Optické rozpoznávání textu

Optické rozpoznávání textu (ale také dalších objektů), často označované jako OCR (Optical character recognition) má velký význam pro digitalizaci knih, rukopisů a dalších dokumentů, ve kterých vystupuje text – psaný buď strojově či volnou rukou. Oblast také silně interaguje s dalšími oblastmi počítačového zpracování přirozeného jazyka či umělou inteligencí.

Digitalizace tištěných dokumentů dnes patří mezi základní činnosti knihoven, ale také dalších institucí. Téměř každému se čas od času hodí, když může oblíbenou knihu fulltextově prohledávat, nebo pokud existuje možnost, analyzovat text na fotografiích. Motivace pro optické zpracování textu tak není v žádném případě jen knihovnická a hojně ji využívají i klasické vyhledávače, jako je Google (převážně v Books). Dnes je již běžné, že součástí softwaru, který je pořizován ke stolnímu sceneru či multifunkčnímu zařízení je právě aplikace, která převod textu tištěného do digitální podoby zajišťuje.

OCR lze využít například na:

- Automatické zadávání dat do systému – například smluv, fakturačních údajů, výkazů práce atp. Čím více jsou data homogenní a opakující se, tím lepší jsou možnosti pro jejich automatické zpracování.
- Digitalizace tištěných objektů a jejich další prohledávání, jazykové a obsahové analýzy atp.
- Převod písma psaného rukou na tabletu či perem do digitální podoby.
- Automatické rozpoznávání SPZ aut v systémech pro měření rychlosti, identifikace slov v jídelním lístku.
- Zpracování a analýzu starých tisků – například převod ze švabachu do moderního písma.
- …

Historie analýzy vytištěných dat automatizovanou metodou je relativně dlouhou. V roce 1914 Emanuel Goldberg vyvinul stroj, který uměl číst písmena a převádět je do telegrafického kódu.[1] Historie je ale mnohem složitější a například již v roce 1900 ruský vědec Tyurin navrhl zařízení na podporu zrakově postižených, která stála právě na rozpoznání jednotlivých znaků v textu. Systematicky se v oblasti počítačového zpracování textu věnuje informatika již od padesátých let minulého století.[2]

---

[1] Optical character recognition. In: *Wikipedia: the free encyclopedia* [online]. San Francisco (CA): Wikimedia Foundation, 2001- [cit. 2013-08-27]. Dostupné z: http://en.wikipedia.org/wiki/Optical_character_recognition

[2] CHAUDHURI, B.B a U PAL. A complete printed Bangla OCR system. *Pattern Recognition*. 1998, vol. 31, issue 5, s. 531-549. DOI: 10.1016/S0031-3203(97)00078-2. Dostupné z: http://linkinghub.elsevier.com/retrieve/pii/S0031320397000782

Je třeba zdůraznit, že OCR je značně komplexní záležitostí a že každé písmo má svá specifika, která musí patřičná aplikace rozpoznávat. Nejde ale jen o písmo, ale pro dobrý překlad jsou třeba také slovníky, takže každý jazyk musí být zpracován ve slovnících, které slouží pro identifikaci jednotlivých slov. Bez nich je efektivita podobných aplikací mizivá.

Často se hovoří o spojení OCR systémů s umělou inteligencí. Cílem je rozpoznat obsah textu a pomocí této znalosti lépe volit slova, která se budou v případě sporných prvků generovat. Jde tak v zásadě o další krok, který umožní přesnější a kvalitnější digitalizaci dokumentů. Samostatným problémem je pak OCR tam, kde se zpracovává ručně psané písmo a kde musí probíhat systém učení, ke kterému se užívají systémy s umělou inteligencí stavějící na skrytých Markovových modelech (HMD). Obecně je tak důležitá nejen slovníková zásoba, ale ideálně také znalost větných struktur či myšlení lidí, kteří jazyk používají.

Proces zpracování

Samotný proces zpracování je rozdělený do několika základních fází. První je předzpracování, což je v zásadě příprava dokumentu pro samotnou analýzu. Pokud je to nutné provede se otočení dokumentu, odstraní se skvrny a šum. Následně proběhne převod do černobílé varianty (binarizace), která zvyšuje kontrast. Systém pak ještě identifikuje různé logické objekty (sloupce, linky, nadpisy) a rozdělí text do jednotlivých písmen (například odstraněním ligatur) a nastaví optimální měřítko. Takový text je pak připraven na zpracování.[3]

Možností jak rozpoznat jednotlivá písmena a jejich význam je více. Nejčastěji se používá převodu písmene do vektorového objektu a jeho porovnání se vzory. Proto je vždy důležité, aby se dobře zvolila znaková sada a jazyk při zadávání dokumentu pro OCR. Jde o metodu rychlou a relativně stabilní. Další možností je analýza různých průsečíků čar písma, podle kterých se určuje, o které písmeno jde. V zásadě vždy jde ale o porovnávání geometrických objektů pomocí k-nearst algoritmů.[4]

Třetí fází je post-processing, který již využívá lexikonů a umělé inteligence. V této fázi je nutné rozhodnout, která z možných slov na základě optické analýzy bude vybráno, identifikovat kontext atp. Čím lepší jsou slovníky, tím dokonalejších je možné dosáhnout výsledků. Opět nejde o práci s jistými pojmy, ale o snahu umístit do dokumentů slova s maximální pravděpodobností. V této fázi dochází také ke kontrole gramatiky a případnému dalšímu grafickému zpracování tak, aby v případě že uživatel chce, existoval dokument i vizuálně co možná nejpodobnější tomu digitalizovanému.

---

[3] BAZZI, Issam; SCHWARTZ, Richard; MAKHOUL, John. An omnifont open-vocabulary OCR system for English and Arabic. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 1999, 21.6: 495-504.

[4] Zjednodušeně řečeno se nehledá stejné písmeno, ale to nejpohodlnější.

Parametry a kritika

OCR není technologií, která by byla čistě bezkonfliktní. První skupina výtek míří proti tomu, aby se písmo upravovalo podle toho, aby s ním mohly stroje co nejlépe pracovat. V USA například vznikl standard OCR-A (1968), který byl v Evropě následování OCR-B (1968).[5] Jde o písma, která se užívají na bankovkách a umožňují snadné a rychlé čtení, identifikaci jednotlivých kusů platidla a jejich třízení. Písmo je tak degradované z vrcholného projevu kultury a intelektu člověka na informaci určenou stroji.

Druhou kritizovanou oblastí je možnost nasazení OCR na prolamování CAPTCHA. Ta je ale konstruovaná tak, že levnější metodou je najmutí lidí na její opisování, než samotná snaha o strojové zpracování uvedeného textu.

OCR přes všechny své možnosti a technologický pokrok má relativně vysokou chybovost, která roste s klesající kvalitou předlohy a exotičností jazyka nemá úplně stoprocentní efektivitu. U dobrých anglicky psaných textů (bez neobvyklých slov) se udává spolehlivost okolo 95-99%, avšak reálně je často značně nižší. Čeština je z tohoto pohledu jazyk relativně exotický a složitý.[6]

V případě studia ručně psaných poznámek je pak schopnost automatického rozpoznávání textu podstatně slabší a silně závisí na dovednostech píšícího. Problematické je také rozpoznávání složitých nebo zcela atypických písem, která užívají některé asijské národy nebo například indiánské kmeny.

Mezi nejčastěji zkoumané parametry se řadí právě spolehlivost a přesnost, které se udávají v procentech. Dalším prvkem je rychlost, která je důležitá například při digitalizaci velkých knihoven nebo archivů. Pro běžného uživatele ale většinou nehraje kruciální roli. Další důležitou schopností je pracovat se strukturou textu – zachovat sloupce, nadpisy nebo například obrázky. Jde o častou slabinu, která může učinit knihy či složitější články v časopise zcela nečitelnými či nepřehlednými. V neposlední řadě je zde pak rozčlenění na online a offline nástroje nebo cenový aspekt.

Příklady aplikací

Jednou z velice populárních aplikací pro online OCR je **Free-OCR**. Podporuje práci s více sloupci, mezi jazyky, které jsou podporovány, nechybí angličtina, čeština ani slovenština a řada dalších. Nahrávané

[5] Typographic Abbreviations Series #1: OCR. MyFonts Musings [online]. 2006 [cit. 2013-08-27]. Dostupné z: http://myfonts.wordpress.com/2006/09/18/typographic-abbreviations-series-1-ocr/

[6] Optical character recognition. In: *Wikipedia: the free encyclopedia* [online]. San Francisco (CA): Wikimedia Foundation, 2001- [cit. 2013-08-27]. Dostupné z: http://en.wikipedia.org/wiki/Optical_character_recognition

soubory jsou omezeny do 2 MB či 5000 PX. Dokumenty je možné nahrávat ve formátech PDF, JPG, GIF, TIFF a BMP. Služba je vhodná pro identifikaci či rozpoznání textu na jedné straně a rychlou úpravu, nikoli pro systematickou činnost.

Velice dobře funguje **Adobe Acrobat XI Standard**, který umožňuje snadnou a rychlou práci s velkými soubory, které mají mnoho stran, obsahují obrázky nebo sloupce. Výstup je možné provést do PDF (editovatelného), RTF či Wordu. Nevýhodou je především cena, která začíná na 299 USD.

ABBYY FineReader 11 je nástroj nabízí funkce pro kontrolu pravopisu, rozeznání čárového kódu, rozdělení obrázků, možnost rychlého prohledávání dokumentů nebo identifikaci vícejazyčného textu. Podporuje češtinu, slovenštinu, angličtinu a další jazyky. Export je možné provádět do formátů PDF, Word, HTML, CSV, DBF atp. Cena je od 619 Kč a na vyzkoušení je 15 denní demoverze.

Google Drive nabízí také možnost identifikovat text v nahraných PDF souborech. Je relativně rychlý, přesný a výsledek je možné editovat podle vlastních představ, snadno sdílet či publikovat. Nevýhodou je, že většinou nezachovává formátování, což může být u řady dokumentů velice problematické.

Závěrem

OCR je oblastí, která se relativně rychle a zajímavě rozvíjí. Jak pro běžného uživatele, tak také pro knihovny bude tato oblast stále důležitější a tak není překvapením, že v ní spatřuje řada vývojářů velký potenciál. Zajímavé budou také její průniky s dalšími oblastmi, které se těsně přimykají ke zpracování textů. Dnes se hovoří o sémantických technologiích a dolování dat, které by měly pomoci zjistit nejen písmennou a slovní strukturu textu, ale také jeho význam a obsah. Také návaznost na text-spech systémy je zřejmá, neboť cílem je opět převod nějaké analogové informace do digitálního systému.