ABBYY FormReader 4.1

Čtenář formulářů

Produkt popisovaný v této recenzi má své potenciální uživatele hlavně ve sféře velkých podniků, firem, kde se zpracovávají data z formulářů ve velkém, zejména však v oblasti státní správy. Nejde tedy jen o pouhé používání nainstalovaného krabicového produktu - hlavním problémem je spíše efektivní zavedení systému.

K přesnější specifikaci závěrů této recenze uvádím hned na začátku konfiguraci počítače, na němž probíhalo testování produktu: stolní počítač CPU Intel Celeron 1 GHz, 256 MB RAM, 20GB HD, skener Mustek 12000 SP Plus připojený přes SCSI rozhraní. Pro normální pracovní nasazení produktu ABBYY FormReader (dále jen FormReader) je vhodnější produkční skener alespoň střední kategorie vybavený automatickým podavačem. FormReader je určen pro zpracování cca 1000 formulářů jednou pracovní stanicí v jedné pracovní směně. Produkt si své budoucí uživatele nalezne hlavně v institucích, ať již státního nebo soukromého sektoru, kde je zapotřebí do počítačového systému zadávat velké objemy dat, která nejsou k dispozici v elektronické formě a jejichž zdroj je mimo organizaci. Typickým příkladem je vyhodnocování zkušebních testů ve školství nebo různých dotazníků ve státní správě, dotazníků o průzkumu trhu, načítání dat z provozních záznamů v průmyslových podnicích, dat z předtištěných smluv pro pojišťovny apod.

Instalace a základní pojmy

Instalace proběhla zcela bez problémů. Požadavky na RAM a prostor na HD jsou celkem skromné, takže je splní prakticky každý běžný počítač. Při skutečném nasazení zabírá program 19,8 MB nad operačním systémem. Plná funkčnost aplikace (možnost exportu rozpoznaných dat) je zajištěna ochranným klíčem zasunutým do USB portu. Při instalaci je potřeba dodržet pořadí - napřed provést instalaci softwaru a ovladače a po restartu počítače zasunout ochranný klíč. Jinak se může stát, že aplikace nebude pracovat správně.

Před vlastním popisem produktu a jeho funkcí si musíme ozřejmit některé pojmy, s nimiž se v dalším budeme zabývat:

Respondent - tímto souhrnným názvem budeme nazývat všechny "vnější" uživatele systému, kteří zadávají data do formulářů, ať již jde o účastníky nějaké ankety, klienty banky vyplňující příkazy k úhradě či návštěvníky velké knihovny vyplňující vstupní listy čtenáře.

Formulář - i když se každý z nás již nesčetněkrát s nejrůznějšími formuláři setkal, neuškodí podívat se na ně podrobněji. Pochopitelně se budeme zabývat pouze formuláři pro strojové snímání, které po vyplnění respondentem do počítačového systému zadává operátor.

Barevné formuláře - z hlediska konečného zpracování jsou nejvýhodnější, protože barevné pozadí definující pole vyplňovaná respondentem při skenování zmizí (je odfiltrováno) a v systému se dále zpracovává jen relevantní datová informace. Tím se do značné míry usnadní proces rozpoznávání a vyloučí některá omezení, zatěžující následující typy formulářů. Zejména není tak výrazná náchylnost k chybám rozpoznání vyplývající z přesahu jednoho znaku do oblasti znaku druhého, která je u čárových formulářů výrazná. Pokud je v barvě pozadí i vysvětlující text, objem naskenovaného obrazu formuláře je výrazně nižší.

Nevýhodou je složitá tvorba takového formuláře. Musí se navrhovat v grafickém editoru a tisknout na profesionální tiskárně, která je schopna zajistit barevnou stálost všech formulářů. Mějme na paměti, že množství formulářů může jít i do mnoha tisíc - v takovém případě je profesionální tisk výhodou. Pokud je vysvětlující text v barvě pozadí, snižuje se jeho čitelnost, a tím se zvyšuje možnost chyby respondenta.

Podtřídou barevných formulářů jsou formuláře šedé, které lze připravovat i s jednodušším vybavením.

Rastrové černobílé formuláře - barvu pozadí a ohraničení políček zde nahrazuje rastr z teček menších než 0,1 mm. Pozadí nemizí ihned při skenování, ale odstraní se až funkcí Despeckle. Podobně jako u barevných formulářů zde přesah znaku z jednoho pole do druhého nepředstavuje tak velký problém. Nevýhodou je nutnost psát při vyplňování tečky natolik tučně, aby nebyly odstraněny spolu s rastrem. Podtřídou jsou formuláře s rastrovým ohraničením polí na bílém pozadí. Příkladem mohou být některé předtištěné obálky s tímto způsobem označení místa k vyplnění PSČ.

Čárové černobílé formuláře - pro jejich přípravu lze použít takřka každý textový editor. Hlavní nevýhodou však je, že při nedbalém vyplnění, kdy dochází k přesahu znaku z jednoho políčka do

druhého, se snižuje kvalita rozpoznávání. To se dá vyřešit tiskem rámečků formuláře v odstínech šedé, takže nejsou při skenování snímány. Vzory takových formulářů jsou na instalačním CD a vytvoření formuláře nepředstavuje pro běžného uživatele programu MS Word větší problém.

Podtříd je hned několik: text nad linkou, text v rámečku, písmena v oddělených rámečcích, písmena v rámečcích, text nad hřebenovou šablonou a text v rámečku s hřebenovou šablonou. Rozbor výhod a nevýhod těchto podtříd je mimo rámec této recenze.

Ke skenování je nejlepší použít černobílý skener s barevnou lampou nebo barevný skener se softwarem na odfiltrování barvy. Z prostředí FormReaderu se dají přímo ovládat skenery připojené přes rozhraní TWAIN.

Pole pro vyplnění a automatické zpracování - část formuláře obsahující shromažďované informace. Mohou to být textová pole, zaškrtávací políčka, čárové kódy, obrazová pole (např. podpis) nebo přepínače (skupina polí, z nichž smí být vyplněno jen jedno).

Vysvětlující informace - jakékoliv textové či grafické informace ve formuláři, které nejsou předmětem rozpoznávání.

Referenční body - jsou zapotřebí k určení polohy polí, kompenzaci eventuálního natočení formuláře při snímání a v případě snímání různých typů formulářů k identifikaci typu formuláře.

Dávka - veškeré zpracování ve FormReaderu se odehrává v dávkách. Každá dávka má své vlastní nastavení a musí obsahovat tyto složky: šablony formuláře, souhrn datových typů, obrazy, výsledky procesu rozpoznávání.

Šablona formuláře - obsahuje obraz prázdného formuláře, popis rozmístění, velikosti a atributy jednotlivých polí formuláře. Dále je zde soubor instrukcí vztahujících se k rozpoznávání, ověřování a exportu dat.

Souhrn datových typů - podchycuje typy dat použitých v šabloně. FormReader již obsahuje předem definované souhrny pro angličtinu, ruštinu a mezi dalšími jazyky i pro češtinu a slovenštinu, kde jsou podchyceny nejčastější formáty adresy, data a čísel.

Pracovní postup

Postup má dvě základní etapy: nastavení aplikace a vlastní práci, tj. naskenování určitého množství vyplněných formulářů. Ještě před nastavením je nutné ve vhodném textovém procesoru nebo v některém z grafických programů vytvořit odpovídající formulář. Pokud má obsahovat velké množství zaškrtávacích políček, stojí za úvahu použití vektorového kreslicího programu, který umožňuje vytváření dvourozměrných polí (čtverečků nebo kroužků). Druhou variantou je naskenování nějakého stávajícího formuláře, který splňuje požadavky na automatizované zpracování. Stejně je však nutné doplnit referenční prvky, takže tento postup asi nebude příliš častý. Možné a výhodné je i využití služeb profesionálních firem, které navrhnou design formuláře a odladí elektronickou šablonu.

Základem úspěšného zvládnutí dávky je vytvoření kvalitní šablony, které začíná naskenováním formuláře nebo jeho načtením ze souboru. Editor šablony nabízí velmi širokou paletu parametrů, pomocí nichž lze vytvářet vícestránkové formuláře či formuláře, které jsou částečně vyplněny rukou a částečně tištěny na laserové či jehličkové tiskárně. Velmi důležitou vlastností jsou vestavěné kontroly jednotlivých vstupních polí. K tomu jsou k dispozici jednak regulární výrazy, jednak výrazy logické. Logické výrazy mohou buď vylučovat nesprávně vyplněný formulář, nebo mohou jen upozornit operátora na to, že údaj ve vstupním poli nesplňuje nastavená kritéria.

Jiným druhem kontroly je porovnání se slovníkem, tj. souborem. Má-li například starosta obce k dispozici databázi trvale hlášených občanů, mohou se při sledování objednávek na odvoz komunálního odpadu konfrontovat zadávaná příjmení a jména a opět jako u chyby hodnoty lze buď záznam vyřadit, generovat upozornění operátorovi, nebo do daného pole nasadit odpovídající obsah (např. "přespolní").

Užitečná je i možnost kontroly orientace stránky ve skeneru. Pokud na šabloně zadáme pět referenčních bodů, lze zabránit chybě obsluhy, která založí formulář "hlavou dolů". Tím je proces odolnější proti chybám a lze používat méně kvalifikovanou obsluhu. K tomu přispívá i to, že rozpoznávání lze provádět buď ihned po naskenování, nebo lze nejdříve vše naskenovat (méně kvalifikovaný operátor) a potom spustit proces rozpoznávání (operátor s vyšší kvalifikací a zkušeností, zejména je-li povolen proces učení se písma). Referenční body zároveň identifikují typ formuláře, takže v jedné dávce je možné zpracovávat různé druhy formulářů, FormReader sám přiřadí příslušnou elektronickou šablonu a vyexportuje data podle nastavení v šabloně.

Skenování formulářů lze tedy oddělit od zpracování, takže je může vykonávat prakticky každý zaškolený pracovník, respektive brigádník. Při dobře naprogramovaných kontrolách zůstane na operátora obsluhujícího proces rozpoznávání jen malé procento ručních zásahů. FormReader umí komunikovat s jinými aplikacemi (soubor EXE nebo knihovna DLL) prostřednictvím OLE, může si s nimi vyměňovat data, takže v exportovaném souboru (formáty viz popis) mohou být data již kompletně připravena pro navazující aplikace (je možný přímý export do existujícího informačního systému).

Nezanedbatelným přínosem je možnost skenovat různé formuláře v jedné dávce - je však nutné, aby byly navzájem jednoznačně rozlišeny. Nejvhodnější je použití pěti referenčních bodů s různým rozmístěním nebo čárového kódu, jehož rozpoznávání FormReader podporuje.

Součástí dodávky jsou dvě příručky. Návod pro tvorbu formulářů je jasně napsaný, formulář se mi podle něj zdařilo vytvořit na první pokus. Uživatelská příručka je zaměřena hlavně na tvorbu kontrol a řešení chybného rozpoznání. Obě příručky jsou v angličtině, což vzhledem k profesím, které je budou používat, nepovažuji za překážku - programátoři i analytici ji v dnešní době musejí ovládat alespoň pasivně.

Je-li instalován na více pracovních stanicích (je nutná plná instalace s ochranným klíčem), umožňuje FormReader zpracování jedné dávky na více počítačích, takže lze kapacitu zpracování snadno přizpůsobit počtu zpracovávaných formulářů.

Poznatky a připomínky

Jako testovací projekt byla vybrána jednoduchá úloha: Starosta kocourkovského okresu se rozhodl upravit poplatky z domácích zvířat. K tomu byl vypracován jednoduchý dotazník ve Wordu, v němž jsou použity pouze jednoduché kontroly polí (v polích udávajících počet - kontrola na zadání čísla, v poli datum - kontrola na formát datum, v poli příjmení - kontrola na výskyt pouze písmen). Při využití vzorů na CD trvá vytvoření takového formuláře v MS Wordu asi 35 minut. Bylo vytištěno a vyplněno 20 formulářů, v některých byla pole úmyslně vyplněna nesprávně (špatné datum, jedno písmeno přesahující nad druhé).

K vytváření formuláře je vhodné mít co nejvhodnější nástroj. Univerzální nástroj (MS Word) není příliš vhodný pro přípravu složitějších formulářů. Vhodnější je vektorový kreslicí program, umožňující snadnou tvorbu řad obdélníčků o daném počtu. Jednoduchý testovací formulář byl vytvořen na první pokus.

Vytváření šablony z formuláře je přímočarou záležitostí, kterou zvládne i středně zkušený pracovník. Jakmile se však přikročí k aplikaci kontrol, je nutná určitá programátorská erudice (vytváření logických nebo regulárních výrazů). K tomuto účelu obsahuje CD příklady vzorových formulářů a elektronických šablon.

U formulářů s textem v obdélníčkových rámečcích se jako nejškodlivější ukázaly případy, kdy se jednotlivé znaky dotýkají nebo přesahují rámeček. Vyplatí se věnovat více pozornosti návrhu formuláře s dodržením některých zásad - výsledek rozpoznání se pak může zlepšit i několikanásobně.

Je-li možné zadat vyplnění jen velkými písmeny, zvýší se značně správnost rozpoznání. Malá tiskací písmena někdy znamenají zhoršení (pletou se c - e, m - n).

Při vytváření šablony je nutné mít na zřeteli, že nastavené rozlišení šablony musí být shodné s rozlišením, kterým se budou skenovat všechny formuláře. Nejvhodnější a jako standardní se používá skenování s rozlišením 300 dpi v černobílém nastavení.

Pro dlouhé dávky je třeba vzít v úvahu i rychlost skeneru. Použitý skener Mustek 12000 SP Plus byl vhodný jen pro testovací dávku, pro použití v praxi by byl pomalý. Naskenování formuláře A4 včetně založení trvalo 22 vteřin (průměr z 10 formulářů skenovaných za sebou, pokud se skenuje po jednotlivých formulářích). V praktickém nasazení se doporučuje skener s ADF podavačem.

Závěr

Při pořízení tohoto produktu je třeba brát v úvahu celou řadu ekonomických parametrů. Základním údajem je hodinová cena a doba práce operátora při ručním zadávání vyplněných formulářů do systému - ta představuje jednu stranu korunové rovnice. Při analýze ručního vkládání je však nutné zahrnout i chybovost takto vkládaných dat a čas nutný na zjištění chyb a následnou opravu.

Na druhé straně rovnice jsou kromě pořizovacích nákladů softwaru i hodinové sazby různých profesí, které se na realizaci projektu podílejí. Bezesporu nejvyšší bude sazba programátora, který může vhodným návrhem formuláře značně ovlivnit efektivnost kontrol a tím snížit pravděpodobnost výskytu chyb, tedy zkrátit čas nutný na ruční úpravy. Programátor zvládající i DTP může navržený formulář realizovat včetně naprogramování kontrol. Pracovník, který řeší případy, kdy se nepodařilo automatické rozpoznání celého formuláře, je v hierarchii na druhém místě. Obsluhu skeneru s ADF podavačem je schopen po jednodenním zaškolení zvládnout i nekvalifikovaný operátor tak, že dokáže naskenovat několik set listů formulářů za hodinu. Vzhledem k nízké potřebné kvalifikaci bude jeho hodinová mzda nejnižší.

Při porovnání nákladů zcela jednoznačně vyplyne limitní počet zpracovávaných formulářů, od něhož se vyplatí použít pro vstup dat ABBYY FormReader. Vedle počitatelných aspektů je nezanedbatelná také stránka časová, protože rychleji zpracované výsledky mohou poskytovat přínosy v jiných oblastech, nepodchycených v přímých nákladech.

ABBYY FormReader je bezesporu velmi výkonným nástrojem, který v rukou zkušeného týmu může organizaci přinést nemalé časové i finanční úspory. Avšak jako u všech produktů zaměřených na

zpracování hromadných dat je jeho pořizovací cena jen částí nákladů na zavedení systému. Při zpracovávání dat z několika tisíc papírových formulářů může zajistit významnou úsporu nákladů.

Miroslav Herold

ABBYY FormReader 4.1
OCR software pro hromadné načítání vyplněných formulářů.
Vyrábí: ABBYY Software House, Moskva, Rusko.
Poskytl: NUPSESO CZ, Praha.
Cena: 67 000 Kč bez DPH.

Infotipy:
www.abbyy.com
www.nupseso.cz

Podporované jazyky pro rozpoznávání

Rozpoznávání rukopisu - angličtina, bulharština, čeština, francouzština, italština, litevština, němčina, polština, ruština, slovenština, turečtina, ukrajinština.
Rozpoznávání tištěného písma - 176 jazyků.
Slovníková podpora - 20 jazyků.

Podporované formáty obrazu:

BMP, PCX, DCX, PNG, JPEG a TIFF (komprimovaný i nekomprimovaný).
Rozpoznaný text lze ukládat v následujících formátech:
XLS, DBF, CSV a TXT v nejrůznějších kódových stránkách Windows a DOS. S použitím ODBC je možný export zpracovaných výsledků do externích databází. Pomocí OLE-Automation může uživatel vytvořit skript pro export dat do libovolné externí aplikace.

Majitel_prijmeni	Majitel_jmeno	Adr_ulice	Adr_misto	Adr_PSC	Pes_jmeno	Pes_datum	Kocka_jmeno	Kocka_datum
VONÁSEK	ALOIS	PŘÍKRÁ 13	KOCOURKOV	34511	ALÍK	10.1.1994	MICKA	5.9.2000
JEBAVÁ	ALENA	ZLATÁ 48	CHLUPÁŇ	33210	FIFI	15.04.84	ANASI	12.03.98
JEŽKOVÁ	ILONA	PRAŽSKÁ 5	PIČÍN	35514	BELA	18.3.1994		
SLON	IGNÁC	ROVNÁ 8	KOZOJEDY	30587			MYŠKA	3.8.1996
HROCH	EMANUEL	SKALNÍ 99	CHLUPÁŇ	33210	FERDA	6.3.1994		
ELEFANT	KAREL	STŘÍBRNÁ 33	PSÁŘE	39601	RALF	3.5.1998		
SYSLOVÁ	IRENA	SOUSKA 3	KOCOURKOV	34511			CELINA	22.2.1994
SÝČKOVÁ	KLÁRA	VEVEŘÍ 5	PSÁŘE	39601	DÁŠENKA	6.7.2002		
NOVÁK	INOCENC	KOZÍ 34	PIČÍN	35514				
NOVOTNÝ	FERRY	BOBKOVA 8	CHLUPÁŇ	33210	BISK	6.6.1997	FRNDA	8.3.1997
DLOUHÁ	JOLANA	HOUSKOVA 6	PSÁŘE	39601				
KRÁTKÝ	DUŠAN	HRUŠKOVÁ 1	PIČÍN	35514	DINA	18.7.1979		
KŘEČEK	PROKOP	JABLOŇOVÁ 6	KOCOURKOV	34511	SOSÁČEK	1.8.1999		
SÝKOROVÁ	ANÁLA	OLIVOVÁ 8	KOZOJEDY	30587				
ŠPAČEK	VÁCLAV	JAHODOVÁ 3	PSÁŘE	39601	ŠMUDLA	5.7.1996	KOS	7.7.2001
STRNAD	JAN	LEVÁ 4	CHLUPÁŇ	33210			KAČENKA	3.4.2001
VRÁNOVÁ	JITKA	TŘEŠŇOVÁ 9	PIČÍN	35514	DUDLÍK	6.8.1995		
KAČER	ROSTISLAV	ŠIŠKOVÁ 8	KOCOURKOV	34511			ROZA	20.7.1999
HUSÁK	JAROSLAV	BANÁNOVÁ 9	PSÁŘE	39601			KELINA	6.7.1999
JELEN	KOSŤA	KOKOSOVÁ 8	KOZOJEDY	30587	ŠEREDA	3.8.1994		
KOS	MÍLA	DATLOVÁ 6	CHLUPÁŇ	33210			STÁZA	2.4.2001