OCR neboli optické rozpoznávání znaků (z anglického Optical Character Recognition) je metoda, která pomocí scanneru umožňuje digitalizaci tištěných textů, s nimiž pak lze pracovat jako s normálním počítačovým textem. Počítačový program převádí obraz buď automaticky, nebo se musí naučit rozpoznávat znaky.

Převedený text je téměř vždy v závislosti na kvalitě předlohy třeba podrobit důkladné korektuře, protože OCR program nemusí rozeznat všechna písmena správně. OCR – zpracování textu z tištěné do elektronické podoby je použitelné pro všechny tištěné výstupy z laserových, inkoustových, termosublimačních a jehličkových tiskáren a samozřejmě pro předlohy vytištěné knihtiskem.

U nevhodných předloh, např. slabě vytištěných jehličkových výtisků nebo dohromady slitých písmen, se z časového hlediska vyplatí spíše přepis textu. wikipedia

OCR aplikace

Kvalitní OCR aplikace jsou tesseract s grafickou nadstavbou YAGF a online OCR.

GOCR

Gocr je zástupce OCR programů dostupných zadarmo, který zvládá rozpoznávání diakritiky a je použitelný pro česky psané texty (umí zpracovat naskenované texty do kódování UTF-8). Má úspěšnost asi 85%. Kvůli této úspěšnosti je dobré využít vhodný nástroj ke kontrole správnosti zpracovaného textu, např. OpenOffice, nebo aspell.

Program podporuje užití databáze známých znaků, nebo vytváření („učení“) nové databáze přímo při rozpoznávání. Tato funkce je ale zatím ve vývojové verzi. Na program existuje několik grafických nadstaveb (např. Kooka,gocr-Gtk - ubuntu 8.0.4)

Příklad použití:

Po mých pokusech s nastavením skeneru jsem zůstal u volby „binary“ s rozlišením 500 dpi. Vyšší rozlišení už se mi jevilo nevhodné. Skenuji v programu xsane, který je počeštěn. Celkem zajímavý je i xscanimage. Gocr podporuje celkem hodně grafických formátů (viz man gocr).

Přejdeme do adresáře kde máme naskenované dokumenty v některém z podporovaných grafických formátů:

```
cd /cesta/k_nasim/obrazkum/ Enter
```

```
ls Enter
```

(skeny jsou ve formátu png, ale můžete použít i jiné formáty viz man gocr) Např.:

```
002.png 006.png  012.png  016.png  020.png  024.png  028.png
000.png 003.png  007.png  010.png  013.png  017.png  021.png  025.png
029.png
001.png 004.png  008.png  014.png  018.png  022.png  026.png  030.png
```

```
005.png 009.png  011.png  015.png  019.png  023.png  027.png  10_600.png
```

```
gocr -f UTF8  -i 001.png -o 001.txt
```

potom pokračujeme i s dalšími soubory až převedeme všechny soubory. Gocr je celkem rychlý.

```
gocr -f UTF8  -i 001.png -o 001.txt
```

atd. Čímž se vytvoří soubory s rozpoznaným textem 001.txt, 002.txt atd.

Můžete také použít jednoduchý příkaz, který rozpozná všechny PNG soubory v adresáři:

```
for f in `ls *.png`; do gocr -f UTF8 -i "$f" -o "$f.txt"; done
```

-f UTF8 = výstup bude v kódování UTF8\ **-i 001.png** = input tj. vstupní soubor\ **-o 001.txt** = output, výstupní soubor - takto se bude jmenovat náš elektronický dokument

Pokud potřebujete výsledné textové soubory převést do jiného kódování, můžete využít například program recode. Pro převod do kódování ISO-8859-2 můžete použít např.:

```
recode UTF-8..ISO-8859-2 001.txt
```

(soubor se překóduje na ISO-8859-2)

```
recode UTF-8..ISO-8859-2 *.txt
```

(překóduje všechny txt soubory ve složce)

Pokud program recode upozorní na špatné konce řádků, můžete použít parametr -f.

```
recode -f UTF-8..ISO-8859-2 *.txt
```

Manuálová stránka (man gocr)

```
GOCR(1)                       Uživatelský manuál
GOCR(1)

JMÉNO
       gocr - konzolový program pro optické zpracování znaků (OCR)

SYNTAXE
       gocr [OPTION] [-i] pnm-file

POPIS
       gocr je program pro optické rozpoznávání znaků, který může být použit
z
       příkazové řádky. Jako vstup přijímá PNM, PGM, PBM, PPM nebo PCX
formát
       a rozpoznaný text vypisuje na standardní výstup. Pokud je namísto
```

```
názvu
       pnm souboru použita pomlčka, jsou data čtena  ze  standardního
vstupu.
       Pokud  jsou  nainstalovány  programy  gzip,  bzip2 a netpbm-progs a
váš
       systém podporuje popen(3) tak jsou jako  vstupní  soubory  (ne
stream)
       podporovány  také  pnm.gz,  pnm.bz2, png, jpg, jpeg, tiff, gif, bmp,
ps
       (pouze jednostránkové) a eps, kde pnm může být nahrazeno jedním z
ppm,
       pgm nebo pbm souborem.


OPTIONS
       -h     vypíše informace o použití

       -i file
              čte  vstup ze souboru "file" (nebo ze standardního vstupu,
pokud
              je jako název souboru uvedena jednoduchá pomlčka)

       -o file
              uloží výstup do souboru  "file"  namísto  výpisu  na
standardní
              výstup

       -e file
              odešle  chyby  do  souboru  "file" namísto na standardní
chybový
              výstup, nebo na standardní výstup, pokud je namísto "file"
uve-
              dena pomlčka

       -x file
              tato  možnost  bude  vypisovat do souboru "file" aktuální
postup
              zpracovávání. "file" může být název souboru,  název
pojmenované
              roury  (viz  man mkfifo), nebo deskriptor souboru 1...255.
Tato
              volba  je  užitečná  pro  GUI  vývojáře,  aby  mohli
zobrazovat
              aktuální  postup  OCR  zpracovávání.  Deskriptor souboru je
dos-
              tupný pouze, pokud bylo gocr zkompilování s definovanou
konstan-
              tou __USE_POSIX.

       -p path
              cesta k databázi včetně ukončovacího lomítka (výchozí je
```

```
./db/).
              Zde budou umístěny obrázky s naučenými znaky

       -f format
              výstupní formát rozpoznaného textu (ISO8859_1 TeX HTML XML
UTF8
              ASCII).   Do  XML  budou  uvedena  také  data  o  pozici znaku
a
              pravděpodobnosti úspěšnosti rozpoznání daného znaku

       -l level (úroveň)
              nastavit úroveň šedé na úroveň "level" (0<160<=255,  default:
0
              pro  autodetekci), tmavší pixely přísluší znakům, světlejší
pix-
              ely jsou interpretovány jako pozadí vstupního souboru

       -d size
              nastavit velikost prachu v pixelech  (částečky  menší  než
tato
              hodnota  budou  odstraněny), 0 znamená autodetekci, výchozí
hod-
              nota je -1 pro autodetekci

       -s num nastavit   šířku   mezery   mezi   slovy   v   jednotkách
bodů
              (typografických terčíků). Výchozí hodnotou je 0 pro
autodetekci.
              Širší mezery jsou chápány jako mezery mezi slovy, užší jako
mez-
              ery mezi znaky.

       -v verbosity
              upovídaný  režim  s výstupem na standardní chybový výstup;
"ver-
              bosity" je bitové pole (specifikace viz. níže)

       -c string
              výpis upovídaných znaků pouze pro znaky z řetězce "string".
Pro
              tyto  znaky  je  generováno  více  informací. Podtržítko
znamená
              neznámé znaky. Tato možnost je vhodná k omezení výstupu
ladících
              informací pouze na potřebné.

       -C string
              rozpoznávat  pouze  znaky z řetězce "string". Toto je
filtrovací
              funkce, která umožní omezení  zpracovávaného  řetězce  pouze
na
```

```
              určité  znaky.  Je možno použít rozmezí 0-9 nebo a-z, pro znak
-
              je potřeba použít --

       -a certainty
              nastavit hodnotu spolehlivosti  (0..100;  výchozí  hodnota
95).
              Znaky  s  větší  hodnotou  spolehlivosti,  než  "certainty"
jsou
              přijaty, znaky s  menší  hodnotou  jsou  považovány  za
neznámé
              (nerozpoznané).   Nastavte  vyšší  hodnotu,  pokud chcete
vypsat
              pouze znaky s větší spolehlivostí na správnost.

       -m mode
              nastavit režim operace; "mode" je bitové pole  (výchozí
hodnota
              je 0)

       -n bool
              pokud  je "bool" nenulový, rozpoznávej pouze čísla. Tato
možnost
              je nyní zastaralá, použijte -C "0123456789"

       Úroveň upovídanosti je specifikován jako bitové pole:

       1         výpis více informací

       2         výpis tvarů z oblastí (viz -c)

       4         výpis vzorů z oblastí (viz -c)

       8         výpis vzorů po rozpoznání (pro ladění)

       16        výpis ladících informací o rozpoznání jednotlivých  řádků
na
                 standardní chybový výstup

       32        vytvoří  outXX.png  s  rámečky (okolo rozpoznávaných znaků)
a
                 jednotlivými rozpoznávanými řádky při každém OCR kroku

       Operační režimy jsou následující:

       2         použij databázi pro znaky, které nejsou rozpoznány
předešlými
                 algoritmy (funkce je v raném vývoji)

       4         přepínání mezi analýzou rozvržení, nebo zónováním (ve
vývoji)
```

```
       8         neporovnávat nerozpoznané znaky s rozpoznanými

       16        nesnažit  se  rozdělit překrývající se znaky na 2 nebo 3
jed-
                 notlivé znaky

       32        neprovádět opravy kontextu

       64        balení znaků: před začátkem rozpoznání jsou prohledány
stejné
                 znaky  a  pouze jeden z těchto znaků je zaslán k analýze.
(ve
                 vývoji)

       130       rozšířit databázi, zeptá se uživatele na nerozpoznané znaky
a
                 rozšíří  databázi  podle  odpovědi uživatele (128+2, funkce
v
                 raném vývoji)

       256       vypnout rozpoznávací engine (má smysl spolu s  parametrem
-m
                 2)



AUTOR
       Joerg Schulenburg (see http://jocr.sourceforge.net/ for EMAIL)
       First version of man page by Tim Waugh <twaugh@redhat.com>

INFORMACE O VERZI
       Tato stránka dokumentuje gocr, verze 0.41

DALŠÍ INFORMACE
       Více  informací  můžete  nalézt  v  /usr/share/doc/gocr-
X.XX/gocr.html.
       Přečtěte si také /usr/share/doc/gocr-X.XX/README abyste  zjistili,
jak
       dosáhnout lepších výsledků

PŘÍKLADY
       gocr -v 33 text1.pbm
              výpis  upovídaných  informací,  out30.png  je  vytvořen, aby
byl
              vidět postup rozpoznávání.

       gocr -v 7 -c _YV text1.pbm
              upovídaný výstup pro neznámé znaky Y and V

       djpeg -pnm -gray text.jpg | gocr -
              převést jpg obrázek do pnm formátu a použít jako vstup
```

```
Linux                                20 Aug 2006
GOCR(1)
```

Domovská stránka http://jocr.sourceforge.net/

gocr-gtk grafická nástavba gocr

Grafická nástavba programu gocr gocr-gtk. Spouští se příkazem „gtk-ocr“

Ve složce **Setting** dopíšeme k příkazu /usr/bin/gocr kódování a uložíme:

```
/usr/bin/gocr -f UTF8
```

Lze vybrat postupně všechny naše naskenované obrázky a pak najednou konvertovat. Soubory, které se vytvoří, budou mít příponu .txt a budou se nacházet ve stejném adresáři jako naskenované obrázky. Jednoduše lze měnit velikost prachu odstín šedi a šířku mezery (přednastaveno je 10/160/0). Pokud potřebujeme překódovat. Např. na ISO-8859-2:

```
recode -f UTF-8..ISO-8859-2 *.txt
```

Kooka

Je kompletní program pro skenování obrázků, využívající jako OCR právě program gocr. Jednoduše lze použít kontrolu pravopisu, která funguje dobře, ale musíme si nainstalovat slovník **aspell-cs**, který je v distribuci.

LOCR

Zatím verze 0.1, je zdarma. http://www.math.northwestern.edu/~mlerma/locr/

ocre

Vypadá zajímavě. Umí i polské znaky, je grafický a autor píše, že ho bude rozšiřovat na další jazyky dle přání uživatelů. Tak pišme!!! http://lem.eui.upm.es/ocre.html

Hebrew OCR

Rozpozná poezii a biblické texty v hebrejštině. http://www.claraocr.org/

ClaraOCR

Výborně vypadající grafický program, ale už se dlouho nevyvíjí. Nepodporuje české znaky, ale je možné naučit jej nové znaky (tip! c + háček = č!) http://www.claraocr.org/

Tesseract OCR

Původně komerční OCR program od firmy HP, nyní vyvíjen společností Google pod licencí Apache. http://code.google.com/p/tesseract-ocr/

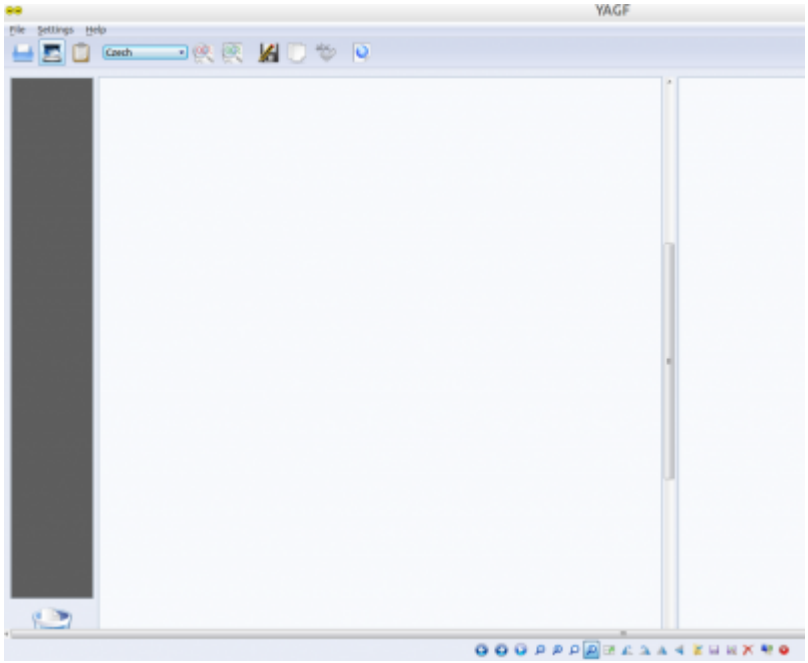
Nainstalujte balík tesseract-ocr a tesseract-ocr-ces

Grafické nadstavby YAGF, gImageReader

Nainstalujte balík yagf

Aplikaci můžete spustit z nabídky **Aplikace → Kancelář → YAGF**, případně příkazem `yagf`.

Nastavení YAGF Spusťte YAGF a ve sloupci **Settings → OCR Settings** vybrat tesseract



Online OCR

Grafická nadstavba teseractu freeware - online OCR

ABBYY FineReader CLI

komerční program

Odkazy

- Původni zdroj
- Linux OCR: A review of free optical character recognition software - gocr, Clara, Ocre, Ocrad, Tesseract, Ocropus, Aspire OCR
- Optical Character Recognition With Tesseract OCR On Ubuntu 7.04

Rozšíření: Tento návod je příliš stručný. Pomozte Ubuntu Wiki tím, že jej rozšíříte. Více...

Grafická úprava: Tento návod potřebuje důležité grafické a stylistické úpravy. Více...

Konvence: Tento návod nesplňuje některé z na Wiki zavedených konvencí. Více...