Obrazová data a strojové učení

Petr Žabička, Ján Bogár, Michal Tran - Moravská zemská knihovna v Brně

Obsah

- Úvod
- Projekt PERO
- Identifikace obrázků na stránce
- Systém VISE

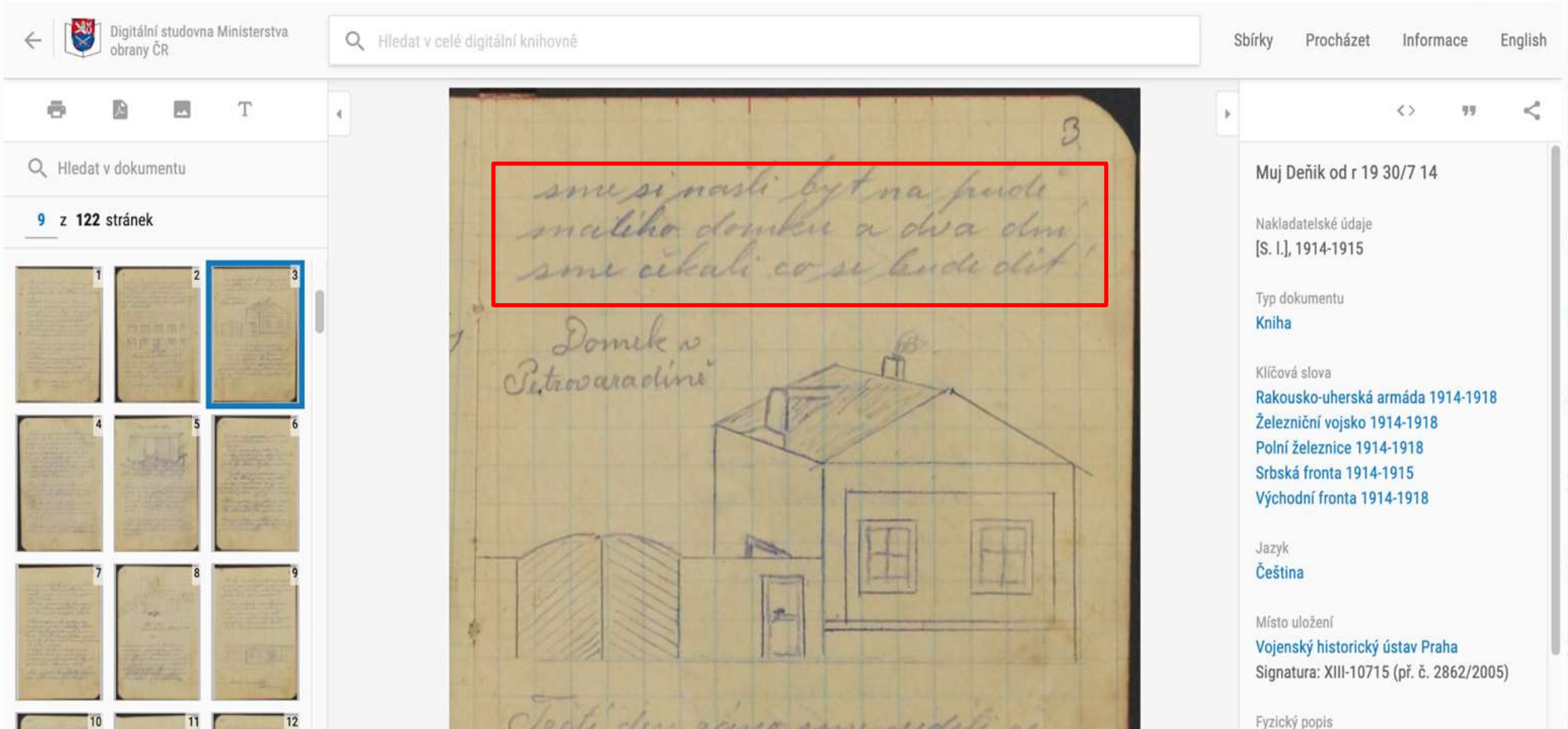
Strojové učení a proces digitalizace

- Ořez, vyrovnání stránky, barevné podání
- Scelování dokumentů
- Struktura dokumentu
- OCR
- ...

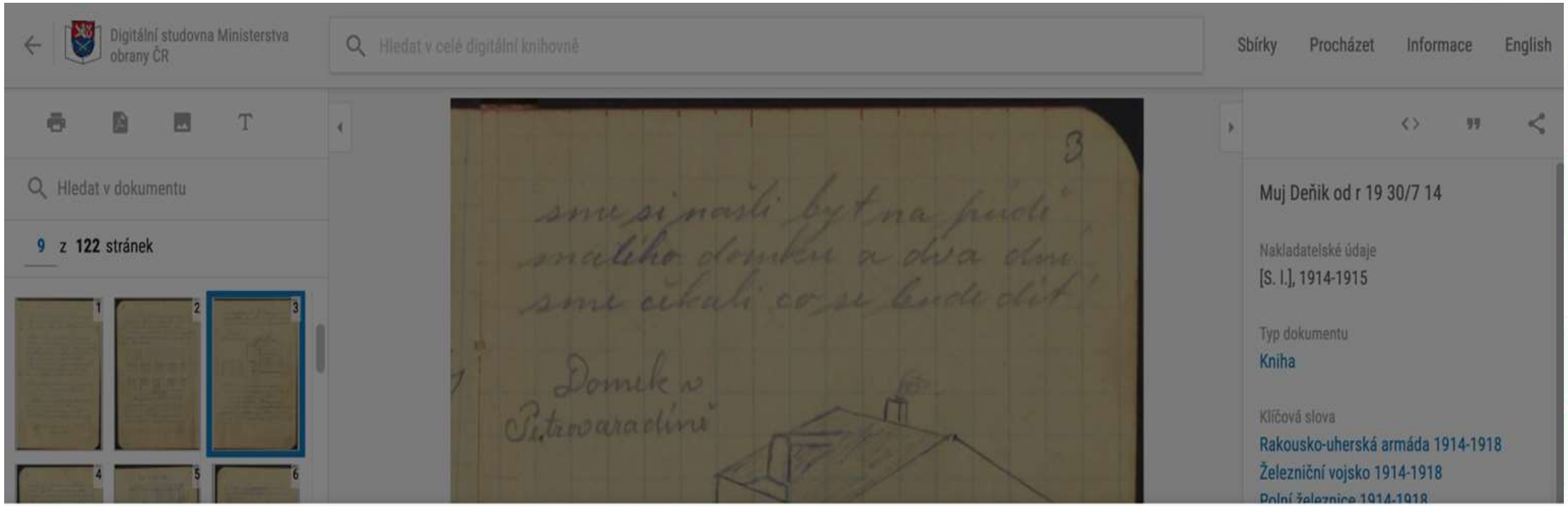
Strojové učení a zpřístupňování

- Vyhledávání obrázků
- Vyhledávání a podobnost částí stránek
- Identifikace obsahu obrázku
- ...

OCR - projekt PERO



OCR - projekt PERO

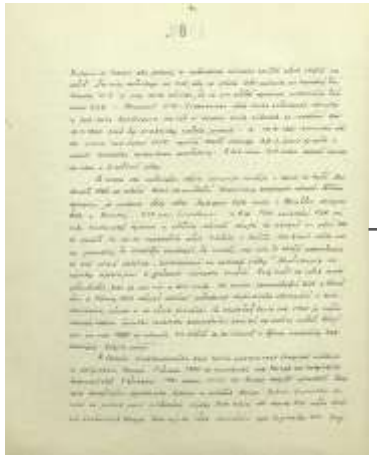


sme si našli byt na půdě malého domku a dva dní sme čekali co se bude dít!

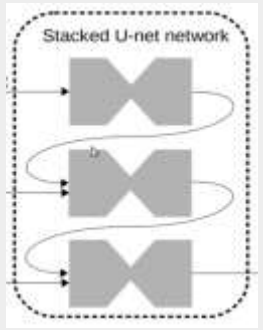
Muj Deňik od r 19 30/7 14. [S. l.], 1914-1915, s. 3. Dostupné také z: http://www.digitalniknihovna.cz/dsmo/uuid/uuid:32f7ba5e-a323-11ea-95c0-001b63bd97ba

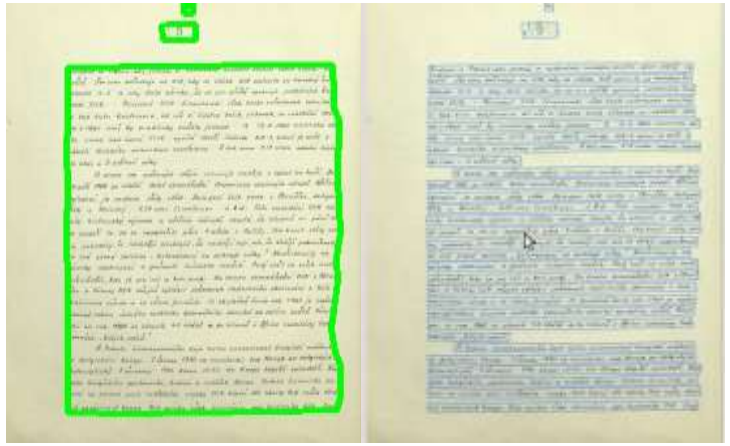
OCR - projekt PERO

Řetězec procesu rozpoznání textu



Layout





Crop line

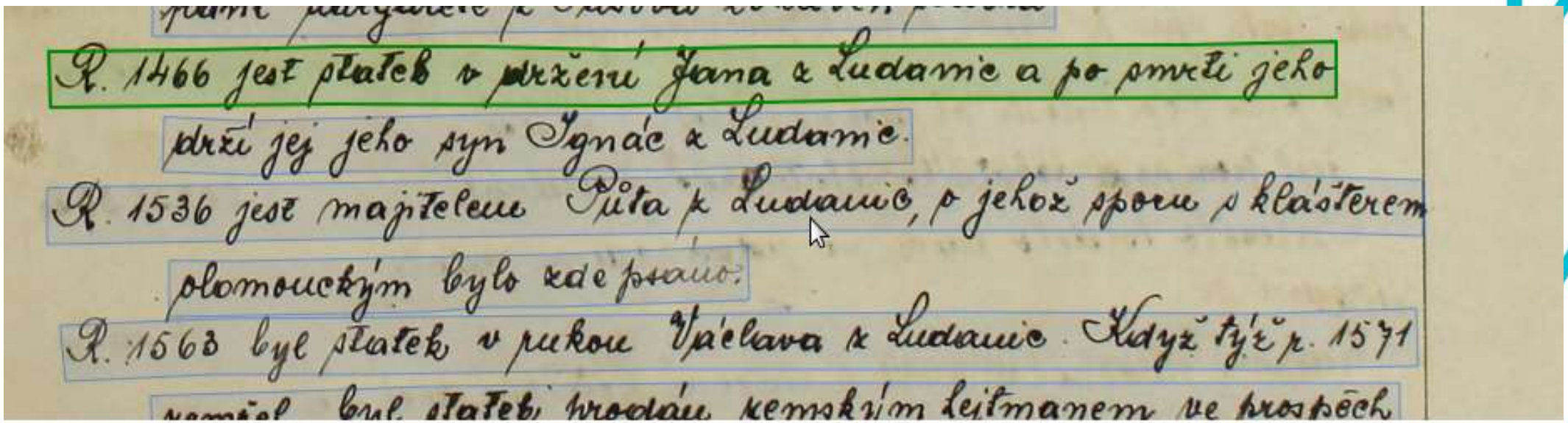
OCR network

Language model

Roku 1906 se stal předsedou spolku přátel školy a podílel se na financování a výstavbě nové školní budovy v Ratibořech nad Pekárkou. Ze své funkce byl odvo-

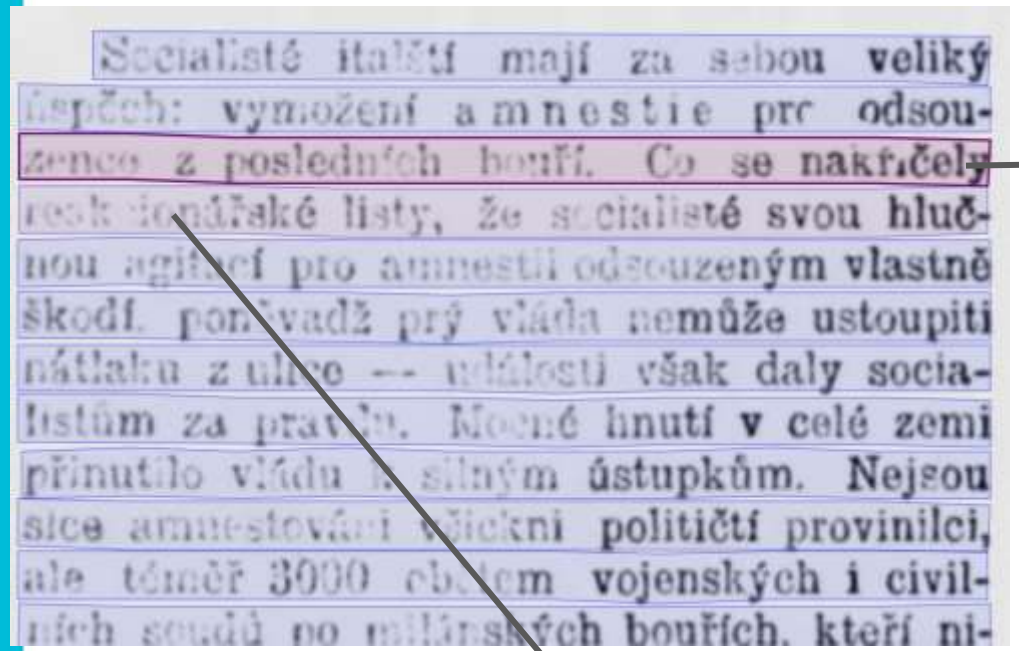
Web application

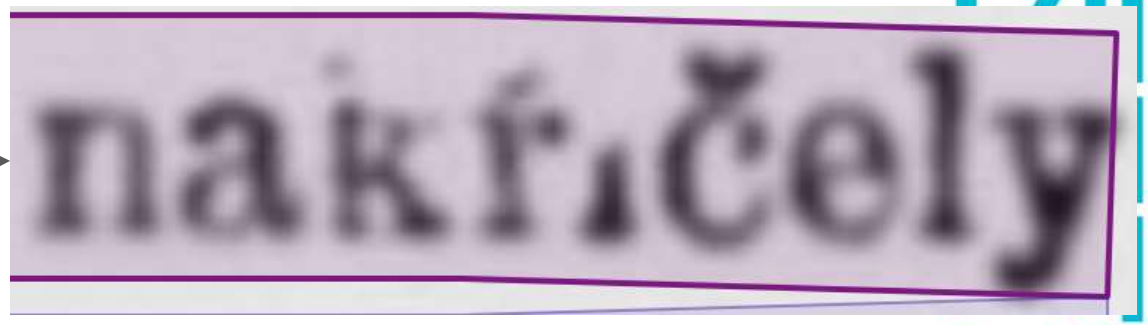
České kroniky 20. století



R. 1466 jest statek v držení Jana z Ludanic a po smrti jeho
drží jej jeho syn Ignác z Ludanic.
R. 1536 jest majitelem Půta z Ludanic, o jehož sporu s klášterem
olomouckým bylo zde psáno.
R. 1563 byl statek v rukou Václava z Ludanic. Když týž r. 1571

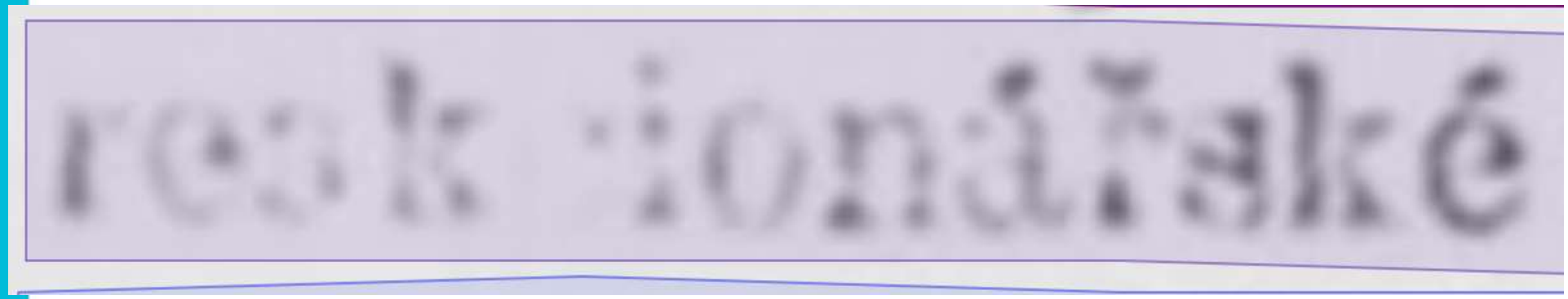
Periodika na mikrofilmech



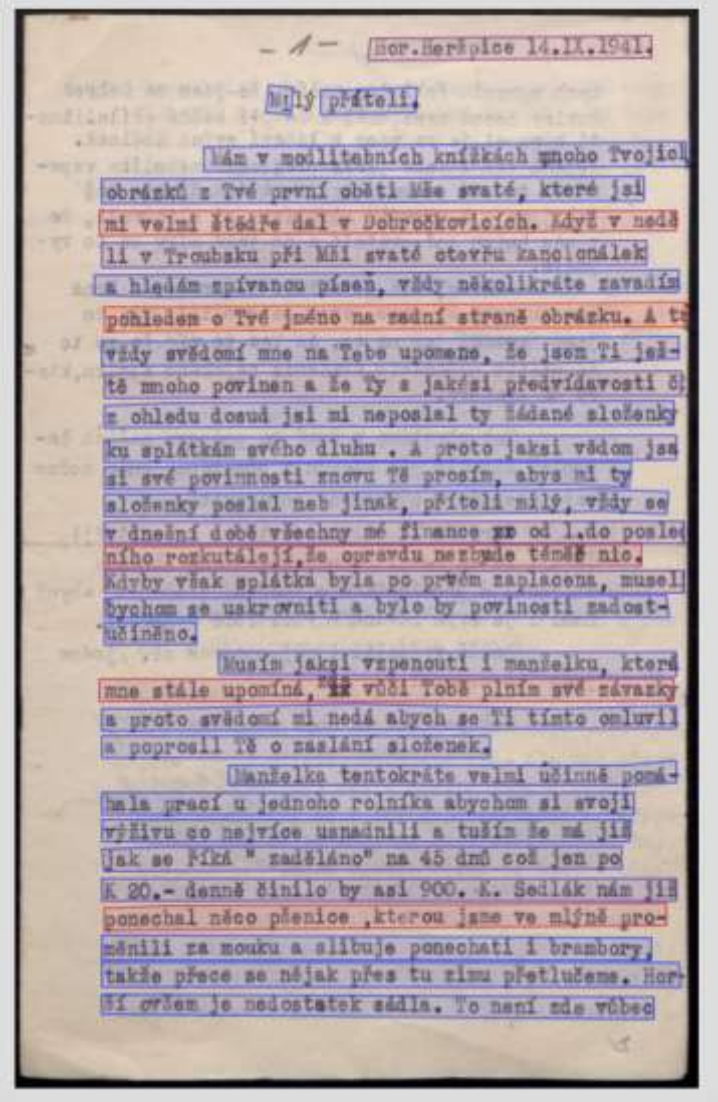


zence z posledních bouří. Co se nakřičely

reskcionářské listy, že socialisté svou hluč-



Strojopis



- 1 - Hor.Heršpice 14.IX.1941.

Milý příteli,

Mám v modlitebních knížkách mnoho Tvojich obrázků z Tvé první oběti Mše svaté, které jsi mi velmi štědře dal v Dobročkovicích. Když v neděli v Troubsku při Mši svaté otevřu kancionálek a hledám zpívanou píseň, vždy několikráte zavadím pohledem o Tvé jméno na zadní straně obrázku. A ty vždy svědomí mne na Tebe upomene, že jsem Ti ještě mnoho povinen a že Ty z jakési předvídavosti či z ohledu dosud jsi mi neposlal ty žádané složenky ku splátkám svého dluhu . A proto jaksi vědom jsa si své povinnosti znovu Tě prosím, abys mi ty složenky poslal neb jinak, příteli milý, vždy se v dnešní době všechny mé finance xx od 1.do posle ního rozkutálejí,že opravdu nezbude téměř nic. Kdyby však splátka byla po prvém zaplacena, museli bychom se uskrovniti a bylo by povinosti zadostučiněno.

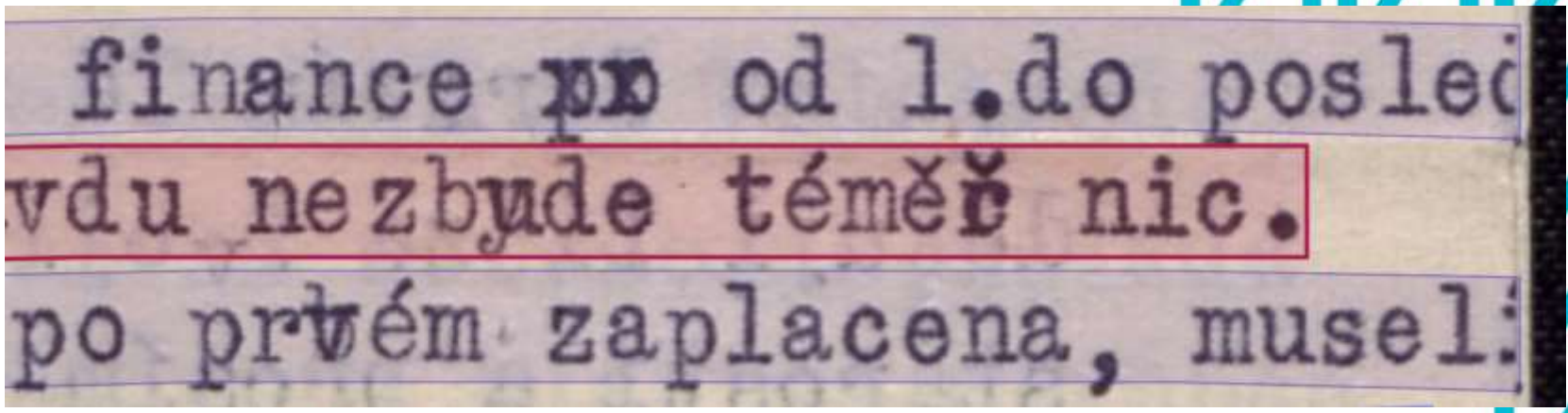
Musím jaksi vzpenouti i manželku, která mne stále upomíná, že vůči Tobě plním své závazky a proto svědomí mi nedá abych se Ti tímto omluvil a poprosil Tě o zaslání složenek.

Manželka tentokráte velmi účinně pomáhala pracovati u jednoho rolníka abychom si svoji výživu co nejvíce usnadnili a tuším že má již jak se říká " zadšláno" na 45 dnů což jen po K 20.- denně činilo by asi 900. K. Sedlák nám již ponechal něco pšenice ,kterou jsme ve mlýně proměnili za mouku a slibuje ponechati i brambory, takže přece se nějak přes tu zimu přetlučeme. Horší ovšem je nedostatek sádla. To není zde vůbec

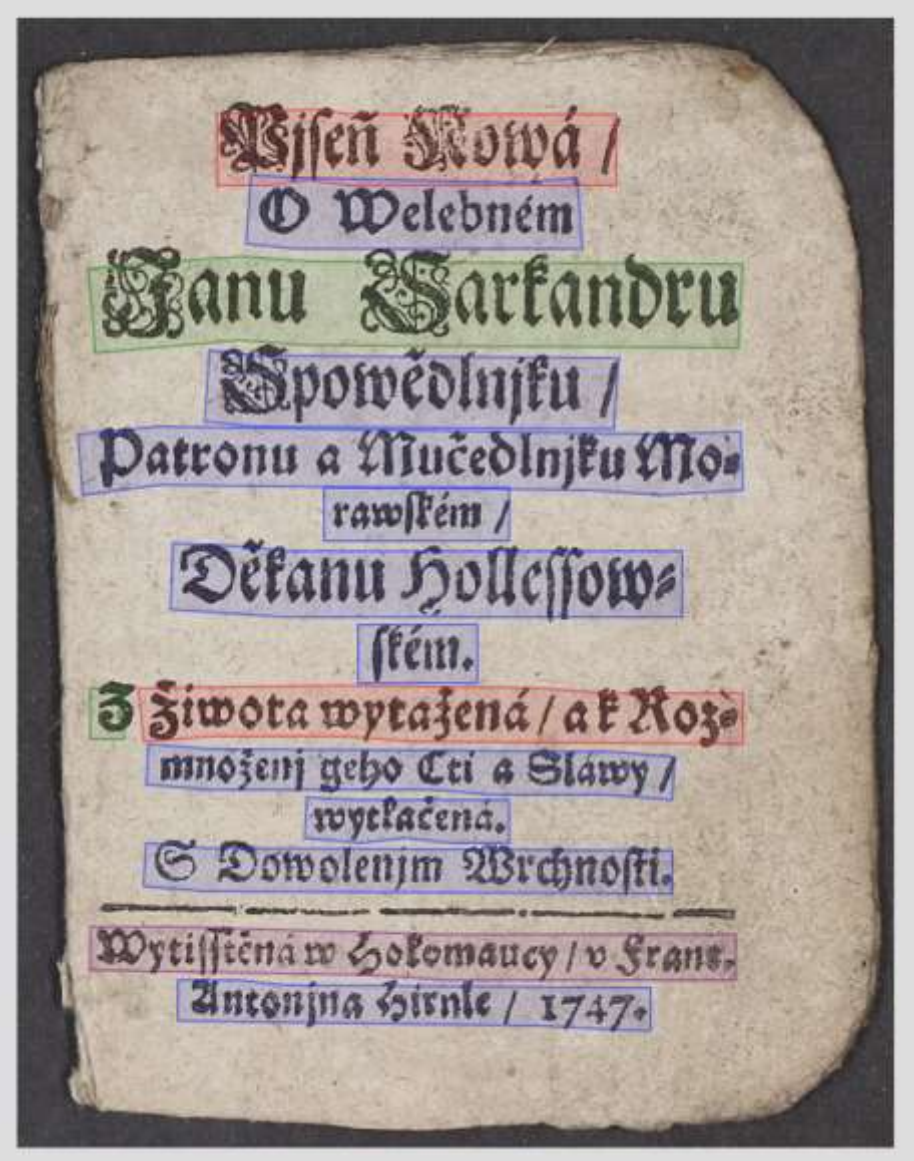
v dnešní době všechny mé finance od 1. do posled

ního rozkutálejí,že opravdu nezbude téměř nic.

Kdyby však splátka byla po prvém zaplacena, museli

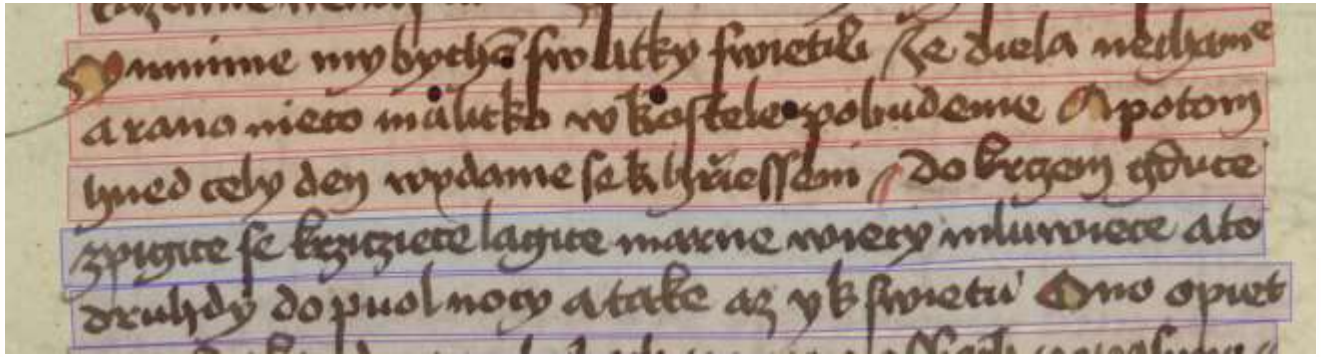


Kramářské tisky

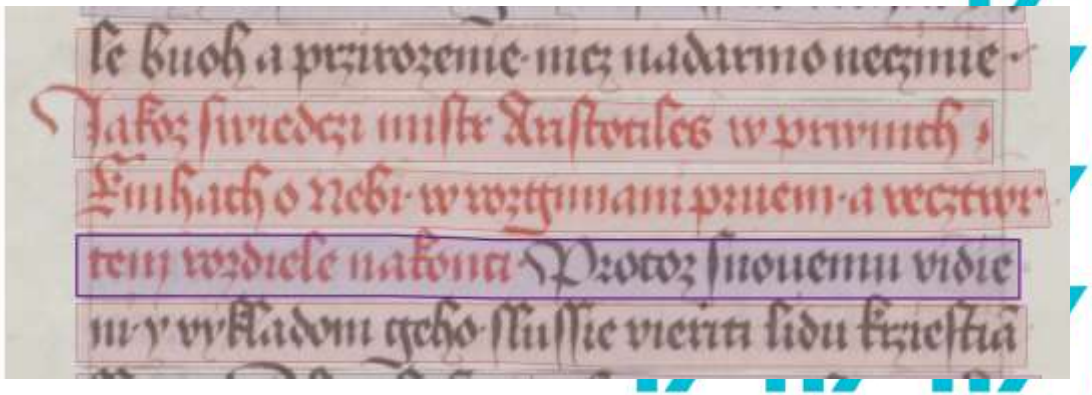


Pjſeň Nowá/
O Welebném
Janu Sarkandru
Spowědlnjku/
Patronu a Mučedlnjku Mo=
rawſkém/
Děkanu Holleſſow=
ſkém.
Z
żiwota wytażená/ a k Roz=
množenj geho Cti a Sláwy/
wytľačená.
S Dowolenjm Wrchnoſti.
Z
Wytiſſtěná w Hoľomaucy/ v Frane.
Antonjna Hirnle/ 1747.

Jan Hus, Vavřinec z Březové

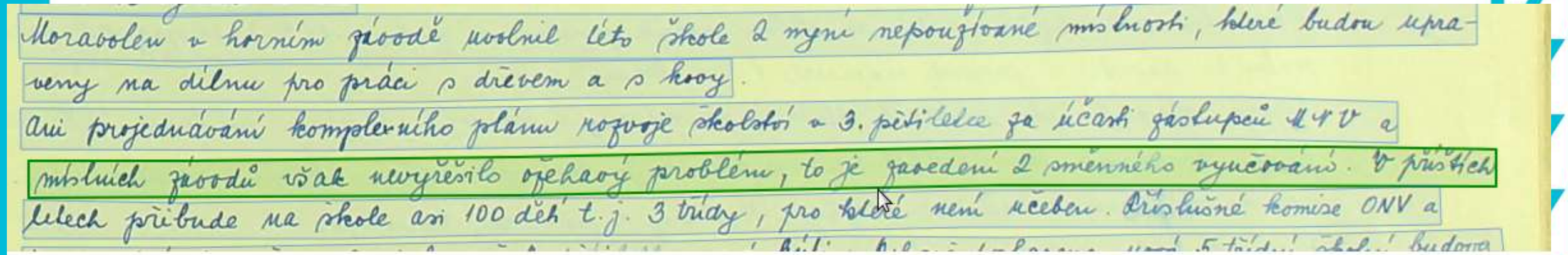


I mnime mybycha ſwatky ſwietili ſe diela nedani
a rano nieto malitko w koſtee pobudeme A potom
hned cely den wydame ſek hrieſſem Do krczem gduce
zpigice ſe krzicziece agite marne wiery mluwiece a to
druhdy do puo nocy a take az ys ſwietu Ono opiet



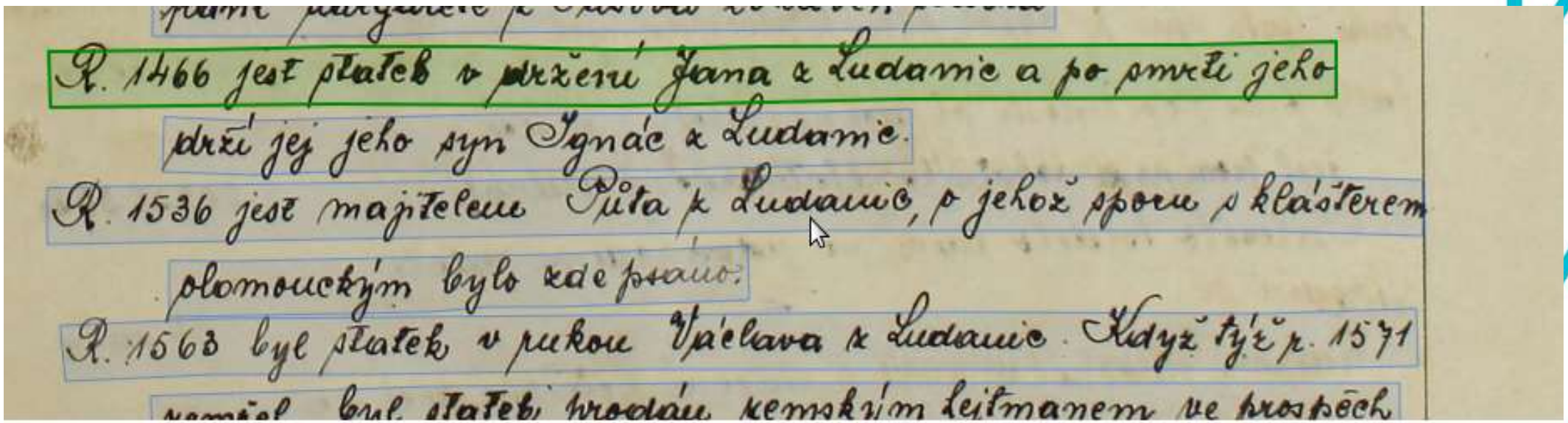
le buoh a przirozenie·niez nadarmo neczinie-
Jakoz ſwiedczi miſtr Ariſtocileś w prwnich
Eiihach o Nebi w rozgiani pruem·a vecztryr
tem rozdiele nakona Protoz ſnouemu vidie
ni·y vykladom geho ſluſſie vieřiti lidu krzieſtia

České kroniky 20. století



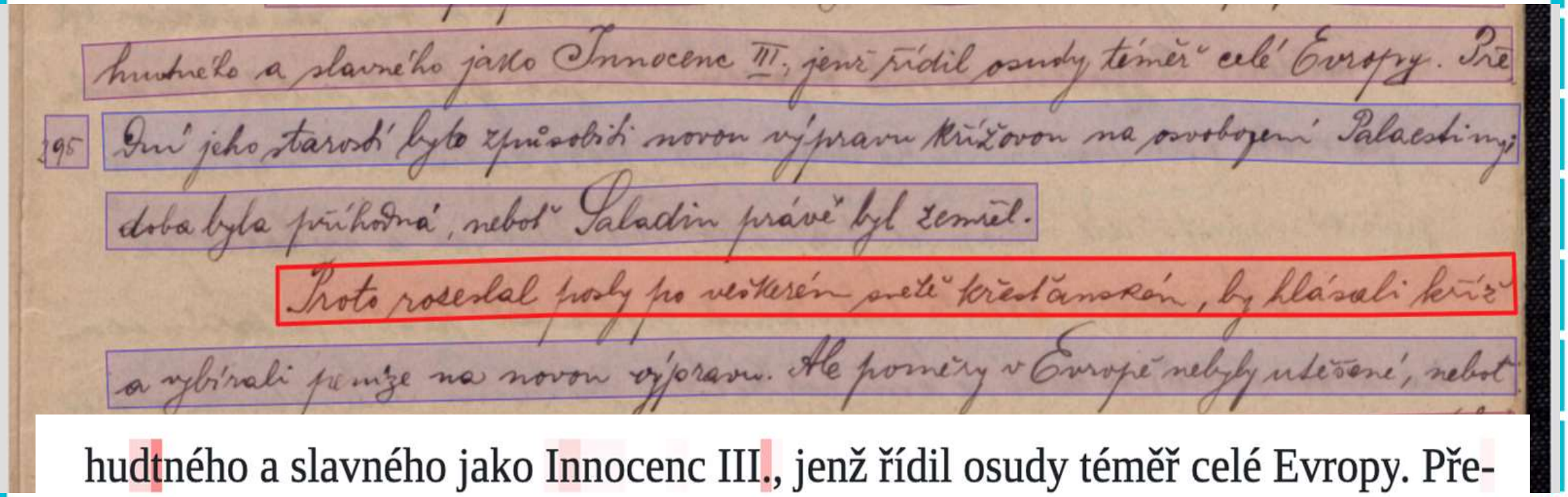
Moravolen v horním závodě uvolnil této škole 2 nyní nepoužívané místnosti, které budou upra-
veny na dílnu pro práci s dřevem a s kovy
ani projednávání komplexního plánu rozvoje školství v 3. pětiletce za účasti zástupců MNV a
místních závodů však nevyřešilo ožehavý problém, to je zavedení 2 směnného vyučování. V příštích
letech přibude na škole asi 100 dětí t. j. 3 třídy, pro které není učeben. Příslušné komise ONV a

České kroniky 20. století



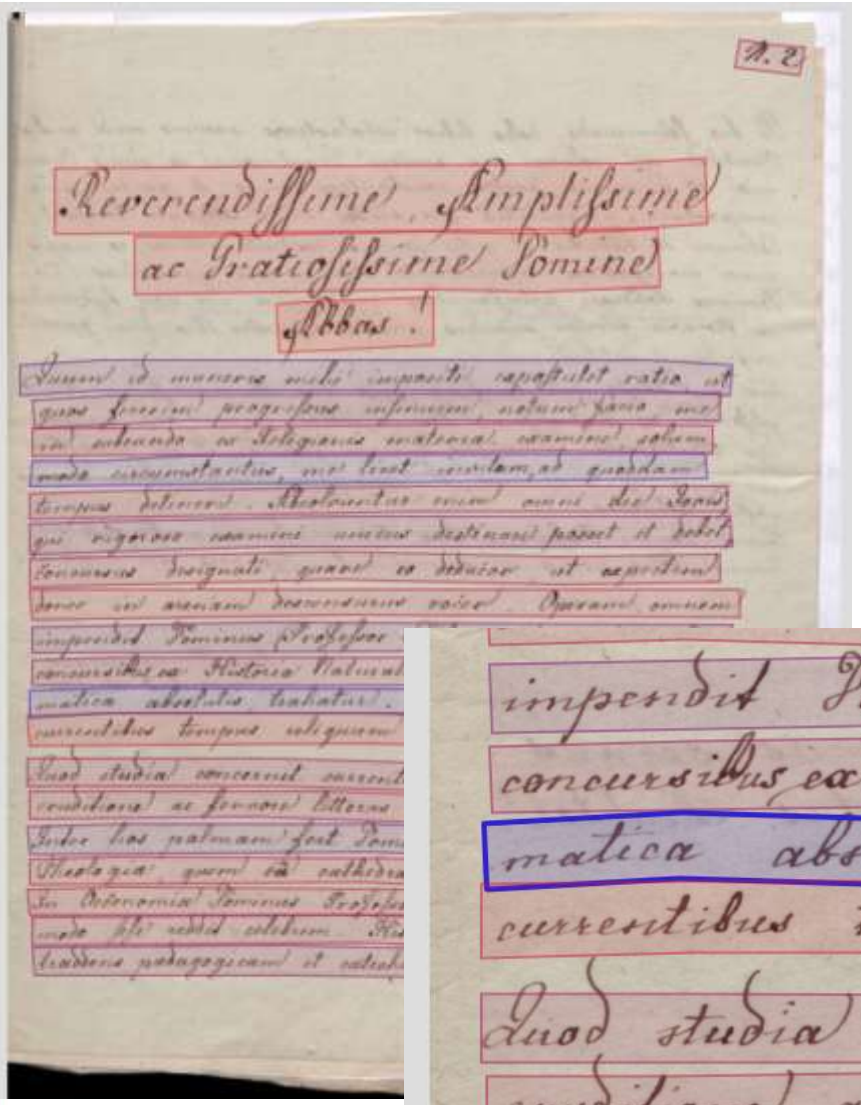
R. 1466 jest statek v držení Jana z Ludanic a po smrti jeho
drží jej jeho syn Ignác z Ludanic.
R. 1536 jest majitelem Půta z Ludanic, o jehož sporu s klášterem
olomouckým bylo zde psáno.
R. 1563 byl statek v rukou Václava z Ludanic. Když týž r. 1571

Rukopisy

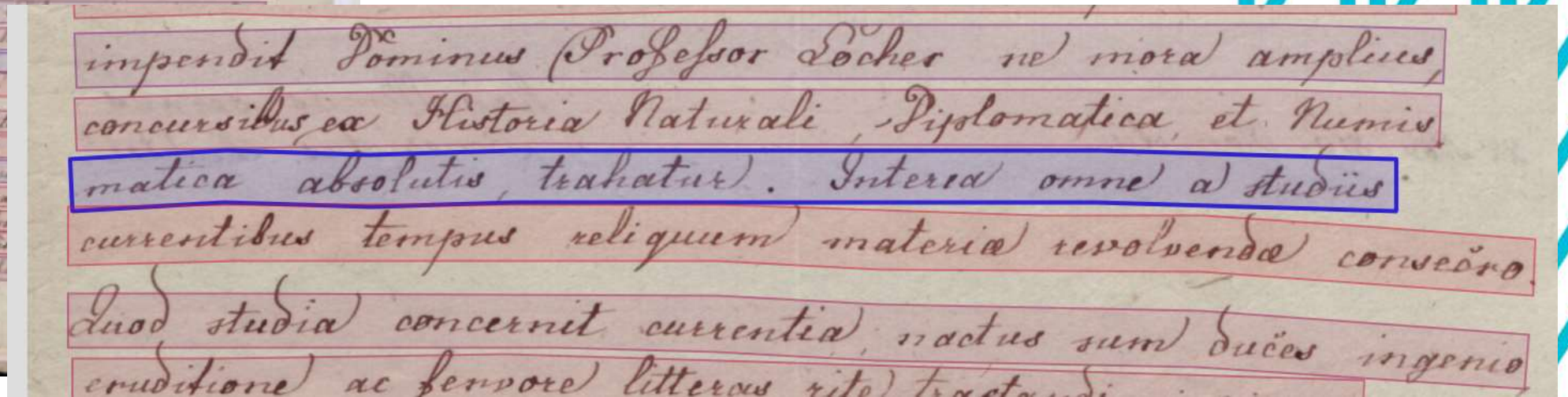


hudtného a slavného jako Innocenc III., jenž řídil osudy téměř celé Evropy. Pře-
dní jeho starostí bylo způsobiti novou výpravu křížovou na osvobození Palaestiny,
doba byla příhodná, neboť Saladin právě byl zemřel.
Proto rozeslal posly po veškerém světě křesťanském, by hlásali kříž
a vybírali peníze na novou výpravu. Ale poměry v Evropě nebyly utěšené, neboť

Rukopisy



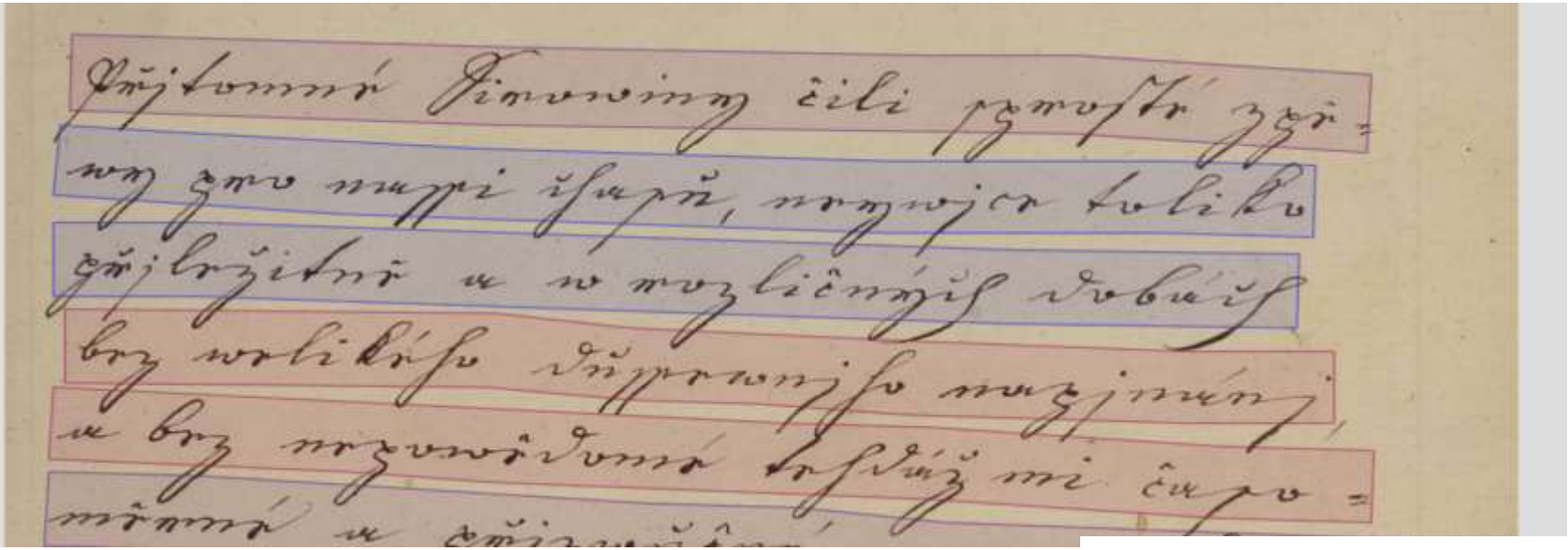
impendit Pominus Proſeſsor Löcher ne mora amplius,
concursilus ex Historia Naturali, Diplomatica, et Nemis,
matica absolutis, trahatur Interea omne a studis
currentibus tempus reliquum materia revolvenda conseoro.
Quod, studia concernit currentia, nactus sum dučès ingenio



Kurent

Die dem Direktor, des philoſophiſchen
Studiums obliegenden Pflichten mit
den übrigen Studiendirektoren
gemeinſchaftlich für den Fler der
Studien zu ſorgen, ſind in der all„

Kurent



Přjtomné Sirowiny čili ſproſté zpě„
wy pro naſſi chaſu, neywjce toliko
přjležitně a w rozličných dobách
bez welikého duſſewnjho napjmáej
a bez nepowědomé tehdáž mi čapo=

PERO - důležité odkazy

- Jádro OCR - pero-ocr python balíček https://github.com/DCGM/pero-ocr
- Webová aplikace pro kontrolu a opravy - pero_ocr_web
  - Běží na https://pero-ocr.fit.vutbr.cz
  - Zdrojové kódy https://github.com/DCGM/pero_ocr_web
- OCR API pro hromadné zpracování
  - https://pero-ocr.fit.vutbr.cz/api
- Informace o projektu - https://pero.fit.vutbr.cz/

Identifikace obrázků na stránce

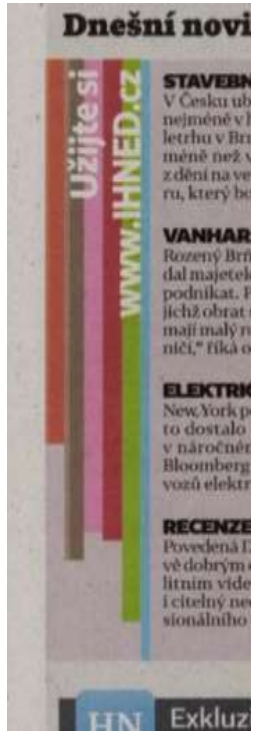
- Cílem je automaticky detekovat pozici obrázků v naskenovaných dokumentech
- **Proč je to těžké?**


O PRÁCI VÝBORŮ ŽEN

Úspěchy a chyby libeňského výboru žen

Co není text ještě nemusí být obrázek

- Grafické elementy, artefakty, pozadí

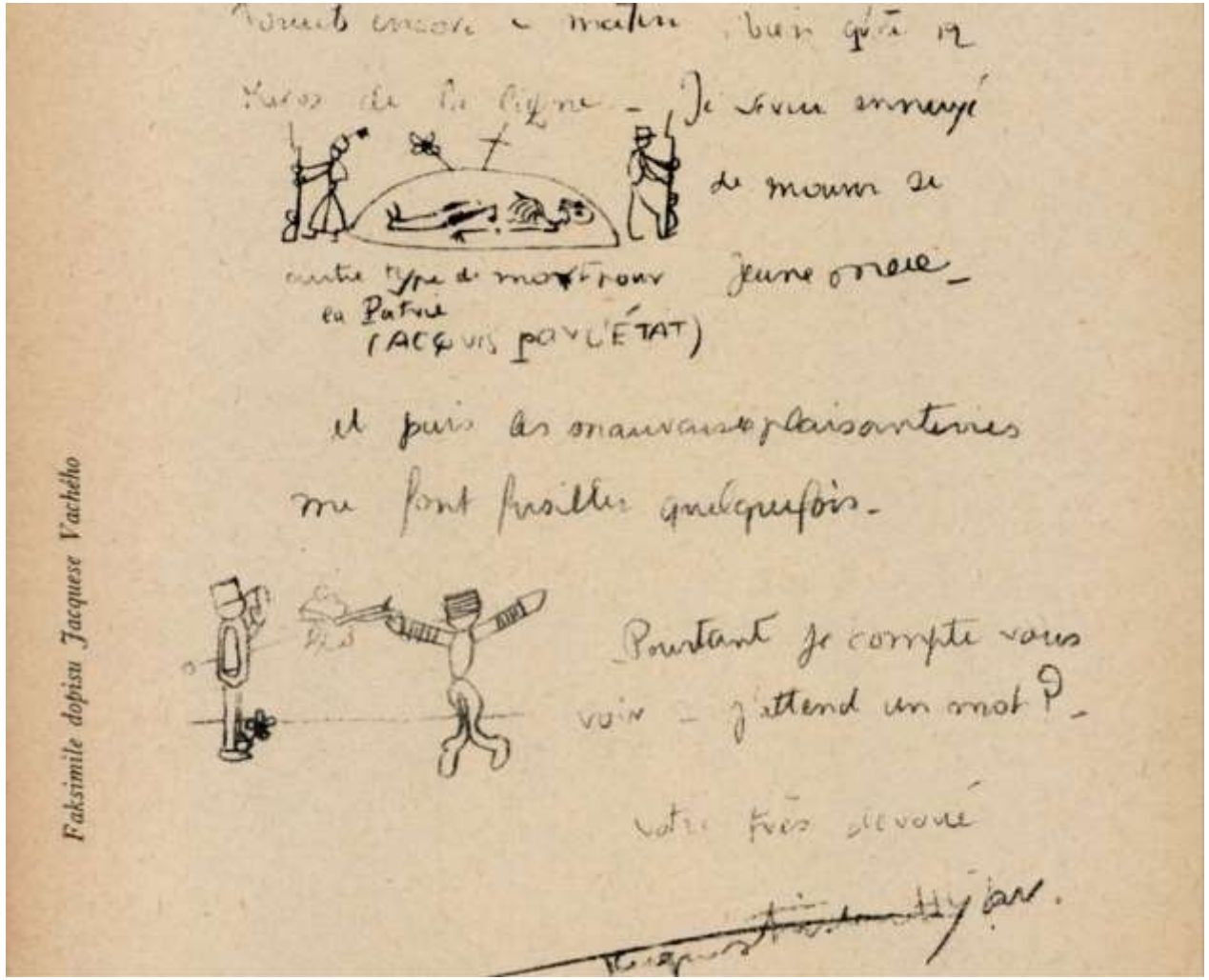


Co obsahuje text stále může být obrázek

- text v obrázcích, překryv

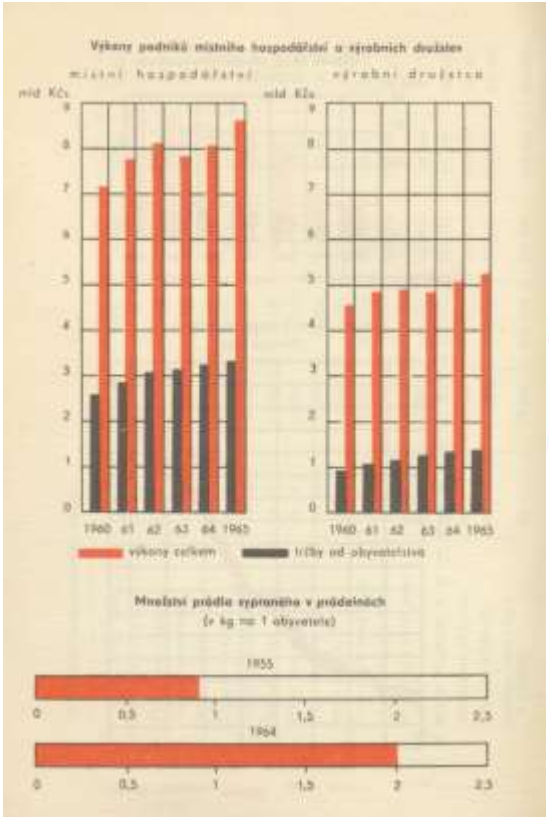
Obrázky nemusí být ohraničeny

Obrázek a text mohou být vizuálně podobné



Kategorie nejsou jasné ani lidem

- Sken textu, typografie, grafy a tabulky



Dostupná řešení - OCR od ABBYY (v ALTO)

- Dataset: 500 náhodných, manuálne oanotovaných stránek
- Výsledek:
  - Celkové IOU*: 0.69
  - Sensitivity**: 0.69
  - Precission**: 0.36
  - 18% obrázků není detekováno vůbec
  - 57% detekovaných obrázků jsou falešná pozitiva***

OCR od ABBYY Ruční anotace

* Intersection over union
**Vypočteno pro práh citlivosti IOU=0.5
***Neobsahují obrázek, nebo obsahují obrázek který už byl detekován, tzn. obrázky se překrývají.

Detekce pomoci vlastního modelu strojového učení

Řešení: Segmentace pomocí konvoluční neuronové sítě

Proč konvoluční sítě?

- Translační invariance: poloha obrázku na stránce není důležitá
- Množství volně dostupných implementací state of the art modelů (napr. ResNet, AlexNet, VGG,...)
- Možnost použít váhy předtrénované na velkém datasetu.

Slepé uličky

- **Newspaper Navigator model**
  - Vyvinut v Library of Congress na detekci obrázků, nadpisů a dalších objektů ve starých novinách.
  - Založen na síti Faster-RCNN od Facebooku.
  - Při evaluaci (bez našeho trénování) měl mnohem horší výkon než OCR od ABBYY.
  - Domníváme se, že model je náchylný k overfittingu a špatně generalizuje na náš dataset.
- **Natrénování UNet sítě (bez předtrénovaných vah)**
  - Architektura využívaná k segmentaci v medicíně a při zpracování satelitních snímků.
  - Nepodařilo se nám dosáhnout dostatečné přesnosti, náš dataset byl zřejmě příliš malý na to, aby se model naučil všechny potřebné vizuální znaky (features).

Neuronová síť dhSegment

- Vyvinuta v 2018 specificky k zpracování skenovaných historických dokumentů
- Použita k detekci ornamentů a fotek na naskenovaných stránkách
- Inspirována segmentační sítí UNet
- Možnost využít předtrénované váhy z ResNet-50
- Implementována v Pythonu pomocí Tensorflow

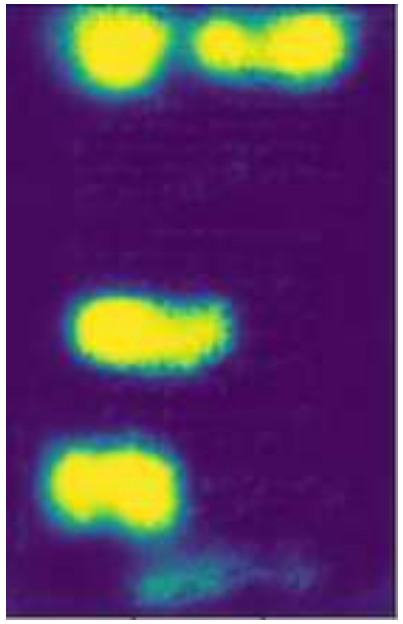

B. C. D.
MICHAELIS ETTMULLERI
INSTITUTIONES MEDICÆ
IUXTA EDITIONEM IN QUARTO.
CAPUT PRIMUM.
DE MEDICINÆ NATURA.

Neuronová síť dhSegment

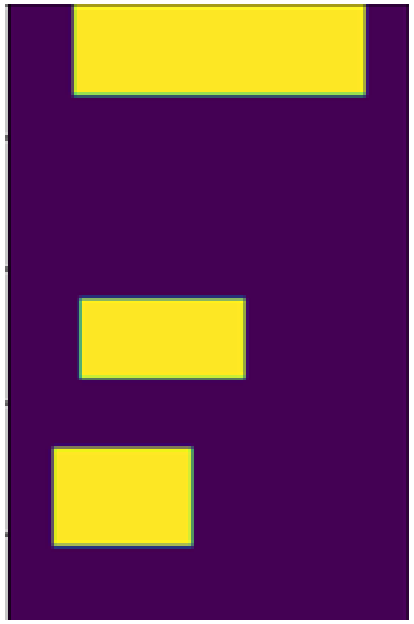
- Natrénována na ~1000 manuálně anotovaných stránkách na notebooku bez GPU

zpracování sítí → binarizace → výstup

vstup



pravděpodobnost pro každý pixel



binární maska pro každý pixel

poloha obrázků

Neuronová síť dhSegment

- Výsledek:
  - Celkové IOU: 0.65
  - Senzitivity*: 0.65
  - Precission*: 0.4
  - 24% obrázků není detekováno vůbec
  - 53% detekovaných obrázků jsou falešná pozitiva**

- Rychlost (notebook bez GPU):
  - 2.87s na stránku
  - 33 dní na milion stránek

- **Při vynaložení relativně malého množství práce je kvalita prakticky identická s OCR od ABBYY**

*Vypočteno pro práh citlivosti IOU=0.5
**Neobsahují obrázek, nebo obsahují obrázek který už byl detekován

Odkazy

- dhSegment
  - Github https://github.com/DCGM/pero-ocr
  - Publikace: Oliveira, Sofia Ares, Benoit Seguin, and Frederic Kaplan. "dhSegment: A generic deep-learning approach for document segmentation." *2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR)*. IEEE, 2018, preprint at https://arxiv.org/abs/1804.10371
- NewspaperNavigator projekt: https://github.com/LibraryOfCongress/newspaper-navigator
- OCR od ABBYY: https://www.abbyy.com/finereader-server/
- Resnet: https://arxiv.org/pdf/1512.03385.pdf
- UNet: https://arxiv.org/pdf/1505.04597.pdf

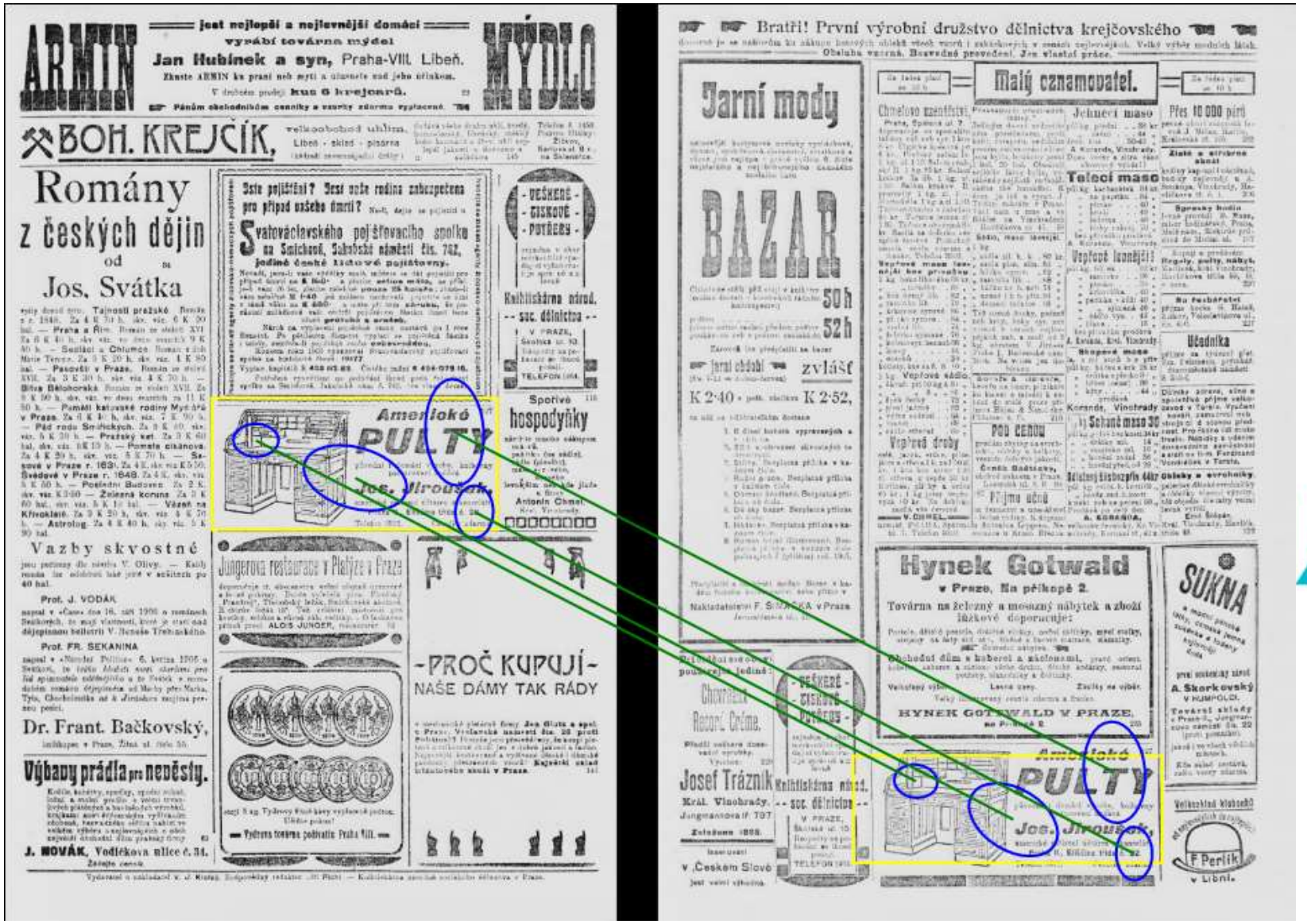
Vyhledávání obrázků podle podobnosti - VISE

- Vyhledávání příbuzných obrázků podle prostorové podobnosti (výřez)
- Periodika, staré ilustrace, grafiky, loga...
- Přesný i pro malé rozlišení (testováno na 1024x1024)
- Možnost kombinace daty identifikujícími obrázky na stránce
- 13 000 obrázků zaindexováno za 7 hodin

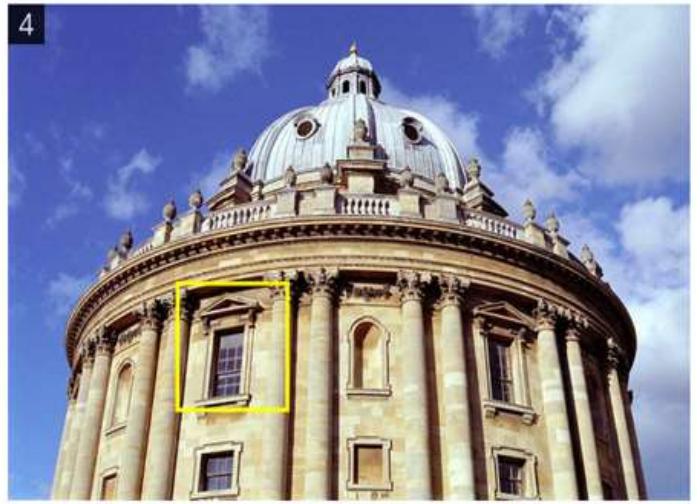
Demonstrace na inzerátech deníku Svobodné slovo

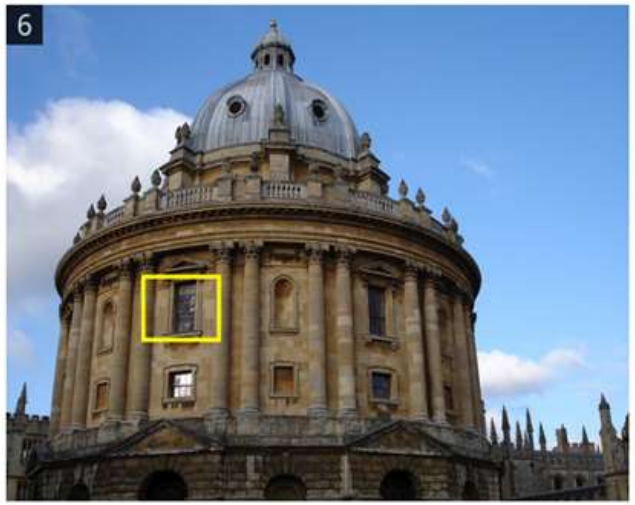
Hledat

Porovnání obrázků inzerce deníku Svobodné slovo



Funkční i pro 3D prostorové obrázky





Výhody VISE

- Velmi dobrá přesnost i pro obrázky s malým rozlišením
- Vyhledávání je rychlé
- Není příliš mnoho konkurenčních systémů
- Není potřeba grafické karty
- Systém je dále rozvíjen

Nevýhody

- Nefunguje pro běžné bloky textů, vhodné spíše pro obrázky, ilustrace nebo větší texty jako tituly, nadpisy atd.
- Může nastat nepřesnost ve vyhledávání, např. pokud je výřez málo detailní nebo je špatná trénovací sada
- Nelze použít obrázky s vysokým rozlišením
- Nelze přidávat nové obrázky již k natrénované datové sadě
- Nelze použít JPEG2000

Odkazy

- Abhishek Dutta, Relja Arandjelović, and Andrew Zisserman. 2021. VGG Image Search Engine. from https://www.robots.ox.ac.uk/~vgg/software/vise/
- Gitlab: https://gitlab.com/vgg/vise
- Oxford Visual geometry group: https://www.robots.ox.ac.uk/~vgg/

Co dál?

- Existuje řada zajímavých projektů, aplikací, modelů
- Není snadné najít hotové řešení
- Specifika reálných datových sad
  - rozsah, variabilita
- Velký prostor pro další rozvoj

Děkuji za pozornost!

Dotazy?

Petr Žabička, Ján Bogár, Michal Tran - Moravská zemská knihovna v Brně