

Příprava obrázků k OCR

Obhajoba maturitní práce

Jakub Ambroz

29. května 2020

Základní pojmy

- ▶ *AI = Artificial Intelligence* = umělá inteligence
- ▶ *OCR = Optical Character Recognition* = optické rozpoznávání znaků

Tesseract

- ▶ OCR program příkazové řádky
- ▶ napsaný v C++ a nyní má otevřený zdrojový kód na Githubu
- ▶ K dispozici jsou už naučené (natrénované) verze
- ▶ Pro zlepšení výsledků doporučuje dokumentace úpravu obrázků

Otázka maturitní práce

Jak (a jestli) je možné upravit obrázek, tak aby byla zvýšena kvalita rozpoznání znaků Tesseractem?

OpenCV

- ▶ *Open Source Computer Vision Library* = Knihovna s otevřeným zdrojovým kódem pro počítačové vidění
- ▶ Psaná v C++, ale má rozhraní i v Pythonu

Binarizace

- ▶ Převeďte obrázek na černé a bílé pixely
- ▶ Práh, který je rozdělí je buď globální nebo adaptivní (bere v potaz okolí bodu)

abandonment and replacement by the limit concept.

Obrázek-1: Gaussovský adaptivní práh

abandonment and replacement by the limit concept.

Obrázek-2: Gaussovský adaptivní práh s jinými parametry

abandonment and replacement by the limit concept.

Obrázek-3: Adaptivní práh

abandonment and replacement by the limit concept.

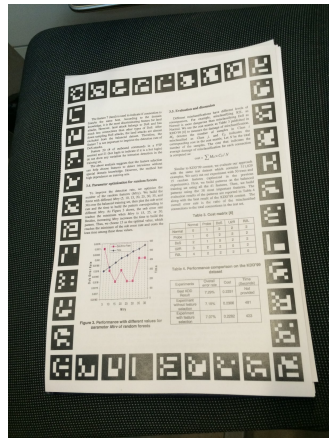
Obrázek-4: Otsuovo práh

Další úpravy

- ▶ Grayscale - převede obrázky do úrovní šedi
- ▶ Rozmazání - nějakým způsobem průměruje okolí pixelu
- ▶ Škálování - změna velikosti obrázku
- ▶ Odstranění šumu

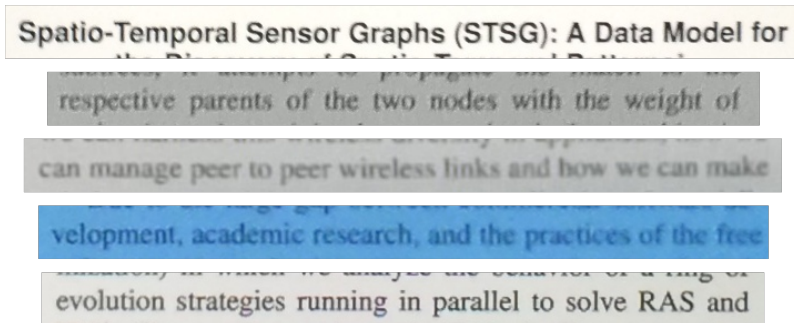
B-MOD

- ▶ *Brno Mobile OCR Dataset* - z VUTBR
- ▶ Rozmanitý: foceno různými zařízeními, za různých světelných podmínek, z různých vzdáleností a z různých úhlů
- ▶ Rozstříhané fotky po jednotlivých řádcích



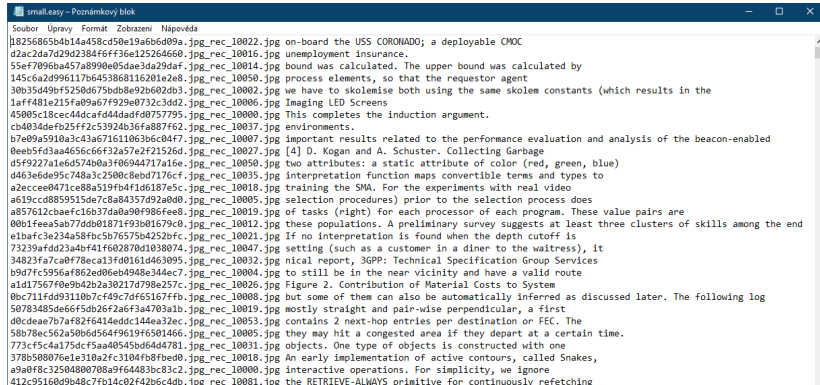
Obrázek-5: Ukázka z datasetu

Mnou použité



Obrázek-6: Po řádcích

Mnou použité 2



Obrázek-7: Textový soubor se správným řešením

Cíle a automatizace

- ▶ Zjistit, jak upravit obrázek, tak aby se zlepšili výsledky Tesseractu
- ▶ Zkusit velké množství úprav obrázků, zkusit rozpoznat Tesseractem a porovnat úspěšnosti

OpenCV

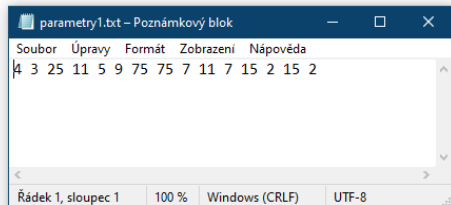
- ▶ Vytvořil jsem si kódy jednotlivým funkcím
- ▶ Program, který je vykoná v takovém pořadí v jakém jsou předloženy
- ▶ Potřebné informace získá program z jedné řádky v textovém dokumentu
- ▶ Program provede operaci pro všechny řádky v souboru

OpenCV parametry

prikazy-pro-opencv3.old – Poznámkový blok

Soubor Úpravy Formát Zobrazení Nápvěda

```
lines-02(4)-11- small.easy parametry3.txt 43
lines-02(4)-11- small.easy parametry1.txt 31
lines-02(4)-11- small.easy parametry1.txt 32
lines-02(4)-11- small.easy parametry1.txt 33
lines-02(4)-11- small.easy parametry1.txt 34
lines-02(4)-11- small.easy parametry2.txt 31
lines-02(4)-11- small.easy parametry2.txt 32
lines-02(4)-11- small.easy parametry2.txt 33
lines-02(4)-11- small.easy parametry2.txt 34
lines-02(4)-11- small.easy parametry3.txt 31
lines-02(4)-11- small.easy parametry3.txt 32
lines-02(4)-11- small.easy parametry3.txt 33
lines-02(4)-11- small.easy parametry3.txt 34
lines-02(4)-11-21(3,25,11)- small.easy parametry1.txt 31 41
lines-02(4)-11-21(3,25,11)- small.easy parametry1.txt 31 42
lines-02(4)-11-21(3,25,11)- small.easy parametry1.txt 31 43
lines-02(4)-11-21(3,25,11)- small.easy parametry1.txt 32 41
lines-02(4)-11-21(3,25,11)- small.easy parametry1.txt 32 42
lines-02(4)-11-21(3,25,11)- small.easy parametry1.txt 32 43
lines-02(4)-11-21(3,25,11)- small.easy parametry1.txt 33 41
```



Obrázek-8: Textový soubor pro OpenCV

Tesseract a pytesseract

- ▶ Je nutné dobře nainstalovat Tesseract
- ▶ Je třeba ho nastavit jako příkaz v cmd
- ▶ A propojit s Pythonem pomocí *pytesseractu*
- ▶ Poté zavolat Tesseract a uložit jeho výstup do textového souboru

Příkazy pro Tesseract

soubory-pro-evaluaci.txt – Poznámkový blok


Šoubor	Úpravy	Formát	Zobrazení	Nápověda
lines-02(4)	-11-21(3,25,11)	-33(11,11,11)	-43-	small.easy
lines-02(4)	-11-21(3,25,11)	-34(11)	-41(9,4)	small.easy
lines-02(4)	-11-21(3,25,11)	-34(11)	-42(9,4)	small.easy
lines-02(4)	-11-21(3,25,11)	-34(11)	-43-	small.easy
lines-02(4)	-11-21(3,25,11)	-31(7,7)	-41(11,10)	small.easy
lines-02(4)	-11-21(3,25,11)	-31(7,7)	-42(11,10)	small.easy
lines-02(4)	-11-21(3,25,11)	-31(7,7)	-43-	small.easy
lines-02(4)	-11-21(3,25,11)	-32(7,60,60)	-41(11,10)	small.eas
lines-02(4)	-11-21(3,25,11)	-32(7,60,60)	-42(11,10)	small.eas
lines-02(4)	-11-21(3,25,11)	-32(7,60,60)	-43-	small.easy
lines-02(4)	-11-21(3,25,11)	-33(9,9,9)	-41(11,10)	small.easy
lines-02(4)	-11-21(3,25,11)	-33(9,9,9)	-42(11,10)	small.easy
lines-02(4)	-11-21(3,25,11)	-33(9,9,9)	-43-	small.easy
lines-02(4)	-11-21(3,25,11)	-34(9)	-41(11,10)	small.easy
lines-02(4)	-11-21(3,25,11)	-34(9)	-42(11,10)	small.easy
lines-02(4)	-11-21(3,25,11)	-34(9)	-43-	small.easy
lines-02(4)	-11-21(3,25,11)	-41(15,2)	-	small.easy
lines-02(4)	-11-21(3,25,11)	-42(15,2)	-	small.easy
lines-02(4)	-11-21(3,25,11)	-43-	-	small.easy
lines-02(4)	-11-21(3,25,11)	-41(9,4)	-	small.easy
lines-02(4)	-11-21(3,25,11)	-42(9,4)	-	small.easy
lines-02(4)	-11-21(3,25,11)	-43-	-	small.easy
lines-02(4)	-11-21(3,25,11)	-41(11,10)	-	small.easy

soubory-pro-tesseract1.old – Poznámkový blok

Soubor	Úpravy	Formát	Zobrazení	Nápověda
lines-02(4)-	small.easy			
lines-02(4)-11-	small.easy			
lines-02(4)-11-21(3,25,11)-	small.easy			
lines-02(4)-21(3,25,11)-11-	small.easy			
lines-02(4)-11-31(5,5)-41(15,2)-	small.easy			
lines-02(4)-11-31(5,5)-42(15,2)-	small.easy			
lines-02(4)-11-31(5,5)-43-	small.easy			
lines-02(4)-11-32(9,75,75)-41(15,2)-	small.easy			
lines-02(4)-11-32(9,75,75)-42(15,2)-	small.easy			
lines-02(4)-11-32(9,75,75)-43-	small.easy			
lines-02(4)-11-33(7,7,7)-41(15,2)-	small.easy			
lines-02(4)-11-33(7,7,7)-42(15,2)-	small.easy			
lines-02(4)-11-33(7,7,7)-43-	small.easy			
lines-02(4)-11-34(7)-41(15,2)-	small.easy			
lines-02(4)-11-34(7)-42(15,2)-	small.easy			
lines-02(4)-11-34(7)-43-	small.easy			
lines-02(4)-11-41(15,2)-	small.easy			
lines-02(4)-11-42(15,2)-	small.easy			
lines-02(4)-11-43-	small.easy			
lines-02(4)-11-41(9,4)-	small.easy			
lines-02(4)-11-42(9,4)-	small.easy			
lines-02(4)-11-43-	small.easy			
lines-02(4)-11-41(11,10)-	small.easy			

Obrázek-9: Textový soubor pro Tesseract

Výstup Tesseractu

 0eeb5fd3aa4656c66f32a57e2f21526d.jpg_rec_l0027.jpg.tess – Poznámkový blok

Soubor Úpravy Formát Zobrazení Nápověda

»~ eer ee Ww AS a Se 2°

[4] D. hogan, and A. Se huster. Collecting Garbage

Obrázek-10: Textový soubor vytvořený Tesseractem

Úspěšnost

- ▶ Převeďte znak nové řádky na mezeru
- ▶ Rozdělí podle mezer
- ▶ Ve vzniklém poli hledá slova v řešení (to také rozdělíme podle mezer)
- ▶ Vypočítá jaké procento slov to nenašlo, tj. byly chybně identifikovány

Výstup programu

sorted.txt – Poznámkový blok

Soubor	Úpravy	Formát	Zobrazení	Nápověda
lines-02(4)-	0.2848375654523957	0.31147177841597873	5059/17761	1958
lines-02(4)-11-31(5,5)-	0.28772635814889336	0.31444496288610835	5148/17892	1970
lines-02(4)-11-31(7,7)-	0.29146799798059125	0.3181398476536851	5196/17827	1966
lines-02(4)-11-32(7,60,60)-	0.29190832817163126	0.3178324795699035	5184/17759	1962
lines-02(4)-11-0.291279554937413	0.3181483729252857	5251/17975	1977	
lines-0.2934296905169085	0.31873574800141735	5319/18127	1995	
lines-11-0.29363382135219607	0.3176051622335382	5355/18237	2003	
lines-02(4)-11-32(9,75,75)-	0.2942031415953938	0.32124600424025984	5263/17889	1968
lines-02(4)-11-33(7,7,7)-	0.2959177943736313	0.3218702065357416	5270/17809	1963
lines-02(4)-11-31(9,9)-	0.29606325519238263	0.3224959704458051	5317/17959	1975
lines-02(4)-11-32(5,90,90)-	0.29661776234740733	0.3239996911754059	5297/17858	1967
lines-02(4)-11-33(9,9,9)-	0.2969534650150653	0.32396897053992785	5322/17922	1974
lines-02(4)-11-34(9)-	0.3009725111023666	0.326157390736683	5354/17789	1963
lines-02(4)-11-34(11)-	0.30341761600444567	0.33111919916946325	5460/17995	1979
lines-02(4)-11-34(7)-	0.30373230373230375	0.33316244704289033	5428/17871	1968
lines-02(4)-11-33(11,11,11)-	0.30629277249207143	0.3374442081132987	5505/17973	1981
lines-02(4)-11-21(3,12,11)-	0.31626912691269127	0.34203520576426194	5622/17776	1961
lines-02(4)-11-21(2,17,7)-	0.3277873932220423	0.35356024444515505	5871/17911	1971
lines-02(4)-11-43-	0.33463437033635635	0.35935778912735766	5830/17422	1927
lines-02(4)-11-43-	0.33463437033635635	0.35935778912735766	5830/17422	1927
lines-02(4)-11-43-	0.33463437033635635	0.35935778912735766	5830/17422	1927
lines-02(4)-11-31(5,5)-43-	0.3408173705451032	0.3643741739331804	5996/17593	1948
lines-02(4)-11-33(7,7,7)-43-	0.34364691161415084	0.3695983695298388	6042/17582	1948
lines-02(4)-11-32(9,75,75)-43-	0.3480196324620477	0.3717901792296982	6098/17522	1942
lines-02(4)-11-34(7)-43-	0.35752383113935543	0.38344194952776406	6301/17624	1946
lines-02(4)-21(3,25,11)-11-	0.3583393622349602	0.3877713976220303	6439/17969	1977

Obrázek-11: Textový soubor vytvořený evaluačním programem

Tabulka

Název složky	Nerozpoznaných slov v %	Průměrný řádek v %
lines-02(4)-	28.48375654523957	31.147177841597875
lines-02(4)-11-31(5,5)-	28.772635814889334	31.44496288610835
lines-02(4)-11-31(7,7)-	29.146799798059124	31.813984765368506
lines-02(4)-11-32(7,60,60)-	29.190832817163127	31.78324795699035
lines-02(4)-11-	29.212795549374132	31.81483729252857
lines-	29.342969051690847	31.873574800141737
lines-11-	29.363382135219606	31.760516223353818
lines-02(4)-11-32(9,75,75)-	29.42031415953938	32.124600424025984
lines-02(4)-11-33(7,7,7)-	29.59177943736313	32.18702065357416

Obrázek-12: Tabulka nejlepších kombinací

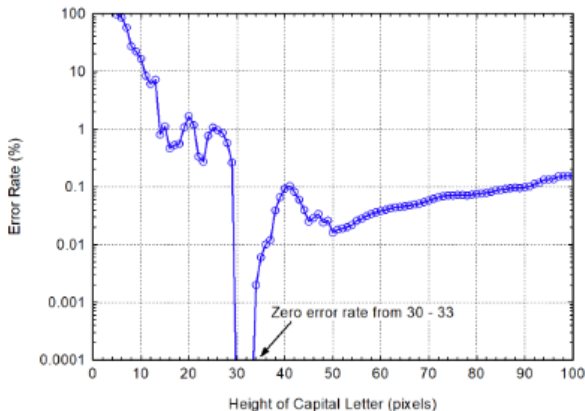
Tabulka



Willus Dotkom



the fact that there is an optimum letter size (in pixels) is quite unexpected



Obrázek-13: Optimální velikost pro Tesseract

Nedostatky

- ▶ Nepřehledné kódování funkcí OpenCV
- ▶ Nesystematické procházení možností

Děkuji za pozornost!

Pokud máte zájem, položte dotazy!

ments 1 and 2 under Conditions 1 and 2.

that, then we excluded those mouse-specific exons found in the

satisfy this condition. Our mining strategy uses different

is evaluated using the fingerprint DB generated from the

density of node identifiers. As a result, the test may result