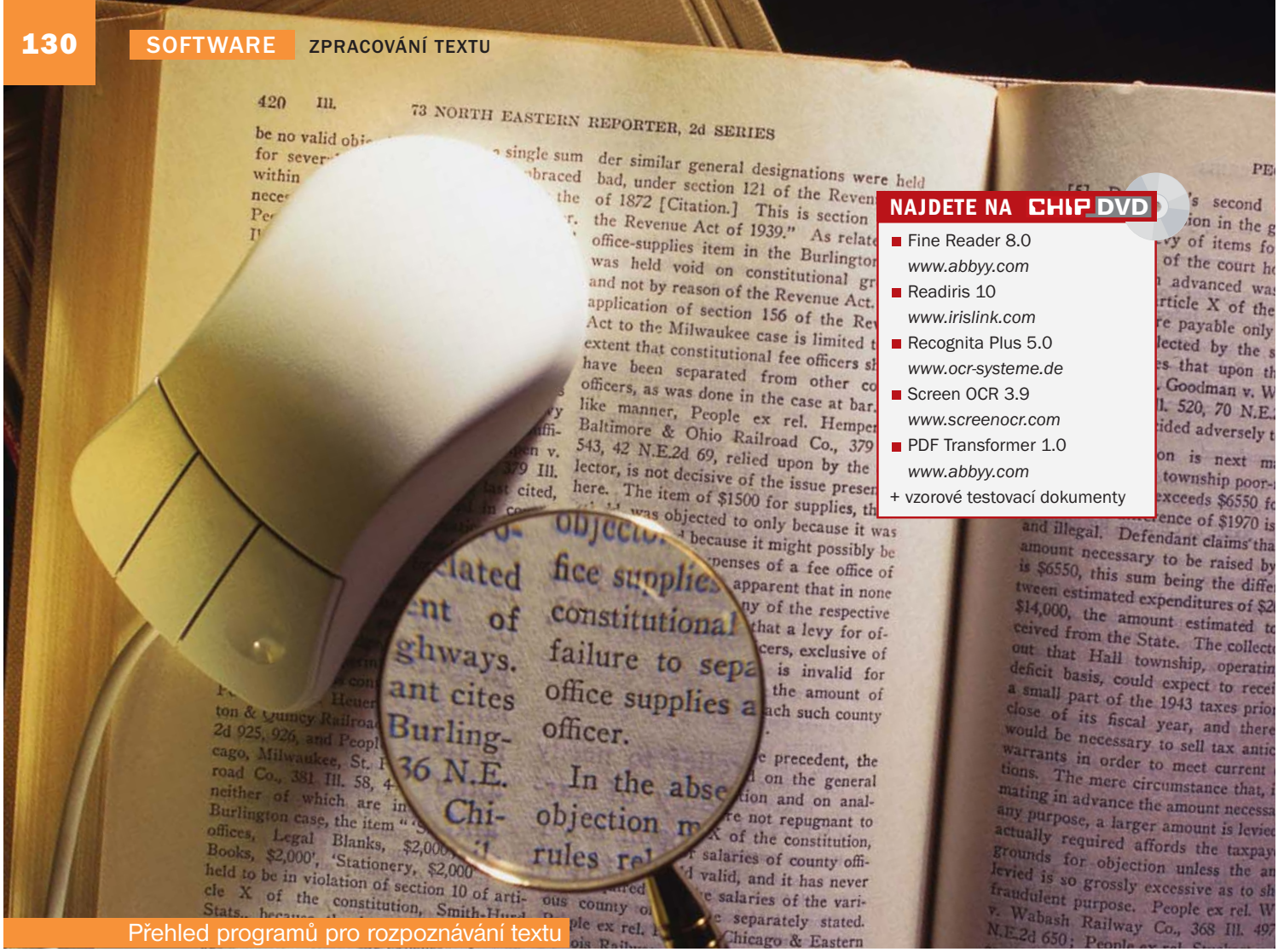


Přehled programů pro rozpoznávání textu

Když počítač čte

LEDEN 2006 CHIP DVD TÉMA MĚSÍCE

NAJDETE NA CHIP DVD

- Fine Reader 8.0
  www.abbyy.com
- Readiris 10
  www.irislink.com
- Recognita Plus 5.0
  www.ocr-systeme.de
- Screen OCR 3.9
  www.screenocr.com
- PDF Transformer 1.0
  www.abbyy.com

+ vzorové testovací dokumenty

Máte digitální fotoaparát nebo skener? Potřebujete přepsat několikastránkový text do počítače? Ukážeme vám, jak naučit počítač číst. Vyzkoušeli jsme některé programy, které převádějí text do počítače, a poradíme vám triky pro co nejlepší využití těchto aplikací.

Text: Vratislav Klega, *vratislav.klega@vogelburda.cz*

Čtení textů počítačem není vůbec nová věc. Vždyť první standard pro optické rozpoznávání textu (OCR) vznikl již v roce 1966, kdy bylo standardizováno i první písmo OCR-A, určené především pro strojové čtení. Rozpoznávání písem se časem propracovalo z velkých strojů až na běžné počítače, ovšem nejlepší systémy si našly své pravé místo zejména u hromadného zpracování dat.

Nabídka programů OCR na trhu není právě největší. Je to způsobeno tím, že své pevné pozice si zajistili největší výrobci, kteří si mohou dovolit investovat prostředky do náročných inteligentních algoritmů, a také tím, že k velké spoustě skenerů se automaticky přikládá odlehčená OEM verze některého z profesionálních nástrojů, takže menší výrobci postupně zkrachovali.

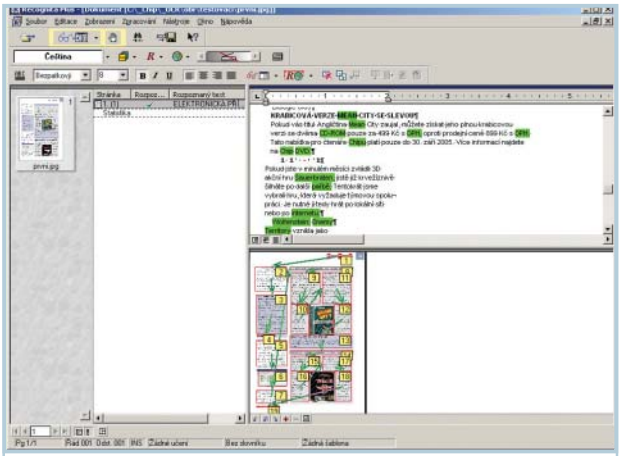
Podívejme se na problematiku strojového zpracování dokumentu. V prvním kroku musí OCR program odlišit text od grafiky, se kterou se dále nepracuje. V dalším kroku dojde k rozpoznávání řádek a k jejich náklonu oproti ideálnímu vodorovnému textu; sleduje se i vzdálenost mezi řádky. Dále přijdou na řadu slova, která nejsou nic jiného než posloupnosti znaků končící prázdným znakem. Rozpoznávání jednotlivých písmen už není tak jednoduché. V praxi se můžete setkat se dvěma druhy písma: s písmem s neproporcionální konstantní šířkou, které lze rozpoznávat snadněji, a s proporcionálním písmem, u nějž různá šířka znaků odpovídá jejich stavbě.

Počítač používá princip velice podobný lidskému rozpoznávání písmen. Analyzuje průběhy křivek, poměry úhlů, kruhové tvary, proporce příčných a podélných čar a otevřené křivky. Některá písmena jsou více jednoznačná a pro jejich rozpoznání stačí pouze několik základních entit, některá písmena jsou si více podobná, a zde tedy musí být rozlišování zpřesněno. Díky této metodě rozpoznávání vůbec nezáleží na velikosti písma.

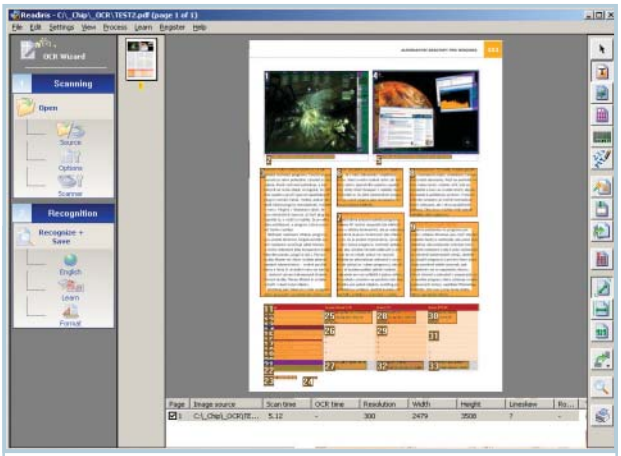
Pouhé rozpoznávání textu je často velice účinné, chybovost se pohybuje kolem 1 %, →

CO JE OCR?

OCR (Optical Character Recognition) je metoda optického čtení dokumentů vytvořených v papírové podobě a následný převod do textové elektronické podoby.



Recognita Plus: učit se, učit se, učit se.



Readiris 10: čtení zamčených PDF dokumentů je o poznání lepší.

i tak však jde o nezanedbatelné číslo, protože prakticky každý druhý řádek na stránce obsahuje chybu. Kvalitnější programy proto používají kontrolu jazyka – porovnávají čtené slovo se slovníkem a dotváří jeho možný tvar. Pro český jazyk však tato metoda není nejvhodnější – především kvůli skloňování slov může snadno dojít k omylu. Díky slovníkům může chybovost klesnout pod 0,1 %, což už je příznivější výsledek.

OCR programy

FineReader 8.0

Výkony hodné ruského bohatýra

Etalonem mezi OCR programy je FineReader od ruské společnosti ABBYY. Při testování jsme začali nejkvalitnějším dokumentem, který vznikl z dokumentu PDF. Program jej zpracoval téměř bezchybně. U 20 znaků si nebyl jistý, ale všechny je interpretoval správně. Třímegapixelová fotografie mu již činila problémy. Kvůli tomu, že text zasahoval k okrajům, nechtěl program tento text akceptovat. Výsledkem bylo 68 špatně rozpoznaných znaků, což tvoří cca 1,5% chybovost. Lepších výsledků jsme dosáhli při použití dvoumegapixelové upravené fotografie s okraji. Vyšší kontrast měl za následek, že program udělal při převodu pouze 43 chyb v rozpoznávání, čímž se chybovost pohybuje na hranici 1 %. Převod faxu byl lepší, než jsme očekávali, přesto nepoužitelný. Porozumět textu nebylo možné. Zamčený dokument PDF program bez znalosti hesla neotevře. Nezamčené dokumenty přečte bez problému, rozpoznání je na vysoké úrovni – nerozpoznány zůstaly pouze jiné než standardní fonty.

Co musíme v případě tohoto programu obzvláště pochválit, je jeho výstup. Ať už je stránka sebevíce komplikovanější, výstup do Wordu či do HTML stránky je velice přesný. Zachovány jsou odstavce, obrázky i tabulky.

Ovládání programu je díky českému prostředí velice jednoduché a intuitivní. Celý program je rozdělen do tří oken a napomáhají i velké ovládací prvky. Český slovník pro kontrolu pravopisu pracuje uspokojivě. Rozpoznávání češtiny v koncovkách by napomohl zejména kontrolér gramatiky, který zatím není k dispozici.

Recognita Plus 5.0

Poslušný žák bez HTML výstupu

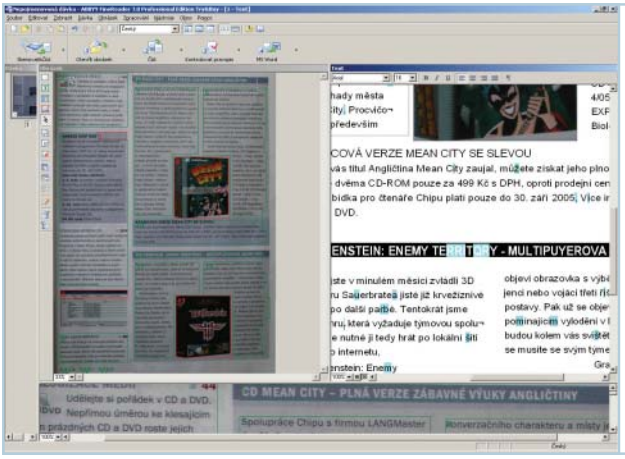
Dalším silným nástrojem v oblasti OCR je program Recognita Plus. Verze 5 je sice již několik let stará, nicméně stále patří k tomu lepšímu, co můžete mít.

Při rozpoznávání nejkvalitnější testovací stránky udělal program 17 chyb (0,4 %), ale ve velkém množství případů si nebyl jistý. Po spuštění české jazykové korekce však byly všechny chyby odstraněny. Při čtení 3Mpx fotografie byl výsledek mnohem horší. Chyb bylo skutečně

ŽÁDNÉ MAZLENÍ

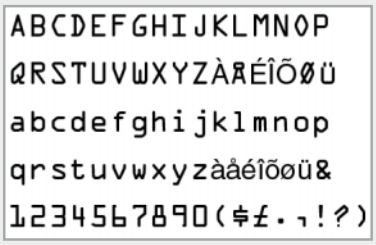
Programy jsme zkoušeli tvrdě, v žádném případě jsme se s nimi nemazlili. Každým z nich prošly desítky dokumentů, abychom otestovali skutečnou funkčnost programu. Typické dokumenty, které jsme programům při testu předkládali, jsme zařadili i na Chip DVD:

1. Stránka, která vznikla z dokumentu PDF. Je tedy zcela bez optických vad a má dostatečné rozlišení.
2. Fotografie v rozlišení 3 Mpx, u které jsme ořezali okraje, ale neprováděli jsme žádné další úpravy.
3. Fotografie v rozlišení 2 Mpx, u které jsme zvýšili jas a kontrast a provedli doostření.
4. Simulace faxu se šumem a rozmazaným písmem.
5. Dokument PDF uzamčený pro kopírování, export a tisk.

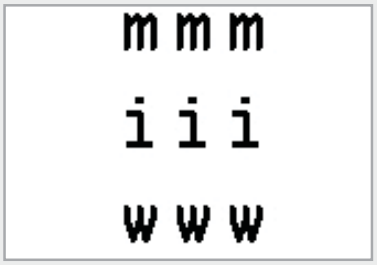


FineReader: ergonomie je na vysoké úrovni.

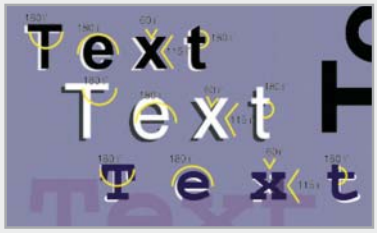
»JAK POČÍTAČ ČTE



Strojově ideální písmo OCR-A nepatří k nejvzhlednějším.



Neproporcionální: Všechna písmena jsou stejně široká.



Každé písmeno má své charakteristické obrazové vlastnosti.

Pro člověka snadný úkol, pro počítač nevyřešitelný problém.

velké množství a ani slovník si se vším neporadil, i když snížil chybovost o řád. Celkem zůstalo špatně rozpoznáno 36 slov. Chybovost se tedy pohybuje kolem 3 %. Velice podobný výsledek byl i u 2Mpx fotografie – chybovost se pohybovala kolem 2,5 %. Nekvalitní faxový dokument program odmítl opakovaně přečíst, zřejmě v obrazu žádný text neviděl. S formátem PDF si program nerozumí – neumí jej otevřít.

Jako jeden z mála OCR programů se Recognita Plus dokáže učit. Pokud nerozpozná písmeno, můžete jí vysvětlit, o které písmeno se jedná, a ona si poznatek uloží do slovníku.

Grafické zpracování výstupu je pouze průměrné. Program sice dokonale automaticky rozpozná odstavce, tabulky i rámečky, ale výstup do DOC pokulhává. Možnost uložit stránku do HTML chybí úplně.

PDF Transformer: Ovládání programu je primitivní

Readiris 10

Prolomí uzamčené PDF

Hlavně v zahraničí je hodně oblíbený program Readiris softwarové společnosti I.R.I.S. Hned na začátku musíme upozornit, že se jedná o program bez podpory češtiny, a tedy i českých znaků. Při převodu kvalitní předlohy program nezaváhal a pracoval výborně, u předloh s horší kvalitou nebyly výsledky tolik přesvědčivé. Program měl výrazné problémy s určením okrajů textu, často chybně určoval začátky odstavce a také směr textu na stránce stanovoval špatně – odstavce na sebe nenavazovaly.

Proč tedy vůbec s tímto softwarem marnit čas? Důvodem je PDF. Readiris si jako jediný ze zde uvedených programů bez jakýchkoliv problémů poradil se zamčeným dokumentem PDF – nepožadoval heslo k odblokování. Převod je zcela bezchybný včetně formátování – ovšem samozřejmě opět bez češtiny.

Další silnou stránkou programu je interaktivní učení programu, které kombinuje přečtení z obrázku a jazykový slovník. Lze tedy zjednodušeně říci, že čím více textů necháte program přečíst, tím dokonalejší bude.

PDF Transformer 1.0

Bleskem do DOC, XLS a HTML

Společnost ABBYY nabízí uživatelům ještě jeden produkt. Nazývá se PDF Transformer, a jak už název napovídá, jedná se o zjednodušený produkt, který převádí dokumenty z PDF. Vstupním souborem je nezamčený PDF dokument a výstupem může být dokument Word, Excel, HTML nebo prostý text. Kvalita převodu je na vynikající úrovni. Program se nezalekne formátování do sloupců, tabulek, infoboxů ani obrázků různých tvarů.

Uživatelské prostředí je vyřešeno průvodcem.

Screen OCR 3.9

Čtení z obrazovky

Existují speciální programy, které sice fungují na principu OCR, ale jejich primární určení a základ fungování jsou zcela jiné. Využijete je třeba tehdy, pokud potřebujete sejmout text z obrazovky, tedy například udělat si přehled o disku a seznam souborů uložit do přehledné tabulky s možností vyhledávání.

Řešení nabízí program Screen OCR. Ten dokáže přečíst cokoli, co je na obrazovce počítače. V praxi to funguje tak, že podrží-

SPRÁVNÉ PÍSMO

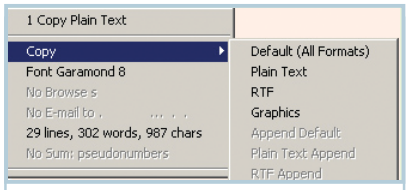
Chcete tisknout dokumenty tak, aby je OCR programy přečetly bez zaváhání? Pak použijte speciální písmo určené pro OCR. Jednotlivé znaky jsou odlišeny do té míry, aby nebyla možná jejich záměna. Díky specifickým tvarům je program dokáže rozpoznat mnohem rychleji a jednodušeji. Vhodné fonty najdete například na *www.developservice.cz.*

te zkratkovou klávesu a ukazatel myši se změní na kříž, kterým uděláte na obrazovce obdélníkový výřez. Po uvolnění se zobrazí kontextové menu, ze kterého vyberete, jakou formou chcete rozpoznaný text vyexportovat.

Program si při instalaci načte všechna písma v systému, podle nichž pak porovnává vybraný text. Tato metoda je velice účinná. Tímto způsobem však není možné rozpoznávat text na obrazovce obsažený v obrázcích.

Gramotné programy

Programy OCR jsou na dosti vysokém stupni vývoje a jejich další zlepšování probíhá spíše v oblasti uživatelského luxusu, dokonalejšího výstupu dat či podpory nových formátů a technologií. Ne každý uživatel také takový program využije (jako například u vypalování nebo u převodu videa). Přesto to tak nemusí být trvale. V nejbližší budoucnosti se předpokládá velký přechod od kapesních počítačů k mobilním, kde klávesnici nahradí dotykové displeje, které budou sloužit pro psaní textu. Podpůrné programy pak budou muset umět bezproblémově rozpoznávat ručně psaný text.



Screen OCR je jednoduchý, ale užitečný.

Co se týká testovaných programů, je patrné, že práci jim dokáže pěkně znepříjemnit naskenovaný nebo nafocený dokument s optickými vadami, který je však pro člověka běžně čitelný. Nejlépe si při čtení podle očekávání vedl program FineReader, který dokáže kvalitní dokument po jazykové korekci přečíst se 100% jistotou. Ostatní produkty tak univerzální nejsou, lze je však použít na specifické úkoly. Převody z PDF dokumentů jsou vcelku bezproblémové včetně formátování tabulek i zachování aktivních odkazů. Redakčním tipem v této oblasti je PDF Transformer, který si nás získal i díky příznivé ceně pod dva tisíce korun. ■ ■ ■

»JAK FOTIT DOKUMENTY?

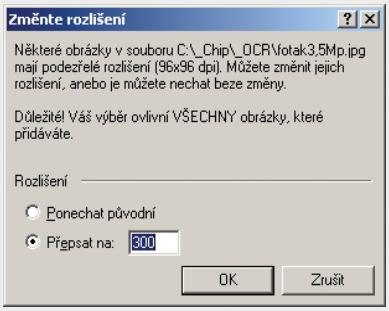
Pro převod papírového dokumentu do digitální podoby existují dva základní způsoby. První je použití skeneru. Tato cesta je velice jednoduchá a přináší kvalitní výsledky. OCR programy dokáží většinou přímo komunikovat se skenerem a správně nastavit optimální parametry skenu. Druhá cesta je pomocí digitálního fotoaparátu, ta však vyžaduje více pozornosti. Náš test prokázal, že i se dvěma megapixely se dá získat dokument vhodný k převodu OCR. Je však potřeba dodržet několik základních zásad:

- Bez blesku. Protože se většinou fotí z velké blízkosti, není možné použít blesk – část dokumentu by byla přesvětlená a část by zůstala tmavá. Lepším řešením je použít lampičku, která má rovnoměrné osvětlení, nebo využít denní světlo, pokud je ho dostatek.
- Kolmo. Fotoaparát držte vždy zcela kolmo na plochu papíru.

Krátce. Doba otevření závěrky by neměla překročit 1/50 sekundy, aby byl obrázek co nejostřejší. Pokud je světla dostatek, není třeba používat stativ.

- Se zoomem. Není vhodné fotografovat z nejkratší vzdálenosti bez zoomu, ale raději s lehkým přiblížením, cca 1,5× – 2×. Důvod: Při nejkratším ohnisku trpí většina fotoaparátů soudkovitostí, která snižuje schopnost čtení. Při mírném přiblížení tento neduh vymizí.
- Ostře. Využijte možnost ostření fotografií přímo ve fotoaparátu.
- S okraji. Programy OCR často ignorují text zcela na kraji obrázku, proto foťte i alespoň centimetrové okolí dokumentu.

Uvedené rady lze většinou následně „dohnat“ i ve fotoeditoru. V něm byste měli provést i změnu rozlišení ze 72 dpi, se kterým pracují fotoaparáty, na OCR programy vyžadovaných 300 dpi.



Programy většinou požadují rozlišení 300 dpi.