Počítač čte z papíru

Před více než rokem jsme v Chipu zveřejnili recenzi páté verze programu FineReader. Dnes se podíváme na nejnovější verzi - 6.0. Ve spektru produktů OCR patří do horní části, určené pro profesionální, respektive podnikové použití. V nové verzi je největším vylepšením schopnost přenést do editovatelného tvaru soubory ve formátu PDF. Tato vlastnost je důležitá zejména pro překladatele, neboť pokud je předloha v tomto formátu, znamená to, že překlad (eventuálně jiné zpracování) musí vycházet z vytištěné podoby. Tím se ztrácí formátování, práce se prodlužuje a není možné použít specializované překladatelské SW systémy (TRADOS, DejaVue).

FineReader OCR Corporate Edition 6.0

Recenzovaný produkt jsme měli k dispozici v jeho nejvyšší verzi - Corporate Edition. Prakticky však byly zkoušeny funkce, které jsou společné s verzí Professional. Netestovali jsme funkce síťové instalace - distribuované zpracování v síti a skupinovou spolupráci nad vlastními jazyky a slovníky. Pro testování jsme využili následující konfiguraci (vzhledem k uváděným časům): stolní počítač s procesorem Intel Celeron 1 GHz, 256 MB RAM, HD 20 GB a skener Mustek 12000 SP Plus, připojený přes SCSI rozhraní.

Instalace

Instalace probíhá celkem bez problémů. Program je chráněn klíčovou disketou. Pro případ, že na počítači není disketová jednotka, je k dispozici odblokování prostřednictvím internetu. Pro počítače, které nemají ani přístup k internetu, lze získat odblokovací sekvenci e-mailem nebo telefonicky.

Velikost prostoru, který instalace na pevném disku zabírá, závisí kromě jiného na počtu nainstalovaných jazyků rozpoznávání. Při 23 jazycích instalace zabírala 94 MB. Soubory podporující rozpoznávání jednoho jazyka jsou různě náročné: počínaje esperantem, které zabere pouhý 1 KB, přes 742 KB bulharštiny až po nejnáročnější finštinu, vyžadující 5,57 MB.

Přehled funkcí

Vzhledem k rozsahu recenze nemohou být probírány všechny funkce. Zaměříme se tedy jen na novinky oproti verzi 5.0 a na funkce důležité z hlediska uživatele.

Rozšířená paleta vstupních a výstupních formátů. Vedle přímého skenování jsou nyní jako vstup do OCR k dispozici formáty BMP, PCX a DCX, JPEG, PNG, TIFF a PDF. Znamená to, že rozpoznávání se nemusí odehrávat na počítači připojeném ke skeneru. Předlohy lze naskenovat kdekoliv a k rozpoznání odeslat prostřednictvím intranetu/internetu na mnohem výkonnější počítač.

Jinou překážku ve zpracování doposud představoval formát PDF. PDF se dá někdy převést s obtížemi do RTF k dalšímu zpracování, v horším případě se dá pouze vytisknout a nic víc. Schopnost převést PDF do množství různých editovatelných formátů sama o sobě představuje velmi žádaný nástroj na zpracování textu. Vzhledem k tomu, že FineReader zachovává formátování předlohy včetně obrázků, odpadá finančně náročné zpracování upraveného textu. Výstupní formáty: MS Word/Excel od verze 95 až po 2002, RTF, TXT a Unicode TXT, HTML a Unicode HTML, DBF a CSV, PDF.

Zachování formátu stránky. Rozpoznávání zachovává vzhled a celkovou formu rozpoznávaného dokumentu, včetně obtékajícího textu, textu svisle orientovaného, sloupců a tabulek, obrázků nepravidelného tvaru a měnícího se typu písma. Pokud definujeme vlastní bloky rozpoznávání, došlo oproti minulé verzi k významnému rozšíření: oblast lze definovat jako logický součet nebo rozdíl více pravoúhlých oblastí.

Rozpoznávání vícejazyčných dokumentů. Buď lze pro každý blok textu zadat jiný jazyk např. pro levý sloupec anglický, pro pravý sloupec český, nebo lze celý dokument deklarovat jako ve více jazycích, např. německý/český.

Integrace s Průzkumníkem Windows. Obrazové soubory (PDF, TIF atd.) a dávky FineReaderu mohou být nyní otevírány přímo z Průzkumníka. Nabídka otevíraná pravým tlačítkem myši má přidánu odpovídající položku.

Podpora vlastních slovníků a vytváření nových jazyků. Funkce nacházející se jen v Corporate Edition, jež dovoluje vytvořit např. nový jazyk, skládající se z češtiny a uvnitř bloků se nacházejících symbolů švabachu, které se používají v některých vědeckých publikacích k označování vektorů a složitějších matematických útvarů.

Odstraňování drobných vad sejmuté předlohy. Kromě předloh snímaných z kvalitního papíru se zejména u starších kopií vyskytují "chlupy" nebo "prach". Jedná se o nežádoucí zdroj informací při rozpoznávání, zejména pokud se vyskytne poblíže písmen náchylných k zaměňování: "a" - "s", "o" - "e", "rn" - "m" apod. Funkce Odstranit skvrny v obrázku v menu Obrázek je schopná zčásti tyto vady odstranit.

Práce s naskenovaným dokumentem. Hlavní okno produktu je přehledně uspořádáno. K dispozici jsou jednotlivé miniatury stránek, umožňující práci na přeskáčku (viz obr. 1), a celkem tři pohledy na dokument, v každém z nich volitelné zvětšení (ikony lupy). Pokud sejmutý dokument nemá příliš komplikovanou strukturu, program si jej po naskenování rozdělí na jednotlivé bloky sám, včetně rozpoznání druhu (text, tabulka, obrázek). V takovémto případě uživatel musí pouze zkontrolovat, popřípadě vyřadit bloky, které nepotřebují rozpoznání - číslování stránek, opakující se záhlaví či zápatí atp. V této fázi můžeme vyřadit i další nepotřebné bloky - obrázky či jiné. Pouze v případě, že z naskenované stránky potřebujeme rozpoznat jen několik bloků, nastupuje ruční definování pomocí nástrojů (viz obr. 1). Tento postup nastoupí také v případě natolik složitého dokumentu, že automatické dělení nedává uspokojivé výsledky. V průběhu naší recenze byla vždy rychlejší první metoda, tj. automatické rozdělení s následným odstraněním nepotřebných bloků.

Kontrola a úpravy rozpoznaného dokumentu. V rozpoznaném textu jsou nerozpoznané znaky a slova nenalezená ve slovníku daného jazyka vyznačeny odlišnými barvami. Při kontrole pracujeme s korektorem klasickým způsobem - buď korektor nabízí možnosti, nebo lze slovo přidat do slovníku. (Počet korektorů viz níže.) Pro úpravy textu před odesláním do cílového formátu (DOC, RTF, HTML a dalších) jsou k dispozici funkce známé z většiny editorů a vyvolávané ikonami obdobnými jako u MS Office - typ a velikost písma, tučné/podtržené/kurziva, horní/dolní index a zarovnání. Zejména není-li po ruce vhodný editor HTML, jsou tyto funkce velmi výhodné.

Šablony. Při práci s větším množstvím stránek se stejným, ale nestandardním uspořádáním (není tedy vhodné používat automatické uspořádání) lze vytvořit šablonu a tu uložit pod jménem. Při dalším snímání stejného typu stránky lze uloženou šablonu znovu vyvolat.

Rozpoznávání s učením masky. Aplikace nemá problémy s většinou písem. Pokud by však uživatel chtěl rozpoznávat netypické/ozdobné písmo, má k dispozici možnost rozpoznávání s učením. Používat ho lze například v případech výskytu symbolů v textu. Dalším případem je rozpoznávání velkého počtu stránek vytištěných nekvalitním písmem stejného řezu a velikosti. Jen v těchto případech se čas vynaložený na naučení uživatelské masky vyplatí.

Rozpoznávané jazyky. Aplikace podporuje ve verzi EU celkem 122 jazyků, ve verzi Cyrillic 177, ve verzi FineReader 5.0 Home Edition pouze 19. Podporu korektoru má 34 jazyků (v Home jen 19). Otestovány byly - pochopitelně - jen některé evropské jazyky.

FormFiller. Jedná se o zcela samostatnou aplikaci, přidanou k oběma verzím zdarma (Corporate i Professional Edition). Využívá snímání předtištěných nebo z internetu či intranetu stažených formulářů (formáty BMP, TIF, JPG, PCX, PNG, DCX). Může sloužit dvěma účelům: pro vyplňování a následný tisk do předtištěných formulářů, nebo pro kompletní tisk včetně rastru formuláře tento režim je však vhodný pro černobílé formuláře. Sejmutí barevného formuláře banky sice proběhlo v pořádku, avšak barevný podklad má za následek několik nedostatků:

1) velký objem souboru - jedna stránka A4 (viz obr. 3) měla 7,5 MB;

2) kvalita a formát reprodukce závisí na použité tiskárně.

Také tisk do typických formulářů FÚ (přeložená A3) dělá jak na laserové, tak na inkoustové tiskárně potíže. Samotné zpracování formuláře probíhá ve dvou průchodech:

(a) sejmutí na skeneru nebo načtení ze souboru;

(b) vytvoření zón pro vyplňování - to je nejsilnější stránka aplikace. Dovede totiž danou zónu rozdělit podle počtu vyplňovaných míst. Známý problém při pokusu o vyplnění formulářů FÚ na psacím stroji nesoudělnost roztečí psacího stroje a políček ve formuláři - je tím odstraněn. Vedle textových polí s šesti různými možnostmi zarovnání jsou i pole typu DATUM a SEZNAM. Poslední jmenovaný typ je obdobou funkce v databázových aplikacích - při vyplňování pouze vybíráme z předem definované množiny. V aplikaci lze slučovat formulář se souborem CSV, který byl připraven předem s využitím veškerých funkcí Excelu.

Lokalizace

Lokalizace představuje zvláštní kapitolu plnou nejrůznějších překvapení. Počet chybných překladů v uživatelském rozhraní se oproti minulé verzi snížil, některé však zůstaly: nabídka Témy pomoci, matricová tiskárna (namísto mozaikové) nebo Otevrit s FineReaderem. Nedůslednosti a chyby, vyskytující se převážně v nápovědě, lze rozdělit do několika skupin:

Chyby, které by odstranil kterýkoliv SW korektor češtiny: Ako: (Vytvořit nový blok); slovenština se vůbec překladatelům často plete s češtinou - Užívateľské jazyky a skupiny jazykov. Pikantní je, že toto

záhlaví odkazuje na stejnojmennou, správně přeloženou stránku. Sem patří i evidentní překlepy (Nnastavení správného jasu) nebo neexistující slova (Novovytvořené pole).

Chyby, jež nenajde SW korektor, ale pouze rodilý mluvčí: k uložení uživatelských jazyků v databáze (nesprávný pád nebo předložka).

Chyba viď seznam se opakuje v nápovědě do omrzení.

V angličtině jsou ponechány i názvy jazyků se zavedeným překladem do češtiny, např. Bashkir. Přitom v seznamu jazyků rozpoznávání je přeloženo správně.

Tabulka jazyků korektoru v uživatelském rozhraní obsahuje Taliansky.

Zapomenuté texty v angličtině - Full-text Search in Recognised ..., Latvian včetně hypertextově připojené popisky. Popiska švédštiny rovněž v angličtině.

O překladech Microsoftu lze diskutovat, ale u velké většiny aplikací jiných dodavatelů platí, že funkce jednou Microsoftem přeložené se již nepřejmenovávají, protože většina uživatelů je prostě na tyto překlady zvyklá. FineReader se takto neomezuje, a tudíž se můžeme setkat s Vystřihnout výběr a vložit ho do schránky = CTRL+X. Někdo jiný zřejmě překládal nápovědu, jiný položky nabídek. Tam je totiž správně Vyjmout. Do stejné kategorie patří Help = Pomoc (místo Nápověda). Manuál je v češtině, na rozdíl od nápovědy je přeložen korektně. Má 98 stran A5 a dostatečně podrobně popisuje práci s aplikací.

Závěr

FineReader ve verzi 6 přinesl uživatelům další velmi užitečné funkce a hlavně - díky automatickému rozpoznání druhu bloku urychlení činnosti. Schopnost rozpoznávat texty uložené v obrazových elektronických formátech, zejména PDF, TIF a JPG, nabízí možnost odděleného pracoviště rozpoznávání a skenování. Pro elektronický přenos mezi těmito pracovišti je limitujícím faktorem už jen přenosová kapacita sítě. Schopnost otevření a následného rozpoznání (tj. převedení do editovatelného formátu) dokumentů PDF s heslem uzamčeným kopírováním bude neocenitelná pro překladatele a všechny, kteří dostanou výchozí podklad v tomto formátu. Dříve totiž nezbývalo nic jiného než pracné opisování. S FineReaderem se vše velmi rychle odbude elektronicky.

Pro českého uživatele bude na produktu nejlákavější široká paleta korektorů pro veškeré jazyky států obklopujících ČR a ještě pro mnoho dalších. Také funkce pro rozeznávání znaků psacího stroje a mozaikové tiskárny (lze použít i na kvalitnější faxy) jsou velmi příjemné. Pro úplně špatné kopie, které jsou alespoň z jednoho psacího stroje s typickými slitky, lze pomocí funkce Učení při delším dokumentu zlepšit efektivitu snímání. Samostatná aplikace FormFiller na vyplňování formulářů může v organizaci, mezi jejíž administrativní povinnosti patří vyplňování četných formulářů, přinést značné časové úspory. Kvalita lokalizace nápovědy se zlepšila, i když některé těžko pochopitelné nedůslednosti a chyby zůstaly. Vzhledem k tomu, že aplikace obsahuje vlastní, velmi slušný korektor češtiny, je těžko pochopitelný výskyt chyb, které by zachytil i korektor podstatně horší. Proč si tak velká organizace, jako je ABBYY, nenajala jako vedoucího projektu lokalizace rodilého Čecha s dostatečnou praxí, je nepochopitelné. Ve srovnání s velmi dobře naprogramovanou aplikací však lze tento nedostatek odpustit.

Ing. Miroslav Herold, CSc., autor@chip.cz

Chcete si vyzkoušet některé z možností programu FineReader?

V říjnovém čísle Chipu jste na přiloženém DVD mohli najít plnou verzi programu FineReader 5 Sprint CZ. Program je zcela lokalizován do češtiny.

FineReader 6.0

OCR software podporující i mimoevropské jazyky.

Vyrábí ABBYY (Rusko)

Poskytl autorizovaný distributor pro ČR Nupseso CZ, www.nupseso.cz

Ceny FineReader 5.0 Home Edition 1858 Kč, FineReader 6.0 Professional 5240 Kč; FineReader 6.0 Corporate Edition 12 495 Kč

Jak jsme testovali - konkrétní výsledky

Testovali jsme spíše problematické dokumenty - texty v méně obvyklých jazycích, strojopisné dokumenty, respektive dokumenty vytištěné na jehličkové tiskárně, což jsou tradičně slabá místa aplikací OCR.

Snímání kvalitní předlohy - knižní výtisk anglicko-českého slovníku. Délka dokumentu byla 2558 slov, 17 889 znaků včetně mezer. Výsledný elektronický dokument, včetně kontroly pravopisu a opravy

některých formátů (tučné znaky se občas sejmuly jako normální), trval 61 minut, čili přibližně 42 slov za minutu. Průměrná znaková chybovost byla lepší než 0,15 %.

Snímání vícejazyčných sloupcových textů - čtyři stránky A4 z dvoujazyčného slovníku, 2295 slov, na každé stránce 2 dvojice sloupců. Nástroje pro segmentování sejmutého obrazu umožňují rozdělení na sloupce a každému přiřazení jiného jazyka - oba jazyky mají svůj korektor, rozpoznávání se tím zrychluje a zpřesňuje. Trvalo 74 minut, čili přibližně 31 slov/min.

Snímání neobvyklých jazyků - při testování této vlastnosti byly zjištěny nedostatky u některých jazyků, jmenovitě u švédštiny a španělštiny. Instalace nabízející výběr jazyků obsahuje chybu (viz obr. 4), název obou jazyků obsahuje špatně kódovaný první znak. Instalační program žádnou chybu nenahlásí, přesto se však uvedené jazyky neobjeví v okénku nabízejícím volbu jazyka rozpoznávání. Ostatní méně obvyklé jazyky, které byly vyzkoušeny: bulharština, dánština, finština, holandština, maďarština, polština, portugalština, rumunština, ukrajinština. U všech korektor fungoval, sejmutí jedné stránky A4 včetně zahřívání skeneru, nastavení jazyka a korektury bylo pod čtyři minuty.

Snímání podkladů z jehličkové tiskárny EPSON LQ 850 v draft modu - čeština, jedna strana A4, velikost 10 bodů, bez problémů, komplet naskenování včetně korektury 2 minuty 10 sekund.

Strojopisný text - jako podklad byla zvolena stará kopie, na nekontrastním průklepovém papíru, písmo perlička, hustě psané. 704 slov, 4750 úhozů. Komplet včetně korektury 16 minut. V tom započteno i upozornění aplikace na vhodnost změnit nastavení kontrastu na skeneru a opakované skenování.

Úpravy rozpoznaného dokumentu - je nutné postupovat opatrně. Je-li ve slově více špatně rozeznaných znaků, musíme opravit všechny najednou. Když totiž opravíme první zvýrazněnou chybu a opravu potvrdíme, chyba třeba až na konci slova již zůstává a musíme se k ní vracet manuálně.

Navrhované úpravy v češtině - obsahují některá slova, jejichž původ je podivný. Např. jedna z nabídek pro nahrazení neznámého jména Oracle zněla Goralce, nerozeznané "p" na začátku slova "pracující" vyvolalo návrhy "vracující", "makující" a další rádobyzábavná slova. Korektor také označil jako neznámé slovo "olupovat".

Přidávání slov do slovníku - velmi dobře zpracováno pro angličtinu; program se táže na slovní druh, způsob psaní velkého písmene (viz obr. 6). Do českého slovníku přidává pouze tvary, bez jakýchkoliv dotazů - pro tvorbu slovníku používá zřejmě odlišných algoritmů.

Otevření naskenovaných souborů jinou aplikací - aplikace si ukládá naskenované stránky do souborů s příponou TIF. Používá však nestandardní metodu kódování, takže soubory nelze otevřít ani v aplikaci PaintShop Pro, ani v Adobe Photoshop. Zdařilo se otevřít pouze v MS Imaging a IrfanView.