

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:

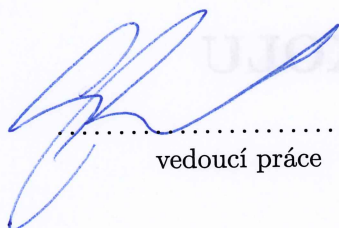
Ing. Adam Novozámský, Ph.D.

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

1	OCR	11
1.1	Text detection	11
1.2	Text recognition	11
2	Software	13
2.1	Scene text detection	13
2.1.1	CRAFT	13
2.2	Scene text recognition	13
2.2.1	CRNN	13
2.3	End-to-end systems	13
2.3.1	Tesseract	14
2.3.2	EasyOCR	14
2.3.3	Keras-ocr	15
2.3.4	Google Cloud Vision	16
2.3.5	Other	16
3	Datasets	17
3.1	SCUT-CTW1500 dataset	18
3.2	KAIST Scene Text Database	18
3.3	Born-Digital Images	20
3.4	Wien TU dataset	20
4	Testing methods	21
4.1	EasyOCR	21
4.2	Tesseract	21
4.3	21

Chapter 1

OCR

1.1 Text detection

1.2 Text recognition

Outline

- What is OCR
- Datasets (types(synthetic, photos, scanned documents),problems(languages, noise, nonhorizontal text))
- Text detection
 - Description
 - Methods (CRAFT)
- Text recognition
 - Description
 - Methods
- End-to-end systems (Annotating tool)
 - 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
 - Bag of words
- Testing methods on free datasets
 - Description of datasets
- Using methods on historical posters
 - Description of dataset

Chapter 2

Software

2.1 Scene text detection

Methods

2.1.1 CRAFT

Character Region Awareness for Text Detection (CRAFT) is framework for scene text detection. It uses a Convolutional Neural Network. It performs good also on curved or differently deformed texts. Its methodology is to localize individual characters then characters belonging to the same word (based on distance) can be connected into word box or polygon. After that bounding box is created around it and output contains the rectangle coordinates.[3]

2.2 Scene text recognition

Methods

2.2.1 CRNN

Convolutional Recurrent Neural Network

2.3 End-to-end systems

End-to-end system when given an image with text it can both detect and recognize the text and produces a string output of the text. The supply of such tools is wide, ranging from open source libraries for various programming languages to commercial softwares with modern GUI. New methods are still being developed as there is always

space for improvement. New methods can come from commercial background or are developed for international OCR competitions. In the next section some of the free available tools are described.

2.3.1 Tesseract

Tesseract is an open source text recognition engine. It supports over 160 languages. Originally Tesseract was created by Hewlett-Packard in late 1980s, from 2006 it is developed and maintained by Google. As it does not have a built-in GUI direct use is via command line. However, there exist a significant number of GUIs for Linux, Windows, Mac for computer usage and also for Android and iOS to use on mobile phones and few online OCR services. Another way how to use the engine is via libraries for computer languages, namely for example they exist for Java called tess4j, Python called pytesseract, R, Ruby and others. [17]

Tesseract is mainly used as tool for recognizing documents (with both computer font text or handwritten text). Best results are obtained on preprocessed images. The preprocessing includes noise reduction, horizontal alignment of text, elimination of dark borders around text region, conversion to binary black and white picture and other adjustments depending on the nature of the picture. Thus when used on scene text images it gives generally worse results than other OCR softwares.

Computations with Tesseract are supported for GPU and also CPU. Tesseract uses for recognition Long Short Term Memory (LSTM) model (kind of RNN).

Page Segmentation models By default Tesseract expects a page of text – black letters on white background grouped in horizontal lines, where font type and font size vary only slightly. To deal with differently distributed text over an image Tesseract provides thirteen page segmentation models (PSM). When selecting the right model Tesseract performance can increase from zero up to almost perfect results. Description of all the PSMs can be find directly via Tesseract help command in console application. <https://pyimagesearch.com/2021/11/15/tesseract-page-segmentation-modes-psms-explained-how-to-improve-your-ocr-accuracy/>

2.3.2 EasyOCR

EasyOCR is a product of Jaded AI for both image text detection and recognition. it supports over 80 languages and various scripts such as Latin, Chinese, Arabic etc. The company offers software with web interface for free and also prepaid version which enables usage of a new model for custom data. However, in addition to the web interface, the company also created a python package under the same name.[1]

The product is still in development and aims for wider functionality. A future idea of EasyOCR package is to provide an easy-to-use tool where one can plug-in already created state-of-art models and use them for annotating. Pipeline of EasyOCR behavior is shown in the image 2.1. As it can be seen in this image, default detection model is CRAFT and for recognition is used CRNN (Convolutional Recurrent

Neural Network) which model is composed of following components: feature extraction (Resnet is used) and VGG (Convolutional Neural Network), sequence labeling (LSTM is used) and decoding (CTC is used).[2]

EasyOCR package by default computes annotation on GPU, however there is a possibility for CPU computations (provided that the selected model supports it).

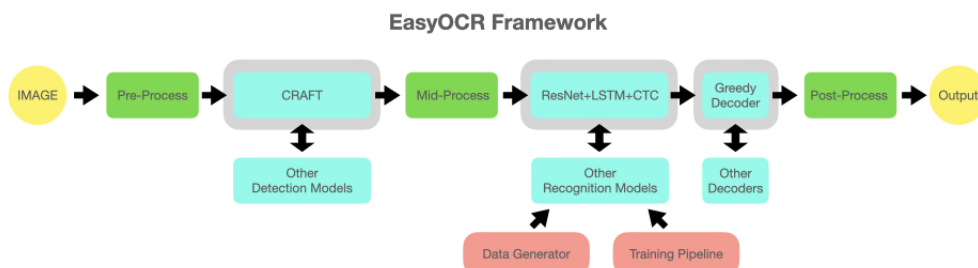


Figure 2.1: Diagram of EasyOCR pipeline. Grey slots are placeholders for models. The mentioned models are the ones used as default. [2]

2.3.3 Keras-ocr

keras-ocr is a python library used for detecting and recognizing text in images created by Fausto Morales. It works with variety of languages and with different writing scripts. It allows computing on CPU as well as on GPU. It unites the CRAFT text detection model and an implementation in Keras python library of CRNN for recognizing text, worth mentioning this is a different implementation of CRNN than in EasyOCR.[9]

On the official website¹ 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. [9]

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.[14] 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.[?] To use these models in the keras-ocr library one either doesn’t specify anything and

¹<https://pypi.org/project/keras-ocr/>

use the defaults, or pass the value `clovaai-general` for the CRAFT pretrained model or `kurapan` for the CRNN model.

Keras-ocr offers preprocessing for four public datasets though any text image dataset can be examined using this tool. These four datasets are: BornDigital dataset, COCO-Text dataset, ICDAR 2013 dataset, ICDAR 2019 dataset (only Latin-only scripts).[8]

2.3.4 Google Cloud Vision

Google Cloud Vision is software from Google which consist of two products: AutoML Vision and Vision API. Vision API detects objects, faces and text from images with already pretrained model. With AutoML Vision user can train custom model from own data. It has free and paid version with a GUI.[5] API for python

2.3.5 Other

AWS Rekognition

Chapter 3

Datasets

Optical Character Recognition requires data as any other machine learning task. Data are usually divided in two main types - scene and synthetic. Scene datasets contains photographs of real world objects and sceneries where some text occurs, for example shop signs, road signs or car plates. Synthetic dataset are automatically generated images where words are chosen from a extensive dictionary, a font is picked for each word and some sort of deformation is applied. It can be a text distortion to make the text curved or projectively altered, as well as blurring or lighting changes that make the text less obvious for the detector.

Synthetic datasets are usually used for training the models, because it is easier to generate millions of synthetic images rather than to take even one hundredth of a such number of photographs. Needless to say that when generating an image, the ground truth is known and can be saved during the generation process while photographs have to be manually labeled which takes time and might be inaccurate or automatically labeled which also often leads to many mistakes. Scene datasets are then used for testing purposes or for fine tuning a pretrained model.

In the following section sample datasets are introduced. To begin with synthetic dataset MJSynth is a very important dataset because it consists of almost 9 million images covering 90k English words. It includes data only for recognition which means that one image has only one word and border of the image represents the word bounding rectangle.[11]. Another synthetic dataset is called SynthText and contains 800 thousand images with approximately 8 million word instances written on the images [16]. Three scene text datasets were created for International Conference on Document Analysis and Recognition (ICDAR) competition. Sets ICDAR03, ICDAR13 and ICDAR15 were used in competitions in years 2003, 2013 and 2015, respectively. First two ICDAR datasets include only horizontal text, text of various orientation appears in set from 2015 [13]. Another widely used dataset is The Street View Text (SVT) which contains images with text harvested from Google Street View [15]. SVT and most other scene text datasets offers mainly frontal text with minimal perspective distortion. However, perspective text is frequent in real life applications of OCR for example previously mentioned street photographs, where it is impossible to capture every visible text from frontal view. Thus Phan et. al [12] created a new StreetViewText-Perspective derived from SVT, it shows the same

places as SVT but from different perspective. Another dataset CUT80 focuses on curved text as well as CTW1500 dataset. For text recognition there exist for example IIIT5k dataset containing 5000 cropped images harvested from Google image search. It combines both scene text images and born-digital images [6]. One of the most widely used dataset is COCO-Text

Datasets that were used for comparison of detection and recognition methods, namely, SCUT-CTW1500 dataset, Kaist Scene Text Database, Born-Digital Images, are described in more detail below.

3.1 SCUT-CTW1500 dataset

SCUT-CTW1500 dataset contains exactly 1500 images of real-world, scene text in English language. Sample images can be seen in Figure 3.1. The key feature of this dataset is that each image contains horizontally aligned text, multioriented text and curved text. There are cases where the curvature is only slight and cases where text forms a circle with letter upside down. Recognizing multi-oriented and curved text is more of a challenge than pure horizontal text. This dataset is split to train and test data. Two thirds of dataset thus one thousand images for training and five hundred for testing. According to the description of this dataset on relevant GitHub repository dataset was manually labeled and lately corrected, therefore labels seem to be very accurate. However for example ground truth for image 1313.jpg misses all occurrences of letter I, as the depicted font was probably misread.[18, 10]

The ground truth for train data are in XML format and each file carries information about the file name of respective image file, text information – i.e., words in a text line, 14 coordinates of a bounding polygon and coordinates, height and width of a circumscribed rectangle. Later the authors added coordinates of center point of each English letter to be used as detection ground truth. The ground truth of test data is in simple text file (TXT) and contains only 14 coordinates of the bounding polygon and a text which is within that region. There is a minor issue with labels that it usually contains a full text line with multiple words and coordinates are not assigned to individual words but to text region as whole. Most end-to-end system detect words rather than groups of corresponding words. This fact needs to be taken into account when evaluating results.

3.2 KAIST Scene Text Database

This dataset contains 3000 images of photographed text. It can be divided into three major categories – text of Korean language, English language and mixed languages. As I concentrate on text in latin script in this paper further information relates to the English language dataset. The number of images is then reduced to less than four hundred images. Figure 3.2 shows few samples of this dataset. Photographed objects are mostly shop banners or parts of magazine front pages. Photographs were either taken by a high-resolution digital camera or a low-resolution mobile phone



Figure 3.1: Sample images of SCUT-CTW1500 dataset.

camera.[7] Each photography has a ground truth description and a bitmap image. In the bitmap file only text is highlighted (by white or red color) and everything else apart from text is set as black. Ground truth files are in XML format and includes a name of an image, its resolution and bounding box for each word and also a bounding box for each letter of the word.

To use this dataset for testing and training the XML ground truth needed to be converted to string and int values. I wrote a parser, that combines letters to form a word that is within a given bounding box. I changed the notation of bounding boxes from one coordinate, width and height attributes to two top left and bottom right coordinates.

Unfortunately this dataset has few errors in filenames of corresponding files or in the content of XML files. Usually these are only typos, however they prevent automatic preprocessing of dataset. Due to this problem these mistakes need to be found and manually corrected. Also there is a small number of ground truth XML file with fully missing data. Despite these shortcomings this dataset is useful because of the bitmap files. This allows to compare results of both images affected by shooting conditions and images dependent only on font and position.



Figure 3.2: Sample images of KAIST Scene Text Database dataset.

3.3 Born-Digital Images

Born-Digital Images contains data of images with text that can be found on various websites. Samples of this dataset can be found in Figure 3.3. There are mostly advertisements, company logos or website headers. Such pictures cannot be classified neither as real scene dataset, neither as synthetic one. On one hand this dataset shares with scene datasets the variability in font styles and sizes, different text orientations and complex colour placement. On the other hand it differs in size because low resolution is significant in smooth and fast loading on websites. Also no noise is present due to lighting conditions. Geometrical deformations that result when capturing a real scene with camera also do not appear here. However compression to lower resolution can lead to artefacts and aliasing. In general we can say that letters are more clearly visible than in photographed text as easy readability is crucial in successful advertising.[4]

The dataset is available for download from the website of Robust Reading Competition. First version was published in 2011 and revised two years later, it contains separate dataset for text localization, segmentation and then for word recognition. In 2015 they published an end-to-end dataset with ground truth for all tasks. The dataset is split in training and testing data. However, ground truth for testing data contains only a possible vocabulary of words in images and no coordinates. This might be due to the fact that the competition might be still ongoing or there was not a sufficient demand for complete ground truth. As for training data, each image has a corresponding TXT file with coordinates of four vertices of bounding rectangle and a word. Text lines are separated and the text within rectangle is always one word. Unfortunately, there are quite a few missing words, usually words that have two or less characters. This can affect the evaluation when the model finds such a short, missing word.[4]



Figure 3.3: Sample images of Born-Digital Images dataset.

3.4 Wien TU dataset

Chapter 4

Testing methods

4.1 EasyOCR

4.2 Tesseract

Potreba cernobílé a threshold tesseract to má radši, problém dělat centralně - otsu a po částech ale stejné težké u spousty obrázků s různým osvětlením. Tesseract zkousíme u bitmap obr PSM 11 na Kaist a je o hodně horší než PSM 6, s 11 je to 9.6. proc. jinak u normálních obrázků je lepší zase 11, taky třeba o 9 procent rozdíl.. pak ještě 4 to je občas lepší U CTW datasetu je to s 6 horší.

ještě hrani s barvama nebo upravený , jak kde

verze tesseract 4.0.0-beta.1 leptonica-1.75.3 libgif 5.1.4 : libjpeg 8d (libjpeg-turbo 1.5.2) : libpng 1.6.34 : libtiff 4.0.9 : zlib 1.2.11 : libwebp 0.6.1 : libopenjp2 2.3.0

Found AVX2 Found AVX Found SSE

4.3

Bibliography

- [1] JAIDED AI. **Jaied AI - Distribute the benefits of AI to the world**. Last accessed 20 May. 2022. [Online]. Available from: <https://www.jaied.ai/>.
- [2] JAIDED AI. **Jaiedai/EasyOCR**. Last accessed 13 Jun. 2022. [Online]. Available from: <https://github.com/JaiedAI/EasyOCR>.
- [3] YOUNGMIN BAEK, BADO LEE, DONGYOON HAN, SANGDOO YUN, AND HWALSUK LEE. **Character region awareness for text detection**. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9365–9374, 2019.
- [4] ROBUST READING COMPETITION. **Overview - born-digital images (web and email)**. Last accessed 21 Jun. 2022. [Online]. Available from: <https://rrc.cvc.uab.es/?ch=1>.
- [5] **Vision AI — derive image insights via ML — cloud vision API google cloud**. Available from: <https://cloud.google.com/vision/>. [Online].
- [6] **The IIIT 5K-word — Graviti**. Last accessed 22 Jun. 2022. [Online]. Available from: <https://gas.graviti.com/dataset/graviti/IIIT5KWord>.
- [7] **Kaist scene text database**. Last accessed 10 Jun. 2022.. [Online]. Available from: http://www.iapr-tc11.org/mediawiki/index.php/KAIST_Scene_Text_Database.
- [8] **keras-ocr documentation**. Last accessed 29 Apr. 2022. [Online]. Available from: <https://keras-ocr.readthedocs.io/en/latest/index.html>.
- [9] **keras-OCR**. Last accessed 29 Apr. 2022. [Online]. Available from: <https://pypi.org/project/keras-ocr/>.
- [10] YULIANG LIU, LIANWEN JIN, SHUAITAO ZHANG, CANJIE LUO, AND SHENG ZHANG. **Curved scene text detection via transverse and longitudinal sequence connection**. *Pattern Recognition*, **90**:337–345, 2019.
- [11] **Text recognition data - Visual Geometry Group - University of Oxford**. Last accessed 21 Jul. 2022. [Online]. Available from: <https://www.robots.ox.ac.uk/~vgg/data/text/#sec-synth>.

- [12] TRUNG PHAN, PALAIAHNAKOTE SHIVAKUMARA, SHUANGXUAN TIAN, AND CHEW LIM TAN. **Recognizing Text with Perspective Distortion in Natural Scenes**. pages 569–576, 12 2013.
- [13] ZUBEIR RAISI, MOHAMED A NAIEL, PAUL FIEGUTH, STEVEN WARDELL, AND JOHN ZELEK. **Text detection and recognition in the wild: A review**. *arXiv preprint arXiv:2006.04305*, 2020.
- [14] CLOVA AI RESEARCH. **Clovaai/Craft-pytorch: Official implementation of character region awareness for text detection (CRAFT)**. Last accessed 29 Apr. 2022. [Online]. Available from: <https://github.com/clovaai/CRAFT-pytorch>.
- [15] **The street view text dataset**. Last accessed 22 Jul. 2022. [Online]. Available from: <http://vision.ucsd.edu/~kai/svt/>.
- [16] **Synthtext in the wild dataset - Visual Geometry Group - University of Oxford**. Last accessed 21 Jul. 2022. [Online]. Available from: <https://www.robots.ox.ac.uk/~vgg/data/scenetext/>.
- [17] TESSERACT-OCR. **Tesseract-OCR/tessdoc: Tesseract documentation**. Last accessed 6 May. 2022. [Online]. Available from: <https://github.com/tesseract-ocr/tessdoc>.
- [18] YULIANG-LIU. **Yuliang-Liu/curve-text-detector**. Last accessed 13 Jun. 2022. [Online]. Available from: <https://github.com/Yuliang-Liu/Curve-Text-Detector>.