Segmentace historických obrazových dokumentů

Miroslav Liška[1]

1 Úvod

Rozvoj počítačových technologií umožňuje stále snadněji převádět papírové dokumenty na digitální. Díky tomu vznikají projekty, které si kladou za cíl s využitím digitalizace zachovat kulturní dědictví lidstva. Pouhým naskenováním však práce s dokumentem nekončí. Dalším návazným krokem je analýza dokumentu. Ta zahrnuje detekci bloků obsahujících informace. Může se jednat například o bloky obsahující text, obrázky, tabulky a jiné. Tento proces se nazývá segmentace a tato práce je jí věnována.

Historické dokumenty jsou často různě deformované, obsahují šum a mají nepravidelnou strukturu a tak je úspěšná segmentace těchto dokumentů velkou výzvou. Cílem této diplomové práce bylo prozkoumat a vyzkoušet současné přístupy a následně vyzkoušet vybrané metody na dodané datové sadě. Analýza ukázala, že vynikajících výsledků dosahují plně konvoluční neuronové sítě.

2 Tvorba datové sady

Dodaná datová sada Portafontium neobsahovala metadata s informacemi, definujícími očekávaný výstup segmentace. Tyto informace jsou však nutné pro trénování modelů a vyhodnocení výsledků. Bylo tak potřeba datovou sadu o tyto informace doplnit. V rámci práce byly vyzkoušeny současné nástroje pro tvorbu těchto metadat a způsoby jejich uložení.

Celkem bylo označeno 17 obrázků. V nich byly označeny bloky obsahující text a čáry, které tyto bloky oddělují. Deset obrázků bylo použito pro trénování, 3 pro validaci a 4 pro testování. Deset obrázků pro trénování klasifikátorů je obecně málo. Klasifikátory tak byly nejprve natrénovány na větší datové sadě Europeana (76 obrázků pro trénování), která je dodané datové sadě podobná a navíc obsahuje potřebná metadata. Vytvořená trénovací sada byla následně použita pro dotrénování modelů.

3 Segmentace

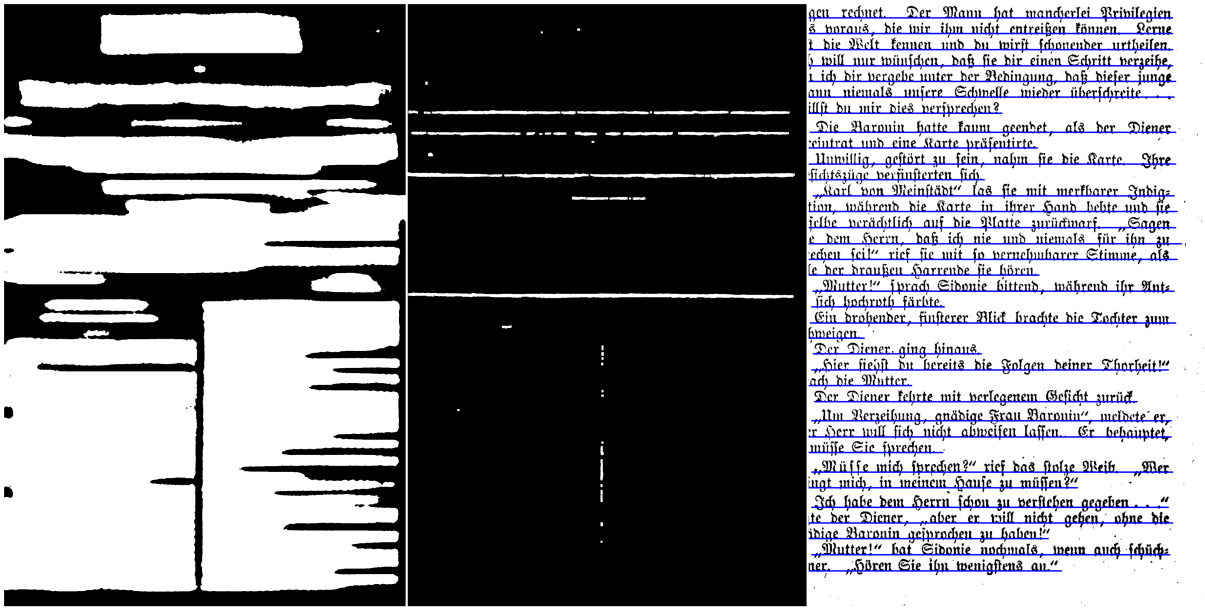
Grüning, Tobias, et al. (2018) navrhli plně konvoluční síť **ARU-NET**, která se osvědčila pro detekci řádků textu. Pro segmentaci textových bloků v historických dokumentech byla navržena plně konvoluční síť (Wick, C., and Puppe, F. (2018)), která je upravenou verzí sítě **U-Net**, použité pro segmentaci biomedicínských snímků.

Výstupem těchto metod je klasifikace na úrovni pixelů – v našem případě se jedná o rozdělení pixelů do dvou tříd: pozadí a oblast zájmu. S využitím získané segmentace textu je možné původní obrázek vymaskovat, čímž docílíme toho, že v původním obrázku zůstane pouze text. Jednotlivé textové bloky je možné v obrázku nalézt použitím algoritmů pro analýzu

[1] student navazujícího studijního programu Aplikované vědy a informatika, obor Softwarové inženýrství, e-mail: topiker@students.zcu.cz

spojených komponent. Může však dojít k jevu, kdy textové bloky, které se nachází vertikálně těsně vedle sebe, spojí klasifikátor do jednoho celku. Aby bylo zachováno **pořadí čtení**, je nutné takové bloky rozdělit. Z toho vyplývá, že je potřeba v obrázku nalézt i oddělovače (separátory). Jakmile jsou textové bloky správně označeny, je již snadné aplikací algoritmů rozdělit bloky na jednotlivé řádky textu. Nalezené řádky pak mohou být vstupem OCR.

Z těchto požadavků vyplynulo, že je nutné připravit klasifikátory pro 3 typy úloh: pro segmentaci textových bloků, pro segmentaci oddělovačů a pro hledání řádků v textu. Provedené experimenty ukázaly použitelnost jednotlivých typů plně konvolučních neuronových sítí pro zmíněné úlohy. Pro označování textových bloků se pro dodanou datovou sadu nejlépe hodí síť upravený U-Net a pro hledání oddělujících čar síť ARU-Net. Pro detekci řádků byla využita síť ARU-Net pro tyto účely předtrénovaná.



Obrázek 1: Dosažené výsledky. Vlevo je segmetace textových bloků, uprostřed segmentace oddělovačů a vpravo jsou označené řádky ve vybraném bloku.

4 Závěr

Na základě analýzy byly vybrány nejvhodnější metody pro dílčí úlohy, jejichž kombinací byl vytvořen program, který je schopen nalézt řádky textů a to ve správném pořadí čtení. Tyto řádky mohou být vstupem dalšího zpracování OCR.

Literatura

Grüning, Tobias, et al. (2018) A two-stage method for text line detection in historical documents. arXiv preprint.

Wick, C., and Puppe, F. (2018). Fully convolutional neural networks for page segmentation of historical document images. In 2018 *13th IAPR International Workshop on Document Analysis Systems (DAS)* (pp. 287-292). IEEE.