# 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

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

misendomer of and replacement be seed

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 conseque.

Obrázek-3: Adaptivní práh

abandonment and replacement by the limit concept.

Obrázek-4: Otsuovo práh

## Další úpravy

- Grayscale převede obrázek 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 zařízením, za různých světelných podmínek, z různých vzdáleností a z různých úhlů
- Rozstřihané fotky po jednotlivých řádcích



Obrázek-5: Ukázka z datasetu



## Mnou použité

Spatio-Temporal Sensor Graphs (STSG): A Data Model for

respective parents of the two nodes with the weight of

can manage peer to peer wireless links and how we can make

velopment, academic research, and the practices of the free

evolution strategies running in parallel to solve RAS and

Obrázek-6: Po řádcích

## Mnou použité 2

```
small.easy – Poznámkový blok
Soubor Úpravy Formát Zobrazení Nápověda
18256865b4b14a458cd50e19a6b6d09a.jpg rec 10022.jpg on-board the USS CORONADO; a deployable CMOC
d2ac2da7d29d2384f6ff36e125264660.jpg rec 10016.jpg unemployment insurance.
55ef7096ba457a8990e05dae3da29daf.jpg rec 10014.jpg bound was calculated. The upper bound was calculated by
145c6a2d996117b6453868116201e2e8.jpg rec 10050.jpg process elements, so that the requestor agent
30b35d49bf5250d675bdb8e92b602db3.jpg rec 10002.jpg we have to skolemise both using the same skolem constants (which results in the
laff481e215fa09a67f929e0732c3dd2.jpg rec 10006.jpg Imaging LED Screens
45005c18cec44dcafd44dadfd0757795.jpg rec 10000.jpg This completes the induction argument.
cb4034defb25ff2c53924b36fa887f62.jpg rec 10037.jpg environments.
b7e09a5910a3c43a671611063b6c04f7.jpg rec 10007.jpg important results related to the performance evaluation and analysis of the beacon-enabled
@eeb5fd3aa4656c66f32a57e2f21526d.lpg rec 10027.lpg [4] D. Kogan and A. Schuster, Collecting Garbage
d5f9227a1e6d574b0a3f06944717a16e.jpg rec 10050.jpg two attributes: a static attribute of color (red. green, blue)
d463e6de95c748a3c2500c8ebd7176cf.jpg_rec_l0035.jpg interpretation function maps convertible terms and types to
a2eccee0471ce88a519fb4f1d6187e5c.jpg rec 10018.jpg training the SMA. For the experiments with real video
a619ccd8859515de7c8a84357d92a0d0.jpg rec 10005.jpg selection procedures) prior to the selection process does
a857612cbaefc16b37da0a90f986fee8.jpg rec 10019.jpg of tasks (right) for each processor of each program. These value pairs are
00b1feea5ab77ddb01871f93b01679c0.jpg rec 10012.jpg these populations. A preliminary survey suggests at least three clusters of skills among the end
elbafc3e234a58fbc5b76575b4252bfc.jpg rec 10021.jpg If no interpretation is found when the depth cutoff is
73239afdd23a4bf41f602870d1038074.jpg rec 10047.jpg setting (such as a customer in a diner to the waitress), it
34823fa7ca0f78eca13fd0161d463095.jpg rec 10032.jpg nical report, 3GPP: Technical Specification Group Services
b9d7fc5956af862ed06eb4948e344ec7.jpg rec 10004.jpg to still be in the near vicinity and have a valid route
ald17567f0e9b42b2a30217d798e257c.jpg rec 10026.jpg Figure 2. Contribution of Material Costs to System
0bc711fdd93110b7cf49c7df65167ffb.jpg rec 10008.jpg but some of them can also be automatically inferred as discussed later. The following log
50783485de66f5db26f2a6f3a4703a1b.jpg_rec_10019.jpg mostly straight and pair-wise perpendicular, a first
d0cdeae7b7af82f6414eddc144ea32ec.jpg_rec_10053.jpg_contains 2 next-hop entries per destination or FEC. The
58b78ec562a50b6d564f9619f6501466.jpg rec 10005.jpg they may hit a congested area if they depart at a certain time.
773cf5c4a175dcf5aa40545bd64d4781.jpg rec 10031.jpg objects. One type of objects is constructed with one
378b508076e1e310a2fc3104fb8fbed0.jpg rec 10018.jpg An early implementation of active contours, called Snakes,
a9a0f8c32504800708a9f64483bc83c2.jpg rec 10000.jpg interactive operations. For simplicity, we ignore
412c95160d9h48c7fb14c02f42h6c4dh.ing rec 10081.ing the RETRIEVE-ALWAYS primitive for continuously refetching
```

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-opency3.old – Poznámkový blok
Soubor Úpravy Formát Zobrazení Nápověda
lines-02(4)-11- small.easy parametry3.txt 43
lines-02(4)-11- small.easy parametry1.txt 31
                                                        parametry 1.txt – Poznámkový blok
lines-02(4)-11- small.easy parametry1.txt 32
                                                        Soubor Úpravy Formát Zobrazení Nápověda
lines-02(4)-11- small.easy parametry1.txt 33
lines-02(4)-11- small.easy parametry1.txt 34
                                                       4 3 25 11 5 9 75 75 7 11 7 15 2 15 2
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
                                                        Řádek 1. sloupec 1
                                                                         100 %
                                                                               Windows (CRLF)
                                                                                               UTF-8
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

Praxe: OpenCV
Praxe: Tesseract
Praxe: Měření úspěšnost

## Příkazy pro Tesseract

```
soubory-pro-evaluaci.txt – Poznámkový blok
                                                              soubory-pro-tesseract1.old – Poznámkový blok
Soubor Úpravy Formát Zobrazení Nápověda
                                                             Soubor Úpravy Formát Zobrazení Nápověda
lines-02(4)-11-21(3,25,11)-33(11,11,11)-43- small.easy
                                                             lines-02(4)- small.easy
lines-02(4)-11-21(3,25,11)-34(11)-41(9,4)- small.easy
                                                             lines-02(4)-11- small.easv
lines-02(4)-11-21(3,25,11)-34(11)-42(9,4)- small.easy
                                                             lines-02(4)-11-21(3,25,11)- small.easv
lines-02(4)-11-21(3,25,11)-34(11)-43- small.easy
                                                             lines-02(4)-21(3,25,11)-11- small.easv
lines-02(4)-11-21(3,25,11)-31(7,7)-41(11,10)- small.easy
                                                             lines-02(4)-11-31(5,5)-41(15,2)- small.easv
lines-02(4)-11-21(3,25,11)-31(7,7)-42(11,10)- small.easy
                                                             lines-02(4)-11-31(5,5)-42(15,2)- small.easy
lines-02(4)-11-21(3,25,11)-31(7,7)-43- small.easv
                                                             lines-02(4)-11-31(5,5)-43- small.easy
lines -02(4) - 11 - 21(3, 25, 11) - 32(7, 60, 60) - 41(11, 10) small.eas lines -02(4) - 11 - 32(9, 75, 75) - 41(15, 2) small.easy
lines-02(4)-11-21(3,25,11)-32(7,60,60)-42(11,10)-small.eas lines-02(4)-11-32(9,75,75)-42(15,2)-small.eas
lines-02(4)-11-21(3,25,11)-32(7,60,60)-43- small.easy
                                                             lines-02(4)-11-32(9.75.75)-43- small.easv
lines-02(4)-11-21(3,25,11)-33(9,9,9)-41(11,10)- small.easy
                                                             lines-02(4)-11-33(7,7,7)-41(15,2)- small.easv
lines-02(4)-11-21(3,25,11)-33(9,9,9)-42(11,10)- small.easy
                                                             lines-02(4)-11-33(7,7,7)-42(15,2)- small.easy
lines-02(4)-11-21(3.25.11)-33(9.9.9)-43- small.easv
                                                             lines-02(4)-11-33(7,7,7)-43- small.easy
lines-02(4)-11-21(3,25,11)-34(9)-41(11,10)- small.easv
                                                             lines-02(4)-11-34(7)-41(15,2)- small.easy
lines-02(4)-11-21(3,25,11)-34(9)-42(11,10)- small.easy
                                                             lines-02(4)-11-34(7)-42(15,2)- small.easy
lines-02(4)-11-21(3,25,11)-34(9)-43- small.easy
                                                             lines-02(4)-11-34(7)-43- small.easv
lines-02(4)-11-21(3,25,11)-41(15,2)- small.easy
                                                             lines-02(4)-11-41(15,2)- small.easv
lines-02(4)-11-21(3,25,11)-42(15,2)- small.easy
                                                             lines-02(4)-11-42(15,2)- small.easv
lines-02(4)-11-21(3,25,11)-43- small.easy
                                                             lines-02(4)-11-43- small.easy
lines-02(4)-11-21(3,25,11)-41(9,4)- small.easy
                                                             lines-02(4)-11-41(9,4)- small.easy
lines-02(4)-11-21(3,25,11)-42(9,4)- small.easy
                                                             lines-02(4)-11-42(9,4)- small.easy
lines-02(4)-11-21(3,25,11)-43- small.easy
                                                             lines-02(4)-11-43- small.easv
lines-02(4)-11-21(3,25,11)-41(11,10)- small.easy
                                                             lines-02(4)-11-41(11,10)- small.easy
```

Obrázek-9: Textový soubor pro Tesseract



Praxe: OpenCV
Praxe: Tesseract
Praxe: Měření úspěšnosti

### Výstup Tesseractu

0eeb5fd3aa4656c66f32a57e2f21526d.jpg\_rec\_l0027.jpg.tess – Poznámkový blo

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

- Převede 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

iii sorted.txt – Poznámkový blok					
Soubor Úpravy Formát Zobrazení Nápo	věda				
lines-02(4)- 0.28483756545239	957 0.3114717	7841597873 505	9/17761 199	8	
lines-02(4)-11-31(5,5)- 0.28772	535814889336 0	.3144496288610835	5148/17892	1970	
lines-02(4)-11-31(7,7)- 0.29146	799798059125 0	.3181398476536851	5196/17827	1966	
lines-02(4)-11-32(7,60,60)-	0.291908328171631	26 0.317832479	5699035 518	4/17759	1962
lines-02(4)-11- 0.2921279554937	413 0.3181483	729252857 525	1/17975 197	7	
lines- 0.2934296905169085	0.318735748001417	35 5319/18127	1995		
lines-11- 0.29363382135219			5/18237 200	13	
lines-02(4)-11-32(9,75,75)- lines-02(4)-11-33(7,7,7)-	0.294203141595393	8 0.321246004	24025984 526	3/17889	1968
				0/17809	1963
lines-02(4)-11-31(9,9)- 0.29606	325519238263 0	.3224959704458051	5317/17959	1975	
lines-02(4)-11-32(5,90,90)-	0.296617762347407			7/17858	1967
lines-02(4)-11-33(9,9,9)-	0.296953465015065	3 0.323968970	53992785 532	2/17922	1974
lines-02(4)-11-34(9)- 0.30097					
lines-02(4)-11-34(11)- 0.30341					
lines-02(4)-11-34(7)- 0.30373					
lines-02(4)-11-33(11,11,11)-	0.306292772492071			5/17973	1981
lines-02(4)-11-21(3,12,11)-				2/17776	1961
lines-02(4)-11-21(2,17,7)-				1/17911	1971
lines-02(4)-11-43- 0.334634		.35935778912735766		1927	
lines-02(4)-11-43- 0.334634					
lines-02(4)-11-43- 0.33463		.35935778912735766			
			9331804 599	6/17593	1948
lines-02(4)-11-33(7,7,7)-43-				2/17582	1948
lines-02(4)-11-32(9,75,75)-43-	0.348019632462047	7 0.371790179		8/17522	1942
lines-02(4)-11-34(7)-43-		43 0.383441949		1/17624	1946
lines-02(4)-21(3,25,11)-11-	0.358339362234960	2 0.387771397	6220303 643	9/17969	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

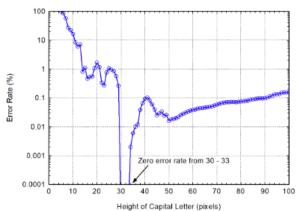


#### **Tabulka**



#### Willus Dotkom

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



Obrázek-13: Optimální velikost pro Jesseract

## 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.

1st then we excluded those mouse-specific exons found in the satisfy this condition. Our mining strategy uses different is evaluated using the hogerytim 1th generaled turn (1) there