Posudek bakalářské práce

Matematicko-fyzikální fakulta Univerzity Karlovy

Autor práce	Lucia Tódová
Název práce	OCR for tabular data
Rok odevzdání	2019
Studijní program	Informatika
Studijní obor	IPSS

Autor posudku	Miroslav Kratochvíl	Vedoucí
Pracoviště	Katedra softwarového inženýrství	

K celé práci	lepší	OK	horší	nevyhovuje
Obtížnost zadání		X		
Splnění zadání		X		
Rozsah práce *... textová i implementační část, zohlednění náročnosti*		X		

Práce je praktičtějšího charakteru, hlavním cílem je rozšířit funkcionalitu dostupného OCR software tak, aby byl použitelný na skenované tabulkové dokumenty. Práce tento cíl splňuje, výsledek je vyhodnocen na vlastním datasetu.

Textová část práce	lepší	OK	horší	nevyhovuje
Formální úprava *... jazyková úroveň, typografická úroveň, citace*		X		
Struktura textu *... kontext, cíle, analýza, návrh, vyhodnocení, úroveň detailu*		X		
Analýza		X		
Vývojová dokumentace		X		
Uživatelská dokumentace		X		

Práce v první kapitole poskytuje poměrně široký (ač povrchní) přehled o běžných algoritmech používaných pro rozpoznávání textu z bitmapové grafiky. Autorka navazuje popisem běžných algoritmů pro segmentaci stránek a svého algoritmu na rozpoznávání tabulkovitých struktur v textu.

Vyhodnocení se věnuje především praktickým výsledkům a vlivu předzpracování obrazu na rychlost a schopnost použitých knihoven rozpoznat text, což (nepřímo) ovlivňuje celkovou schopnost algoritmu rozpoznávat tabulky. Přímé porovnání vlastního zlepšení s alternativním software (TableFind z knihovny Tesseract) je bohužel poměrně krátké. Chybějící srovnání obou přístupů podle nějaké metriky jde ale částečně opodstatnit tím, že pro tento typ dat nejsou k dispozici anotované benchmarkovací datasety, podle kterých by srovnání bylo možné provést.

Angličtina práce je na relativně dobré úrovni a text je srozumitelný, občas se vyskytují mírné stylistické a syntaktické neobratnosti.

Implementační část práce	lepší	OK	horší	nevyhovuje
Kvalita návrhu *... architektura, struktury a algoritmy, použité technologie*		X		
Kvalita zpracování *... jmenné konvence, formátování, komentáře, testování*		X		
Stabilita implementace		X		

Program je implementovaný v C++, využívá běžné knihovny pro práci s obrazem a rozpoznávání písmen (Tesseract, Leptonica) a v dokumentovaných případech je stabilní. Implementace je rozsahem spíše menší, čímž ale zhruba odpovídá podobným pracím aplikujícím existující algoritmy strojového učení na nové problémy.

Implementovaný algoritmus je poměrně jednoduchý, ale díky výběru odlišné heuristiky většinou poskytuje srovnatelné nebo lepší výsledky než existující TableFind. Zajímavou výhodou výsledku oproti použití TableFind je podpora pro korektní rozdělení tabulky na jednotlivé buňky s obsahem. Výsledek je následně možné exportovat včetně této struktury (v JSONu), díky čemuž jde jednoduše převést na libovolný jiný tabulkový formát.

Celkové hodnocení Výborně

Práci navrhuji na zvláštní ocenění Ne

Datum Podpis