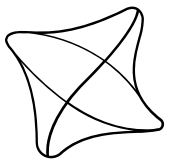


ŽILINSKÁ UNIVERZITA V ŽILINE
Fakulta riadenia
a informatiky

OCR riešenia pre rozpoznávanie textu v reálnych scénach a štruktúrovaných dokumentoch

Bakalárska práca

MATEJ ZACHVEJA

Študijný program: informatika
Študijný odbor: informatika
Školiace pracovisko: Žilinská univerzita v Žiline,
Vedúci: Ing. Peter Tarábek, PhD.
Žilina 2025

ŽILINSKÁ UNIVERZITA V ŽILINE, FAKULTA RIADENIA A INFORMATIKY.

ZADANIE TÉMY BAKALÁRSKEJ PRÁCE.

Študijný program: Informatika

Meno a priezvisko	**Osobné číslo**
Matej Zachveja	561073

Názov práce v slovenskom aj anglickom jazyku

OCR riešenia pre rozpoznávanie textu v reálnych scénach a štruktúrovaných dokumentoch

OCR Solutions for Text Recognition in Real-World Scenes and Structured Documents

Zadanie úlohy, ciele, pokyny pre vypracovanie (Ak je málo miesta, použite opačnú stranu)

Cieľ bakalárskej práce:
Cieľom tejto bakalárskej práce je preskúmať existujúce voľne dostupné OCR riešenia zamerané na rôzne aplikácie, najmä na rozpoznávanie textu v reálnych scénach (či už prirodzene vyskytujúcich sa alebo digitálne vložených) a v štruktúrovaných dokumentoch. V prípade štruktúrovaných dokumentov ide o strojovo generované texty s konzistentným formátovaním. Práca bude zahŕňať kategorizáciu rôznych typov dokumentov a aplikácií, na ktoré sa OCR využíva. Dôležitou súčasťou bude výber vhodných OCR technológií a metód pre tieto aplikácie, ako aj identifikácia relevantných datasetov na testovanie. Na základe analýzy a vyhodnotenia výkonu týchto riešení sa vyberú vhodné prístupy, ktoré budú následne integrované do demonštračného nástroja schopného načítať obrázky, spracovať text a uložiť výsledky v textovej forme.

Obsah:
1. Preskúmanie a analýza aktuálnych OCR systémov, najmä tých zameraných na rozpoznávanie textu v reálnych scénach a štruktúrovaných dokumentoch
2. Identifikácia a výber vhodných datasetov pre testovanie riešení v rôznych scenároch
3. Výber vhodných metód na základe ich použiteľnosti pre rôzne typy dokumentov
4. Implementácia testovacej metodiky a vyhodnotenie riešení na základe ich presnosti, rýchlosti a flexibility
5. Vývoj jednoduchého demonštračného nástroja, ktorý bude schopný načítať obrázky, spracovať text pomocou vybraných OCR riešení a uložiť výsledky v textovej forme
6. Záver a diskusia výsledkov

Meno a pracovisko vedúceho BP: Ing. Peter Tarábek, PhD., KMMOA, ŽU
Meno a pracovisko tutora BP:

garant štud. programu
(dátum a podpis)

Zadanie zaregistrované dňa ..... pod číslom 3233/2024 podpis _____

ČESTNÉ VYHLÁSENIE

Vyhlasujem, že som zadanú bakalársku prácu vypracoval samostatne, pod odborným vedením vedúceho práce a používal som len literatúru uvedenú v práci.

Žilina 30. apríla 2025

Matej Zachveja

POĎAKOVANIE

Chcel by som poďakovať vedúcemu práce a konzultantovi Ing. Petrovi Tarábkovi, PhD. za cenné rady a pripomienky, ako aj za jeho seriózny prístup počas celej tvorby tejto práce. Zároveň som vďačný za príležitosť venovať sa písaniu práce v tejto zaujímavej a aktuálnej oblasti.

ABSTRAKT

ZACHVEJA, Matej: *OCR riešenia pre rozpoznávanie textu v reálnych scénach a štruktúrovaných dokumentoch* [Bakalárska práca] – Žilinská univerzita v Žiline. Fakulta riadenia a informatiky. KMMOA. – Vedúci práce: Ing. Peter Tarábek, PhD. – Žilina 2025 - 40 s.

Cieľom bakalárskej práce je preskúmať existujúce open-source OCR riešenia so zameraním na ich využitie, a to najmä na rozpoznávanie textu v reálnych scénach a štruktúrovaných dokumentoch. Súčasťou práce je výber relevantných datasetov pre testovanie vybraných riešení, ako aj hodnotenie ich presnosti pomocou štandardných metrík. Na základe hodnotenia bude realizované porovnanie a následne odporúčanie riešení. Jednotlivé riešenia budú následne integrované do demonštračného nástroja schopného načítať obrázky, spracovať text a uložiť výsledky v textovej forme, ktorý je zároveň schopný otestovania datasetu podľa vlastnej voľby.

Kľúčové slová: OCR, textová extrakcia, textové rozpoznávanie, štruktúrované dokumenty, reálne scény.

ABSTRACT

ZACHVEJA, Matej: OCR Solutions for Text Recognition in Real-World Scenes and Structured Documents [Bachelor thesis] – University of Žilina, Faculty of Management Science and Informatics. KMMOA. – Thesis supervisor: Ing. Peter Tarábek, PhD. – Žilina 2025 - 40 p.

The aim of this bachelor thesis is to explore existing open-source OCR solutions, focusing primarily on their application in text recognition in real-world scenes and structured documents. The thesis includes the selection of relevant datasets for testing the chosen solutions, as well as the evaluation of their accuracy using standard metrics. Based on the evaluation will be produced comparison and final recommendation of the solutions. The individual solutions will be integrated into demonstration tool capable of loading images, processing text and saving the results in text format, and capable of testing any freely chosen dataset.

Key words: OCR, text extraction, text recognition, structured documents, real-world scenes.

OBSAH

ZOZNAM OBRÁZKOV

ZOZNAM TABULIEK

ZOZNAM SKRATIEK

Skratka	Anglický význam	Slovenský význam
OCR	Optical Character Recognition	Optické rozpoznávanie znakov
DAR	Document Analysis and Recognition	Analýza a rozpoznávanie dokumentov
UNICODE	Universal Character Encoding Standard	Univerzálny štandard pre kódovanie znakov
ANPR	Automatic Number Plate Recognition	Automatické rozpoznávanie evidenčných čísel vozidiel
CAPTCHA	Completely Automated Public Turing test to tell Computers and Humans Apart	Úplne automatizovaný verejný Turingov test na rozlíšenie počítačov a ľudí
LVLM	Large Vision-Language Models	Veľké vizuálno-jazykové modely
CLIP	Constrastive Language-Image Pre-Training	Kontrastívne jazykovo-obrazové pred trénovanie
CRAFT	Character-Region Awareness For Text detection	Vnímanie oblastí znakov pre detekciu textu
CRNN	Convolutional Recurent neural network	Konvolučná Rekurentná neurónová sieť
CNN	Convolutional neural network	Konvolučná neurónová sieť
RNN	Recurent neural network	Rekurentná neurónová sieť
WER	Word Error Rate	Chybovosť slov
CER	Character Error Rate	Chybovosť znakov

ÚVOD

V súčasnosti, keď množstvo informácií existuje v papierovej, alebo obrazovej podobe, je optické rozpoznávanie znakov (OCR – Optical Character Recognition) kľúčovou technológiou umožňujúcou efektívny prechod z hmotného sveta do digitálneho sveta. Jeho význam narastá nielen v administratíve a digitalizácii archívov, ale aj v oblasti umelej inteligencie, mobilných aplikácií, autonómnych systémov či v priemyselnej automatizácii. V posledných rokoch dochádza k výraznému pokroku v oblasti OCR najmä vďaka rozvoju strojového učenia a hlbokých neurónových sietí, ktoré umožnili vznik nových a výkonných riešení.

Cieľom tejto bakalárskej práce je preskúmať a porovnať viaceré open-source OCR nástroje z hľadiska ich schopnosti rozpoznávať text v rôznych typoch vstupov, od reálnych scén s prirodzeným pozadím až po štruktúrované dokumenty s formátovaním. V práci sa budú posudzovať nielen výkonnostné parametre ako presnosť rozpoznávania, schopnosť spracovania zložitejších vstupov a rýchlosť riešení, ale aj jednoduchosť použitia a možnosti integrácie do vlastných systémov.

Dôležitou súčasťou práce je výber vhodných datasetov, ktoré umožnia objektívne otestovanie vybraných riešení, spolu s aplikovaním štandardných vyhodnocovacích metrík. Výstupom práce bude tiež návrh a implementácia demonštračného nástroja, ktorý umožní používateľovi možnosť načítať obrázok, spracovať ho pomocou vybraného OCR nástroja a získať výstup v textovej forme, alebo možnosť otestovať implementované riešenia na vlastnom datasete.

Táto práca má taktiež za cieľ lepšie pochopiť funkcionalitu OCR riešení, ich limity a prispieť k lepšiemu pochopeniu a praktickému využitiu pri riešení úloh súvisiacich s extrakciou textu z obrazových vstupov.

1 ZÁKLADY ROZPOZNÁVANIA A SPRACOVANIA TEXTU

1.1 Čo je optické rozpoznávanie textu

OCR (Optical Character recognition) je technológia, ktorá prevádza obrázky naskenovaného, alebo vytlačeného textu, ako aj ručne písaného textu, do upravovateľnej digitálnej podoby na ďalšie spracovanie [1]. Hoci sa OCR systémy tradične spájajú so spracovaním naskenovaných dokumentov, rovnako dôležité je ich využitie pri digitálne vzniknutých dokumentoch. Existuje množstvo prípadov, kde je potrebné rozpoznať text v digitálnych formátoch ako sú PDF dokumenty obsahujúce naskenované stránky, snímky obrazoviek aplikácií, digitálne fotografie s textovým obsahom a pod.

OCR umožňuje zariadeniam automaticky rozpoznávať text v dokumentoch. Hlavnou úlohou, ale nie je text pochopiť, iba ho rozpoznať. Na pochopenie tejto technológie môžeme použiť analógiu s ľudským telom: oko dokáže detegovať, vidieť a extrahovať text z obrázka, ale je to práve ľudský mozog, ktorý spracuje detegovaný alebo extrahovaný text. [2]

OCR technológia však ešte nedosahuje úroveň ľudskej schopnosti rozpoznávania textu. Jej výkonnosť a presnosť sú priamo závislé na kvalite vstupných dokumentov. Podobne, aj u človeka závisí schopnosť rozpoznávania textu na kvalite vstupu, ktorý oči prečítajú a mozog spracuje. [2]

Pri vývoji OCR systémov sa môžu vyskytnúť rôzne problémy. Napríklad, niekedy je na vstupnom dokumente veľmi malý viditeľný rozdiel medzi niektorými písmenami a číslicami, čo môže spôsobiť ich nesprávne rozpoznanie. Ďalším problémom môže byť extrakcia textu z tmavého pozadia alebo textu, ktorý sa nachádza v blízkosti iných prvkov. Jedným z hlavných zameraní výskumu OCR bolo rozpoznávanie kurzívy a ručne písaného textu vzhľadom na jeho širokú oblasť využitia. [2]

1.2 Typy OCR systémov

Delenie podľa jazyka

Existujú rôzne typy OCR systémov v závislosti od jazyka a spôsobu písania textu, ktorý obsahujú. Dokumenty môžu byť písané rukou, tlačené alebo skenované a môžu obsahovať jeden alebo viac jazykov. Preto môžeme OCR systémy kategorizovať ako jednojazyčné alebo viacjazyčné. Jednojazyčný OCR systém dokáže rozpoznať iba jeden jazyk. Na druhej strane existujú systémy, ktoré vykonávajú úlohy rozpoznávania a extrakcie pre viacero jazykov – tieto sa nazývajú viacjazyčné OCR systémy. [3]

Delenie podľa režimu spracovania

OCR systémy môžeme rozdeliť aj podľa režimu spracovania na off-line a online OCR systémy. Off-line OCR systém spracováva vstupné dokumenty vo forme skenovaných, tlačených alebo rukou písaných textov. Online OCR systémy ponúkajú služby v reálnom čase a využívajú sa na rôzne účely, ako napríklad čítanie bankových šekov, overovanie podpisov a podobne. [3]

Tradičné a LVLM-driven prístupy

Tradičné OCR systémy označované ako OCR-1.0, fungujú na modulárnom princípe, kde sa spracovanie textu rozdeľuje na viaceré kroky, ako analýza rozloženia stránky, detekcia textu, extrakcia regiónov a samotné rozpoznávanie obsahu. Tento prístup bol nevyhnutný, pretože rozpoznávanie textu nedokázalo efektívne pracovať s celými dokumentmi naraz. Síce existujú pokročilé modely, ktoré dokážu spracovať celé stránky naraz, často sú obmedzené na špecifické úlohy a vyžadujú použitie rôznych modelov podľa konkrétnych OCR potrieb. Nevýhodou OCR-1.0 je aj vysoká náročnosť na údržbu a možnosť vzniku chýb pri prenose informácií medzi modulmi.

Druhým prístupom je OCR využívajúce LVLM (Large Vision-Language Models), ktoré využívajú pokročilé modely na priame prepojenie vizuálnych a textových reprezentácií. Pre súčasné LVLM, ktoré disponujú komplexnou schopnosťou vnímania a uvažovania, sa schopnosť OCR stala veľmi aktuálnou témou vzhľadom na rastúci dopyt po vizuálnom porozumení založenom na texte. Väčšina OCR schopností LVLM pochádza z modelov CLIP (Contrastive Language-Image Pre-Training), alebo podobných technológií pre základné prepojenie obrazu a textu. [4]

CLIP je neurónová sieť trénovaná na rôznych pároch obrázku a textu. Dá sa inštruovať v prirodzenom jazyku, aby predpovedala najrelevantnejší textový úryvok k danému obrázku bez toho, aby bola priamo optimalizovaná na túto konkrétnu úlohu, podobne ako GPT-2 a GPT-3. CLIP je vyvíjaná OpenAI. [5]

1.3 Postupy OCR riešení pri analýze a rozpoznávaní dokumentov

Analýza a rozpoznávanie dokumentov DAR (Document Analysis and Recognition) zahŕňa rôzne úlohy spracovania textových a obrazových dokumentov, ktoré vedú k ich plne čitateľnej podobe. [6] Všeobecne sa tieto úlohy rozdeľujú do troch hlavných fáz:

Predspracovanie (Preprocessing)

Cieľom tejto fázy je odstrániť vizuálne chyby, zlepšiť kvalitu obrazu a pripraviť dáta na ďalšie spracovanie. [6] Táto fáza sa dá rozdeliť na ďalšie pod fázy:

- **Oprava obrazu** – redukcia šumu, vyrovnávanie sklonu písma, zjednotenie veľkostí znakov, korekcia osvetlenia, rotácia a narovnávanie deformovaných riadkov
- **Vylepšenie obrazu** - Ostrenie a vylepšenie kontrastu
- **Kompresia** - odstránenie nepotrebných pixelov podľa intenzity, redukcia znakov na ich kostry
- **Binarizácia** - zmena obrazu na čisto čiernobiely text

Segmentácia (Segmentation)

Táto fáza sa zameriava na rozdelenie dokumentu do menších častí, napríklad blokov textu, tabuliek, obrázkov, riadkov a znakov. [6] Pozostáva z niekoľkých menších fáz:

- **Segmentácia strán** – odlíšenie rôznych prvkov na strane (text, obrázky, tabuľky, ...)
- **Segmentácia textu** – rozdelenie textu na riadky, slová a znaky
- **Identifikácia a lokalizácia oblastí záujmu**

Rozpoznávanie textu (Text Recognition)

V tejto fázy sa vykonáva samotná extrakcia znakov z obrázkov a premena na digitálny text. Ide o klasifikačné algoritmy založené buď na báze štatistiky, strojového učenia, alebo hlbokých neurónových sietí. Môže dôjsť k rozpoznávaniu tlačeného písma, alebo ku rozpoznávaniu rukou písaného textu. [6]

Post spracovanie (Postprocessing)

Konečná fáza tradičných OCR systémov. V tejto fáze prebieha ku korekcii chýb rozpoznávania, prevodu do UNICODE (Universal Character Encoding Standard) znakov, pravopisná a gramatická kontrola. [6]

Jednotlivé fázy sa môžu líšiť v závislosti od jednotlivých OCR riešení, zvyčajne ale využívajú aspoň väčšinu z vyššie spomenutých fáz.

2 TAXONÓMIA TEXTOV A VÝZVY OCR

2.1 Taxonómia textov

Štruktúrovaný dokument je dokument založený na štandardizovanej šablóne. Táto šablóna poskytuje prísne pravidlá potrebné pre zabezpečenie, aký obsah bude obsiahnutý, v akom poradí a za akých podmienok. Šablóna taktiež poskytuje štandardizovaný typ písma pre názvy a určité paragrafy, zoznamy a tabuľky. [7]

Reálne scény sú obrázky prírodných scenérií, často zobrazujú zložité geometrické usporiadania. Text v takýchto obrázkoch poskytuje ďalšie informácie o prostredí a pomáha lepšie pochopiť dané miesto. Textové zobrazenia krajiny a prírodných scenérií nám môžu odhaliť dôležité detaily a poskytnúť cenné informácie o lokalite. [7]

2.2 Oblasti využitia OCR

Dôležitou oblasťou nasadenie OCR systémov zameraných na štruktúrované dokumenty, je napríklad bankovníctvo, kde sú tieto systémy použité na spracovanie šekov, bez ľudského zasahovania. Šek môže byť vložený do prístroja, kde systém rozpozná a spracuje sumu, ktorá má byť vydaná a správna suma peňazí bude prevedená. Metóda vloženia do prístroja a následne spracovanie šeku, bola zdokonalená pre tlačené šeky a je veľmi presná aj pre ručne písané šeky, čo šetrí čakací čas v bankách. [8]

Ďalšou oblasťou využitia OCR riešení je Zdravotníctvo. Zdravotníctvo zaznamenalo nárast využívania OCR technológie na spracovanie lekárskej dokumentácie. Zdravotníci musia spracovávať veľké množstvo formulárov pre každého pacienta. Vďaka použitiu OCR systémov sa ich práca zjednoduší a zrýchli. Aby bolo možné spracovávať všetky tieto informácie, je užitočné zadávať relevantné údaje do databázy, ku ktorej je možné neustále pristupovať. Nástroje pomocou OCR technológie dokážu extrahovať informácie z formulárov, digitalizovať ich a následne ich uložiť do databáz, aby boli údaje každého pacienta ihneď zaznamenané. [8]

OCR systémy v oblasti reálnych scén nachádzajú uplatnenie v rôznych oblastiach, kde automatizácia spracovania textových údajov zvyšuje efektivitu a presnosť. Medzi niektoré z týchto oblastí patria napríklad oblasti spomenuté nižšie. [8]

ANPR (Automatic Number-Plate Recognition) sa používa ako metóda hromadného sledovania využívajúca optické rozpoznávanie znakov na obrázkoch, ktorá slúži na identifikáciu poznávacích značiek vozidiel. Je schopná rozpoznať nie len evidenčné číslo, ale aj uchovať snímky zachytené kamerami, vrátane čísiel získaných z poznávacích značiek. [8]

Rozpoznanie rukopisu pridáva počítaču schopnosť prijímať a interpretovať text z rôznych zdrojov ako sú napríklad papierové dokumenty, fotografie, dotykové obrazovky a iné zariadenia. [8]

Špeciálnou oblasťou využitia systémov OCR je napríklad aj CAPTCHA systém, ktorý dokáže vytvárať a vyhodnocovať testy, ktoré človek dokáže prejsť, ale súčasná softvérová technológia by toho nemala byť schopná. V rámci jedného z typov CAPTCHA sa generuje obrázok obsahujúci kombináciu písmen a číslic s rôznymi veľkosťami a typmi písma, rušivým pozadím a šumom, aby nebolo možné text rozpoznať pomocou OCR systémov. Súčasné OCR systémy však dokážu odstrániť šum a segmentovať obrázok tak, aby bol čitateľný aj pre automatizované systémy a tým dokážu obísť aj ochranné mechanizmy CAPTCHA. [8]

2.3 Výzvy OCR

Pre dosiahnutie vysokej kvality a presnosti OCR metód sú potrebné vstupy vysokej kvality, alebo obrázkov vysokého rozlíšenia. OCR systémy zvyčajne dosahujú najlepšie výsledky pri spracovaní naskenovaných dokumentov. Na druhej strane, obrázky zachytené kamerami zvyčajne nie sú takým dobrým vstupom ako skenované obrázky. Tento rozdiel môže byť spôsobený viacerými faktormi, ktoré sú popísané nižšie. [2]

Komplexnosť reálnej scény v OCR

V reálnom prostredí obsahujú obrázky zachytené kamerami množstvo rôznych objektov vytvorených človekom, ako sú budovy, symboly alebo obrazy. Tieto objekty majú zložité štruktúry a rôznorodé vzhľady textov, čo výrazne sťažuje rozpoznávanie textu v takýchto obrázkoch. [2]

Vplyv nekonzistentného osvetlenia v OCR

V prirodzenom prostredí dochádza často krát ku zachyteniu obrazu s nekonzistentným, alebo zlým osvetlením. Tento faktor predstavuje výzvu pre OCR systémy, pretože obmedzuje kvalitu vstupného obrázka, ktorá je pre správne rozpoznávanie textu nevyhnutná. [2]

Šikmo umiestnené dokumenty

Pre systémy OCR nie je uhol pohľadu vstupného obrazu, ktorý je zachytený kamerou ručného zariadenia, alebo iných zariadení, pevne daný ako pri skeneri. To môže viesť ku skresleniu textových riadkov a odchýleniu od pôvodnej orientácie. Ak je takýto obraz predložený OCR systému, môže dôjsť k výrazne horším výsledkom. [2]

Vplyv typov písma na kvalitu v OCR

Kurzíva a ozdobné písma môžu spôsobovať prekrytie znakov, čo komplikuje niektoré základné procesy OCR, ako je segmentácia. Rôzne typy písiem vykazujú veľké variácie a vytvárajú mnoho vzorov, čo sťažuje presné rozpoznávanie najmä pri veľkom počte znakových tried. [2]

Spracovanie viacjazyčných dokumentov

Hoci jazyky s latinskou abecedou obsahujú veľa znakov, jazyky ako japončina, čínština a kórejčina majú obrovské množstvo znakových tried. V arabských jazykoch sa znaky v závislosti od kontextu menia a prepájajú, čím sa ich tvar prispôsobuje písanému textu. V hindčine sú slabiky tvorené kombináciou znakov do tisícov rôznych tvarov. V OCR systémoch pre viacjazyčné prostredia, zostáva spracovanie skenovaných dokumentov veľkým problémom, pretože spracovanie textu je v takýchto prípadoch oveľa náročnejšie. [2]

3 PREHĽAD SÚČASNÝCH OPEN SOURCE OCR SYSTÉMOV

Cieľom tejto práce je analyzovať súčasné open-source OCR systémy a porovnať ich pri rôznych typoch použitia. Prvým krokom je podrobná analýza vybraných OCR riešení, ktoré pokrývajú určité spektrum aplikácií, ako sú štruktúrované dokumenty a text v reálnych scénach.

V tejto kapitole sa zameriame na 4 významné a voľne dostupné systémy. Každý z nich má svoje špecifické vlastnosti a využitie. Tesseract je dlhodobo najpopulárnejšie open-source riešenie s vysokou presnosťou pri štruktúrovaných dokumentoch. OCRmyPDF rozširuje Tesseract o možnosť spracovania PDF súborov. EasyOCR ponúka jednoduché a rýchle riešenie pre širokú škálu jazykov, zatiaľ čo PaddleOCR sa vyznačuje vysokou presnosťou a podporou pokročilých modelov. Podrobný popis jednotlivých systémov je nižšie.

3.1 Tesseract OCR

Tesseract OCR je open-source systém používaný v oblasti OCR, ktorý pracuje na binárnych obrazoch s voliteľne definovanými textovými oblasťami. Využíva tradičný krokový pipeline prístup so špecifickými fázami spracovania. [9]

Spracovanie textu prebieha v niekoľkých nasledovných fázach [10]:

Analýza komponentov a segmentácia textu

V tejto fáze sa ukladajú obrysy znakov a analyzuje sa vnorenie obrysov. Rozpoznáva sa inverzný text a text sa organizuje do riadkov a slov. Pri pevnej šírke znakov sa delí priamo, pri proporcionálnom texte sa analyzujú medzery medzi znakmi.

Dvojfázové rozpoznávanie textu

V prvej fáze sa analyzujú jednotlivé slová a úspešne rozpoznané slová sa používajú na trénovanie adaptívneho kvalifikátora. V druhej fáze sa menej presne rozpoznané slová znovu analyzujú s využitím naučených vzorov.

Hľadanie a organizácia riadkov

Algoritmus umožňuje rozpoznávanie šikmých strán bez ich narovnávania, čím sa zabraňuje strate kvality obrazu. Použitím výškového filtra sa odstraňujú veľké písmená, diakritika a šum. Skupiny znakov sa zoradia a priradia k textovým riadkom s ich sklonom.

Prispôsobenie základnej línie textu

Metóda quadratic spline upravuje zakrivenie textu, čím umožňuje lepšie rozpoznávanie aj na skenovaných dokumentoch so zvlnenými riadkami.

Rozdelenie slov a detekcia šírky znakov

Pri pevnej šírke znakov sa text rozdelí rovnomerne. Pri proporciálnom texte sa analyzujú medzery medzi znakmi, pričom sa nejasné segmenty riešia až v záverečnej fáze rozpoznávania.

Rozpoznávanie slov a optimalizácia výstupu

Ak rozpoznávanie slova nie je dostatočne presné, systém sa pokúsi rozdeliť znaky s najnižšou presnosťou a znovu ich analyzuje. V prípade nejednoznačnosti využíva A* algoritmus na hľadanie najlepšej kombinácie znakov a optimalizuje výstup.

3.2 OCRmyPDF

OCRmyPDF je open-source nástroj napísaný v jazyku Python, ktorý pridáva vrstvu s rozpoznaným textom do naskenovaných PDF súborov, čím ich robí preskúmateľnými a umožňuje kopírovanie textu. Tento nástroj je ideálny pre tých, ktorí potrebujú konvertovať naskenované dokumenty do formátu PDF/A s možnosťou vyhľadávania. [11]

Spracovanie textu prebieha v nasledovných fázach [12]:

Extrahovanie obrázkov z .pdf

Systém najprv analyzuje vstupný súbor a následne extrahuje obrázky pre ďalšie spracovanie.

Predspracovanie obrázkov

Tento krok je voliteľný. Automaticky sa deteguje orientácia a otočenie strán na správny smer. Ak je text mierne naklonený, tak sa narovná. Odstráni sa šum a zlepší kontrast pre lepšiu čitateľnosť.

OCR pomocou Tesseract

Systém používa Tesseract OCR, ktorý rozpoznáva text v obrázkoch. Identifikuje a extrahuje text zo strán ako čitateľný textový obsah.

Vytvorenie PDF s textovou vrstvou

Pôvodný obrázok sa ponechá nedotknutý, ale pod neho sa pridá neviditeľná textová vrstva. Používateľ tak vidí pôvodný dokument, čo mu umožní kopírovať text a vyhľadávať v ňom.

Rozdiel OCRmyPDF oproti TesseractOCR spočíva v účele, funkcionalite a použití. OCRmyPDF spracováva celé PDF súbory a nie iba obrázky. Umožňuje taktiež

otočenie nesprávne orientovaných strán, narovnanie nakloneného textu a kompresiu obrázkov na zníženie veľkosti výstupného PDF súboru. [12]

3.3 EasyOCR

EasyOCR je open-source nástroj, umožňujúci extrahovať text z obrázkov a dokumentov. Využíva hlboké neurónové siete a je optimalizovaný pre vyše 80 jazykov. Bol vyvinutý spoločnosťou Jaided AI a je postavený na frameworkoch PyTorch a OpenCV. Na podrobné spracovanie vstupu sú využité modely hlbokého učenia ako Resnet a VGG. [13]

Fungovanie EasyOCR je možné popísať v nasledovných krokoch [14]:

Predspracovanie text

EasyOCR upravuje vstupný obrázok, odstráni šum a zvýrazní text pre lepšie rozpoznanie. Používa na to techniky ako binarizácia, normalizácia jasu a kontrastu.

Detekcia textu

Pomocou modelu CRAFT (Character-Region Awareness For Text detection) identifikuje oblasti obsahujúce text. Tento model je schopný detegovať text aj v rôznych orientáciách a na zložitých pozadiach.

Rozpoznávanie znakov

Použitím kombinácie CNN (Convolutional Neural Network) a RNN (Recurent Neural Network) prevádza rozpoznané oblasti na čitateľný text.

Dodatočné spracovanie

Oprava chýb, odstránenie nepresností a optimalizácia výsledkov na zvýšenie presnosti. EasyOCR tiež podporuje opravu gramatických chýb a validáciu výsledku na základe výskytu slov.

3.4 PaddleOCR

PaddleOCR je open-source systém na rozpoznávanie znakov, založený na hlbokom učení. Využíva framework PaddlePaddle na efektívne detegovanie a rozpoznávanie textu v obrázkoch a dokumentoch. Podporuje viac ako 80 jazykov. Tento systém používa mnoho rôznych modelov, spomeniem len zopár ako PP-OCRv4_server_rec_doc na rozpoznávanie textu, PP-LCNet_x0_25_textline_ori pre klasifikáciu orientácie riadku textu. [15]

Funkčnosť systému pozostáva z troch hlavných komponentov [15]:

Detekcia textu

Používa model DBNet na rýchlu a presnú detekciu textových oblastí. Funguje na princípe segmentácie a dokáže efektívne rozpoznať text aj v náročných podmienkach (šikmý, zakrivený text).

Rozpoznávanie textu

Používa CRNN (Convolutional Recurent neural network) na čítanie textu v detegovaných oblastiach. Podporuje viacero jazykov a modely môžu byť trénované na špecifické abecedy. Implementuje Attention Mechanism pre lepšiu presnosť pri komplikovanom texte.

Klasifikácia textu

Používa Angle Classification Network, ktorá koriguje orientáciu textu pred samotným rozpoznaním. Tento krok je voliteľný, ale zlepšuje presnosť pri textoch otočených v rôznych uhloch.

Systém na začiatku konvertuje obrázok na vhodný formát a následne normalizuje veľkosť obrázka. Pomocou modelu DBNet identifikuje oblasti s textom a vygeneruje binárnu masku na lokalizáciu textu. Model CRNN prečíta text a prevedie ho na reťazec znakov. V konečnej fáze sa vykoná možná korekcia chýb v texte pomocou jazykových modelov.

3.5 Záver a porovnanie OCR systémov

Hoci existuje mnoho ďalších open-source OCR systémov, ako napríklad Kraken OCR, ktoré sa špecializuje na historické, nie latinským písmom písané rukopisy [16], alebo Calamari OCR, nadväzujúci na OCRopy a Kraken [17], výber systémov v tejto kapitole bol cielený. Kraken OCR nebol vybraný, lebo si nemyslíme, že by bol vhodnou voľbou pre oblasť štruktúrovaných dokumentov a reálnych scén. Calamari OCR riešenie sa nám nepodarilo implementovať do demonštračného nástroja a preto nemohlo byť použité. Zohľadňovala sa popularita, dostupná dokumentácia, jednoduchosť ovládania a možnosť jednoduchého nasadenia do vlastných aplikácií.

Kľúčovým kritériom výberu bola aj doména využitia, teda typ dokumentov, na ktoré boli tieto nástroje zamerané. Zároveň analyzované systémy majú predstavovať rôzne technologické prístupy: od tradičného pipeline prístupu v Tesseract systéme, cez rozšírené spracovanie pomocou PDF súborov v OCRmyPDF systéme, až po moderné riešenia založené na hlbokom učení ako systémy EasyOCR a PaddleOCR.

Licenčné podmienky vybraných riešení:

- Tesseract: Apache License 2.0 [10]
- OCRmyPDF: Mozilla Public License 2.0 [12]
- EasyOCR: Apache License 2.0 [14]
- PaddleOCR: Apache License 2.0 [15]

Tieto licencie umožňujú flexibilné použitie v komerčných aj akademických projektoch, pričom zabezpečujú transparentnosť a možnosť prispievania ku kódu.

V nasledujúcej tabuľke sú popísané domény systémov, ich výhody a nevýhody.

Tabuľka 1 Tabuľka porovnaní vybraných systémov

Systém	Doména využitia	Výhody	Nevýhody
Tesseract	Štruktúrované dokumenty	Overená technológia, vysoká presnosť	Slabšia podpora pre neštruktúrovaný text
OCRmyPDF	Štruktúrované dokumenty	Nadstavba Tesseract, dodatočné funkcie	Menej flexibilné pre iné vstupy ako Tesseract
EasyOCR	Text v rôznych jazykoch a scénach	Podpora 80+ jazykov, hlboké učenie	Menej presné výsledky pri komplikovaných štruktúrach
PaddleOCR	Pokročilé aplikácie	Vysoká presnosť, modulárnosť	Vyššia komplexita nasadenia

Výber vhodného OCR nástroja tak závisí nielen od jeho technických parametrov a presnosti rozpoznávania, ale aj od konkrétneho použitia, jazykovej podpory, ich rôznorodosti na základe výhod a nevýhod a náročnosti nasadenia do reálneho prostredia.

4 IDENTIFIKÁCIA A VÝBER VHODNÝCH DATASETOV PRE TESTOVANIE RIEŠENÍ V RÔZNYCH SCENÁROCH

4.1 Identifikácia a typy datasetov

Pri testovaní open-source OCR systémov je dôležitý správny výber datasetu, ktorý bude reprezentovať rôzne scenáre, s ktorými sa OCR systémy stretávajú v praxi. Rozhodli sme sa, že nebudeme testovať na rozsiahlych datasetoch, ale skôr na rôznorodých datasetoch, ktoré zachytávajú rôzne podmienky. Vybrané datasety by mali byť odlišné v type textu a rôznych podmienkach snímania.

OCR systémy sa môžu používať na spracovanie rôznych typov textových vstupov. Hlavný rozdiel medzi datasetmi pre štruktúrované dokumenty a datasetmi z reálnych scén spočíva v podmienkach, v ktorých sa text nachádza.

Datasety pre štruktúrované dokumenty obsahujú a spĺňajú:

- Pokladničné doklady, emaily, formuláre, záznamy obrazoviek, rôzne orientované a zakrivené texty
- Text býva dobre formátovaný, často horizontálny a s minimálnymi deformáciami
- Farby a kontrast bývajú optimalizované, vo vysokom rozlíšení, aby bol text dobre čitateľný
- Výzvy zahŕňajú rozpoznávanie tabuliek, rukopisov a rôznych typografických štruktúr, všeobecne dobre definované textové bloky s jasnou hierarchiou

Datasety pre reálne scény obsahujú a spĺňajú:

- Fotografie s textom zo skutočného sveta, fotky billboardov a záznamy obrazoviek z aplikácií
- Text je často deformovaný, rôzne orientovaný a môže byť čiastočne zakrytý. Osvetlenie je variabilné a môže spôsobovať tieňovanie alebo odlesky, ku ktorým môže dôjsť z dôsledku textu nachádzajúceho sa na rôznych povrchoch
- Obrázky často obsahujú komplikované pozadie a množstvo textúr
- Výzvy zahŕňajú odstránenie šumu, korekciu perspektívy a rozpoznanie textu na nehomogénnom pozadí

OCR systém by mal byť schopný pracovať s oboma typmi datasetov, pričom pre každý scenár môžu byť vhodné rôzne metódy predspracovania a modely strojového

učenia. Avšak môže dôjsť k rozdielom vo výsledkoch napríklad z dôvodu, trénovania modelov. Systém zameraný na štruktúrované dokumenty by mohol zlyhávať pri rozmazaných, alebo perspektívne skreslených textoch. Naopak, systém určený na reálne scény by mohol mať problém s presnou extrakciou údajov z rôznych typografických štruktúr v dokumentoch.

4.2 Anotácie datasetov

Anotácia datasetu predstavuje kľúčový krok v príprave datasetu na testovanie OCR systémov. Ide o proces, pri ktorom sa vstupným dátam priraďujú presné a overené textové popisy, tzv. pravdivé hodnoty, anglicky ground truth. Tieto texty slúžia ako referenčný základ, s ktorým sa následne porovnáva výstup OCR systému. [17]

Anotácie sú potrebné najmä z nasledovných dôvodov:

- **Objektívne hodnotenie presnosti –** Bez anotácie nie je možné presne určiť, do akej miery OCR systém správne rozpoznal text. Anotované dáta umožňujú vypočítať metriky ako sú CER (Character Error Rate) a WER (Word Error Rate), ktorými kvantifikujeme výkon systému.
- **Porovnateľnosť výsledkov –** Vďaka jednotnej anotácií môžeme porovnávať rôzne OCR systémy medzi sebou za rovnakých podmienok. Ak by boli dáta neanotované, alebo nejednotné, výsledky jednotlivých testov by boli neporovnateľné a neobjektívne.
- **Zabezpečenie kvality datasetu –** Počas anotácie sa často odhalia problémy v samotnom datasete, ako napríklad nejasné, alebo nečitateľné texty, chýbajúce súbory či nesprávne zaradenie vstupov. Anotácia tak slúži ako kontrola kvality dát a dodatočné upravenie obrázkov.

Anotácia teda nie je len technickým medzi krokom, ale zásadným krokom v procese testovania OCR systémov. Zabezpečuje transparentnosť, opakovateľnosť a spoľahlivosť hodnotenia, pričom umožňuje analýzu a porozumenie správania systému.

4.3 Popis použitých datasetov

Nižšie spomenuté a popísané datasety sme použili na testovanie vyššie spomenutých OCR systémov. Naše datasety sú oproti pôvodným upravené tak, že obsahujú iba malú časť obrázkov. Tieto obrázky boli filtrované, aby boli na prvý pohľad zrozumiteľné pre ľudské oko. Taktiež boli niektoré časti obrázku zakryté farbou, ak obsahovali nejasný text. Link ku pôvodným datasetom bude na príslušnom čísle v zozname použitej literatúre pri datasete. Použité upravené datasety budú prístupné v elektronickej prílohe.

ICDAR 2003

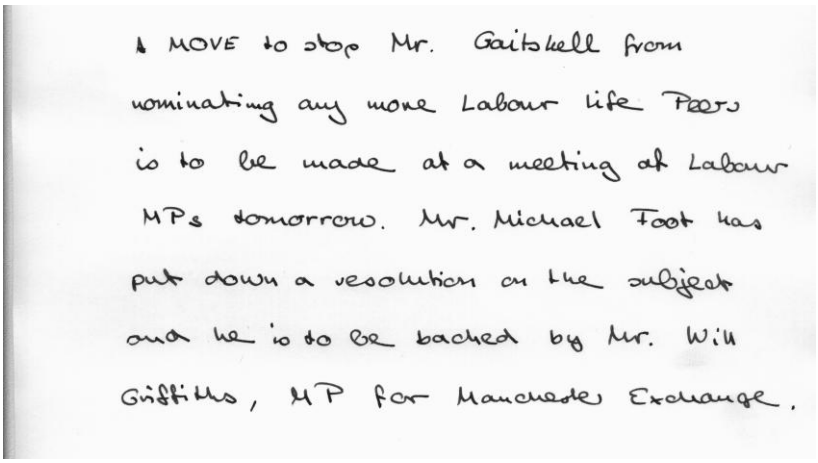
Tento dataset je súčasťou ICDAR 2003 Robust Reading Competitions a zameriava sa na čítanie textu v prirodzených scénach. Obsahuje fotografie z reálneho sveta (ulice, obchody, ...), kde sa vyskytuje text v rôznych formách a podmienkach. [18]

Obrázok 1 Príklad ICDAR 2003 Datasetu

Upravený dataset obsahuje 17 obrázkov, 434 písmen a 68 slov.

IAM Handwritten Forms

Dataset obsahuje ručne písané formuláre zo známeho IAM Handwriting datasetu. Využíva sa na trénovanie modelov pre rozpoznávanie rukopisu. [19]



Obrázok 2 Príklad IAM Handwritten Forms Datasetu

Upravený dataset obsahuje 10 obrázkov, 2 520 znakov a 501 slov.

SROIE

Dataset zameraný na extrakciu údajov z pokladničných dokladov a dokumentov so štruktúrovaným textom. Obsahuje pokladničné doklady v rôznych formách a rozlíšeniach. [20]


CROSS CHANNEL NETWORK SDN. BHD.
47, JALAN MERANTI 1, SEK. 3,
BANDAR UTAMA BATANG KALI,
44300 BATANG KALI, SELANGOR.
Tel : 03-6057 9688 Fax : 03-6057 9678
GST ID : 001151500288
Tax Invoice
No. : BTG-060036

	Qty	RM Tax
20MM HERO PVC CONDUIT PIPE @ 1.7000	1	1.80 SR
15A PVC CONNECTOR @ 1.8000	1	1.91 SR
R & B-C/CLIP 8MM X 100PCS CABLE CLIPS @ 3.5000	1	3.71 SR
3" X 3" 2862 A1 PVC NUT BOX @ 0.5000	1	0.53 SR

Total Amt Incl GST @ 6% : 7.95
Rounding Adjustment :
Total Amt Payable : 7.95
Paid Amount : 10.00
Change : 2.05
Total Qty Tender : 4.00

GST Summary	Amount (RM)	Tax (RM)
SR @ A	7.50	0.45
Total	7.50	0.45

THANK YOU
For any enquiry, please e-mail us:
gowa9688@gmail.com
31/12/2017 10:33:59 AM

Obrázok 3 Príklad SROIE Datasetu

Upravený dataset obsahuje 9 obrázkov, 5 742 znakov a 1237 slov.

Street View Text Dataset

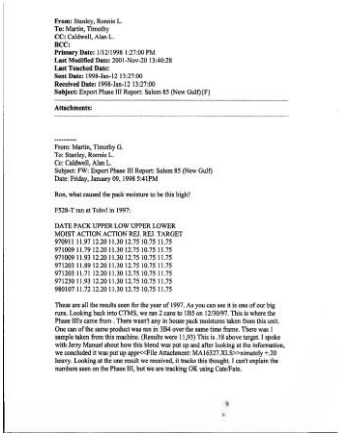
Dataset obsahuje textové nápisy z ulíc, billboardov často názvy firiem, reštaurácií alebo obchodov. Zameraný na extrakciu krátkych slov alebo názvov z reálneho sveta. [21]

Obrázok 4 Príklad Street View Text Datasetu

Upravený dataset obsahuje 13 obrázkov, 231 znakov a 41 slov.

Scanned images dataset for ocr and vlm finetuning

Dataset obsahuje naskenované dokumenty, emailové hlavičky, štruktúrovaný a čiastočne štruktúrovaný text. Výborný na trénovanie OCR metód v dokumentoch s hlavičkami. [22]



From: Stanley, Ronnie L.
To: Martin, Timothy
CC: Caldwell, Alan L.
BCC:
Primary Date: 1/12/1998 1:27:00 PM
Last Modified Date: 2001-Nov-20 13:40:28
Last Touched Date:
Sent Date: 1998-Jan-12 13:27:00
Received Date: 1998-Jan-12 13:27:00
Subject: Export Phase III Report: Salem 85 (New Gulf)(F)

Attachments:

From: Martin, Timothy G.
To: Stanley, Ronnie L.
Cc: Caldwell, Alan L.
Subject: FW: Export Phase III Report: Salem 85 (New Gulf)
Date: Friday, January 09, 1998 5:41PM

Ron, what caused the pack moisture to be this high?

F528-T ran at Tolivl in 1997:

DATE PACK UPPER LOW UPPER LOWER
MOIST ACTION ACTION REJ. REJ. TARGET
970911 11.97 12.20 11.30 12.75 10.75 11.75
971009 11.79 12.20 11.30 12.75 10.75 11.75
971009 11.93 12.20 11.30 12.75 10.75 11.75
971203 11.89 12.20 11.30 12.75 10.75 11.75
971203 11.71 12.20 11.30 12.75 10.75 11.75
971230 11.93 12.20 11.30 12.75 10.75 11.75
980107 11.72 12.20 11.30 12.75 10.75 11.75

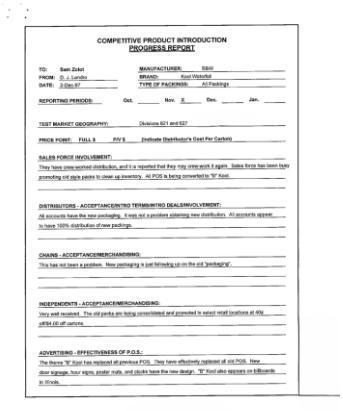
These are all the results seen for the year of 1997. As you can see it is one of our big runs. Looking back into CTMS, we ran 2 cans to 1I85 on 12/30/97. This is where the Phase III's came from . There wasn't any in house pack moistures taken from this unit. One can of the same product was ran in 3B4 over the same time frame. There was 1 sample taken from this machine. (Results were 11,93) This is .38 above target. I spoke with Jerry Manuel about how this blend was put up and after looking at the information, we concluded it was put up appr<<File Attachment: MA16327.XLS>>imately +.20 heavy. Looking at the one result we received, it tracks this thought. I can't explain the numbers seen on the Phase III, but we are tracking OK using CuteTrate.

Obrázok 5 Príklad Scanned images dataset for ocr and vlm finetuning Datasetu

Upravený dataset obsahuje 10 obrázkov, 12 597 znakov a 2 524 slov.

FUNSD

Dataset pre porozumenie formulárov v zašumených naskenovaných dokumentoch, ktorej cieľom je extrahovať a vytvoriť štruktúru pre textový obsah formulárov. [23]



COMPETITIVE PRODUCT INTRODUCTION
PROGRESS REPORT

TO: Sam Zidel
FROM: D. J. Landis
DATE: 3 Dec 87

MANUFACTURER: Stihl
BRAND: Kool Wisteful
TYPE OF PACKINGS: All Packings

REPORTING PERIODS: Oct. ____ Nov. X Dec. ____ Jan. ____

TEST MARKET GEOGRAPHY: Divisions 621 and 627

PRICE POINT: FULL $ RW $ (Indicate Distributor's Cost Per Carton)

SALES FORCE INVOLVEMENT:
They have crew-worked distributors, and it is reported that they may crew-work it again. Sales force has been fully promoting old style packs to clean up inventory. All POS is being converted to "R" Kool.

DISTRIBUTORS - ACCEPTANCE/INTRO TERMS/INTRO DEALS/INVOLVEMENT:
All accounts have the new packaging. It was not a problem obtaining new distribution. All accounts appear to have 100% distribution of new packings.

CHAINS - ACCEPTANCE/MERCHANDISING:
This has not been a problem. New packaging is just following up on the old "packaging".

INDEPENDENTS - ACCEPTANCE/MERCHANDISING:
Very well received. The old packs are being consolidated and promoted in select retail locations at 40¢/ pk vs$64.00 off cartons.

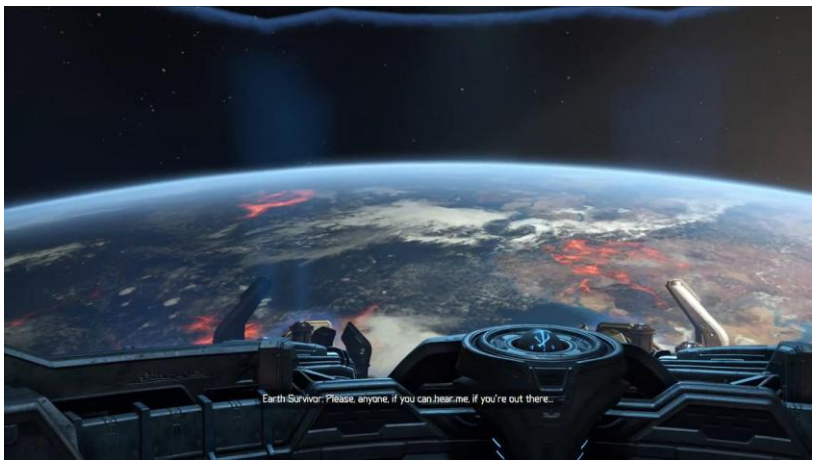
ADVERTISING - EFFECTIVENESS OF P.O.S.:
The theme "R" Kool has replaced all previous P.O.S. They have effectively replaced all old POS. New door signage, hour signs, poster mats, and clocks have the new design. "R" Kool also appears on billboards in Illinois.

Obrázok 6 Príklad FUNSD Datasetu

Upravený dataset obsahuje 10 obrázkov, 7 840 znakov a 1 457 slov.

GameplayCaptions

Dataset pozostáva zo záznamov obrazovky z hier spolu s textovými popismi alebo titulkami. Vhodné na testovanie OCR metód, alebo viac modálnych modelov s textom v hernom kontexte. [24]



Obrázok 7 Príklad GameplayCaptions Datasetu

Upravený dataset obsahuje 12 obrázkov, 1 774 znakov a 382 slov.

Screenshots-dataset

Dataset záznamov obrazovky z rôznych zariadení a aplikácií (napr. plocha, mobil, webové stránky), obsahuje rôzne formy textu (menu, hlavičky, tlačidlá). [25]

Obrázok 8 Príklad Screenshots-dataset Datasetu

Upravený dataset obsahuje 17 obrázkov, 4 538 znakov a 887 slov.

Multi - oriented

Dataset zameraný na rozpoznávanie textu v rôznych orientáciách a uhlových natočeniach. Dataset bol vlastnoručne vytvorený.

Obrázok 9 Príklad Multi - oriented Datasetu

Dataset obsahuje 10 obrázkov, 215 znakov a 38 slov.

CAPTCHA

Dataset obsahujúci CAPTCHA obrázky. Vytvorený v 1997 ako spôsob na odlíšenie používateľov od botov, s cieľom zabrániť spamu, DDoS útokom a podobne. Odvtedy bol tento systém nahradený systémom reCAPTCHA, pretože sa dá prelomiť pomocou umelej inteligencie. [26]

Obrázok 10 Príklad CAPTCHA Datasetu

Upravený dataset obsahuje 10 obrázkov, 50 znakov a 10 slov.

4.4 Zhrnutie a záver ku datasetom

V tejto kapitole sme popísali prečo sme zvolili rôznorodé datasety, ktoré reprezentujú široké spektrum vstupov, od čisto štruktúrovaných dokumentov, cez ručne písané texty, až po zložité vizuálne scény z reálneho či dokonca herného prostredia. Dôležitým aspektom pri výbere bola nielen rôznorodosť obsahu, ale aj variabilita vo veľkosti, formáty textu, orientácia a kvalita textu na obrázkoch.

Niektoré datasety ako FUNSD alebo Scanned Images for OCR and VLM finetuning, poskytujú komplexné štruktúrované dokumenty vhodné pre extrakciu údajov, zatiaľ čo datasety ako Street View Text, alebo GameplayCaptions testujú robustnosť modelov voči rušivým vizuálnym prvkom a nehomogénnemu pozadiu. Mimoriadne náročné sú aj datasety ako CAPTCHA či Multi-oriented, ktoré skúšajú limity OCR nástrojov pri práci s deformovaným, alebo neštandardne orientovaným textom.

Z praktických dôvodov sme pre testovanie zvolili menšie výseky z pôvodných datasetov, čo nám umožnilo rýchlejšiu a detailnejšiu analýzu s lepšou kontrolou kvality anotácií. Napriek tomu však datasety pokrývajú viac než 30-tisíc znakov a niekoľko tisíc slov, čo poskytuje dostatočný základ pre objektívne porovnanie výkonu jednotlivých OCR systémov.

Táto diverzia a systematický prístup k anotáciám umožnili vytvoriť testovacie prostredie, ktoré odráža skutočné výzvy pri nasadzovaní OCR technológií v praxi. V ďalších kapitolách sa preto môžeme oprieť o tieto výsledky a analyzovať silné a slabé stránky hodnotených systémov v realistických podmienkach.

5 METRIKY TESTOVANIA OCR

Vyhodnocovacie metriky ako WER (Word Error Rate) a CER (Character Error Rate), popísané nižšie, sú rozsiahlo používané a zohrávajú veľmi dôležitú úlohu pri hodnotení výkonnosti systémov OCR oblasti. Merajú presnosť rozpoznaného textu porovnaním s referenčným textom, čím poskytujú objektívne ukazovatele presnosti OCR systému. [27]

5.1 WER

Miera chybovosti slov (WER) je založená na Levenshteinovej vzdialenosti (Levenshtein distance), o minimálnom počte substitúcií, vymazaní a vložení, ktoré musia byť vykonané na konverziu rozpoznaného textu hypotézy na referenčný text. [28] Nevýhodou WER je skutočnosť, že neumožňuje zmenu poradia slov, pričom poradie slov v hypotéze môže byť odlišné od poradia slov v referencii, aj keď ide o správny preklad. Čím je výsledná hodnota nižšia, tým je správnosť systému lepšia. Výsledná hodnota rovná nule reprezentuje bezchybný výsledok. [28]

5.2 CER

Miera chybovosti znakov (CER) funguje rovnako ako WER, ale pracuje na úrovni znakov a nie na úrovni slov.

CER trestá rôzne chyby odlišne. Drobné preklepy, ako sú napríklad chýbajúce diakritické znamienka, by pri WER viedli k úplnej chybe substitúcie, zatiaľ čo pri CER by išlo len o jednu chybu substitúcie znaku, To však nemusí byť vždy výhoda, pretože aj jednopísmenové chyby môžu meniť význam. Rovnako ako pri WER, čím je výsledná hodnota nižšia, tým je správnosť systému lepšia. Výsledná hodnota rovná nule reprezentuje bezchybný výsledok. [30]

Táto metrika je pomalšia na výpočet, pretože vzdialenosť úprav sa musí počítať na oveľa dlhšej sekvencii. [31]

Obe z týchto metrík môžeme vypočítať nasledovným vzťahom, v ktorom používame počet slov pri WER, alebo počet znakov pri CER.

$$CER, WER = \frac{S + D + I}{N} = \frac{S + D + I}{S + D + C} \quad (1)$$

Kde *S* je počet substitúcií,
D je počet vymazaní,
I je počet vložení,
C je počet správnych slov (WER) / znakov (CER),
N je počet slov / znakov v referencii (*N* = *S* + *D* + C*)*.

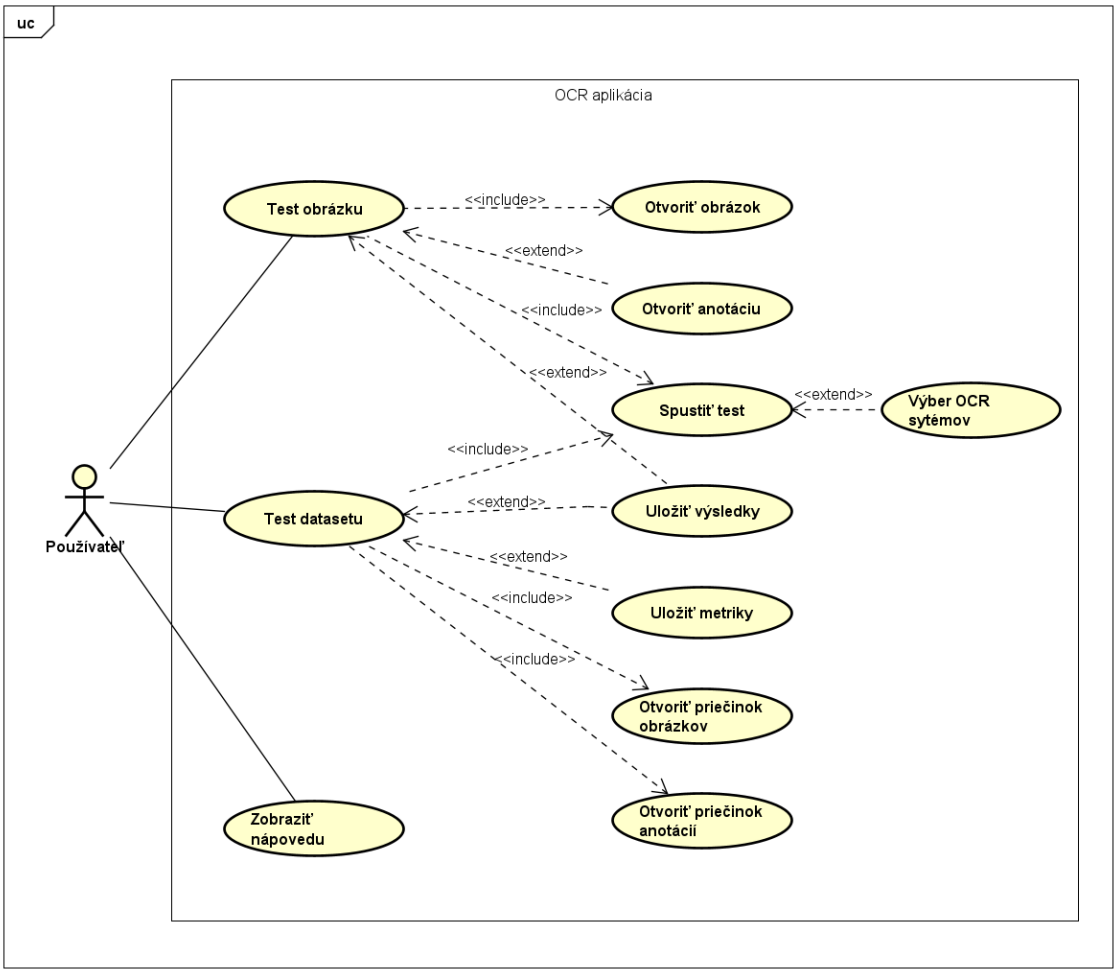
6 DEMONŠTRAČNÝ NÁSTROJ

6.1 Popis aplikácie

Aplikácia bola vyvinutá ako súčasť tejto práce a slúži ako demonštračný a testovací nástroj pre OCR systémy. Jej cieľom je poskytnúť jednoduchý a používateľsky prístupný nástroj, ktorý umožňuje načítavanie obrázkov, spracovanie textu pomocou vybraných OCR riešení a uloženie výsledkov v textovej forme.

Používateľ má možnosť nahrať jeden obrázok, alebo celý súbor obrázkov na hromadné spracovanie. Rovnako je možné nahrávať aj anotácie a to jednotlivo, alebo ako súbor anotácií. Po vykonaní testov, alebo extrahovaní textu z obrázkov, je možné si výsledky jednotlivých OCR systémov exportovať v textovej podobe. Podobne je možné exportovať aj metriky testu. Aplikácia obsahuje jednoduché grafické používateľské rozhranie pre jednoduché navigovanie. Je prístupná v elektronickej prílohe.

6.2 Spôsob obsluhy aplikácie z používateľského pohľadu



Obrázok 11 Diagram prípadov použitia z používateľského pohľadu

Po spustení aplikácie sa zobrazí hlavné menu, kde je uvedený názov aplikácie a tri tlačidlá. Dve z nich slúžia na presun používateľa do ďalších častí aplikácie. Pod týmito tlačidlami sa nachádza krátky popis, ktorý vysvetľuje jeho funkciu.

V ľavom hornom rohu sa nachádza malé tlačidlo, s ikonou otáznika, ktoré po kliknutí otvorí informačné okno so stručným návodom na používanie aplikácie. Toto tlačidlo je prístupné z ktoréhokoľvek okna aplikácie, pričom si zachováva svoju pozíciu.

Po stlačení prvého tlačidla v hlavnom menu sa otvorí rozhranie určené na spúšťanie OCR a testovanie jedného obrázku. V tomto rozhraní sa nachádza viacero ovládacích prvkov:

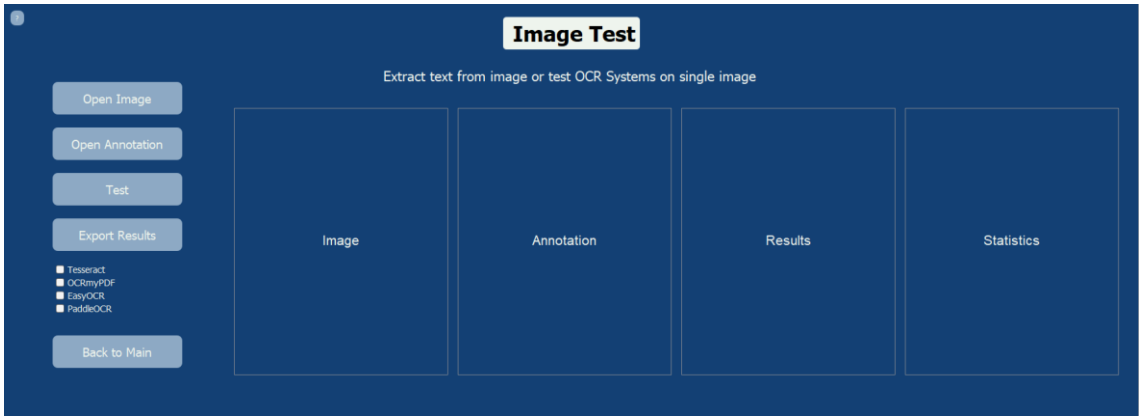
- **Open Image** – otvorí systémový prehliadač súborov, pomocou ktorého si používateľ vyberie vstupný obrázok.
- **Open Annotation** – umožňuje výber referenčného (anotačného) textu, ktorý sa použije pri testovaní presnosti OCR.
- **Test** – spustí samotný test. Ak nie je načítaná anotácia, vykoná sa len extrakcia textu. Ak je anotácia dostupná, vykoná sa porovnanie výstupu s referenciou.
- **Export Results** – otvorí dialóg pre výber umiestnenia, kam sa majú výsledky uložiť. Výstup bude uložený v textovom formáte ako image_{NázovSystému}.txt.

Ďalej nasledujú štyri zaškrtávacie tlačidlá reprezentujúce jednotlivé OCR systémy. Používateľ si môže vybrať, ktoré systémy sa majú použiť pri spracovaní. Ak nie je vybraný žiadny systém, aplikácia spustí všetky dostupné systémy.

- **Back to Main** – vráti používateľa späť do hlavného menu aplikácie.

Rozhranie tejto časti obsahuje aj štyri vizuálne oblasti:

1. **Oblasť obrázka** – zobrazí sa na ňom načítaný obrázok s možnosťou priblíženia a oddialenia.
2. **Oblasť anotácie** – zobrazuje nahratý referenčný text.
3. **Oblasť výsledkov** – zobrazuje výstup OCR systémov.
4. **Oblasť metrík** – zobrazuje vyhodnotené metriky.



Obrázok 12 Menu aplikácie pre testovanie jedného obrázku

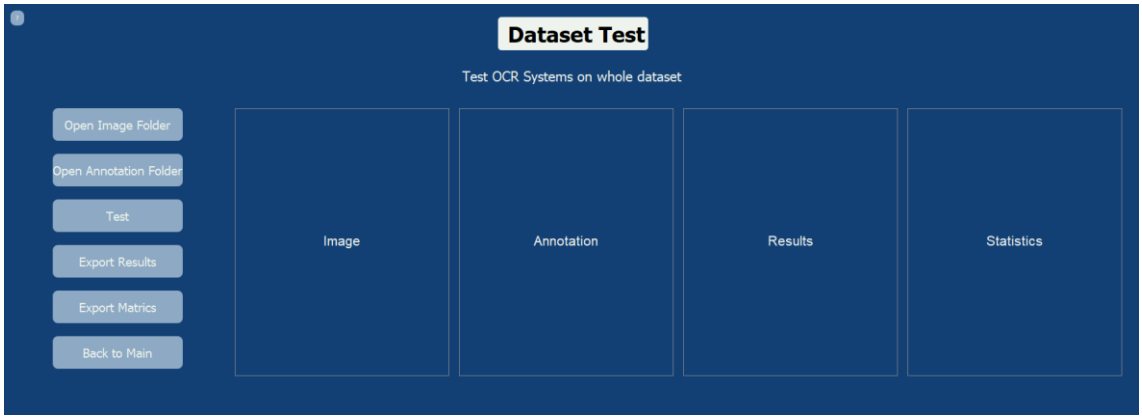
Po stlačení druhého tlačidla v hlavnom menu sa používateľovi zobrazí rozhranie určené na testovanie OCR systémov na datasete.

- **Open Image Folder** – otvorí prehliadač súborov, pomocou ktorého si používateľ vyberie priečinok obsahujúci obrázky. Po načítaní sa názvy obrázkov zobrazia v zozname pod sebou.
- **Open Annotation Folder** – umožňuje výber priečinka s anotáciami, ktoré musia byť vo formáte TXT a musia mať rovnaké názvy ako zodpovedajúce obrázky. Po nahratí sa zobrazí zoznam načítaných anotácií.
- **Test** – spustí testovanie na všetkých načítaných obrázkoch s priradenými anotáciami. Táto operácia môže v závislosti od veľkosti datasetu trvať niekoľko minút.
- **Export Results** – otvorí dialógové okno na výber umiestnenia, kam sa uložia výsledky. Výstupy sú generované pre každú kombináciu obrázku a OCR systému vo forme súborov s názvom: NázovObrázka_NázovSystému.txt.
- **Export Metrics** – umožňuje uložiť výsledné metriky do CSV súboru. Tento súbor obsahuje štyri stĺpce: Názov obrázka, Názov systému, CER a WER. Pre každý obrázok sú uvedené hodnoty všetkých systémov.
- **Back to Main** – vráti používateľa späť do hlavného menu.

Rozhranie tejto časti obsahuje aj štyri vizuálne časti:

1. **Zoznam obrázkov** – zobrazuje načítané obrázky určené na testovanie.
2. **Zoznam anotácií** – zobrazuje načítané referenčné texty.
3. **Výpis výsledkov** – obsahuje výstupy jednotlivých OCR systémov pre všetky obrázky.

4. **Tabuľka metrík** – prehľadne zobrazuje metriky (CER a WER) pre každý obrázok a každý systém. Na konci tabuľky sa nachádza riadok overall_average pre každý systém, ktorý obsahuje priemerné hodnoty metrík naprieč celým datasetom. Tabuľku je možné zoraďovať podľa jednotlivých stĺpcov zostupne, alebo vzostupne, čím sa uľahčuje analýza výkonnosti systémov.



Obrázok 13 Menu aplikácie pre testovanie datasetu

6.3 Použité knižnice a frameworky v aplikácií

V nasledujúcej časti sú popísané hlavné knižnice, frameworky a dodatočné úpravy ktoré boli použité a nutné pri implementácií aplikácie.

Python

Celá aplikácia bola vyvíjaná v jazyku Python, konkrétne vo verzii 3.10.0. Táto verzia bola zvolená kvôli kompatibilite s vybranými OCR riešeniami a dostupnosti všetkých potrebných knižníc.

Frontend

Grafické rozhranie aplikácie bolo vytvorené pomocou knižnice PyQt5, ktorá poskytuje nástroje na tvorbu grafického používateľského rozhrania v jazyku Python. Návrh rozhrania bol realizovaný v nástroji Qt Designer, ktorý natívne podporuje PyQt5. Výsledné návrhy boli exportované ako UI súbory, ktoré boli následne načítané a integrované do aplikácie prostredníctvom modulu PyQt5.uic.

OCRmyPDF funkcionalita

OCRmyPDF je nástroj určený na spracovanie PDF súborov, čo si vyžiadalo doplnenie špecifickej funkcionality do aplikácie. Pred samotným použitím OCRmyPDF bolo potrebné každý obrázok skonvertovať do PDF formátu, čo bolo zabezpečené knižnicou img2pdf. Po ukončení OCR procesu bolo potrebné z výsledného PDF získať text, čo sa realizovalo pomocou knižnice PyPDF2. Na

manipuláciu s dočasnými súbormi, ako napríklad kopírovanie alebo mazanie, bola použitá knižnica shutil.

Spracovanie obrázkov

Na prácu s obrázkami bola použitá knižnica PIL (Python Imaging Library), konkrétne jej verzia Pillow. V aplikácii sa využívala najmä na otváranie obrázkov ako vstupov pre OCR systémy, konverziu farebných módov (napr. RGBA do RGB) na zabezpečenie kompatibility, ukladanie a základné úpravy obrázkov.

Výpočet metrík

Pre výpočet hodnotiacich metrík CER a WER bola použitá knižnica Levenshtein. Táto knižnica poskytuje implementáciu Levenshteinovej vzdialenosti, ktorá bola využitá aj pri porovnávaní podobnosti medzi textovými reťazcami.

Github

Na správu verzií a priebežné ukladanie vývoja aplikácie bol použitý systém Git a vzdialený repozitár na stránke GitHub. Repozitár k aplikácií je prístupný na adrese [32].

7 TESTOVANIE A VYHODNOTENIE SYSTÉMOV

V tejto kapitole sú podrobne popísané postupy testovania jednotlivých OCR systémov, ktoré boli spomenuté vyššie v práci. Testovanie bolo realizované pomocou popísanej aplikácie vyššie, ktorá nám umožňuje automatizované spúšťanie OCR systémov nad nami zvolenými datasetmi a zber metrík. Vyhodnotenie testov je následne vizualizované a analyzované.

7.1 Postup testovania

Pri testovaní sme použili našu aplikáciu, ktorá poskytuje rozhranie pre spúšťanie OCR systémov. Každý z týchto systémov bol otestovaný na rovnakej množine dát, ktorú predstavujú nami zvolené datasety.

Prvým krokom testovania bolo otvorenie obrázkov datasetu v aplikácii, následne načítanie anotácií (referenčných textov) prislúchajúcim k jednotlivým obrázkom. Po tejto príprave bol spustený samotný test, počas ktorého aplikácia automaticky spracovala všetky obrázky v datasete.

Pre každý obrázok aplikácia spustila spracovanie pomocou všetkých testovaných systémov. Výstupy jednotlivých systémov boli uložené a následne porovnané s referenčnými anotáciami daného obrázku. Na základe tohto porovnania boli pre každý systém a každý obrázok vypočítané príslušné metriky.

Po spracovaní celého datasetu boli všetky metriky zozbierané a zobrazené v prehľadnej tabuľke priamo v aplikácii. Táto tabuľka bola následne exportovaná vo formáte vhodnom na ďalšie spracovanie a vizualizáciu výsledkov.

Pri testovaní sme pre jednotlivé datasety spúšťali meranie času, ktorý nám hovorí o orientačnom čase potrebnom na spracovanie jednotlivých datasetov. Pred uvedením jednotlivých časov je však dôležité uviesť nastavenia, s ktorými boli OCR systémi spustené.

Všetkým systémom bol prednastavený jazyk na angličtinu. V prípade EasyOCR a PaddleOCR systémov bolo zakázané využívanie grafickej karty. Navyše pre PaddleOCR systém bola aktivovaná funkcia klasifikácie natočenia textu počas detekcie. Ďalšie systémové nastavenia zostali v predvolenej konfigurácií a neboli nijakým spôsobom upravované.

Špecifickým prípadom bol OCRmyPDF systém, ktorého meraný čas je ovplyvnený niekoľkými dodatočnými faktormi. Tento systém nepracuje priamo s obrazovými súbormi, ale s formátom pdf. Preto bolo pre každý test potrebné vytvoriť pdf dokument, nahrať doň príslušný obrázok, spustiť systém, extrahovať text zo spracovaného dokumentu a následne vymazať dočasne vytvorený dokument.

7.2 Vyhodnotenie testov

Výsledky testov boli analyzované a porovnané na základe vyššie uvedených metrík. Výsledky sú zobrazené v tabuľkách, ktoré znázorňujú priemerné hodnoty metrík pre každý OCR nástroj pre dataset, alebo samostatný obrázok. Spomenuté sú taktiež prípady, pri ktorých sa výrazne líšia výsledné metriky.

V tabuľke nižšie sú zobrazené priemerné časy, ktoré ukazujú, koľko trvalo jednotlivým systémom spracovať obrázok v datasete. Tieto hodnoty sú v jednotkách sekúnd. Na základe týchto údajov možno pozorovať, že dĺžka spracovania je ovplyvnená zložitosťou dokumentov, samotnou architektúrou systému a počtom slov, alebo znakov v jednotlivých obrázkoch. Upozorňujeme, že namerané hodnoty boli vykonané iba v jednom behu programu, to znamená bez replikácií a teda predstavujú iba orientačné časy.

Tabuľka 2 Tabuľka priemerných časov spracovania obrázku pri testovaní

Názov Datastu	Tesseract	OCRmyPDF	EasyOCR	PaddleOCR
CAPTCHA	**0.16**	1.44	2.20	1.15
FUNSD	**0.52**	1.91	13.04	14.42
GameplayCaptions	**0.42**	2.55	14.60	4.32
IAMHandwrittenForms	**1.31**	4.75	48.02	6.38
multi-oriented	**0.27**	1.59	10.07	1.78
Scanned Images	**1.05**	3.45	41.91	16.90
ICDAR_2003	**0.30**	2.41	13.17	2.05
Screenshots	**0.53**	2.46	16.33	5.96
SROIE	**0.66**	2.74	16.56	13.35
Street View Text	**0.24**	1.95	8.87	2.04

Na obrázku nižšie sú vizuálne znázornené priemerné hodnoty metriky CER spolu so smerodajnými odchýlkami pre každý OCR systém pre jednotlivé datasety. Farebné odtiene od svetlých po tmavé reprezentujú výšku chybovosti – čím tmavšia farba, tým vyššia chybovosť.

Systémy OCRmyPDF a Tesseract vykazujú výrazne vyššie hodnoty pri datasetoch, ktoré obsahujú texty z reálneho prostredia, často skreslené, zakrivené a so zložitou štruktúrou. Naopak, v štruktúrovaných a pravidelných dokumentoch, ako napríklad SROIE, Scanned images dataset for ocr and vlm finetuning či Screenshots, dosahujú tieto systémy nízku chybovosť, čo potvrdzuje ich efektivitu v oblasti OCR v štruktúrovaných dokumentoch.

Systémy EasyOCR a PaddleOCR vykazujú stabilnejší výkon naprieč viacerými datasetmi. PaddleOCR sa ukazuje ako najúčinnejší systém pri vizuálne zložitejších datasetoch ako napríklad GameCaptions, ICDAR alebo StreetViewText.

EasyOCR má celkovo tiež dobrú výkonnosť, pričom v niektorých prípadoch dosahuje najnižšiu mieru chybovosti, avšak vo vizuálne náročnejších prípadoch mu klesá presnosť.

Dataset IAMHandwrittenText, ktorý obsahuje iba ručne písaný text, je zrejme jedným z najťažších, keďže žiaden systém nedosahuje nízku chybovosť. Naopak datasety Screenshots a Scanned Documents sa zdajú byť najjednoduchšie na spracovanie, keďže tam všetky systémy dosahujú veľmi nízke hodnoty CER, pod 0.1.

Väčšina systémov vykazuje nízke smerodajné odchýlky, čo svedčí o konzistentnosti výsledkov v rámci jednotlivých datasetov. Výnimkou sú datasety ako napríklad ICDAR kde je hodnota odchýlky dosahuje až 0.46, čo poukazuje na variabilitu výsledkov medzi vzorkami.

	EasyOCR	OCRmyPDF	PaddleOCR	Tesseract
CAPTCHA	0.46±0.35	1.00±0.35	0.34±0.35	1.00±0.35
FUSND	0.22±0.17	0.53±0.17	0.13±0.17	0.27±0.17
GameCaptions	0.26±0.27	0.69±0.27	0.18±0.27	0.67±0.27
IAMHandwrittenText	0.69±0.08	0.53±0.08	0.59±0.08	0.53±0.08
ICDAR	0.25±0.46	0.97±0.46	0.14±0.46	0.99±0.46
Multi-Oriented	0.65±0.12	0.88±0.12	0.65±0.12	0.85±0.12
SROIE	0.18±0.04	0.12±0.04	0.15±0.04	0.20±0.04
ScannedImages	0.08±0.02	0.04±0.02	0.05±0.02	0.04±0.02
Screenshots	0.04±0.02	0.07±0.02	0.03±0.02	0.07±0.02
StreetViewText	0.31±0.41	0.99±0.41	0.19±0.41	0.92±0.41

Obrázok 14 Priemerné hodnoty CER a smerodajné odchýlky naprieč datasetmi

	EasyOCR	OCRmyPDF	PaddleOCR	Tesseract
82250337_0338.png	0.23	0.70	0.07	0.10
82252956_2958.png	0.24	0.50	0.20	0.26
82253058_3059.png	0.10	0.50	0.03	0.12
82253245_3247.png	0.27	0.61	0.13	0.25
82253362_3364.png	0.35	0.70	0.13	0.56
82491256.png	0.05	0.55	0.03	0.55
82562350.png	0.29	0.56	0.22	0.31
83443897.png	0.22	0.51	0.04	0.09
83553333_3334.png	0.36	0.36	0.35	0.36
83624198.png	0.11	0.28	0.05	0.06

Obrázok 15 Hodnoty CER v datasete FUNSD

Pri teste datasetu FUNSD si môžeme všimnúť, že metriky veľmi kolíšu, pričom obrázky v tomto datasete sa od seba príliš neodlišujú a všetky predstavujú štruktúrovaný dokument. Napríklad pri piatom a šiestom obrázku sa presnosť Tesseract systému výrazne zhoršila oproti ostatným obrázkom. Piaty obrázok, síce obsahuje zložitejšie štruktúry ako sú napríklad tabuľky, zaškrtávacie polia, alebo zle podfarbené textové polia, ale šiesty obrázok neobsahuje žiadne štruktúry ako tabuľky, alebo deliace čiary navyše - je jednoducho štruktúrovaný.

Taktiež si môžeme všimnúť, že OCRmyPDF systém mal výrazné problémy oproti Tesseract systému, pričom OCRmyPDF sa štruktúrou a funkcionalitou oproti Tesseract systému až tak veľmi nelíši.

Obrázok 16 Testovacia vzorka – GameplayCaptions

Pri spracovaní vyššie zobrazeného obrázku z datasetu GameplayCaptions došlo k výraznému zhoršeniu výkonnosti OCR systémov Tesseract, OCRmyPDF a EasyOCR. Hodnoty metriky WER dosiahli 9.55, 6.45 a 8.27, zatiaľ čo CER sa pohybovala na úrovni 1.28, 0.87 a 1.11 pre Tesseract, OCRmyPDF a EasyOCR systémov.

Naopak, systém PaddleOCR vykázal výrazne lepšie výsledky s hodnotami WER = 2.00 a CER = 0.27, čo svedčí o jeho lepšej adaptabilite na typ textu a vizuálnu štruktúru obrázka v tomto datasete. Referenčný text obrázku znel nasledovne:

EARTH
DEMONIC PRESENCE
SEARCHING
VEGA: Attempting to acquire Hell Priest signal...

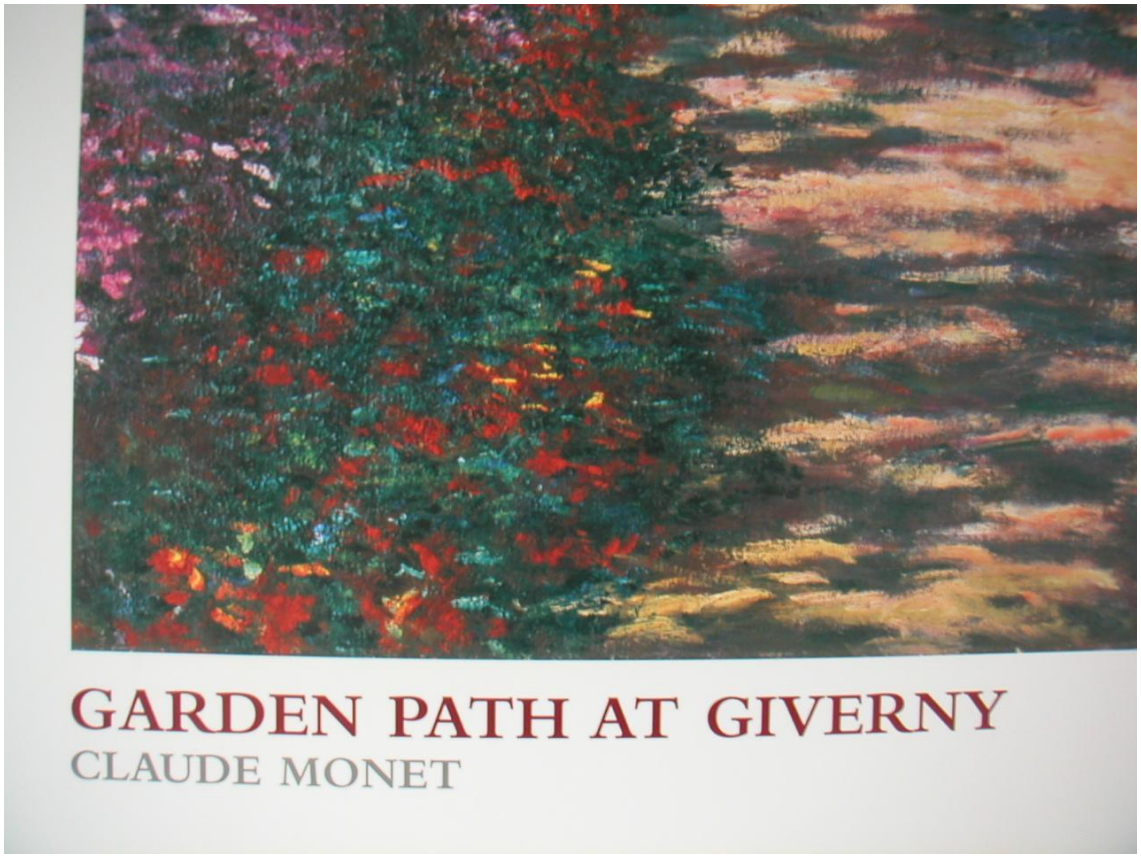
Výsledky systémov sú:
Tesseract - EARTH. Nata in Lyiglager = ane Tyoime dd Bens bo Tio : jhispim Cargye Bhai me Fes, Copan] NG, iml iw DEMS! "PRESENCE SEARCHING

EasyOCR - EARTH J4b Ilti DEMONIC PRESENCE $ 341-641 inslyu d4p* Jush n "45 Jm SEARCHING Kioime $ Ri:' VEGA: Attempting tg acquire Hell Priest signal iLutjm Dulite jwsy Yull, Ligvj

OCRmyPDF - EARTH. Nathan in insu ayer = ak Tyoim ed Bens bo TaoPani Syste Fes, ipa] Naim inSEARCHING

PaddleOCR - EARTH ain DEMONIC PRESENCE jliyim SEARCHING nyy lly VEGA: Attempting to acquire HellPriest signal

Tieto výstupy ukazujú vysokú mieru skreslenia rozpoznaného textu, predovšetkým v dôsledku kombinácie náročného fontu, vrstveného pozadia a vymysleného jazyka typického pre herné svety. Po bližšom skúmaní zlyhania systémov sme prišli k výsledku, že sa systémy snažili rozpoznať text aj na miestach kde sa nenachádza. PaddleOCR systém ako jediný systém dokázal v tomto prípade zachovať zrozumiteľnosť výstupu v rámci akceptovateľnej chybovosti.



Obrázok 17 Testovacia vzorka – ICDAR2003

Pri spracovaní vyššie zobrazeného obrázku z datasetu ICDAR2003 došlo k výraznému zhoršeniu výsledkov OCR systémov Tesseract a OCRmyPDF. V prípade systému Tesseract dosiahla metrika CER hodnotu 4.37 a WER až 25.5. Systém OCRmyPDF zaznamenal podobne nepriaznivé výsledky, a to CER = 3.86 a WER = 22.5. Referenčný text pre tento obrázok je nasledovný:

GARDEN PATH AT GIVERNY CLAUDE MONE

Výsledky systému **Tesseract**:
> ‘ Ae eae w ” ae, - ™* — - . ~~ aye ed ~y ps xs ——— Re AS Ne ae ee Oe Oo A al is “

ai ~ tr Pi , “a - <4 4 ¢ —— unit Oe – eae ae Dm 4 “ rs - + % a ae a5 GARDEN PATH AT GIVERNY CLAUDE MONET

Výsledky systému **OCRmyPDF**:

. =—¥e age dea – be Hn artim, we ei —. Ma 43 = aa * i* id Sab . the 5 ad “ bien ye ° fing nite a“iy Sy, =| P40) we Pe . <P 5 GARDEN PATH AT CLAUDE MONET

Výstupy zvyšných systémov odpovedajú anotácií. Ako je z výsledkov zrejmé, systémom Tesseract a OCRmyPDF sa síce podarilo rozpoznať koniec textu, no úvodné časti boli značne zdeformované. Tieto chyby mohli vzniknúť v dôsledku obrázku na obale knihy.

Obrázok 18 Testovacia vzorka – Multi-oriented

V prípade jedného z obrázkov z datasetu Multi-Oriented došlo k úplnému zlyhaniu väčšiny testovaných OCR riešení. Takmer všetky dosiahli najhoršie možné skóre v oboch hodnotených metrikách – CER 1.00 a WER 10.00, čo zodpovedá úplnému zlyhaniu rozpoznávania textu. Jedinou výnimkou bol systém EasyOCR, ktorý ako jediný poskytol výsledok. Tento výsledok vyzeral nasledovne:

~Trosai+

Systém EasyOCR tak dosiahol hodnoty 0.9 CER a 9.00 pre WER, ostatné systémy dosiahli hodnoty 1.00 CER a 10.00 WER.

Možno predpokladať, že výrazné zlyhanie systémov bolo spôsobené vysokým sklonom textu a rotáciou textu. Tieto faktory predstavujú výzvu najmä pre OCR systémy, ktoré nemajú robustne implementovanú klasifikáciu uhla natočenia textu, alebo nedisponujú pokročilými krokmi vo fáze pred pripravenia.

7.3 Záver a odporúčania

Na základe vykonaného testovania možno konštatovať, že medzi jednotlivými OCR systémami existujú významné rozdiely v presnosti, robustnosti aj rýchlosti spracovania a to v závislosti od typu a náročnosti vstupu.

Systémy EasyOCR a PaddleOCR sa ukázali ako najstabilnejšie naprieč rôznorodými datasetmi, pričom PaddleOCR vo viacerých prípadoch dosiahol najlepšie výsledky pri zložitejších obrázkoch, ako sú herné scény, alebo texty z reálneho prostredia. Naopak systémy Tesseract a OCRmyPDF, napriek ich výhodám pri štruktúrovaných dokumentoch, nedosahovali až také presné výsledky pri dokumentoch s deformáciami, zakriveným textom, netradičným fontom, alebo iným šumom.

Testovanie taktiež ukázalo, že žiaden z testovaných OCR riešení nie je univerzálne najlepší. Každý z nich má svoje silné a slabé stránky. Niektoré vynikajú v spracovaní jednoduchých skenovaných dokumentov, iné sa lepšie prispôsobujú vizuálne náročnejším, alebo menej štruktúrovaným obrazovým vstupom.

Čo sa týka časovej náročnosti, rozdiely medzi systémami boli výrazné. OCRmyPDF riešenie nebolo úplne možné testovať pri rovnakých podmienkach, ako zvyšné riešenia a to kvôli nutnosti prípravy konverzie vstupov. Po zoradení od najpomalšieho systému k najrýchlejšiemu dostaneme poradie EasyOCR, PaddleOCR, OCRmyPDF a Tesseract. Musíme ale zobrať do úvahy, že systémy EasyOCR a PaddleOCR majú architektúru založenú na robustných neurónových sieťach.

Celkovo možno z testovania vyvodiť záver, že výber vhodného OCR nástroja by mal byť podmienený typom spracovávaných dokumentov a očakávanou mierou presnosti. Rovnako je vhodné pri práci s náročnejšími typmi vstupov zvážiť nasadenie krokov pred spustením systému, ako sú napríklad korekcia orientácie textu, binarizácia, alebo filtrácia pozadia.

Z osobných skúseností by sme odporúčali používať PaddleOCR systém kvôli jeho rýchlosti pri spracovaní zložitých vstupov, ale treba brať do úvahy, že tento systém je podporovaný v čase písania práce len staršou verziou Pythonu 3.10. Ak by na vstup boli vkladané iba štruktúrované dokumenty, odporúčame použiť Tesseract riešenie, aj keď pri niektorých dokumentoch nevyprodukoval výsledok. K tomuto odporúčaniu sme dospeli v dôsledku jeho jednoduchosti nasadenia, veľkej podpory komunity a rýchlosti spracovania vstupov.

ZÁVER

Cieľom práce bolo preskúmať a porovnať open-source OCR riešenia so zameraním na ich využitie pri rozpoznávaní textu v štruktúrovaných dokumentoch a v reálnych scénach.

V rámci práce sme identifikovali vhodné testovacie datasety, ktoré zahŕňali rôzne výzvy pre OCR systémy, napríklad zníženú kvalitu obrazu, zakrivenie textu, nekonzistentné osvetlenie či prítomnosť šumu (napr. datasety FUNSD, GameplayCaptions a SROIE), ako aj komplexnosť reálnych scén (ICDAR 2003, Street View Text Dataset) či vplyv typu písma na presnosť rozpoznávania (IAM Handwritten Forms).

Boli analyzované štyri populárne OCR systémy a to Tesseract, OCRmyPDF, EasyOCR a PaddleOCR. Pri testovaní sme použili rôznorodé datasety, ktoré zahŕňali viacero oblastí možných vstupov. Analýza a výsledné porovnanie prebehlo pomocou štandardných metrík CER a WER a ich výsledkov.

Na základe týchto výsledkov bolo možné identifikovať silné a slabé stránky jednotlivých riešení. Tesseract a OCRmyPDF systémy sa preukázali ako spoľahlivé a rýchle riešenia pri spracovaní štruktúrovaných dokumentov, zatiaľ čo PaddleOCR a EasyOCR dosiahli lepšie výsledky pri rozpoznávaní textu v zložitejších scénach vďaka ich robustnej architektúre. Všetky tieto riešenia boli zahrnuté do demonštračného nástroja, ktorý umožňuje používateľom vykonanie extrakcie textu, alebo testovanie vlastných datasetov.

Získané poznatky z tejto práce je možné využiť pri ďalšej analýze OCR riešení. Zároveň ich je možné použiť ako základ pre ďalšiu implementáciu jednotlivých systémov pre špecifickú oblasť.

ZOZNAM POUŽITEJ LITERATÚRY

[1] **PATEL, C., PATEL, A. a PATEL, D.**, 2012. *Optical character recognition by open source OCR tool tesseract: A case study*. International Journal of Computer Applications, roč. 55, č. 10. ISSN 0975-8887.

[2] **KAYA, M. a YILDIZ, A.**, 2015. *A study on the effects of the use of smart boards in mathematics teaching on students' academic achievement and attitudes*. The Turkish Online Journal of Educational Technology, roč. 14, č. 1, s. 1–9. Dostupné z: https://dergipark.org.tr/en/download/article-file/236939 [cit. 2025-04-19].

[3] **FAIZULLAH, S., AYUB, M. S., HUSSAIN, S. a KHAN, M. A.**, 2023. *A Survey of OCR in Arabic Language: Applications, Techniques, and Challenges*. Applied Sciences [online]. Roč. 13, č. 7, s. 4584. Dostupné z: https://doi.org/10.3390/app13074584 [cit. 2025-04-19].

[4] **WEI, H., LIU, C., CHEN, J., WANG, J., KONG, L., XU, Y., GE, Z., ZHAO, L., SUN, J., PENG, Y., HAN, C. a ZHANG, X.**, 2024. *General OCR Theory: Towards OCR-2.0 via a Unified End-to-end Model* [online]. arXiv preprint, cs.CV. Dostupné z: https://arxiv.org/abs/2409.01704 [cit. 2025-04-19].

[5] **RADFORD, A., KIM, J. W., HALLACY, C., RAMESH, A., GOH, G., AGARWAL, S., SASTRY, G., ASKELL, A. et al.**, 2021. *Learning Transferable Visual Models From Natural Language Supervision* [online]. arXiv preprint, cs. Dostupné z: https://arxiv.org/abs/2103.00020 [cit. 2025-04-19].

[6] **NIGAM, S., VERMA, S. a NAGABHUSHAN, P.**, 2023. *Document Analysis and Recognition: A survey* [online]. Dostupné z: https://doi.org/10.36227/techrxiv.22336435.v1 [cit. 2025-04-19].

[7] **PRECIADO, R.**, 2023. *The difference between a structured document and structured content authoring* [online]. Content Rules. Dostupné z: https://contentrules.com/the-difference-between-a-structured-document-and-structured-content-authoring/ [cit. 2025-04-19].

[8] **SINGH, A., BACCHUWAR, K. a BHASIN, A.**, 2012. *A Survey of OCR Applications*. International Journal of Machine Learning and Computing, s. 314–318. ISSN 2010-3700.

[9] **SMITH, R. W.**, 2013. *History of the Tesseract OCR engine: what worked and what didn't*. In: ZANIBBI, R. a COÜASNON, B. (eds.). Proceedings of the 2013

Document Recognition and Retrieval Conference. Burlingame, California, USA, s. 86580. Dostupné z: https://doi.org/10.1117/12.2010051 [cit. 2025-04-19].

[10] Tesseract OCR, Navštívené [2025-02-24], Dostupné na: https://github.com/tesseract-ocr/tesseract

[11] OCRmyPDF documentation,, Navštívené [2025-01-02] Dostupné na: https://ocrmypdf.readthedocs.io/

[12] OCRmyPDF, Navštívené [2025-02-05], Dostupné na: https://github.com/ocrmypdf/OCRmyPDF

[13] **MAHAJAN, Aditya,** 2022. *EasyOCR: A comprehensive guide* [online]. Medium. Dostupné z: https://medium.com/@adityamahajan.work/easyocr-a-comprehensive-guide-5ff1cb850168.

[14] EasyOCR, Navštívené [2025-01-02], Dostupné na: https://github.com/JaidedAI/EasyOCR

[15] PaddleOCR Documentation, Navštívené: [2025-02-20], Dostupné na: https://paddlepaddle.github.io/PaddleOCR/main/en/index.html

[16] Kraken, Navštívené: [2025-04-17], Dostupné na: https://kraken.re/main/index.html

[17] Calamari, Navštívené: [2025-04-17], Dostupné na: https://github.com/Calamari-OCR/calamari

[18] **ALI, Hasmot, RABBY, Akm Shahariar Azad, ISLAM, Md Majedul, MAHAMUD, A.K.M., HASAN, Nazmul a RAHMAN, Fuad**, 2023. *Gold Standard Bangla OCR Dataset: An In-Depth Look at Data Preprocessing and Annotation Processes*. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track. Singapore: Association for Computational Linguistics, s. 460–470. Dostupné z: https://aclanthology.org/2023.emnlp-industry.44/ [cit. 2025-04-19].

[19] ICDAR 2003 Robust Reading Competitions, Navštívené [2025-03-13], Dostupné na: http://www.iapr-tc11.org/mediawiki/index.php?title=ICDAR_2003_Robust_Reading_Competitions

[20] IAM Handwritten Forms Dataset, Navštívené: [2025-04-05], Dostupné na: https://www.kaggle.com/datasets/naderabdalghani/iam-handwritten-forms-dataset?resource=download

[21] SROIE datasetv2, Navštívené: [2025-03-20], Dostupné na: https://www.kaggle.com/datasets/urbikn/sroie-datasetv2

[22] The Street View Text Dataset, Navštívené: [2025-03-20], Dostupné na: https://www.kaggle.com/datasets/nageshsingh/the-street-view-text-dataset

[23] Scanned Images Dataset for OCR and VLM finetuning, Navštívené: [2025-03-23], Dostupné na: https://www.kaggle.com/datasets/suvroo/scanned-images-dataset-for-ocr-and-vlm-finetuning

[24] FUNSD-Form Understanding Noisy Scanned Documents, Navštívené: [2025-03-20], Dostupné na: https://www.kaggle.com/datasets/aravindram11/funsdform-understanding-noisy-scanned-documents

[25] GameplayCaptions, Navštívené: [2025-03-23], Dostupné na: https://huggingface.co/datasets/asgaardlab/GameplayCaptions

[26] Screenshots Dataset, Navštívené: [2025-03-23], Dostupné na: https://www.kaggle.com/datasets/patzold/screenshots-dataset

[27] CAPTCHA Images, Navštívené: [2025-04-14], Dostupné na: https://www.kaggle.com/datasets/fournierp/captcha-version-2-images

[28] **TAMANNA, Tam**, 2023. *Deciphering Accuracy Evaluation Metrics in NLP and OCR: A Comparison of Character Error Rate (CER)* [online]. Medium. Dostupné z: https://medium.com/@tam.tamanna18/deciphering-accuracy-evaluation-metrics-in-nlp-and-ocr-a-comparison-of-character-error-rate-cer-e97e809be0c8 [cit. 2025-04-19].

[29] **HLAING, Thin Thin, PHYO OO, May a ZARLI MYINT, Thaint**, 2019. *Analyzing Word Error Rate on Optical Character Recognition (OCR) for Myanmar Printed Document Image*. International Journal of Computer Trends and Technology [online]. Roč. 67, č. 8, s. 51–57. Dostupné z: https://doi.org/10.14445/22312803/IJCTT-V67I8P109 [cit. 2025-04-19].

[30] **ALKENDI, Wissam, GECHTER, Franck, HEYBERGER, Laurent a GUYEUX, Christophe**, 2024. *Advancements and Challenges in Handwritten Text Recognition: A Comprehensive Survey*. Journal of Imaging [online]. Roč. 10, č. 1, s. 18. Dostupné z: https://doi.org/10.3390/jimaging10010018 [cit. 2025-04-19].

[31] Metrics for Speech Recognition, Navštívené: [2025-03-17], Dostupné na: https://speechbrain.readthedocs.io/en/v1.0.2/tutorials/tasks/asr-metrics.html

[32] OCR_app, Dostupné na: https://github.com/Zachve4037/OCR_App

PRÍLOHY

ZOZNAM PRÍLOH

Príloha A | Obsah pamäťového média

Priložené médium obsahuje:

- Dokumentácia práce v elektronickej podobne (formát PDF)
- Aplikácia
- Datasety použité pri testovaní