FAKULTA
APLIKOVANÝCH VĚD
ZÁPADOČESKÉ
UNIVERZITY
V PLZNI

odb. as. Ing. Kamil Ekštein, Ph.D.
Katedra informatiky a výpočetní techniky
Fakulta aplikovaných věd Západočeské univerzity
Univerzitní 22, CZ-30614 Plzeň
Tel.: (+420) 377 63 2406
E-mail: kekstein@kiv.zcu.cz

Hodnocení vedoucího diplomové práce

Bc. Patrik Patera

Extrakce údajů z heterogenních dokumentů pomocí šablon

Diplomová práce Bc. Patrika Patery vychází z průniku odborného zájmu vedoucího práce a aktuálního vývoje R&D společnosti Palaxo Development s. r. o., která dlouhodobě spolupracuje s Katedrou informatiky. Řešené téma představuje drobnou modifikaci již dříve neúspěšně řešeného zadání, které bylo nyní přizpůsobeno tak, aby byl jeho výstup přímo použitelný jako součást digitální platformy Circularo vyvíjené Palaxem.

Úkolem diplomanta bylo vytvořit grafické uživatelské rozhraní, ve kterém může obsluha systému vytvářet šablony heterogenních dokumentů tím, že na naskenované předloze referenčního dokumentu daného typu vyznačí, kde se nachází jaký konkrétní typ detekovatelné komponenty (např. textové pole, logo, lineární grafika, apod.), a modul pro extrakci údajů z naskenovaných dokumentů podle vytvořených šablon. Když pak systém zpracovává neznámý vstupní dokument, pokusí se nejprve aplikací technik strojového vidění identifikovat, které z uložených připravených šablon dokument odpovídá, a pak s využitím strukturálně-sémantické informace uložené v šabloně extrahovat údaje z příslušných oblastí.

Jedná se o poměrně teoreticky náročné zadání. Ačkoliv se může na první pohled zdát, že lze řešit pouze na implementační úrovni za použití frameworků a knihoven (např. OpenCV pro podúlohy z oblasti počítačového vidění, Qt 5 pro tvorbu aplikace s GUI, atp.), není tomu tak zcela. Je třeba značného množství kódu, které užité algoritmy konzistentně propojí, zajistí transformace dat, rozhodovací logiku, apod. Nemluvě o přípravě aplikace pro tvorbu šablon, která z principu musela být implementována zcela od základu bez možnosti využít komplexní prefabrikáty.

Autor nástrahy zadání zvládl zdatně, čímž vrchovatou měrou prokázal své programátorské kvality a také způsobilost k výzkumně-vývojové činnosti v oblasti informatiky (nebo spíš přesněji Computer Science). Vyvinuté programové produkty již firma testuje a budou zřejmě nasazeny do ostrého provozu v rámci platformy Circularo.

Autor práce je velmi schopný, inteligentní a kreativní student. K práci přistoupil velice aktivně a s velkým počátečním nadšením, které bylo v průběhu řešení úkolu modulováno (ne)zájmem nadřízeného pracovníka z Palaxa, který měl autorovi dodat detailní specifikaci požadavků na výsledný produkt, jeho use-casy a také testovací data.

Autor musel nastudovat řadu pokročilých technik z oblasti zpracování digitalizovaných obrazových dat, a to i takových, které nejsou v celé své šíři pokryty výukou v předmětech navazujícího magisterského studia informatiky. Vyhledal a přečetl řadu odborných publikací, včetně značně teoreticky náročných konferenčních příspěvků zabývajících se tématy restaurace zašuměných snímků, rektifikace snímku a odstraňování deformací, detekce specifických druhů objektů, apod.

Spolupráci s autorem práce hodnotí vedoucí jako téměř vzornou: Na konzultace sice nedocházel zcela pravidelně, ale vždy, když vedoucí zavelel, ochotně se dostavil. Vždy byl výborně připraven, a diskuse tedy byly věcné a efektivní.

Největším problémem v průběhu řešení úkolu byla bezpochyby neschopnost zástupce společnosti Palaxo přesně definovat, jaké typy dokumentů by měl být systém schopen automaticky zpracovat a jaké druhy údajů je třeba z dokumentů extrahovat, příp. posléze dodat vzorová data. Nicméně diplomant v této nepříjemné situaci úspěšně improvizoval a do představ zadavatele se – i díky precizně provedené analýze – víceméně trefil.

Na připomínky vedoucího práce reagoval okamžitě a velmi ochotně požadované úpravy ihned zapracovával do software, resp. posléze do textu práce. Průvodní text práce byl konzultován včas, avšak s ohledem na jeho značný rozsah bych spíše uvítal začít dříve s více iteracemi nad menšími objemy textu.

Práce je původní. Autor při řešení zadání vycházel z dostupných materiálů, z literatury, diskusí s vedoucím práce a z ukázek řešení jednotlivých konkrétních úloh prostředky knihoven OpenCV a Tesseract-OCR. Přestože existují volně dostupné nástroje pro sémantické značkování dokumentů (což je vlastně analogie tvorby šablony), jejich funkcionalita ani zdaleka neodpovídá požadavkům kladeným na vytvářenou aplikaci – autor se tedy nemohl ani nijak výrazněji inspirovat jiným produktem a musel se spolehnout výhradně na své vlastní schopnosti a dovednosti.

Celý poměrně rozsáhlý zdrojový kód díla je původním dílem autora, k implementaci využívá především aplikační framework Qt 5, knihovny OpenCV na počítačové vidění a Tesseract-OCR na rozpoznání textu. Jak framework Qt 5, tak knihovny OpenCV a Tesseract-OCR představují de facto technologické standardy a jejich použití při řešení úlohy takovéhoto rozsahu je proto nejen vhodné, ale v podstatě i nezbytné.

Citace v textu i bibliografie na konci práce jsou provedené v souladu s požadavky (až na drobné technikálie, vypsané níže). Uvedené zdroje literatury (46) jsou dostatečné a relevantní. Většinou jde o články z oblasti počítačového vidění a zpracování obrazových dat publikované na významných oborových konferencích, dokumentace a tutoriály k frameworku Qt 5 a knihovně OpenCV, příp. k technikám, které jsou v nich implementovány. Výběr literatury považuji za naprosto adekvátní řešené problematice.

Implementační část předloženého díla je plně funkční, vytvořená aplikace pro návrh šablon a modul extrakce údajů pracují správně a jsou stabilní. Dosažené hodnoty spolehlivosti detekce vhodné šablony, přesnosti rektifikace snímku dokumentu, nastavení prahu binarizace a následně přesnosti extrakce údajů ze snímku technikami OCR jsou velmi dobré a zajišťují dostatečnou užitnou hodnotu celku.

K vývoji byl použit jazyk C++ bez orientace na konkrétní platformu (autor při vývoji používal GNU/Linux). Implementace je poměrně rozsáhlá (zhruba 24 000 řádek zdrojového kódu), zdrojový kód programového řešení je zapsán celkem přehledně, za dodržení běžných doporučení a zvyklostí. Bohužel ale není téměř vůbec komentovaný! Dokumentační komentáře často obsahují jen obsah nagenerovaný vývojovým prostředím, komplexní konstrukce také nejsou komentovány, když už se někde objeví komentářová závorka, nachází se za ní obvykle jen zakomentovaný kód, který autor z nějakého důvodu ponechal ve zdrojovém souboru. V kódu se také na mnoha místech objevují magická čísla, což by bylo možné prominout u modulu extrakce údajů, který lze považovat do jisté míry za experimentální prototyp, ale určitě ne u aplikace pro tvorbu šablon – ta by mohla a měla být zapsána kódem finální kvality.

Textová část díla je mimořádně rozsáhlá – včetně příloh 131 stran. Autorovo vyjadřování je na vysoké úrovni, text je psán dobrou odbornou češtinou; drobné problémy lze najít v interpunkci a stavbě vět, ale na srozumitelnost a čtivost textu to vliv nemá. Autor své myšlenky formuluje a předává spolehlivě. Gramatické chyby se v textu prakticky nevyskytují, stejně jako překlepy či odchylky od typografických zvyklostí.

Grafická úroveň dokumentu je vynikající a úprava působí profesionálním dojmem. Dokument je vysázen v LaTeXu, očividné technické chyby neobsahuje. Při bližším zkoumání lze nalézt pár drobností (např. v zápisu některých desetinných čísel je za desetinnou čárkou mezera navíc, na str. 125 v Seznamu obrázků naopak mezery chybí).

Struktura textu odpovídá typu a rozsahu práce. Práce je dobře logicky členěná a poměr jednotlivých částí je velmi dobře vyvážený. Text je vhodně doplněn obrázky, grafy, schématy, výpisy zdrojového kódu a vzorci, které jej žádoucím způsobem obohacují a jsou vysázené v odpovídající kvalitě.

Drobnou výhradu mám také k bibliografii, kde lze na str. 117, položka [10], úspěšně pochybovat o tom, že je skutečně autorem publikace jistý „Company, Q." Jinak je ovšem dokument vzhledem k jeho rozsahu až překvapivě chyb prostý.

Autorem implementovaný software bude v průběhu následujících měsíců integrován do digitální platformy Circularo společnosti Palaxo, tedy do prémiového komerčního produktu, čímž je – domnívám se – o využitelnosti řečeno téměř vše. Nepochybně se jedná o produkt velmi dobře nasaditelný v praxi a zároveň i o stabilní základnu pro případné experimenty v oblasti automatické a automatizované digitalizace heterogenních dokumentů.

Všechny body zadání byly splněny. Práce je bez jakékoliv pochybnosti solidním inženýrským dílem. Autor prokázal nejen výborné programátorské schopnosti, ale i značnou míru sociálních dovedností při interakci s budoucím uživatelem produktu.

I přes výše uvedené výhrady – a také proto, že jsou spíše formálního druhu – práci bez váhání **doporučuji k obhajobě** a hodnotím klasifikačním stupněm

„výborně".

Ing. Kamil Ekštein, Ph.D.
KIV FAV ZČU

V Plzni dne 5. června 2020