

České vysoké učení technické v Praze
Fakulta jaderná a fyzikálně inženýrská

Katedra softwarového inženýrství

Obor: Aplikace informatiky v přírodních vědách



**Optické rozpoznávání znaků
na naskenovaných historických
plakátech pomocí
nejmodernějších metod**

**Optical Character Recognition
on Scanned Historical Posters
Using the State-of-the-Art
Methods**

VÝZKUMNÝ ÚKOL

Vypracoval: Anna Gruberová

Vedoucí práce: Ing. Adam Novozámský, Ph.D.

Rok: 2022

České vysoké učení technické v Praze

Fakulta jaderná a fyzikálně inženýrská

Katedra softwarového inženýrství

Akademický rok 2021/2022

ZADÁNÍ VÝZKUMNÉHO ÚKOLU

Student: Bc. Anna Gruberová

Studijní program: Aplikace informatiky v přírodních vědách

Název práce česky: Optické rozpoznávání znaků na naskenovaných historických plakátech pomocí nejmodernějších metod

Název práce anglicky: Optical Character Recognition on Scanned Historical Posters Using the State-of-the-Art Methods

Pokyny pro vypracování:

1. Seznamte se s problematikou optického rozpoznávání znaků. Na základě rešerše vyberte několik metod, se kterými budete dále pracovat a vyhodnocovat úspěšnost jejich detekce.
2. Stáhněte několik volně dostupných datasetů, které jsou využívány v literatuře k porovnání jednotlivých metod na OCR. Dále vytvořte svůj vlastní dataset z obdržených dat.
3. Nastudujte techniky porovnání OCR výstupů s ground-truth.
4. U vybraných metod prostudujte jejich chování na jednotlivých datasetech při různém nastavení parametrů.
5. Navrhněte také možnosti filtrování výstupů jednotlivých metod za účelem snížení falešných detekcí.

Doporučená literatura:

- [1] R. C. Gonzalez, R. E. Woods, Digital Image Processing (4th ed.). Pearson, 2018. ISBN 9353062985.
- [2] GOODFELLOW, Ian, Yoshua BENGIO a Aaron COURVILLE. Deep learning. Cambridge, Massachusetts: The MIT Press, [2016]. ISBN 0262035618.
- [3] SMITH, R. An Overview of the Tesseract OCR Engine. In: Ninth International Conference on Document Analysis and Recognition (ICDAR 2007) Vol 2 [online]. IEEE, 2007, 2007, s. 629-633. ISBN 0-7695-2822-8. ISSN 1520-5363. Dostupné z: doi:10.1109/ICDAR.2007.4376991
- [4] CHEN, Xiaoxue, et al. Text Recognition in the Wild. ACM Computing Surveys [online]. 2021, 54(2), 1-35 [cit. 2021-10-2]. ISSN 0360-0300. Dostupné z: doi:10.1145/3440756

Jméno a pracoviště vedoucího práce:

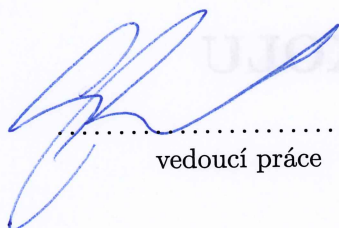
Ing. Adam Novozámský, Ph.D.

Computer Vision Lab, Institute of Visual Computing & Human-Centered Technology,
TU Wien - Faculty of Informatics

Datum zadání výzkumného úkolu: 15. 10. 2021

Termín odevzdání výzkumného úkolu: 31. 8. 2022

V Praze dne 15. 10. 2021



.....
vedoucí práce

.....
vedoucí katedry

Prohlášení

Prohlašuji, že jsem svou bakalářskou práci vypracovala samostatně a použila jsem pouze podklady (literaturu, projekty, SW atd.) uvedené v příloženém seznamu.

V Praze dne

.....

Anna Gruberová

Poděkování

.

Anna Gruberová

Název práce:

Optické rozpoznávání znaků na naskenovaných historických plakátech pomocí nej

Autor: Anna Gruberová

Obor: Aplikace informatiky v přírodních vědách

Druh práce: Výzkumný úkol

Vedoucí práce: Ing. Adam Novozámský, Ph.D.
Computer Vision Lab, Institute of Visual Computing & Human-Centered Technology, TU Wien - Faculty of Informatics

Konzultant: –

Abstrakt: .

Klíčová slova: .

Title:

Optical Character Recognition on Scanned Historical Posters Using the State-of-t

Author: Anna Gruberová

Abstract: .

Key words: .

Contents

0.1	Scene text detection	10
0.1.1	CRAFT	10
0.2	End-to-end systems	10
0.2.1	EasyOCR	10
0.2.2	Keras-ocr	10
0.2.3	tesseract	11

- What is OCR
- Text detection
 - CRAFT
- Text recognition
- End-to-end systems
 - Reading scanned documents
 - EasyOCR
 - keras-ocr
 - Tesseract (PyTesseract)
 - (Google Cloud Vision free) paid
 - (AWS Recognition) paid
 - (Kili) paid
- Results evaluation
 - Comparison of output and ground-truth
 -
- Testing methods on free datasets
 - Description of datasets
- Using methods on historical posters
 - Description of dataset

0.1 Scene text detection

Methods

0.1.1 CRAFT

0.2 End-to-end systems

0.2.1 EasyOCR

0.2.2 Keras-ocr

Keras-ocr is a python library used for detecting and recognizing text in images created by Fausto Morales. It unites the CRAFT text detection model¹ and an

¹hereinafter referred to as CRAFT

implementation in Keras python library of CRNN for recognizing text². [1]

On the official website³ of the package there is a comparison of this method with two other OCR APIs – Google Cloud Vision and AWS Rekognition. Their performance was tested on 1,000 images from the COCO-Text validation set using a basic pre-trained model of each method. None of the investigated methods performed poorly; however, AWS Rekognition had the worst precision and recall results. Google’s method and keras-ocr has similar results. It is important to mention that no tuning parameters were used in any of these methods. Another candidate for comparison was Tesseract but it performed on very badly on given data, most likely due to the fact that Tesseract is suitable for scanned documents rather than for photos of real life scenery and objects with text. [1]

CRAFT already provides a pretrained model which can be used directly without modification for text detection or it is used as initial model for training a new model on new data. This model was trained on three datasets (SynthText, IC13, IC17) and supports English and multi language text detection. [3] Similarly for recognition, CRNN also has a pretrained model This model was trained on the synthetic word dataset which consists of 9 million images with vocabulary of 90K English words. [2] To use these models in the keras-ocr library one either doesn’t specify anything and use the defaults, or pass the value `clovaai-general` for the CRAFT pretrained model or `kurapan` for the CRNN model.

0.2.3 tesseract

²hereinafter referred to as CRNN

³<https://pypi.org/project/keras-ocr/>

Bibliography

- [1] **Keras-OCR**. Last accessed 2022. Available from: <https://pypi.org/project/keras-ocr/>.
- [2] **Text recognition data - Visual Geometry Group - University of Oxford**. Last accessed 2022. Available from: <https://www.robots.ox.ac.uk/~vgg/data/text/#sec-synth>.
- [3] CLOVA AI RESEARCH. **Clovaai/Craft-pytorch: Official implementation of character region awareness for text detection (CRAFT)**. Available from: <https://github.com/clovaai/CRAFT-pytorch>.