KATEDRA INFORMATIKY A VÝPOČETNÍ TECHNIKY

Bc. Patrik Patera

Diplomová práce

Inženýrská informatika
Softwarové inženýrství
2019/2020

Vedoucí práce:
Ing. Kamil Ekštein, Ph.D.

FAKULTA APLIKOVANÝCH VĚD ZÁPADOČESKÉ UNIVERZITY V PLZNI

Extrakce údajů z heterogenních dokumentů pomocí šablon

Abstrakt

Diplomová práce se zabývá problémy z oblasti počítačového vidění k automatizované extrakci užitečných informací z naskenovaných dokumentů (obrazových dat) dle uživatelsky definovaných šablon. Cílem bylo analyzovat používané techniky a nástroje zaměřující se na zpracování digitálních snímků s následným optickým rozpoznáním znaků (OCR) z textových oblastí. Na základě analýzy byl navržen a implementován software pro tvorbu šablon dokumentů s grafickým uživatelským rozhraním a modul pro práci s naskenovanými dokumenty, který podle příslušné šablony extrahuje oblasti s užitečnými informacemi a ty předá OCR systému. Implementované algoritmy byly podrobeny evaluačním testům k získání přehledu o jejich funkčnosti a robustnosti s ohledem k zamýšlenému účelu.

Úvod

Zadání práce vzešlo z potřeb firmy Palaxo Development s. r. o. k rozvoji její digitální platformy CIRCULARO™ pro zpracování dokumentů. Dosavadní řešení je schopné extrahovat a analyzovat obsah pouze u dokumentů textového charakteru. Obsah naskenovaného dokumentu je pro platformu takřka nedostupný. Proto bylo hlavním cílem této práce navrhnout a implementovat modul pro práci s naskenovanými dokumenty a pomocí metod počítačového vidění extrahovat oblasti s užitečnými informacemi podle uživatelsky definovaných šablon, které budou předány systému pro rozpoznávání textu (OCR). Dále implementovat software pro vytváření příslušných šablon s funkcí nalezení vhodné šablony pro zpracovávaný dokument. Implementované řešení bude podrobeno evaluačním testům ověřujícím stabilitu algoritmů a úspěšnost shody nalezené šablony, ale také přesnost extrahovaného obsahu vůči referenčnímu dokumentu z testovací množiny dat.

Východiska, analytická část

Analyzovány byly techniky a nástroje používané pro předzpracování obrázku, ověření shody vzoru analyzovaného dokumentu se šablonou a extrakci významných oblastí. Prostudovány byly i volně dostupné OCR systémy, které je možné intergrovat do modulu počítačového vidění a zpracování dokumentů.

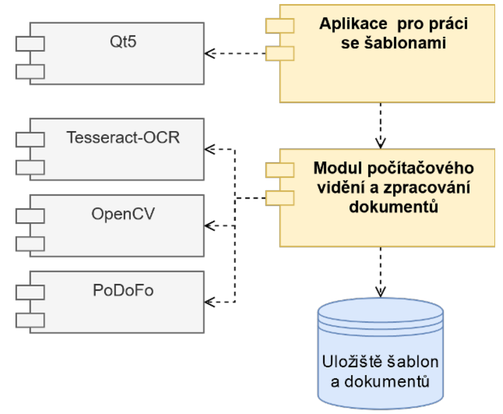
Analyzované techniky počítačového vidění pro práci s obrázkem:

- prahování (binarizace),
- morfologické operace,
- detekce hran,
- odšumění,
- škálování,
- detekce a korekce natočení,
- hledání vzoru.

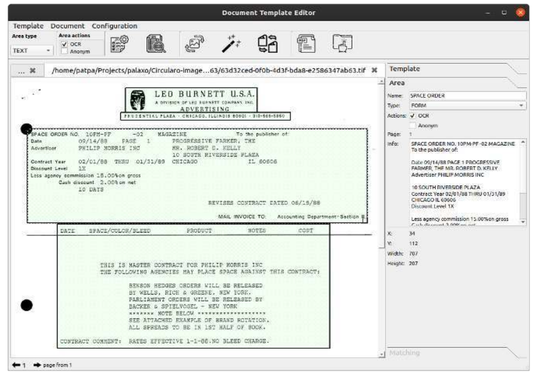
Hlavní aspekty realizace

Na základě analýzy byly pro vývoj softwaru a modulu vybrány následující nástroje:

- **Nástroje použité v modulu počítačového vidění a zpracování dokumentů:**
  - Implementované techniky předzpracování obrázku zmíněné v analýze,
  - OpenCV — algoritmy počítačového vidění,
  - PoDoFo pro zpracování dokumentů PDF.
- **Nástroje použité v softwaru pro vytváření a manipulaci se šablonami:**
  - Qt5 pro grafické uživatelské rozhraní.
- **Integrovaný OCR systém:**
  - Tesseract-OCR s využitím natrénovaných jazykových modelů neuronové sítě LSTM.



UML diagram implementované aplikace a modulu s použitými nástroji.



Snímek obrazovky vyvinuté aplikace.

Dosažené výsledky

U algoritmů předzpracování obrázku byly empiricky nalezeny ideální vstupní parametry. Pro hledání shody se vzorem dosáhla nejvyšší přesnosti **97,9 %** metoda *normalizované vzájemné korelace.* OCR systém dosáhl nejvyšší přesnosti **92,9 %** s využitím *korekce natočení obrázku* a jeho zvětšením *lineární interpolací*.

Závěr

Konečným výsledkem je software pro vytváření a manipulaci se šablonami společně s modulem počítačového vidění a zpracování dokumentů. Výsledný software umožňuje také export dokumentu do formátu PDF s textovou vrstvou, která překrývá výchozí obrázek. Vzhledem ke kvalitě dokumentů v testovací sadě (cca 90 dpi pro rozměr A4) lze konstatovat, že výsledné přesnosti nalezení ideální šablony i OCR jsou uspokojivé.