CORE147

DIGITÁLNÍ DATA V HUMANITNÍCH A SOCIÁLNÍCH

Zuzana Nevěřilová
Hana Žižková

2024/25

POČÍTAČOVÉ VIDĚNÍ

Aktivity

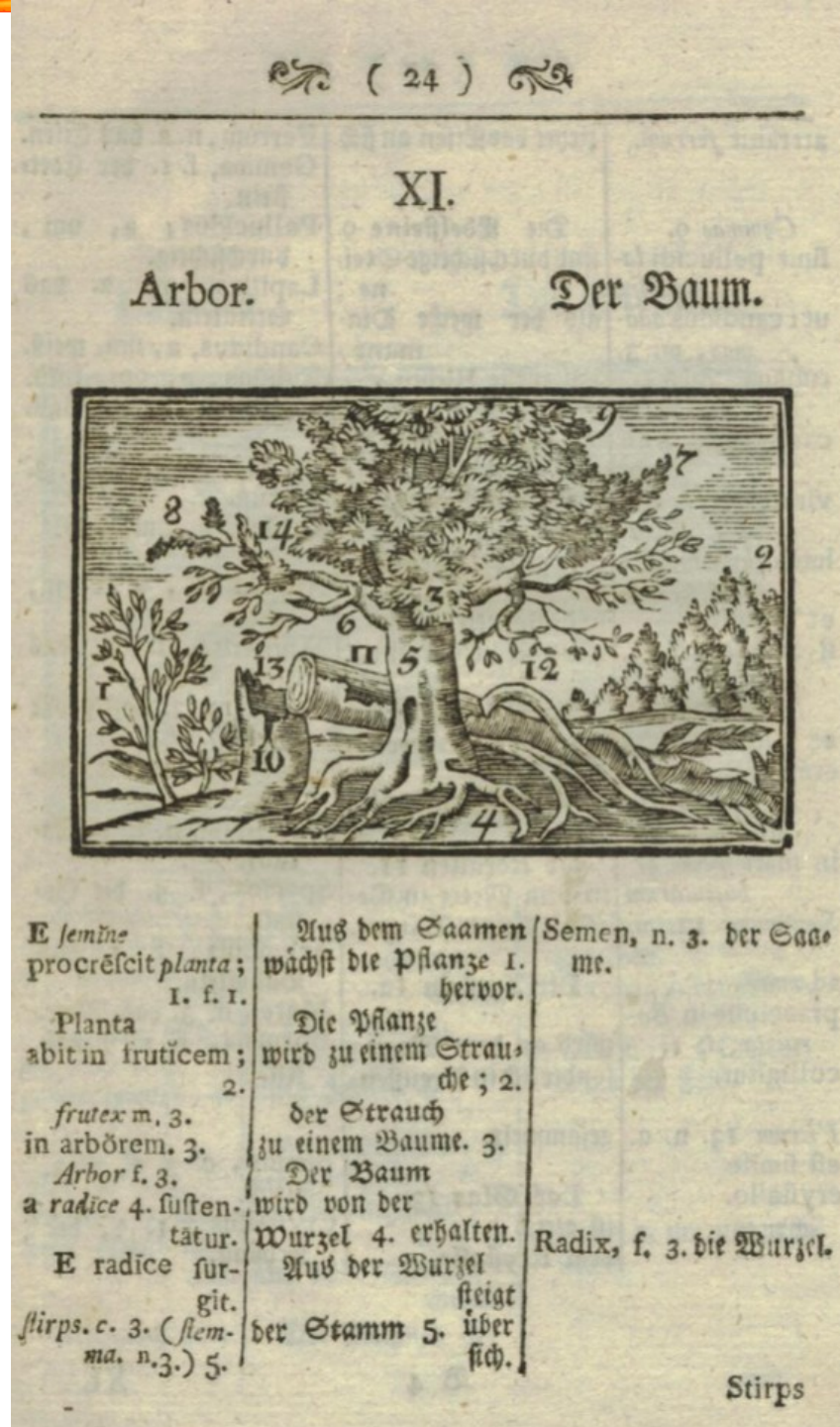
- Detekce objektů
- Popis scény
- Detekce stylu (fotografie, kresba)
- Detekce znaků (OCR)
  - Skeny
  - OCR in wild
- Rozpoznání místa (reverse image search)

Aplikace

- Stejné jako pro textový obsah (např. fulltextové hledání)
- Metadata (popis obrázků)
- Rozpoznání anomálií
- Rozšířená realita
- Součást storytellingu

ZPRACOVÁNÍ OBRAZU

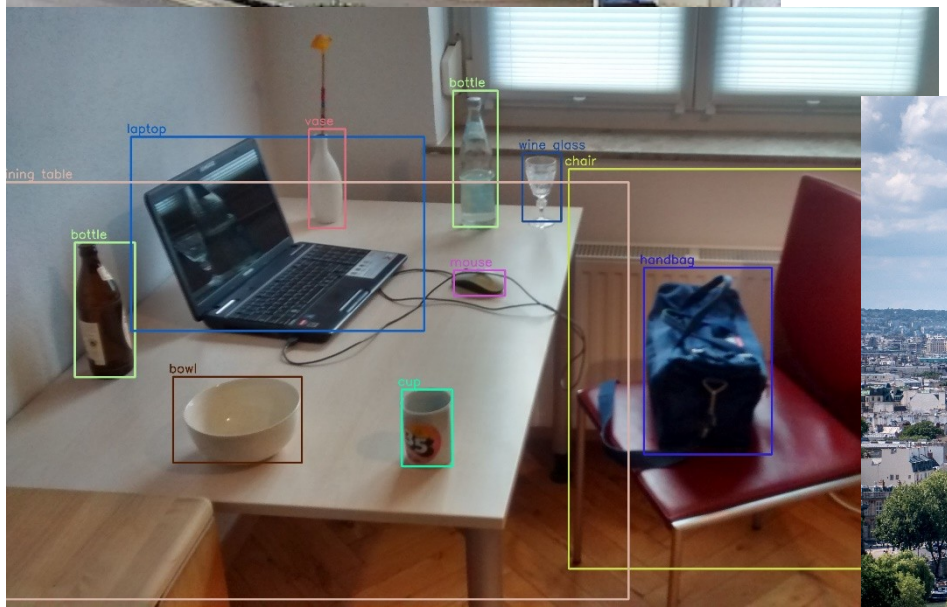
- Získat informaci z obrazu
- Vylepšit (vyčistit) obraz
- Přidat související informace (propojení s jinými objekty, obohacení metadaty)

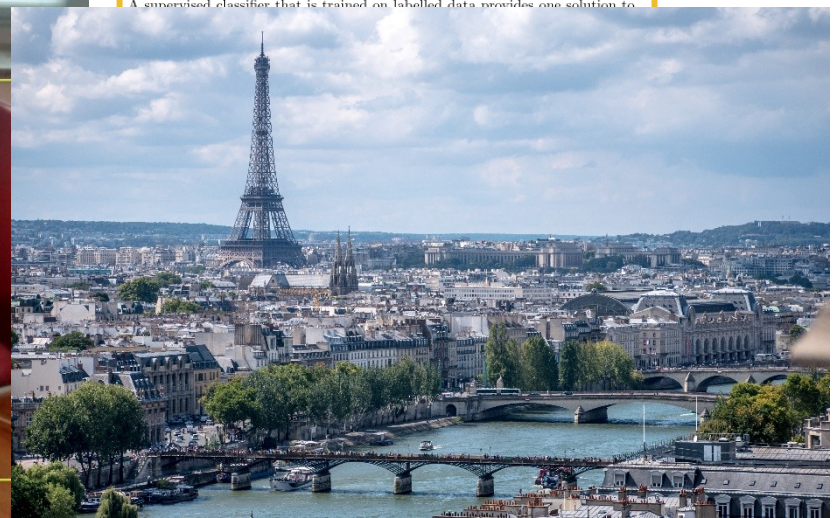


Comenius, J.A., 1778. Orbis pictus : die Welt in Bildern, in zwey und achtzig Abschnitte zum Gebrauche der kleinsten studirenden Jugend in den kaiserl. königl. Staaten zusammengezogen / Johann. Amos Comenii. Trattner, Wien.
https://doi.org/10.24355/dbbs.084-200510210200-0
https://publikationsserver.tu-braunschweig.de/receive/dbbs_mods_00000250?q=orbis%20pictus

INFORMACE V OBRAZE

- Kde je text?
- Co se tam píše?
- Jaký je layout?
- Co je na obrázku?
- Odkud byl obrázek pořízen?
- Kdy byl obrázek pořízen?





A Supervised Learning Approach For Heading Detection

Sahib Singh Budhiraja and Vijay Mago
(sbudhira,vmago)@lakeheadu.ca
Lakehead University, 955 Oliver Rd, Thunder Bay, ON P7B 5E1

Abstract. As the Portable Document Format (PDF) file format increases in popularity, research in analysing its structure for text extraction and analysis is necessary. Detecting headings can be a crucial component of classifying and extracting meaningful data. This research involves training a supervised learning model to detect headings with features carefully selected through recursive feature elimination. The chosen classifier has an accuracy of 95.83%, sensitivity of 0.981 and a specificity of 0.946. This research into heading detection contributes to the field of PDF based text extraction and can be applied to the automation of large scale PDF text analysis in a variety of professional and policy based contexts.

Keywords: Heading Detection · Text Segmentation · Supervised Approach.

1 Introduction

As the amount of information stored within PDF documents increases worldwide, the opportunities for large scale text based analysis requires increasingly automated processes, as the amount of document processing is time consuming and labour intensive for human professionals. Systematic processing and extraction of textual structure is increasingly necessary and useful as demonstrated in El-Haj et al.'s work involving 1500 financial statements[7]. Categorizing data into seperate sections is quite easy for humans, as they rely on visual cues such as headings to process textual information. Machines, despite being able to process large amounts information at high speeds, require effort to classify and interpret text based data. This paper explores the application of supervised classifiers to operationalize a system that would aid in the identification of headings. PDF documents are a visually exact digital copy that displays text by drawing characters on a specific location [10] and present a challenge in analysis because the files do not provide enough information on how the text is organized and formatted.
A supervised classifier that is trained on labelled data provides one solution to

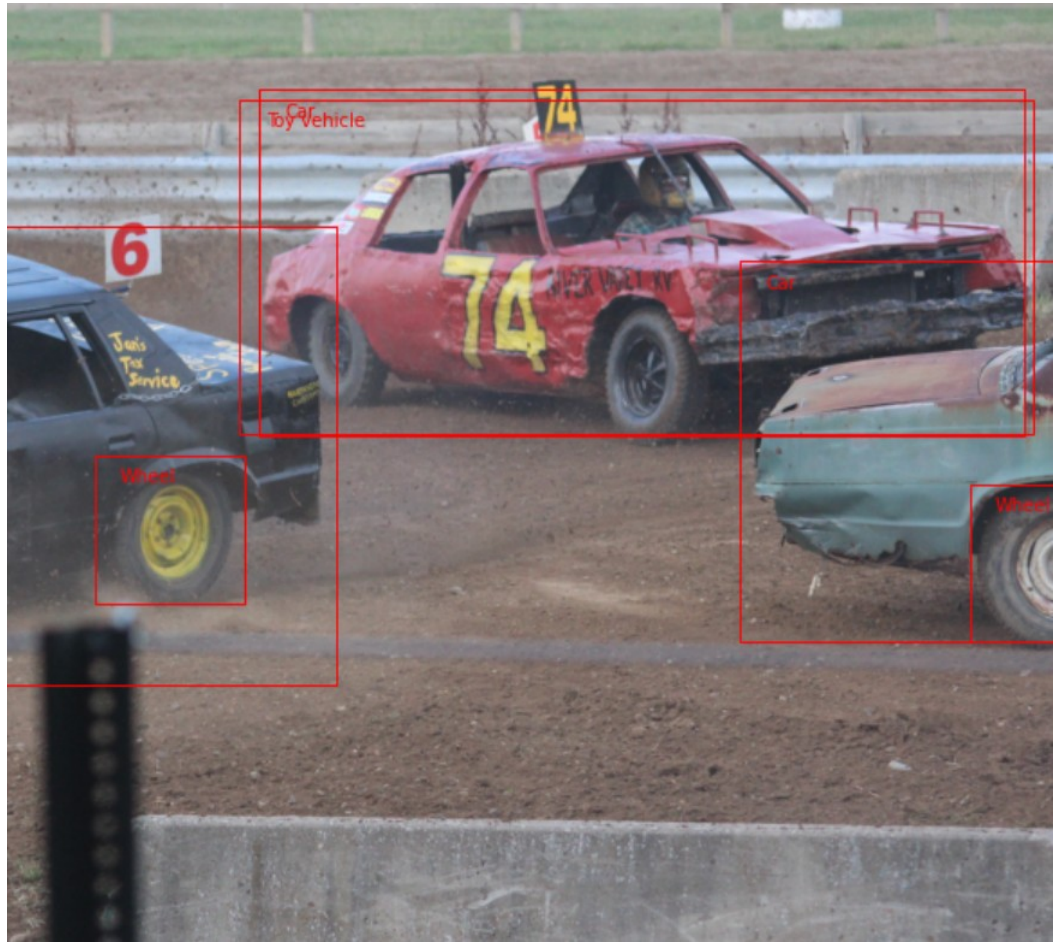
EXISTUJÍCÍ NÁSTROJE

Cloudové služby

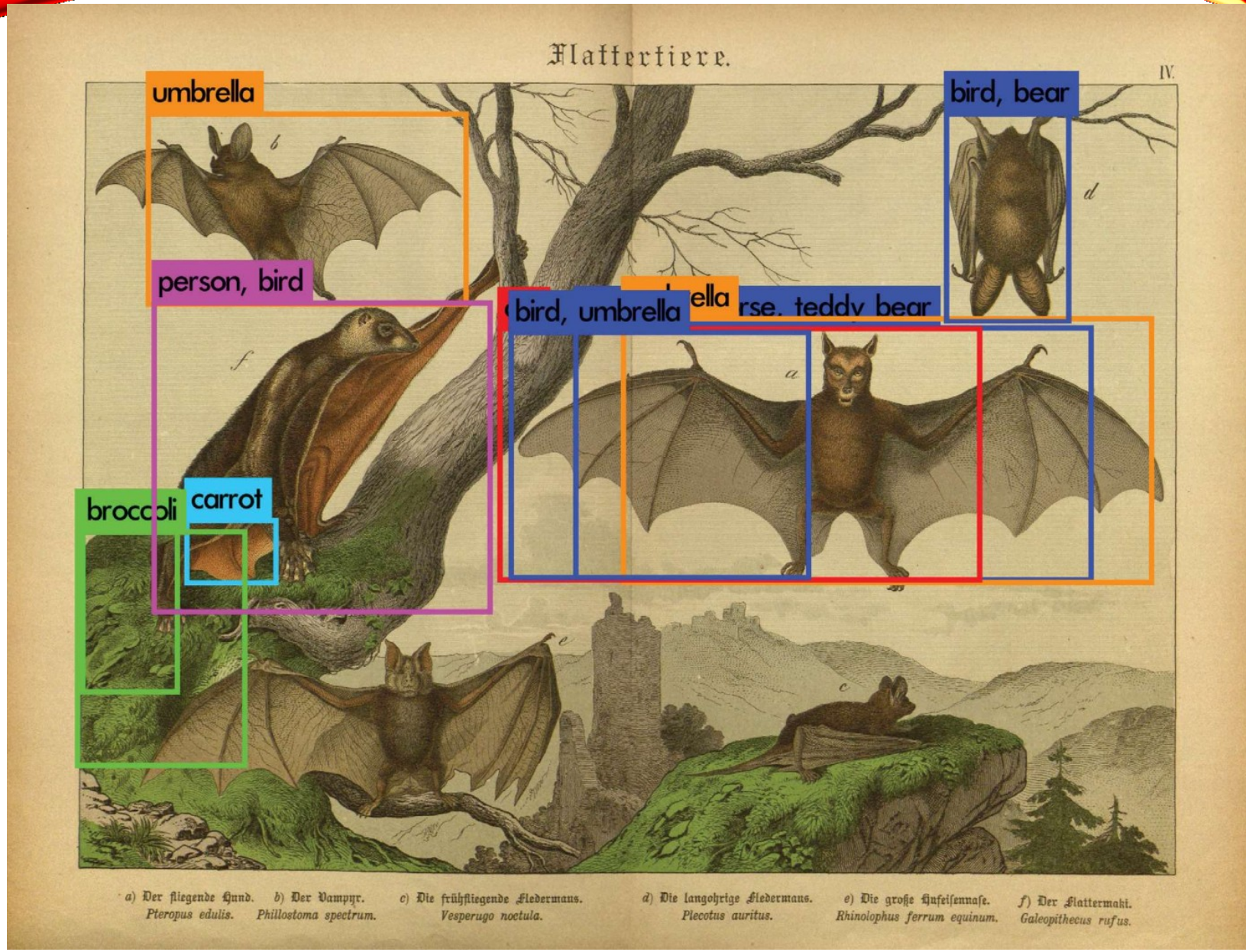
- Google Vision API
- Microsoft Computer Vision
- Amazon Rekognition
- Clarifai
- IBM Watson Visual Recognition

```
client = vision.ImageAnnotatorClient()

google_image = vision.Image()
google_image.content=image_as_bytes(image)
objects = client.object_localization(
        image=google_image).
        localized_object_annotations
```



PROBLÉMY S EXISTUJÍCÍMI NÁSTROJI



Thomas Mandl, Im Chanjong, Sebastian Schmideler, Wiebke Helm: *Automatic image processing in the Digital Humanities: A pre-study for Children`s Books in the 19th Century*. In ISCHE 40 pre-conference workshop. 3rd workshop "Pictura Paedagogica Online: educational knowledge in images" ISCHE International Standing Conference for the History of Education. 2018

MINISTER
SPRAW WOJSKOWYCH

Warszawa d. 31 maja 1926 r.

39

Panie Marszałku!

Dziękuję Zgromadzeniu Narodowemu za wybór. Po raz drugi w mem życiu mam w ten sposób zalegalizowanie moich czynności i prac historycznych, które - niestety dla mnie - spotykały się przedtem z oporem i niechęcią obecnej większości. Tym razem dziękuję wszystkim Panom, że wybór mój nie był jednomyślnym tak, jak to było w Listopadzie 1919 roku. Mniej może będzie w Polsce zdrad i faszu.

Niestety, przyjąć wyboru nie jestem w stanie. Nie mogłem wyzwolić w sobie zapomnienia,

ROZPOZNÁNÍ RUKOPISU

- Různá písma
- Osobní písmo
- Velká variabilita

MINISTER
SPRAW WOJSKOWYCH

Warszawa dn. 31 maja 1926 r.

39

Panie Marszałku!

Dziękuję Zgromadzeniu Narodowemu za wybór.

Po raz drugi w mem życiu mam w ten sposób zale-
galizowanie moich czynności i prac historycznych,
które niestety dla mnie spotykały się przedtem
z oporem i niechęcią słowy srebrnemi. Tym razem
dziękuję wszystkim Panom, że wybór mój nie
był jednomyślnym tak, jak to było w chwili
1919 roku. Mniej może gorzkie w dolnej zdrowej
i jaśniej.

Niestety, przyjąć wybór nie jestem w stanie.
Nie mogłem wyzwolić w sobie zapomnienia,

ROZPOZNÁNÍ RUKOPISU

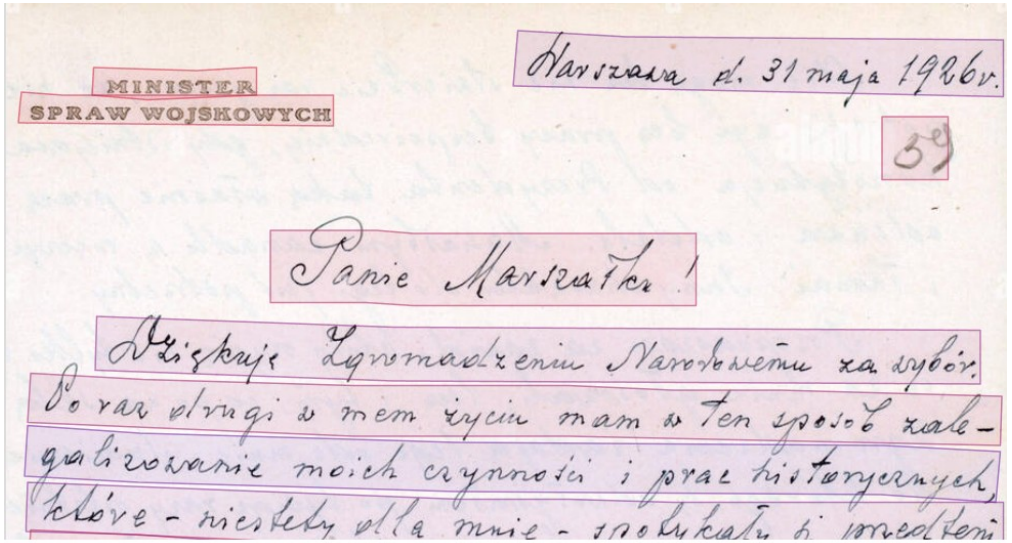
- Co je jeden řádek?
- Co je písmo a co šum?
- Kombinace tištěného a rukou psaného textu

SOFTWARE PRO ROZPOZNÁNÍ RUKOPISU

HANDWRITTEN TEXT RECOGNITION (HTR)

- Transkribus
- Pero OCR (VUT Brno)
- Pen2Text
- Google Cloud Vision (umí i jiná písma než latinku)
- Amazon Textract
- PyLaia – knihovna pro Python (lze trénovat)

PERO A PEN2TEXT


MINISTER
SPRAW WOJSKOWYCH

Warszawa dn. 31 maja 1926r.

59

Panie Marszałku!

Dziękuję Zgromadzeniu Narodowemu za sybór.
Po raz drugi w mem życiu mam a ten sposób zale-
galizowanie moich czynności i prac historycznych,
które - niestety dla mnie - polegały na przedtem i

FLINISTER
SPRAVY WVOJSKOWYCEA
Warszaka d. 31 maja 1926v.
59)
Panie Marsaška!

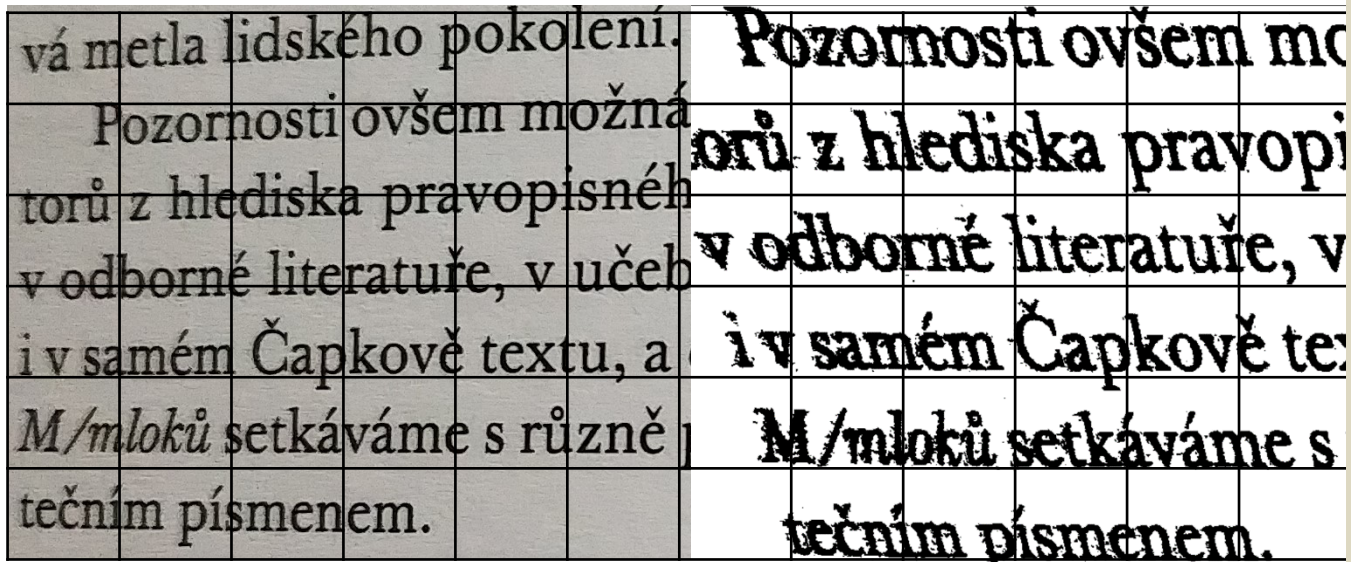
Harsrana, dnia 31 maja 1926 roku. MINISTER SPRAW WOJSKOWYCH Panie Marseater! Dziękuję Zgromadzeniu Narodowemu za sybor. Do wielu druków a mem życiu mam a ten sposób walidacji moich czynności i prac historycznych, niestety – głównie dla mnie – polegały na przedtem i aporém; mechanik dorgó nerola. Tym razem daję wszystkim Panom znać, że Asfor możliwych nic być jednomyślne z Trym Rak, jak to było i w 1919 roku notowane. Mniej, bardziej berkie a Volver adwad i faTien. Niestety, pomimo tego nie jestem w stanie. Nie mogę appalenié we sobie zapomnienia, nie mogę wydobyć z siebie atcru zanfaria i do siebie a fej pracy. Ichóra już van krypiTemi, ani też do tych, co mnie na ten unoo

PŘEDZPRACOVÁNÍ OBRÁZKŮ

Napravení deformací:

- descrew, dewarp

- Snížení šumu
- Úprava barev (jedna barva-jeden význam)



https://www.topocr.com/extract.html

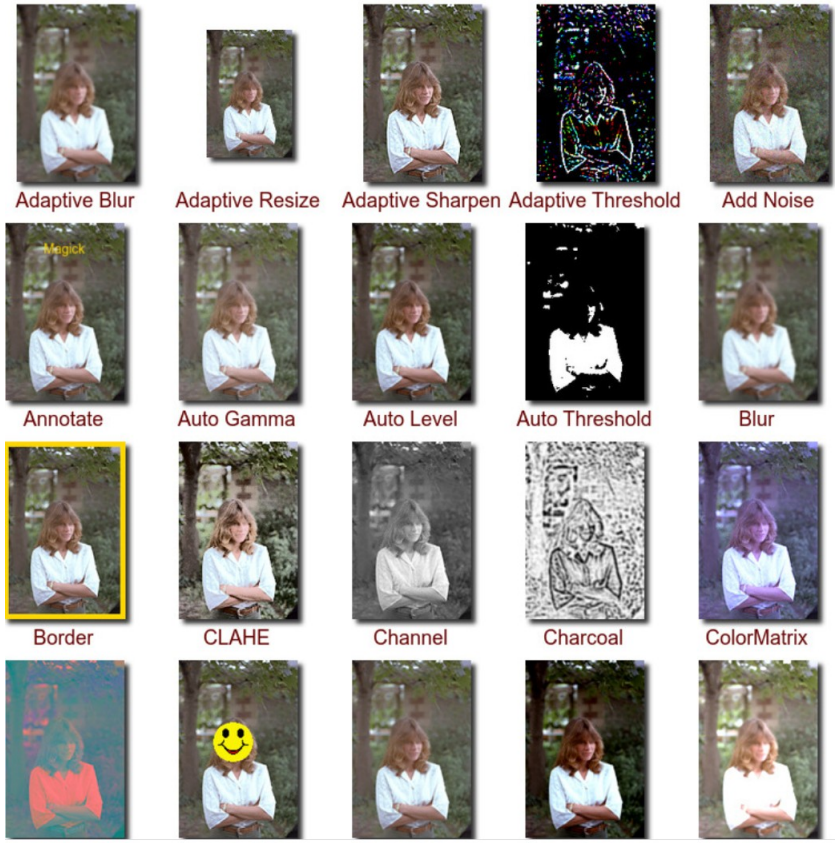
LZE OBRÁZKY VYLEPŠIT?

Obrázky s textem (před-OCR):

- Manuálně: grafické programy
- Automaticky: neuronové sítě

Grafické programy:

- Vizuální (a zdarma): GIMP
- Příkazová řádka: ImageMagick
- Program v Pythonu: balíčky PythonMagick, scikit-image

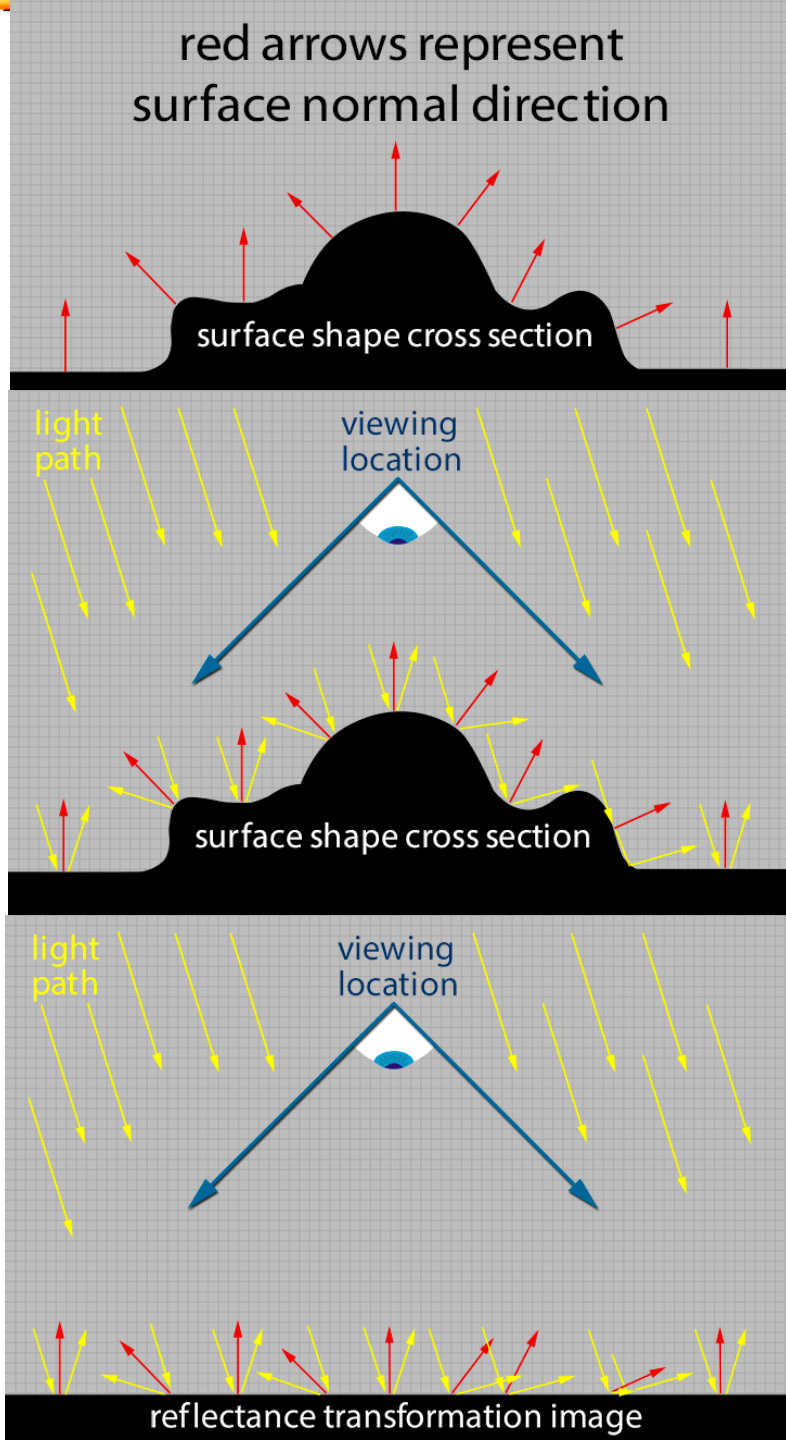


https://imagemagick.org/script/examples.php
https://scikit-image.org/docs/stable/auto_examples/

SPECIÁLNÍ SNÍMKY

Reflectance Transformation Imaging (RTI)

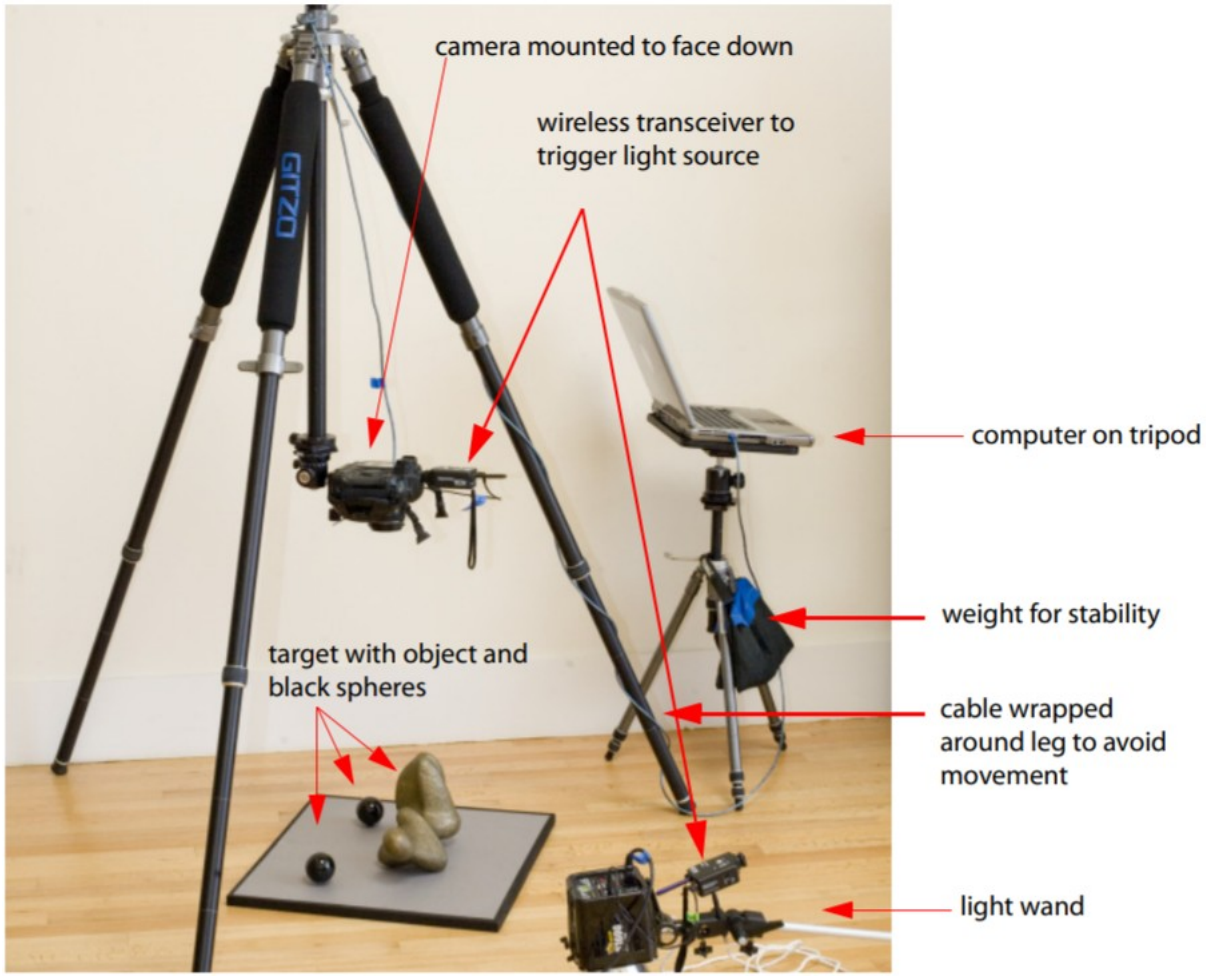
- Zobrazí 3D pomocí 2D a speciálního software



Jak funguje Reflectance Transformation Imaging (RTI)

- Využití obvyklého vybavení (fotoaparát, stativ, osvětlení)
- Použití **dvou odrazivých koulí** (známé velikosti)
- Porovnání odrazu na kouli s odrazem na fotografovaném objektu: ***směr dopadajícího světla***
- Modelování v další vrstvě obrazu

http://culturalheritageimaging.org/Technologies/RTI/



HROMADNÉ ZPRACOVÁNÍ

The Vogue Collection

- Statistické informace
- Barevné histogramy
- Klastrování podle fotografů a modelek
- Vývoj technologií ve fotografii i polygrafii
- Identita značky (Brand identity)

http://dh.library.yale.edu/projects/vogue/

VOGUE ARCHIVE

Every issue. Every page. 1892 to today.

ODBOČKA K OSINT

OPEN SOURCE INTELLIGENCE

Volně dostupná data

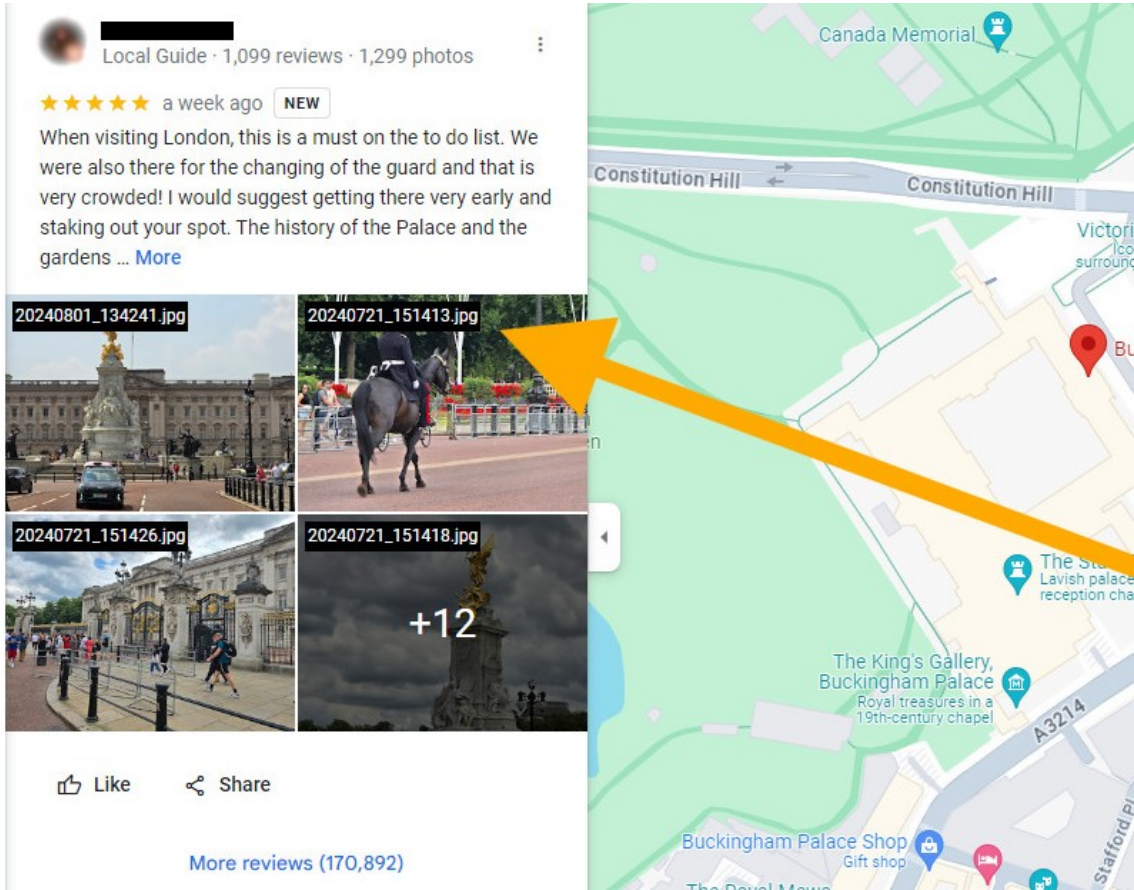
- Klasická média (radio, TV, noviny)
- Internetová média (blogy, diskuzní skupiny, sociální média, občanská média)
- Data veřejné správy
- Akademické publikace (články, kvalifikační práce, konference)
- Komerční data (finanční zprávy)
- Grey literature (technické zprávy, preprinty, patenty, newletters)

Pokročilé nástroje

- Mapy, satelity, geolokace
- Obraz, video
- Lidé
- Weby
- Společnosti a finance
- Doprava
- Životní prostředí, příroda
- Archivace
- Organizace a analýza dat

NÁSTROJE BELLINGCAT
BCATTOOLS

- Zpětné hledání v obrazech (reverse image search)
  - Google Lens
  - RootAbout
  - VISE
- Rozpoznání obličejů (facial recognition)
  - PimEyes
- Metadata v obrazech (image metadata)
  - Filename Finder
- Jiné
  - GeoHints (reklamy, značky, osvětlení, řízení, čísla domů, semafory, ...)
  - Forensically (detekce úprav obrazu)



PRO INSPIRACI

- pictura paedagogica online http://opac.bbf.dipf.de/virtuellesbildarchiv/
- Children's book collection https://www.tu-braunschweig.de/en/ub/search/special-collections/childrens-book-collection
- Digital Humanities at Yale University Library http://dh.library.yale.edu/projects/vogue/
- E-Science and Ancient Documents http://esad.classics.ox.ac.uk/
- ImageNet http://www.image-net.org/