnyttelse af indscannede billeder, tekster, tegninger mm. Jeg må erkende, at jeg ikke har nået at få nogen større erfaring med billedredigerings-programmerne. Det er også muligt, at de fleste kan klare sig med de muligheder, der ligger i de medfølgende programmer ved køb af en scanner.

Det, der har overrasket mig mest, er den størrelse, billedfilen får ved indscanning. Fra ca. 100 Kbyte ved små billeder og op til ca. 1 Mbyte ved større billeder eller tekster. Det kan derfor være hensigtsmæssigt efter brugen at gemme filerne på disketter, således at de ikke optager for meget plads på harddisken. Det vil så være hensigtsmæssigt at lave et lille arkiv over indscannede billeder. Det kan f.eks, udmærket laves i Windows kartotekskort.

Jeg havde forventet at kunne indscanne næsten ubegrænset i en bevægelse, men p.g.a. en opdeling af harddisken i flere drev, var der kun godt 1 Mbyte til rådighed, hvor programmet var installeret.

Dette medførte, at jeg kun kunne indscanne korte stykker, før pladsen var brugt op. Brugere skal her være opmærksom på, at indscanningen bliver indlæst i en temporær fil, hvor pladsen på harddisken sætter begrænsning på filens størrelse og dermed også på hvor meget, der kan scannes ind ad gangen. Der skal helst være 2 - 3 Mbyte ledig plads på det område af harddisken, hvor filen gemmes under indscanningen. Så vidt jeg husker, spørger programmet ved installering, hvor man ønsker, at den temporære lagring skal ske.

Sammenfatning

Efter et års brug kan jeg kun sige, at jeg ikke har fortrudt, at jeg anskaffede en scanner. Jeg kunne ønske mig, at det havde været muligt at indbygge scannerkortet i min notebook, således at jeg kun skulle medbringe scanneren ud over notebooken. Jeg er glad for, at jeg i sin tid valgte en notebook i stedet for en bordmodel. Således kan jeg medbringe computer og scanner til de forskellige arkiver, hvor jeg skal hente mine oplysninger.

Jeg vil dog anbefale EDB-brugere grundigt at overveje, hvor meget de har behov for en scanner, før de evt. anskaffer en. Især hvis den kun skal tilsluttes bordmodellen derhjemme. Det er jo begrænset. hvor meget man kan tage med hjem af materiale som slægtsforsker.

Selvfølgelig kan man godt scanne billeder, selv om man kun bruger DOS styresystem, men anvendelsesmulighederne er langt større ved anvendelse i Windows. Jeg har ikke kendskab til OS-2, hvorfor jeg ikke kender mulighederne heri. Jeg håber hermed at have bidraget lidt til forståelsen vedrørende mulighederne for anvendelse af en håndscanner.

Jørgen Rasmussen
Erik Skeels Vej 24 A
5700 Svendborg

Jeg - en scanner

Af Bent Pilgaard

I sommeren 1993 fik jeg chancen for at udskifte min „gamle 8086" Commodore med en '386-er. Og skulle der være gilde, så la' der blive gilde, så jeg købte også en sort/hvid håndscanner.

Hvis jeg tidligere har haft mulighed for at kede mig, så var den tid i hvert fald forbi nu. I min kamp med programmet og i et forsøg på at udnytte scannerens muligheder forsvandt den ene time efter den anden. Lidet overraskende blev jeg overladt til mig selv med en diskette og en engelsk manual, og så måtte jeg selv finde ud af at komme igang. Jeg tror i øvrigt, at EDB-branchen er den eneste branche, der i den grad overlader brugerne til sig selv -uden i tilstrækkelig grad at være i stand til at yde hjælp, eller give mulighed for indlæring af brugeren.

Men videre til min kamp! Lad det være sagt med det samme, mine engelskkundskaber lader meget tilbage - min engelsklærer nikker sikkert bekræftende - og alene dette giver problemer. Scanneren, jeg købte, var af et mærke i den billige ende af skalaen, men den kunne dog scanne op til 400 DPI. Manualen viste sig egentlig at være skrevet ganske brugervenligt til forskel fra så mange andre. Valgmulighed angives på rullegardiner og omtales meget godt. Alligevel fik jeg mine problemer, hvilket betød, at resultatet ikke altid stod mål med ønskerne.

Installation af scannerprogram

I første omgang blev programmet installeret ved hjælp af det medfølgende setup, hvor der skulle angives videokode, printervalg og kontrol af fileantal i Config.sys. Som i så mange andre tilfælde, blev der knas med printeren. Min printer er ikke HP-konvertibel, så det endte med, at der blev valgt en 24-nåls matrixprinter. Herefter var det bare at taste „EHS" og vips fremkom et skærmbillede med den kendte menubjælke. Tilbage til manualen for at finde ud af selve scanneren.

Her var en pæn tegning, der viste de forskellige funktioner på scanneren. Det støttede de manglende sprogkundskaber. Efter gennemgangen og en prøve af de forskellige knapper, skulle der nu scannes. Op på menubjælken - et klik på SCAN, og valg af SCAN SETUP, hvor indstillingerne på scanneren blev kontrolleret. Derefter kvittering på SCAN, og så skulle den være der, men ak. Der skulle vælges og bestemmes mere endnu.

Klik på scan-rullegardinets NEW PICTURE gav mulighed for på forhånd at begrænse den flade, der skulle scannes. Størrelsen kunne angives både i tommer (engelske eller danske?) og centimeter. Jeg valgte det sidste. Bredden kunne blive op til 10,5 cm, medens en længere formel, som jeg opgav at løse, beregnede hvor meget hukommelse, der var nødvendig. Nu skulle det være, og sandelig om ikke en gengivelse af scanneren på skærmen bad mig om at begynde.

Min første indscanning

Scanneren blev langsom ført ned over „billedet" samtidig med, at det inscannede kunne ses rulle frem på skærmen. Det var bare sagen. Desværre lå en stak bøger i vejen for min albue, så jeg nåede ikke det sidste uden at billedet blev forskubbet. Altså forfra igen. Et godt råd, sørg for at der er plads nok; der skal virkelig kunne „arbejdes". Det var godt, at skærmen kunne slettes. Forfra igen.

Endnu engang blev der kvitteret for de angivne størrelser, og jeg var klar igen. Billedet rullede fint ind på skærmen og jeg sluttede, da hele billedet var „inde".

Glad over at det var lykkedes, kunne jeg læne mig tilbage og betragte resultatet. Men hvad var det? Det var jo et skævt billed. Det hældede slemt til siden, og et grimt, sort felt kunne ses i højre side. Om igen.

Næste resultat var bedre. Billedet sås pænt lodret og den sorte flade var væk. Jeg havde bl.a. forøget bredden af det scannede areal. Men ak, trængslernes tid var ikke forbi endnu. Den nederste del af billedet buede fælt til højre. Jeg måtte ubevidst have drejet hånden under scanningen. Endnu engang forfra igen. For at undgå den sidste fejl, var jeg en tur i kælderen og kom op med en træliste ca. 1x3 cm og en længde på ca. 30 cm. Den anbragte jeg parallelt med billedets længste side, så jeg kunne styre efter denne. Nu skulle jeg blot holde scanneren vinkelret på listen, så skulle problemet ikke opstå igen.

Efter et par forsøg lykkedes det langt om længe. Nu sad billedet der på skærmen og var flot. Så var det bare med at få det gemt. På menubjælken blev valgt „FILE" og på rullegardinet „SAVE AS", og en ny skærm bad mig vælge Filnavn, Bibliotek og udvidelse samt Filtype, hvor der var mange forskellige at vælge imellem. Hvilken skulle det være? Det havde jeg ikke tænkt på, men mit tekstprogram kunne vist godt bruge TIFF, så jeg valgte denne. Endelig kunne jeg klikke på OK og billedet var gemt. Eller var det.

På dette tidspunkt var der forsvundet flere timer. Frokost var det ikke blevet tid til, og det føltes kun som om, der bare var gået en halv times tid. Sådan kan det jo gå, men skidt, jeg havde billedet i kassen.

Efter en vel overstået „glemt frokost", blev billedet hentet ind på skærmen igen. Resultatet skulle beundres af den øvrige del af familien. Jo, far kunne bare det der.

„Hvad er det for nogle prikker der sidder derude, og den streg der? Hører den med til billedet?" Ja, det var deres reaktion.

Billedmanipulation

Men hvad var det der? Ved nærmere sammenligning med billedet viste det sig, at de nævnte ting intet havde med billedet at gøre. Hvad nu? Skulle det hele gøres om igen? Fat i manualen og under menuen „CUT" viste der sig mulighed for at benytte et viskelæder i form af en „fantom"-ramme. Uden at trængslerne med dette fænomen skal med her, lykkedes det at beskære billedet for de fleste „forkerte" prikker. Men der var stadig nogle tilbage tæt ved billedet, der godt kunne undværes.

På menubjælken var der et valg der hed „ZOOM". Det blev prøvet, og nu delte skærmen sig i to dele. Den til venstre indeholdt en masse sorte prikker uden særligt sammenhæng, medens den til højre gengav billedet på samme måde som før. En lille firkant på skærmen til højre kunne flyttes med musen, og samtidig ændredes billedet til venstre. Det, der var i denne lille firkant blev altså gengivet på skærmen til venstre. Alle tiders. Nederst på skærmen var der mulighed for, at skifte mellem en hvid og en sort cursor, og nu kunne der retoucheres i det indscannede billede. Klik på klik fulgte nu, og billedet blev klarere og klarere. Selv inde i billedet blev der fjernet uønskede prikker. På et sted så det ud til, at billedet manglede noget. Dette kunne nu afhjælpes med cursoren og efter et par forsøg kunne det faktisk ikke ses, at det var repareret. Igen forsvandt nogle timer før et flot, flot billede kunne ses på skærmen, og nu kunne det virkeligt accepteres - også af resten af familien.

Billedindsættelse i tekstfil

Meningen med det indscannede billede var, at det skulle placeres i en tekstfil. Altså blev tekstprogrammet hentet ind og stedet i teksten blev

fundet. Et grafikfelt blev dannet, og billedfilen blev hentet ind. Resultatet sås straks på skærmen, men billedet sad ikke i centrum af feltet. Hvad nu? Skærmen blev slettet, og grafikdelen i tekstbehandlingsprogrammet blev kaldt frem på skærmen, og billedfilen hentet ind, hvorefter billedet blev beskåret, så kun det areal, der skulle bruges, blev tilbage. Igen tilbage til tekstfilen, hvor grafikfeltet nu blev fyldt med billedfilen, og resultatet var godt.

På denne måde var der gået adskillige timer. Resultatet var nået, og jeg var adskillige erfaringer rigere.

Siden disse trængsler er det blevet til en del billeder - stregbilleder. Jeg har forsøgt mig med rigtige billeder, men jeg synes ikke resultatet har været helt acceptabelt, men det kommer måske en anden dag. En del af de mange muligheder, der ligger i programmet for at behandle et indscannet materiale har jeg forsøgt mig med. Noget er godt, andet mindre godt, men det skyldes jo nok mine evner til at læse og forstå en manual.

Denne beretning må endelig ikke afholde andre fra at gå igang; den skal blot vise, at der skal øvelse til. Når rutinen er opnået, opvejes trængslerne lang ved glæden over slutresultatet.

God fornøjelse.

Bent Pilgaard
Randersvej 29
8800 Viborg

Eksempel på en vellykket indscanning. Det er Karl d. 1. af Flanderns segl.

Scanning og OCR

af Søren H. Sørensen

Dansk Data Arkiv iværksatte for et par år siden en række forsøg med Scanning og OCR. Formålet var primært at undersøge hvad der var på markedet, og hvad man kunne forvente at opnå med scanning og OCR; det vil sige, om man i forskellige situationer kunne spare tid ved at slippe for manuel indtastning. Forsøget er ikke afsluttet, da vi blandt andet afventer en ny version af OmniPage. Men allerede nu kan argumenterne for OCR siges at være både for og imod.

Hvad er OCR?

OCR står for Optical Character Recognition, hvilket på dansk vil sige optisk tegngenkendelse, og dækker over det, at computeren ved hjælp af software er i stand til at oversætte en grafisk repræsentation af tekst - typisk et indscannet bitmap - til et format computeren umiddelbar kan „læse", for eksempel ASCII eller et givent tekstbehandlingsformat. OCR kræver naturligvis OCR-software, men også at man har adgang til en scanner. Hvor OCR-programmer er i stadig udvikling til forskellige formål, må teknologien omkring scannere siges at være nået et stade, hvor vi så at sige ikke behøver at nå længere: Scannere kan i dag købes endog meget billigt, og i kvaliteter, som er tilfredsstillende i OCR-sammenhæng.

Hvordan virker OCR?

Selve OCR-processen kan fremstilles således: Originalt dokument - Scanning - Selve OCR-delen - Tilretning - Maskinlæsbar fil.

Det originale dokument, man ønsker bearbejdet, bør være af en vis kvalitet. Mange faktorer bør tages i betragtning, ikke mindst papirets fysiske stand. Der må ikke være for mange krøller eller pletter, som kan generere „støj", og samtidig bør skriften være tydelig for at undgå for mange fejlfortolkninger. Ofte vil en fotokopi af det, man ønsker indscannet, være at foretrække, da fotokopimaskiner ofte genererer en kopi, der er tydeligere end originalen, eller ihvertfald kan indstilles til det. Hertil kommer, at det er lettere at indscanne et stykke papir, fremfor for eksempel en bog, specielt hvis man skal passe på ikke at ødelægge ryggen på bogen.

Scanneren

Der findes groft sagt 2 typer scannere. Håndscannere og flatbedscannere. Flatbedscannere er absolut at foretrække - ikke mindst fordi de ofte kan forsynes med Sheet-feeder, hvilket kan spare megen tid, og fordi det har vist sig meget besværligt at frembringe gode bitmaps af A4-sider med håndscannere. Scanneren producerer et digitalt billede af det scannede dokument. Scanneren „oversætter" det scannede til millioner af punkter ved at projicere lys på dokumentet og ved hjælp af censorer at registrere det reflekterede lys, hvorved hvert enkelt punkt tildeles en værdi, sort eller hvid. Dette „kort" gemmes i computeren som en bit-map fil, der umiddelbart kan hentes frem i et billedbehandlingsprogram.

Antallet af punkter i scanningen defineres som dpi, dots per inch, dvs. punkter per tomme. I OCR-sammenhænge vil 300 dpi (90.000 punkter pr. kvadrattomme, dvs. 14.400 punkter pr kvadratcentimeter) normalt være tilstrækkeligt, men ved meget små skrifttyper (8 punkts skrifttyper og ned) vil det være nødvendigt med en højere opløsning, typisk 400 dpi.

OCR-programmet

Selve OCR delen kan igen fremstilles således:

Lokalisering af tegnet - Isolering af tegnet - Klassificering af tegnet - Oversættelse af tegnet.

Det lyder enkelt, men ved en gennemgang af processen fremgår det tydeligt, hvor mange steder det kan gå galt. Ved lokaliseringen af tegnet kan programmet nemt komme i tvivl, om der rent faktisk er tale om et tegn. Måske drejer det sig om en flueklat, en kaffeplet eller en sort plet genereret af scanneren, fordi papiret var krøllet netop på det sted. Ved isoleringen af tegnet kan programmet igen komme i tvivl, for det kan være vanskeligt at bestemme, hvor langt tegnet egentligt går. Når et givent tegn forstørres, vil man ofte se, at tegnet ikke er helt sammenhængende, men godt kan have 'huller'. En typisk fejl i denne kategori er fejllæsning af 'm' som 'rn' eller omvendt. Ved gennemgangen af programmets klassifikation af tegnet er det nødvendigt først at vide lidt om, hvordan OCR-programmer 'genkender' tegn.

Der findes to typer tegngenkendelse: Den ældst kendte er den såkaldte matrix-match, hvor bitmap-billedet af hvert enkelt bogstav sammenlignes med tegn fra et bibliotek indeholdende alle kendte tegn. Hvis formen passer (nøjagtigt) med et kendt tegn, er det genkendt. Denne type må siges at være den mest nøjagtige, men den kræver naturligvis, at biblioteket indeholder samtlige tegn, som programmet kan komme ud for, og vel at mærke i den korrekte størrelse og form, så det kræver gigantiske computere at lagre den store mængde tegntyper, og det tager alt for langt tid at matche dem. Derfor anvender de fleste OCR-programmer den såkaldte karakteristika-match, hvor bitmapbilledets formkarakteristika analyseres, og sammenlignes med kendte former. Denne type er langt mindre krævende, da det er uden betydning, for eksempel hvilken størrelse tegnet har, og i mange tilfælde, hvilken font der er tale om. De fleste OCR-programmer i dag kan således læse de almindeligste latinske fonte, uanset hvilken størrelse de har. Hvis tegnet ligner noget kendt, oversættes det så, for eksempel til en ASCII-kode.

Fejl

Hvis programmet ikke genkender et tegn, erstattes det med et specialtegn - hos OmniPage for eksempel '~'. Dette letter tilretningen, da man i den efterfølgende fase kan søge og erstatte på dette specialtegn, og er det en gennemgående fejl (tidlige versioner af OmniPage har således svært ved de danske karakterer æ, ø og å), kan man 'træne' OCR-programmet til genkendelse af disse tegn,(Train Recognition) og køre processen igen. Ud over denne type fejl, hvor OCR-program met afviser et tegn, fordi det ikke er kendt, er der også de lidt mere farlige fejl: Fejlfortolkning, hvor output-tegnet er forkert: 'S' bliver til '5' eller omvendt, 'm' bliver til rn' fordi isoleringen af tegnet gik galt, og de fejl, hvor et tegn slet og ret bliver ignoreret, fordi OCR-programmet troede der var tale om 'støj', eller hvor 'støj' faktisk afføder et tegn. Denne type fejl kan være svære at fange uden en intensiv korrekturlæsning, hvorfor det er nødvendigt at planlægge sin efterbearbejdning af det OCR'ede materiale. Man kan spørge sig selv, hvorfor maskinen har så svært ved at skelne disse ting fra hinanden, når det for et menneske er så let at læse trykte bogstaver. Men her er det vigtigt at tænke på, at mennesker ofte læser ord som billeder: Det vil sige at man normalt ikke staver sig igennem ordene, men genkender dem som hele ord, og det går vel at mærke lynende hurtigt. Man kunne forestille sig, at man også kunne få maskinen til at læse sådan, altså hele ord, og det er da også teoretisk muligt, men i praksis vil det kræve så stor maskinkraft, at det ikke kan lade sig gøre.

Efterbearbejdning

Som nævnt er efterbearbejdningen vigtig. Her er det til stor hjælp med et tekstbehandlingssystem med en god stavekontrol. Efter at man har rettet de fejl, hvor OCR programmet gav helt op, kører man sin stavekontrol og fanger således fejl, hvor et tegn er misfortolket; men naturligvis ikke alle, hvorfor en manuel korrek-

turlæsning også vil være nødvendig. Samtidig kan OCR-programmet have store problemer, hvis det indscannede har et specielt format, som man ønsker bibeholdt i sit output. Her tænker jeg specielt på tekst i kolonner eller skemaer, hvor det ofte er nødvendigt manuelt at angive kolonnernes format, og i hvilken rækkefølge de ønskes oversat. Dette kan være en ret tidskrævende proces.

Hardware- og Software-krav

Man kan så spørge, hvad der kræves af soft- og hardware. Der findes i dag en mængde OCR-software, der efter min mening groft kan opsplittes i to grupper: Forsknings-/industri-software og Off-the-Shelf. I denne forbindelse er det mest interessant med den sidste gruppe. Her kan man købe ret professionelle systemer for mellem 2.000 og 10.000 kroner. Jeg vil nok personligt sætte min lid til systemer i prislaget fra 4.000 kroner, og det er i denne kategori OmniPage falder. Jeg skal dog ikke afvise, at der findes billigere og lige så gode programmer, men det er OmniPage, som er blevet testet på DDA, da dette system var og er state of the art.

Hertil kommer behovet for effektivt at kunne efterbearbejde det indscannede. Her er det som sagt nødvendigt med et godt tekstbehandlingssystem med en god dansk stavekontrol. Med OmniPage følger en teksteditor, som er udmærket, men langt fra lever op til seriøse krav. Et andet godt hjælpemiddel er en grafisk editor eller en godt (avanceret) tegne-/billedbehandlingsprogram som en hjælp i forbehandlingen af det indscannede, det vil sige efter indscanningen og før selve OCR-delen. Navnlig hvis man ikke har gode muligheder for at indstille parametre som kontrast og brightness på sin scanner, vil man være glad for at kunne justere disse ting i bitmappet, samtidig med at man vil kunne 'rense' det scannede dokument -dvs. fjerne pletter og anden støj inden man kører billedet gennem OCR-programmet.

I hardware-enden kræves naturligvis en scanner. Dansk Data Arkiv råder over en HP Scan Jet II c, som i sin tid blev erhvervet i prislaget ca. 7.000 kr. Dette er en flatbed scanner med mulighed for montering af en ark-føder. Scannere kan i dag erhverves billigt, og kravene er som nævnt ikke store, men har snarere karakter af magelighed - dvs. hvor meget tid man vil spare i scanneprocessen. Samtidig stilles der også krav til pc'en. Bit-map filer har det med at blive endog meget store. Typisk vil en A4 side fylde cirka 1/2 Megabyte i en 300 dpi opløsning. At man er i stand til at pakke sine grafikfiler i forskellige formater er kun med til at gøre processen mere tidskrævende. Så god plads på harddisken og rigeligt med RAM er nødvendigt. OmniPage siger 4 Megabyte RAM, men 8 Megabyte er ønskeligt. Hertil er en hurtig PC rar at have - hvor hurtig, er igen et spørgsmål om, hvor lang tid processen må tage. For mennesker, som har en rimelig PC i forvejen med den nødvendige for- og efterbehandlingssoftware, altså en investering på fra cirka 4.000 - 5.000 kroner og op. Jeg skal dog ikke forsværge, at det kan gøres billigere, men erfaringen er, at kvaliteten følger prisen.

Pålidelighed

OCR's pålidelighed er, som det ses, et samspil mellem mange faktorer. Kort ridset op kan man sige, at pålideligheden afhænger af følgende faktorer:

1. Kvaliteten på det, man ønsker inscannet:
- papirets fysiske tilstand (helst ingen krøller)
- det tryktes tydelighed (font-type og -størrelse, utydelig kopi, men også om det er sat op i et specielt format, som man ønsker bibeholdt, f. eks. skemaer eller kolonner)
2. Scannerens kvalitet:
- Opløsning
- Muligheder for manuelt at styre scanne-optioner, såsom kontrast/brightness, filtre etc.

3. Selve OCR programmet:
- Yderligere filtrering
- Definering af det originale format
- Muligheder for indlæring (Train Recognition)

Normalt siger man, at man vil acceptere en akkuratesse på 96%. Det vil sige 4 fejllæsninger pr. 100 læste tegn, men efter min mening skal man op på en nøjagtighed på 98 - 99%, for at resultatet er tilfredsstillende. Man skal i den forbindelse være opmærksom på, at producenter af OCR-software, når de opgiver programmets fejlprocenter, kun medregner fejl, hvor programmet har givet helt op, dvs. erstattet et tegn med specialtegn, men ikke fejl, hvor f.ex. 'denne' blev oversat med 'derme' etc., og kun fejlprocenter efter en eventuel kørsel af Train Recognition.

OCR vs. manuel indtastning

Når man har opnået et tilfredsstillende resultat med scanning og OCR, er det interessant at sammenligne med manuel indtastning. Her har man på DDA og andre steder forsøgt at lave nogle tidsstudier. Vores erfaringer har vist, at en standard-side kan scannes og OCR'es i løbet af rundt regnet 2. min. - naturligvis afhængigt af, hvor hurtig maskinen er. Heri er indregnet tid til at skifte papiret manuelt i scanneren, og jeg skal ikke forsværge, at det kan gøres hurtigere. En god sekretær kan vel opnå en hastighed på 200 tegn i minuttet, svarende til 4,4 sider i timen, forudsat at en normal-side er 2700 tegn.

Hertil kommer, at der under OCR-processen kan spares tid ved at indscanne store mængder tekst, og dernæst lade maskinen OCR're det indscannede natten over. Omvendt kræver tekster til OCR som nævnt en vis standard, og ovenstående beregninger kræver en god og ensartet kvalitet af det som ønskes scannet.

Det er klart, at både OCR behandlede og manuelt indtastede filer kræver en vis efterbehandling f.ex. korrekturlæsning. Og her kommer helt andre konkurrenter ind i billedet: Store firmaer i Østen tilbyder afskrifter af alle mulige slags dokumenter, og til en meget billig pris, naturligvis på grund af meget lave lønninger. Hertil vil nogle indvende, at det ikke kan nytte noget, når afskriverne ikke forstår, hvad det er, de skriver, men det viser sig, at man opnår den bedste afskrift, hvis kopisten ikke ved, hvad der rent faktisk står.

Konklusion

Konklusionen må være, at der er tid og arbejde at spare ved OCR. Dog bør man ved hver enkelt opgave vurdere det hensigtsmæssige i at anvende OCR fremfor manuel indtastning. Selve scanner-teknologien har længe været på et tilfredsstillende stade. Derfor har mange ment, at det kunne være en fordel at indscanne dokumenter som billeder og vente på den 'perfekte' OCR-teknologi. Men ved OCR af normale fonte og i rimeligt ukomplicerede layouts forløber OCR tilfredsstillende, forudsat at originaldokumenterne er af rimelig karakter. Og da OCR-software er kommercielle produkter - det vil sige, at de skal kunne konkurrere og sælges - kan man roligt regne med store forbedringer samtidig med betragtelige prisfald, hvorfor jeg mener det er en teknologi med gode fremtidsudsigter.

Søren H. Sørensen
Dansk Data Arkiv.

Scan - scan ikke

Af John Thomsen

Et af de nye „frække" fremmedord, vi er begyndt at støde på i vores omgang med computere, er scanning, og for ganske mange har det en helt magisk klang. Der er imidlertid hverken noget nyt, frækt eller fremmed over begrebet, der ligger bag. Det er i familie med det gode gamle danske „skimme", og står faktisk blot for en systematisk afsøgning (af et eller andet).

Der er således tale om en skimming eller scanning, når vi afsøger telefonbogen spalte for spalte, eller når vi lader blikket glide hen over en kasse øl - række for række - for at se, hvor mange der er hhv. med og uden kapsler. Der er også tale om en slags scanning, når en radar-skærm viser, hvor og hvor mange skibe eller fly, de nærmeste omgivelser byder på. Drejer man skala-knappen på sin radio hele skalaen igennem, foretager man også en scanning, og det findes der modtagere, der kan gøre automatisk; de scanner frekvenserne.

I den sammenhæng, hvorom det følgende handler, fungerer scannere som en art kopimaskiner. Den mest kendte af slagsen er de stadigt mere udbredte telefax-maskiner. En lysbjælke bevæger sig hen over/under en original og afsøger - linie for linie -hvor originalen er hhv. sort og hvid. Resultatet bliver et digitalt signal og/eller en elektronisk fil, bestående af 0 og 1 („nuller og et-taller").

Dpi-betegnelsen

Forestiller man sig et stykke kvadreret papir eller milimeterpapir, hvor en hel masse celler er sorte og en masse andre er hvide, så har man et godt begreb for,hvorledes en scanning opfattes af scanneren. På det kvadrerede papir vil fem celler udgøre en tomme, mens milimeterpapiret har 25. På milimeterpapiret kan man „tegne" noget mere nøjagtigt, navnlig i kurver, end på det kvadrerede. Dette kaldes opløsning, og benævnes „dpi" - dots per inch - prikker pr. tomme.

Når man møder dpi-betegnelser, står der oftest to tal foran, fx 75x75, 200x300, 600x600 etc. Som oftest er betegnelsen „kvadratisk", altså to lige store tal, men der er intet til hinder for at arbejde med uens størrelser. Bredden (bør) nævnes før højden. Det er vel ret indlysende, at 600x600 ikke er dobbelt så fint som 300x300, men fire gange så fint. Det gælder så_også, at signalet eller filen af det scannede bliver fire gange så stor.

Nøjagtig de samme begreber gør sig gældende i laserprintere og i telefax-maskiner. For få år siden var det ren luksus at arbejde i 300x300, men i dag er 600x600 dpi-printere ved at overtage førerpositionen, 1200x1200 lurer i kulissen, og der er ikke langt op til de ca. 2400x2400 som fotosatsmaskiner arbejder med. På sidstnævnte har man haft en anden målestok, nemlig raster-linier pr. cm, hvor det var ligesom et netværk, svarende til selve stregerne på milimeterpapiret, der udgjorde opløsningen.

Hvad kan det så bruges til? Ja, i telefax-maskinen drejer det sig kun om at give modtagermaskinen besked om at sværte et stykke papir tilsvarende. De første telefaxmaskiner blev da også benævnt „tele-copier", fordi de fungerede som fjern-kopimaskiner.

Til computerbrug kan man forestille sig tre forskellige former for udgangsmateriale til scanning:
- Et billede, som ren halvtone (foto) eller som raster (fra avis, blad eller brochure)
- En stregtegning
- Tekster, der principielt også er en slags stregtegning.

Billeder og stregtegninger

Billedet vil i scanneren og dens program blive opløst i prikker, alt efter hvilken opløsning man har valgt. Er det et rasteret billede, skal man være opmærksom på, at der kan opstå moire („moaræ") i scanningen, dvs. at den oprindelige rasterfinhed ikke „går op" med den opløsning, man har valgt i scanningen. Moire kan te' sig på flere måder, men for at give et billede af resultetet, så prøv fx at holde to stykker nylonstrømpe op foran ansigtet, i god afstand bag hinanden, og i hver sin grad af udspænding.

En stregtegning er så afgjort meget letter at få til at blive god, og i den sammenhæng er det værd at erindre, at de gamle stålstik, kobberstik osv. også er stregtegninger, ligesom den gamle tegnestil. På dette sted skal det også nævnes, i forbindelse med vor specielle hobby, at scanninger fra fx gamle kirkebøger, folketællinger og lignende også er at opfatte som stregtegninger, med håndskriften som „tilfældige" streger.

Billeder og stregtegninger kan man gøre brug af ved siden hen at lægge dem ind i, hvad vi kan kalde dokumenter. Har man et rimeligt nyt tekstbehandlingsprogram, kan man på selvvalgte steder oprette en ramme, hvortil det ønskede billede importeres. Vi kan altså scanne billedet (på væggen, i albummet, fra avisen) af oldeforældrene og lægge det ind i den slægtskrønike, vi er ved at skrive på. Vi kan også scanne et gammelt dokument (dåbsattest, skøde, arveskifte, diplom fra fugleskydning etc.) og lægge det ind i vor krønike, men denne gang som en stregtegning.

De scannede billeder og stregtegninger kan selvfølgelig også efterbehandles. Der findes en række tegnings- og billedbehandlings-programmer, hvor man kan tage en scanning ind, og fjerne eller indsætte helt ned til enkelte prikker (som her kaldes pixels i stedet for dots. men det er en anden historie). Vi kan fjerne en distraherende baggrund, jævne vorten på oldemors næse og give oldefar et par ekstra hentehår på issen. På stregtegningerne kan vi sågar lave dokumentfalsk; ændre tredjepladsen i fugleskydning til en vinder!

Når scanningerne skal lægges ind i et dokument, er man ikke nødvendigvis bundet af scanningens faktiske størrelse. Dette være sagt med forbehold for de enkelte programmer. Men sædvanligvis kan man godt gøre sit „image" større eller mindre. Hvis man ikke gør det linært, dvs. ændrer bredde og højde lige mange procent, skal man være opmærksom på, at resultatet enten kan blive klemt eller strukket på den ene led. Har man derimod bedt sit program om at opretholde skalering, vil scanningen blive inden for det mindste af de to mål.

Tekstgenkendelse (OCR-scanning) Det „nyeste dyr i åbenbaringen" hedder tekstscanning. Fra starten af er der, som nævnt, tale om scanning af en stregtegning. Dernæst kan man så lede sin scanning gennem et program, som så kan „læse" billedet og omsætte det til ren tekst. Grundbegrebet her er OCR - Optical Character Reading. Heller ikke dette er specielt nyt.

De ældste former opererer med nogle specielle tegnsæt, som specialmaskiner kunne læse. Bl.a. såkaldte OCR-A og OCR-B alfabeter. Vi har stort set alle set dem og deres slægtninge. OCR-A kendes bl.a. på sine kantede tal, hvor specielt 1-tallet skilte sig ud med sin store „hæl", mens OCR-B ser noget mere „normal" ud. OCR-B bruges bl.a. nederst på giro-taloner. En OCR-type er også de lodret stribede tal, vi kan se nederst på vore checks.

For nogle år siden fremkom de første programmer, der kunne omsætte scannede billeder af tekst - til tekst. Forudsat altså, at skriften var en som programmet kendte. Dvs., at der i programmet lå en art „spejling" af en skrifttype, som scanningen blev sammenlignet med. De

første programmer/versioner kunne kun „læse" skrivemaskineskrift, og højst to typografier, hvoraf Courier gerne var den ene. Følgelig var der heller ikke nogle problemer med skriftstørrelsen!

Senere programmer/versioner blev i stand til at genkende (faktisk et mere præcist udtryk end „læse") op til 16 slags skrift, men stadig kun i et behersket antal størrelser. Udviklingen går dog i retning af, at også trykt skrift skal kunne genkendes, og at der skal kunne ses forskel på fed skrift, kursiv, understreget osv. En ting er stadig påkrævet: der må kun være en slags skrift - en font - i hvert dokument.

Kurzweil-scanninger og ICR

En meget dygtig, ung amerikaner af tysk afstamning, Kurzweil, fandt ud af, at han ville konstruere en læsemaskine for blinde. Et delresultat blev hans ICR -Intelligent Character Reading - maskiner. Han bygger her på et princip om bogstavernes topografi, altså bestanddele såsom stammer, buer, skråstreger etc. Ud fra dette kan maskinen lære en hvilken som helst skrift at kende, så længe bogstaverne bare ikke hænger sammen (og af og til alligevel). Fx kan man have held til at fortælle maskinen, at „rn" på en dårlig kopi ikke er et „m" men „rn", og lignenede såkaldte ligaturer (sammensmeltninger).

Når man starter på en scanning, vil programmet lige som „smage" på teksten for bl.a. at finde ud af, hvor brede de bredeste bogstaver er (W, Æ, M). Møder den et bogstav, den ikke kan forstå, fx. et „k", der har mistet lidt af fanen (en gammel slidt skrivemaskine), spørger den måske, om det er et „h". Det rettes til „k", den spørger måske et par gange eller flere, men har så lært sig, at sådan ser altså et „k" ud i denne skrift. Dette sker med anvendelse af såkaldt Artificial Intelligence -"kunstig intelligens" - hvad berettiger I'et i ICR.

Kurzweil-scanneren kan læse op til ni forskellige skriftsnit på samme side, med variationer som kursiv, fed, understreget, hævet, sænket etc. i hver af dem - men hvilken typograf ville dog have lavet sådan noget makværk! Den er i stand til at skelne 0 fra O (nul og store o), l, log I (lille L, et-tal og store i), og den kan endog læres at se bort fra orddelings-bindestreger.

Resultatet kan leveres ud i alle de kendte tekstbehandlings-typer, og vel at mærke, medbringende besked om fed, kursiv, indrykning osv. Modtager man resultatet i et rigtigt DTP (DeskTop Publishing) program, vil de forskellige typografier være afmærket til brug som titel, overskrift, underoverskrift, brødtekst og hvad der ellers måtte være.

Kurzweil's ICR er ikke nogen fjern fremtid; jeg har bl.a. scannet en godt 600 sider trykt bog (kopieret op til A4) på under tre uger, så den med efterfølgende redigeringsarbejde kunne gå til sats og tryk, samt (gen)udgivelse inden for tre måneder. Bogforlag regnes med 8-10 måneder for en traditionel udgivelse. Jeg medtager kurzweil i denne omtale af scanning, da det dels kan være praktisk at vide, at det forefindes, dels fordi jeg venter at både OCR-programmerne og PC'erne også med tiden vil nærme sig ICR's muligheder.

Hvad kan man så ikke? Ja, hverken OCR eller ICR må forventes at kunne „læse" vore gamle kirkebøger. Dertil er skriften for snørklet, sammenhængende og uregelmæssig. Skrev man lige så ensartet som mange arkitekter og ingeniører på deres tegninger, var det en smal sag. Men systemer til læsning af håndskrift ligger stadig et stykke ude i fremtiden. På den anden side, så kan man måske i stedet blot læse højt til et program, der forstår, hvad man siger, og derfra kan omsætte til almindelige tekstfiler.

John Thomsen
Søvangs Allé 6
3500 Værløse

ORDLISTE

Algoritme:
Beskrivelse af en fastlagt række regneoperationer, der fører til et ønsket mål. F.ex. fører denne algoritme til en vares udsalgspris: Beregn fortjeneste -> Læg købspris og fortjeneste sammen til salgspris -> Beregn moms -> Læg momsbeløb og salgspris sammen til udsalgspris.

ANSI:
(American National Standard Institute). Et tegnsæt i lighed med ASCII. ANSI består af 256 tegn, og det benyttes i Windowsprogrammer.

ASCII:
(American Standard Code for Information Interchange.) Udtales: "aski". Et tegnsæt med en fastlagt standard for hvilke talværdier, der repræsenterer hvilke tegn. Det originale ASCII-tegnsæt kunne med 7 bit kun repræsentere 128 tegn og indeholdt ikke talværdier for f.ex. de specielle danske/nordiske tegn. Senere udgaver af tegnsættet kan med 8 bit repræsentere 256 tegn, men her er der ikke absolut overensstemmelse mellem forskellige udgavers talmæssige placering af f.ex. danske tegn. Derfor den megen bøvl med æ, ø og å. Der arbejdes med en ny standard på området, som kommer til at rumme mere end 16000 tegn. Hermed skulle problemerne med alle landes særtegn kunne løses.

ASCII tekstfil:
En tekstfil, der kun indeholder tegn fra ASCII-tegnsættet. Et meget anvendeligt format til udveksling af tekster mellem forskellige programmer. Indeholder kun oplysninger om den rå tekst, bogstaver, tal, linieskift og mellemrum - dvs. oplysninger om f.ex. skrifttype og -størrelse i overskrifter kommer ikke med.

Bilevel:
Et udtryk for, at der kun arbejdes med to muligheder - noget eller ingenting, sort eller hvidt. Det er faktisk også en angivelse af, hvor store/små punkter scannerprogrammet skal anerkende som værende acceptable.

Billedbehandlingsprogram:
Edb-program til manipulering af elektroniske/digitaliserede billeder. Kan grundlæggende udføre samme arbejde som ellers i et mørkekammer eller et fotolaboratorium; dvs. ændre kontrast, lysstyrke, farver, størrelse De mere avancerede har mange specielle effekter og raffinerede tegneredskaber som f.ex. at fremhæve kontur, vende og dreje hele billedet eller enkelte dele af det, fjerne og indsætte billedelementer og mulighed for at konvertere et billede fra et fil-format til et andet. Findes i mange udgaver/priser.

Digitalisere:
At tildele en talværdi. F.ex. kan lyd digitaliseres ved at måle tonehøjden i et forløb af ultrakorte tidsenheder. Når lyden er repræsenteret af en mængde tal, kan den manipuleres ved almindelige beregninger. Det samme med farver, hvor kombinationen af grundfarverne rød, blå og grøn kan måles og omsættes til talværdier, og med gråtoner, hvor lysstyrken i hvert punkt omsættes til en talværdi. Farverne/gråtonerne kan så ændres ved at ændre talværdierne.

Dither-teknik:
En billedbehandlingsteknik til udregning af en gennemsnitsfarveværdi for alle pixels i et fastlagt område - f.ex. 3 x 3 pixels. Alle farveværdier i dette område ændres til gennemsnitsfarven, så det ændrede billede nu fremstår som ensfarvede punkter/dots a 9 pixels. I et udprintet sort/hvidt billede bliver resul-

tatet, at et sådant område enten får den sorte farve eller beholder papirets hvide. Dvs. at afstanden mellem de sorte punkter, tætheden ændres, og på udskriften snydes øjet til at opfatte de sorte punkter af varierende tæthed som gråtoner.

DPI:
(Dots Per Inch). Punkter pr. tomme. En angivelse af hvor mange punkter pr. måleenhed f.ex. en scanner eller en printer kan håndtere, et mål for dens opløsningsevne. Et højt tal angiver en stor tæthed af punkter, en højere opløsning og en finere kvalitet.

EPS-format:
(Encapsulated PostScript). Filformat, der bruges af f.ex. DtP-programmerne Ventura Publisher og PageMaker.

Flat-bed-scanner:
En scanner-type, hvor oplægget/billedet placeres på en glasplade som i en kopimaskine.

Fotodiode:
Elektronisk komponent, der kan omsætte lys til elektrisk spænding, som så kan måles med talværdier, digitaliseres.

Gammakorrektion:
En teknik til at ændre lysstyrken i en bestemt farve/gråtone i et billede eller et område af et billede. Bruges specielt til at kalde det frem, der ligger i alt for mørk en skygge.

GEM-format:
Fil-format til billedfiler fra tegneprogrammer dannet i streg/vektor-grafik. Bruges f.ex. af GEM Draw.

GIF-format:
(Graphics Interchange Format) Fil-format til billedfiler. Meget udbredt f.ex. i BBS-verdenen, hvor billeder ofte udveksles i dette format.

Gråtone-grafik:
En billedtype, hvor lysstyrken i hvert punkt/hver pixel måles til en af et i forvejen fastlagt antal gråtoner, ofte 256.

Halvtone-grafik:
En billedtype, hvor der bare måles på hvert punkt/hver pixel, om der er noget eller ej, sort eller hvidt. En billedfil i halvtone fylder langt mindre end en i gråtone, men giver til gengæld ikke de store muligheder for efter behandling. Halvtone bruges specielt om sort/hvid fotografier - halve toner.

Histogram:
En grafisk oversigt i form af et søjlediagram over fordelingen af lysværdierne på alle pixels i billedet eller i et område af et billede. I flere billedbehandlingsprogrammer er det et nyttigt værktøj til justering af lys og kontrast i billedet.

ICR:
ICR (Intelligent Character Reading). Vurderer (genkender) de enkelte bogstaver ud fra deres topografi. Den intelligente del af programmet kan også afgøre logiske sammenhænge og forskelle, f.eks hvor der skal være tal eller bogstaver m.v. Ved ICR-scanning kan alle bogstavtyper i princippet behandles.

Kontrast:
Et mål for antallet af og variationen i forskellige gråtoner i et billede. Et billede med høj kontrast har stor spredning mellem lyseste og mørkeste gråtone. Det giver tydelig forskel mellem lyse og mørke partier i billedet. Ved lav kontrast er de anvendte gråtoner samlet i et midterområde på gråtoneskalaen. Det giver slørede, tågede billeder.

Lysstyrke:
Et mål for indholdet af hvidt i et billede, et punkt eller en pixel. Et helt hvidt billede/punkt/pixel har således højeste lysstyrke.

Notebook:
Bærbar, batteridrevet edb-maskine. Mappedatamat. Har ofte samme kapacitet og ydeevne som en traditionel PC.

OCR:
(Optical Character Recognition). Optisk karaktergenkendelse. Et OCR-program aflæser et indscannet billede af en tekst, område for område i ganske små områder. Når programmet identificerer et sådant område som et af bogstaverne, ændres de grafiske talværdier i området til talværdien for det genkendte bogstav. Således bliver billedet af teksten til en tekst, der kan justeres og manipuleres f.ex. til videreskrivning og opsætning.

Omnifont:
Et udtryk for, at et OCR-program kan genkende tekst fra nogle forskellige fonte.

PCX-format:
Fil-format til billeder. Er ikke helt standardiseret, så forskellige udgaver kan træffes. Er ellers et af de mest anvendte billedformater, som de fleste billedbehandlingsprogrammer kan håndtere. Bruges bl.a. af Windows PaintBrush og programmet Paint Shop Pro.

Pixel:
(PICtureS ELement) En prik eller et punkt på et billede. Den mindste enhed, som et digitalt billede er dannet af. F.ex. dannes skærmbilledet af et antal linier med et antal pixels i hver linie. Hver pixel beregnes for sig med hensyn til farve og lysstyrke. Jo flere linier med jo flere pixels des finere opløsning på skærmen. Svarer til de dots, der også betegner opløsningen i et billede.

Rastergrafik:
Billedet dannes ved at opdele billedfladen i et antal punkter (dots/pixels), der hver for sig kan bestemmes med hensyn til farve og lysstyrke. Jo flere punkter på et bestemt areal, jo finere opløsning og desto skarpere og "naturtro" et billede

Scanner:
Inddataenhed til digitalisering af billeder. Lysfølsomme celler føres i ultrakorte ryk hen over billedet og sender ved hjælp af elektrisk strøm i varierende spænding oplysninger om hvert billedpunkts egenskaber til computeren.

SDF-fil:
(Standard Data Format) Fil-format til udveksling af data mellem f.ex. database- og regnearksprogrammer. Danner en tekst-fil med hver post på hver sin linie og med teksten i felter med fast længde - i modsætning til det kommaseparerede fil-format, der reducerer feltlængden.

Sheet-fed-scanner:
En scanner-type, hvor oplægget/billedet føres ind i scanneren og over en tromle som i en telefax eller i en laserprinter. Denne type har kun en ringe udbredelse i Danmark.

Streggrafik:
Billedet dannes ved at beregne og udfylde rette linier mellem fastlagte punkter, dvs. afstand og retning mellem punkterne. Det er den ældste form for grafisk teknik, der nu oftest er erstattet af rastergrafik. Kaldes også vektorgrafik.

Tekstgenkendelsesprogram:
EDB-program til genkendelse af tegn i indscannet billede af trykt tekst; også kaldet OCR-program. De genkendte tegn/bogstaver bliver digitaliseret, og kan efterfølgende behandles på PC-en som anden indtastet tekst.

TIFF-format:
(Tagged Image File Format) Fil-format til billedfiler. Næsten blevet standard for alle typer af grafik i forbindelse med scanning. Kan håndteres af næsten alle billedbehandlingsprogrammer og indeholder oplysninger om et stort antal egenskaber ved et billede.

LITTERATUROVERSIGT

Alt om DATA.
6/91 *På talefod med scanneme.*
Morten Strunge Nielsen: *Scannernes forunderlige verden.*
Morten Strunge Nielsen: *Fra billede til tekst.*
Torben Hyst & Morten Strunge Nielsen: *Scan - scan ikke.*

Frøslev-Nielsen, Aage: *Processer i digital billedreproduktion.*
(udg. Den Grafiske Højskole, juni 1993).

GENEALOGICAL COMPUTING, January-February-March 1994. Larry Ledden: *Multimedia for Genealogy.*

HISTORY and COMPUTING, Vol 5 No 2, 1993. Special Issue - Scanning and OCR.
René van Horik: *Recent Progress in the Automatic Reading of Printed Historical Documents.*
Gunnar Thorvaldsen: *Making Printed Historical Sources Machine Readable: Some Experiences with Optical Character Recognition*
Eric L. Helsper, Lambert R. Schomaker og Hans-Leo Teulings: *Tools for the Recognition of Handwritten Historical Documents.*

KontorBladet, Nr. 13 - 7. december 1992, 2. sektion: indkøb af kontor- og dataudstyr.

PC WORLD
juni 1991 (produktguide). Sonny B. Andersen: *Scannere - om scanneres opbygning og funktion.*
8/92. Jens Erlandsen: *Scanning af billeder til Multimedia.*

POLITIKEN. Torsdag d. 3. marts 1994, 4. sektion.
Læsning til lavpris.
Sproget er et problem.
Markedsføreren.
Når maskinen får læsebriller. Af Nikolaj Frandsen.

Privat Computer, 3-94. Lennart Svenstrup: *Tegnelærerens mareridt.*

SLÆGT & DATA, Nr. 2 og 3, 1993. Erik Helmer Nielsen: *Scanning af billeder, I-II.*

WINDOWS WORLD
5/1993. Henrik Graversen: *Fremtidens grønne mørkekamre.*
Klaus Møllgaard: *Logitech ScanMan Color.*
6/1994. John Agger. *Scanning for alle os andre.*

DIS-DANMARKS SÆRNUMRE

Siden foreningen for Databehandling I Slægtsforskning (DIS-Danmark) blev stiftet i 1987 er der - foruden de 4 årlige numre af medlemsbladet SLÆGT & DATA - hermed blevet udsendt 3 særnumre, hvor særlige temaer og emner tages op. De er udsendt gratis til medlemmerne af foreningen, men kan også købes i løssalg. Det drejer sig om:

Særnummer 1. **Slægtsforskningsprogrammer.**
Indeholder en generel beskrivelse af, hvilke muligheder programmerne kan give i arbejdet med slægtsforskningen samt en kort beskrivelse af 8 dansk-(nordisk)-sprogede programmer.

Udsendt i maj 1992. 32 sider. Pris: Kr. 37,50

Særnummer 2. **Kildeoversigten.**
Det er en oversigt over afskrevne slægts- og lokalhistoriske kilder (folketællinger, kirkebøger, tingbøger, matrikler m.v.). Omfatter både EDB-baserede, maskinskrevne og håndskrevne afskrifter. Oversigten indeholder oplysninger om ca. 2900 afskrifter.

Udsendt i september 1993. 96 sider. Pris: Kr. 50,00

Kildeoversigten (DIS-Kilde) kan også fås på diskette med et tilhørende læseprogram. Denne opdateres jævnligt og indeholder i skrivende stund oplysninger om ca. 3700 afskrifter. Pris: 50,00.

Særnummer 3. **Scanning.**
En kort beskrivelse af teknikken bag billed/tekst-scanning, og hvilke muligheder det giver i forbindelse med slægtsforskning. Syv brugere af scannere har bidraget med deres erfaringer på området.

Udsendt i september 1994. 44 sider. Pris: Kr. 50,00.

Alle priser er incl. moms og forsendelse.

Ovennævnte særnumre i papirudgave kan købes hos:
Villy Danielsen, Vodroffsvej 27, 1.tv., 1900 Frederiksberg C.

Kildeoversigten på diskette (DIS-Kilde) kan købes hos:
Arne Julin, Hovedgaden 75, 4050 Skibby.

ISBN 87-90044-01-0