SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

ZAVRŠNI RAD br. 1605

SUSTAV ZA OPTIČKO RASPOZNAVANJE TEKSTA STARIJIH KNJIGA NA HRVATSKOME JEZIKU

Dominik Agejev

Zagreb, rujan, 2024.

SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

Zagreb, 4. ožujka 2024.

ZAVRŠNI ZADATAK br. 1605

Pristupnik:	**Dominik Agejev (0036537505)**
Studij:	Elektrotehnika i informacijska tehnologija i Računarstvo
Modul:	Računarstvo
Mentor:	izv. prof. dr. sc. Tomislav Hrkać

Zadatak: **Sustav za optičko raspoznavanje teksta starijih knjiga na hrvatskome jeziku**

Opis zadatka:

Optičko raspoznavanje teksta izazovan je problem računalnog vida koji podrazumijeva pretvorbu skeniranog ili fotografiranog teksta u strojno čitljiv oblik prikazan u nekom od standardnih računalnih tekstovnih formata. Dodatni izazovi javljaju se pri takvoj pretvorbi starijih knjiga, primjerice zbog narušene kvalitete papira, slabije kvalitete tiska, arhaičnog jezika i slično. U okviru ovog završnog rada potrebno je: proučiti najznačajnije slobodno dostupne alate za optičko raspoznavanje teksta i opisati njihove osnovne značajke, ispitati primjenjivost odabranih sustava za raspoznavanje teksta starijih knjiga na hrvatskome jeziku uz primjenu prikladnih tehnika predobrade i naknadne obrade, pripremiti bazu slika za ispitivanje sustava, analizirati ponašanje ostvarenog sustava te prikazati i ocijeniti ostvarene rezultate. Radu priložiti izvorni i izvršni kod razvijenih postupaka, ispitne podatke i rezultate, uz potrebna objašnjenja i dokumentaciju te navesti korištenu literaturu.

Rok za predaju rada: 14. lipnja 2024.

Jakovu, Lovri i Roku, hvala na pomoći!

Sadržaj

1. Uvod

Cilj rada nadići je uspješnost gotovih sustava za optičko raspoznavanje teksta (eng. *Optical Character Recognition* ili *OCR*) na starijim knjigama hrvatskoga jezika koristeći se nenadziranim metodama učenja, odnosno bez označenih podataka za trening modela, uz predobradu i naknadnu obradu.

Iako suvremeni sustavi poput DTrOCR-a [1] postižu gotovo savršene rezultate u raznim primjenama, optičko raspoznavanje teksta nipošto nije riješen problem. Još uvijek i najbolji sustavi, poput gore navedenog, pogrešno prepoznaju više od 10% riječi na fotografijama teksta „u divljini“ i gotovo 20% riječi u rukopisima na kineskom jeziku.

Uz to, zbog ovisnosti o jezičnim modelima, novija rješenja općenito nisu primjenjiva bez dodatne prilagodbe na manje zastupljene jezike, poput hrvatskog, ili se pak oslanjaju na veliku količinu označenih podataka ili na sintetičke podatke, generirane modelima koji za rjeđe jezike nisu dostupni, te zahtijevaju značajne računalne resurse.

Nadalje, zbog suviše uske primjene, specifični problemi, poput predmeta ovog rada, raspoznavanja teksta antikvarnih knjiga i to na jeziku ograničene uporabe, redovito se zaobilaze u prilog doprinosima aktualnim primjenama. [2]

U okviru ovog rada najprije se uvodi u područje, metode i temeljne pojmove koji će se koristiti u radu. Zatim je detaljno izložen zadatak i njegove specifičnosti uz pregled dosadašnjih postignuća unutar područja.

Prelazeći na izvedbu rješenja, razmatraju se najznačajniji slobodno dostupni OCR alati prikladni zadatku, a to su Tesseract [3], OCR sustav opće namjene koji održava Google, te Ocular [4], razvijen specifično za primjenu na antikvarnim dokumentima.

Nakon treniranja i optimiziranja hiperparametara Oculara, uspoređen je s Tesserac-

tom na već predobrađenim ispitnim podacima gdje se pokazuje da usprkos starijoj arhitekturi u bitnome nadjačava Tesseract, ali uz određena ograničenja.

Konačno, izveden je sustav glasanja kojim se postiže veća uspješnost od one samostalnih modela.

2. Uvod u OCR

Optičko raspoznavanje teksta grana je računalnog vida koja se bavi izdvajanjem teksta iz slika, bilo dokumenata, rukopisa ili scenskih fotografija, radi lakog pretraživanja i uređivanja, jednostavnijeg arhiviranja ili pak dostupnosti sadržaja slabovidnima i slijepima.

U odnosu na sadržaj ulaznih slika najčešće govorimo o prepoznavanju teksta tiskanih dokumenata, rukopisa ili teksta „u divljini“, npr. natpisa na pročeljima trgovina, s tim da je potonje uže povezano s drugim granama računalnog vida poput detekcije i klasifikacije objekata.

OCR sustavi često se razvijaju i za još uže definirane zadatke, primjerice prepoznavanje teksta na računima ili antikvarnim dokumentima. Takvi sustavi, kakvima se bavi i ovaj rad, nazivaju se jednonamjenskima (eng. *task-specific*), dok se sustavi prilagođeni raznim uporabama zovu sustavima opće namjene (eng. *general purpose*). [5]

Optičko raspoznavanje teksta podrazumijeva u bitnome pet koraka: predobradu, segmentaciju, izdvajanje značajki, klasifikaciju te naknadnu obradu. [6] U ovom poglavlju izložit će se ugrubo najznačajnije metode i pojmovi koji će se koristiti u ostatku rada.

2.1. Predobrada

Predobrada se odnosi na postupak prilagodbe ulazne slike radi uspješnijeg raspoznavanja znakova. Načela po kojima se ravna predobrada uključuju pojednostavljenje ulaza izostavljanjem nebitnih informacija, što čine binarizacija i eliminacija šuma, te ispravljanje fizičkih nesavršenosti, uzrokovanih bilo tiskom bilo digitalizacijom ulaza, što rade metode poput ispravljanja nagnuća (eng. *skew correction*).

2.1.1. Binarizacija

Cilj binarizacije razlučivanje je između teksta i pozadine. Najjednostavniji način za to postavljanje je praga (eng. *thresholding*) za koji su svi pikseli s RGB ili sivotonskim (eng. *grayscale*) vrijednostima nižim od praga obojani crno, tj. označeni kao tekst, a pikseli iznad praga označeni kao pozadina. [7]

Slika 2.1. Primjer binarizacije [8]

Razlikujemo globalne i lokalne metode binarizacije. Globalne, poput često korištene Otsuove metode [9], postavljaju jedan prag za cijeli dokument, što je vremenski učinkovito i uspješno u idealnom slučaju s jednoličnom pozadinom, no zakazuje pri nejednakom osvjetljenju ili sjeni uslijed loše skeniranog pregiba knjige. Lokalne, poput metode adaptivnog kontrasta [10], temeljenoj na prepoznavanju rubova pomoću kontrasta susjednih piksela, nešto su resursno zahtjevnije, ali zato daju bolje rezultate. [9, 10]

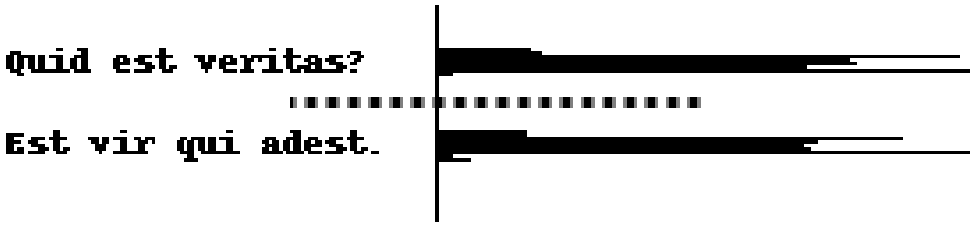
2.1.2. Ispravljanje nagnuća

Cilj ispravljanja nagnuća zaokrenuti je retke teksta tako da su vodoravni. Utvrđivanje kuta nagnuća redaka binarizirane slike može se svesti na pronalazak pravca koji najbolje aproksimira redak, a u tu svrhu najčešće se koristi Houghova transformacija. [11] Konceptualno, Houghova transformacija za svaku rubnu točku slike pronalazi parametre (m, c) pravaca koji se kroz nju mogu provući. Pronađeni skup parametara zapravo je pravac u m, c prostoru, a odredivši pripadajući pravac svakoj točci, ako neki pravci imaju zajedničko sjecište, kroz njima pripadne točke moguće je povući pravac koji u konkretnom slučaju određuje redak teksta.

2.2. Segmentacija

Segmentacija podrazumijeva razlučivanje semantički značajne dijelove slike od kojih su najbitniji retci, riječi i znakovi te koji se obično tim redom i pronalaze: retci u slici, riječi u retku, znakovi u riječi.

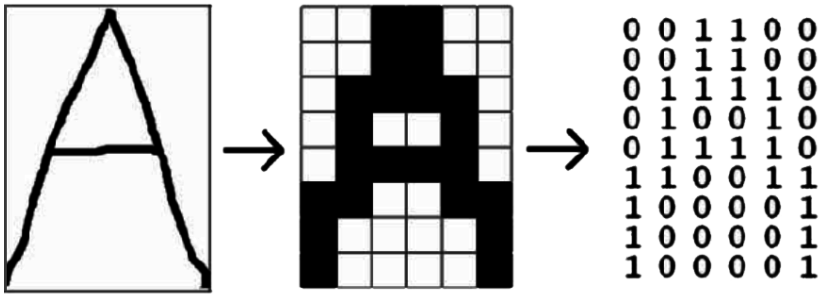
Najjednostavnija metoda segmentacije gradi histogram piksela teksta te postavlja granice između elemenata gdje ima najviše piksela pozadine.



Slika 2.2. Linijska segmentacija histogramom

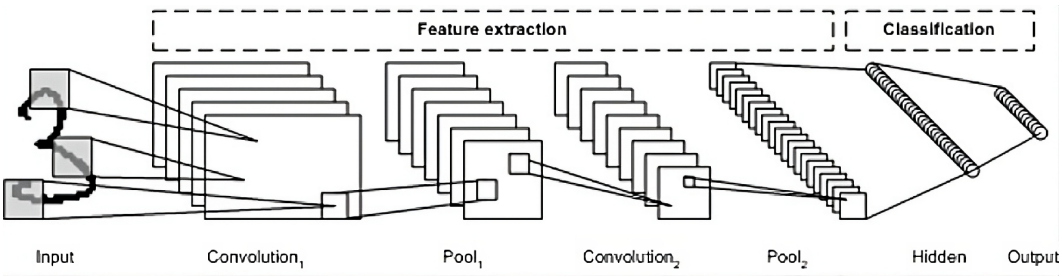
2.3. Izdvajanje značajki

Starije metode OCR-a, prije prelaska na neuronske mreže, koristile su ručno definirane značajke, poput tzv. Granlundovih opisnika temeljenih na Fourierovoj transformaciji [12] ili jednostavnih matrica koje predstavljaju oblik znaka, kao na slici 2.3., dok u suvremenim sustavima poput Tesseracta raniji slojevi neuronske mreže izdvajaju značajke, a kasniji provode klasifikaciju teksta.



Slika 2.3. Tvorba predloška znaka [13]

Za izdvajanje značajki koriste se konvolucijske neuronske mreže (CNN) koje se mogu predočiti kao niz filtera, od kojih prvi provjeravaju npr. je li određeni rub ili ugao prisutan u znaku, a kasniji slojevi zadržavajući najbitnije informacije uče općenitija pravila. [14]



Slika 2.4. LeNet CNN arhitektura [14]

2.4. Klasifikacija

Klasifikacija je najbitniji dio procesa i može se zvati prepoznavanjem teksta u užem smislu. U početku su se OCR sustavi za klasifikaciju oslanjali na uspoređivanje znakova s unaprijed pripremljenim predlošcima (eng. *template matching*), međutim, glavna manjkavost ovog pristupa osjetljivost je na promjenu fonta te šumove poput prelijevanja tinte prilikom tiska, osim očite potrebe za dodatnim ručnim radom.

Razvojem i širenjem neuronskih mreža one preuzimaju ujedno i izdvajanje značajki i njihovu klasifikaciju. Budući da neuronske mreže mogu, bez potrebe za ljudskim radom, u svojim težinama spremiti značajke na više razina apstrakcije, i, što je još bitnije, više samih značajki, nego li ljudi mogu ručno pretočiti u stroju razumljiv zapis, brzo su iskorijenile starije klasifikacijske arhitekture. [15]

Za klasifikaciju znakova koriste se povratne neuronske mreže (RNN) jer omogućuju pamćenje prijašnjih ulaza te stoga mogu modelirati nizove znakova, a ne samo jedan znak poput konvolucijskih mreža.

LSTM mreže [16] (eng. *long short-term memory*) vrsta su povratne mreže koje imaju dulju kratkoročnu memoriju od običnog RNN-a koji pamti samo prethodno stanje.

Transformerska arhitektura [17] isto adresira temporalne podatke, no omogućuje paralelnu obradu ulaza što ujedno ubrzava klasifikaciju uporabom grafičkih kartica i proširuje prostor pamćenja.

2.5. Naknadna obrada

Naknadna obrada podrazumijeva metode koje djeluju na klasificirani tekst. Najčešće korištena metoda naknadne obrade jest ispravljanje riječi pomoću jezičnog modela jer značajno povećava preciznost.

Također se mogu uspoređivati rezultati različitih klasifikatora ili čak više iteracija istog klasifikatora, odlučujući na temelju glasanja koja je riječ vjerojatnija ili pak uspoređujući stupnjeve pouzdanosti transkripcije ako ih klasifikatori podržavaju. [18]

Osim toga, ako se prilikom faze segmentacije nisu razlučivali retci nego izravno zna-

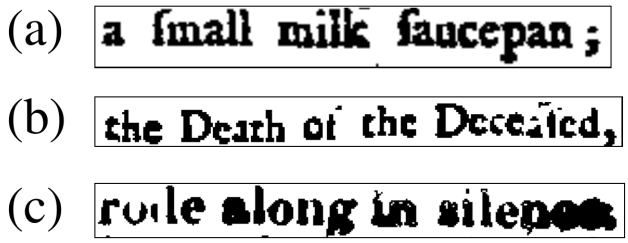
kovi, može se naknadno utvrditi hijerarhijski odnos teksta na slici u ovoj fazi.

3. OCR starijih tekstova

Iz specifičnosti zadatka, tj. starosti knjiga, proizlaze određene poteškoće, naime:

- Čest višak ili manjak tinte pri tisku pojedinih znakova
- Zastarjeli fontovi s neuobičajenim znakovima
- Otežana predobrada zbog spremanja na mikrofilmu
- Arhaičan jezik
- Neravan tisak
- Istrošenost i oštećenja papira

Neke od osobitosti starijeg tiska vidljive su na slici 3.1. Takvim i drugim poteškoćama pristupili su istraživači na razne načine, o čemu ćemo više u idućem odjeljku.



Slika 3.1. Isječci povijesnih dokumenta s (a) nepoznatim fontom, (b) neravnim tiskom, te (c) viškom tinte. [4]

3.1. Pregled literature

Budući da je područje rada dosta usko, ali i zahvaljujući činjenici da napredak općenitijih modernih rješenja, bilo arhitektura OCR sustava bilo obrade prirodnog jezika, u novije vrijeme zasjenjuje specijalizirane sustave, nema mnogo radova na temu transkripcije starih dokumenata.

Prvi značajan i još uvijek relevantan doprinos ujedno je i obuhvaćen ovim radom, a to

su radovi u kojima je prezentiran [4] i unaprijeđen Ocular [19, 20, 21], koji je razjašnjen u poglavlju 5. te se stoga ne će o njemu duljiti ovdje.

Springmann i sur. u svom radu iz 2014. [22] predstavljaju model ispravljanja transkripcije starih latinskih tekstova temeljen na velikom uređenom leksikonu te prilagođavaju Tesseract povijesnom latinskom znakovlju pomoću sintetičkog skupa podataka.

U novijem radu iz 2017. [23], Springmann i sur. čine sličnu stvar. Ovaj put prilagođava se OCRopus ručno označenim slikama, tj. uz *"ground-truth"* tekst. Specifičnost rješenja je u prilagođenosti za dijakronijske tekstove, odnosno uzimanja u obzir promjene jezika kroz dulje vremensko razdoblje.

Christy i sur. [24] pak detaljno opisuju proces masovne digitalizaciju tiskanih tekstova koristeći širok raspon alata te opširno razlažući korake predobrade, naknadne obrade, itd.

Iako je rad Wicka i sur. [25] usmjeren na starije tiskane knjige, bit rada je općenitija: uporaba dublje LSTM mreže uz više označenih podataka te sustav glasanja kako bi se postigao CER ispod čak 0.5%.

Usprkos tome što se ne tiče izričito starih dokumenata te zahtijeva nadzirano učenje, rad *OCR Post Correction for Endangered Language Texts* [26] osobito je obećavajuć. Naime, razvijen je model za ispravljanje transkripcije OCR sustava koji zahtijeva vrlo malo podataka za nadzirano učenje i uz to podržava učenje iz dvojezičnih ili prevedenih dokumenata.

Za razliku od Oculara koji inkorporira jezični model u sam proces transkripcije, najveća prednost ovog rješenja jest njegova modularnost, tj. mogućnost korištenja na tekstu neovisno o OCR modelu.

4. Tesseract

Tesseract [3] je najpoznatiji i najprecizniji slobodno dostupan OCR sustav opće namjene koji podržava i hrvatski jezik, a od 4. inačice temeljen je na LSTM neuronskim mrežama. U ovom poglavlju objasnit će se ugrubo Tesseractov proces prepoznavanja teksta prema zadanim postavkama.

4.1. Predobrada

Tesseract ima ugrađena tri koraka predobrade: binarizaciju, eliminaciju šuma i analizu uređenja stranica.

4.1.1. Binarizacija

Tesseract se koristi Otsuovom metodom, no ne za postavljanje jednog globalnog praga za čitavu stranicu, već rabi implementaciju Leptonica biblioteke [27] koja dijeli stranicu u jednake blokove te na njima postavlja prag. Takvim pristupom nadvladavaju se varijacije u svjetlini na makro razini slike, ali se zadržava i veća resursna učinkovitost uslijed paralelizacije i izbjegavanja složenijih računa lokalnih metoda.

Za slike koje nisu više-manje dvobojne nego pate od većih nejednakosti u osvijetljenju Tesseract podržava i Sauvolinu [28] lokalnu metodu binarizacije koja utvrđuje prag za svaki piksel slike na temelju srednje vrijednosti i standardne devijacije okolnih piksela.

4.1.2. Analiza uređenja stranice

Detekcija slika

Funkcijom `FindImages` Tesseract pronalazi slike koje potom zanemaruje prilikom prepoznavanja teksta.

Detekcija crta

Tesseract rabi Leptonicu za pronalaženje i uklanjanje crta, odnosno razdjelnih linija, na ulaznoj slici što pomaže u odvajanju teksta od grafičkih elemenata poput tablica ili obrazaca.

Analiza povezanih komponenti

Ovaj korak izvodi funkcija `find_components` koja skenira binarnu sliku piksel po piksel, označava povezane crne piksele i grupira ih u povezane komponente koje predstavljaju potencijalne znakove ili dijelove znakova.

Detekcija orijentacije i pisma

Ako je ova opcija omogućena, Tesseract će provjeriti o kojem je pismu riječ (latinično, kinesko,...) i u kojem smjeru se piše (kineski se npr. može pisati odozgo prema dolje ili zdesna na lijevo).

Detekcija stupaca

Koristi se ako je tekst pisan u stupcima poput novinskog članka ili znanstvenog rada.

Pronalazak redaka teksta

Tesseract analizira prostorne odnose između povezanih komponenti kako bi detektirao linije teksta. Ovaj korak koristi statistički pristup temeljen na razmacima između komponenti.

4.1.3. Eliminacija šuma

Razlučivši pozadinu od ostatka prelazi se na brisanje šuma poput razlivene tinte. To se postiže pronalaskom spojenih piksela te usporedbom karakteristika nakupine s tipičnim karakteristikama teksta.

Tijekom analize razmatra se: [29]

Širina poteza

Potezi jednolične širine vjerojatnije pripadaju znaku.

Veličina nakupina

Skupine piksela koje se protežu izvan uobičajene visine retka vjerojatno nisu zna-

kovi.

Obujmljena površina

Gledajući površinu koju skupina piksela okružuje možemo procijeniti je li znak ili nije.

Broj nakupina po retku

Brojeći skupine piksela u retku provjerava se omjer malih nakupina naspram skupina veličine znaka.

Odnos među točkama

Ako se detektira velik broj susjednih točaka na istoj visini ne odbacuju se već su označene kao "vodeće točke" sadržaja.

4.2. Izdvajanje značajki i klasifikacija

Tesseractovu arhitekturu, prikazanu na slici 4.1., dobiva se združivanjem konvolucijske mreže za izdvajanje značajki te LSTM povratne mreže za klasifikaciju znakova.

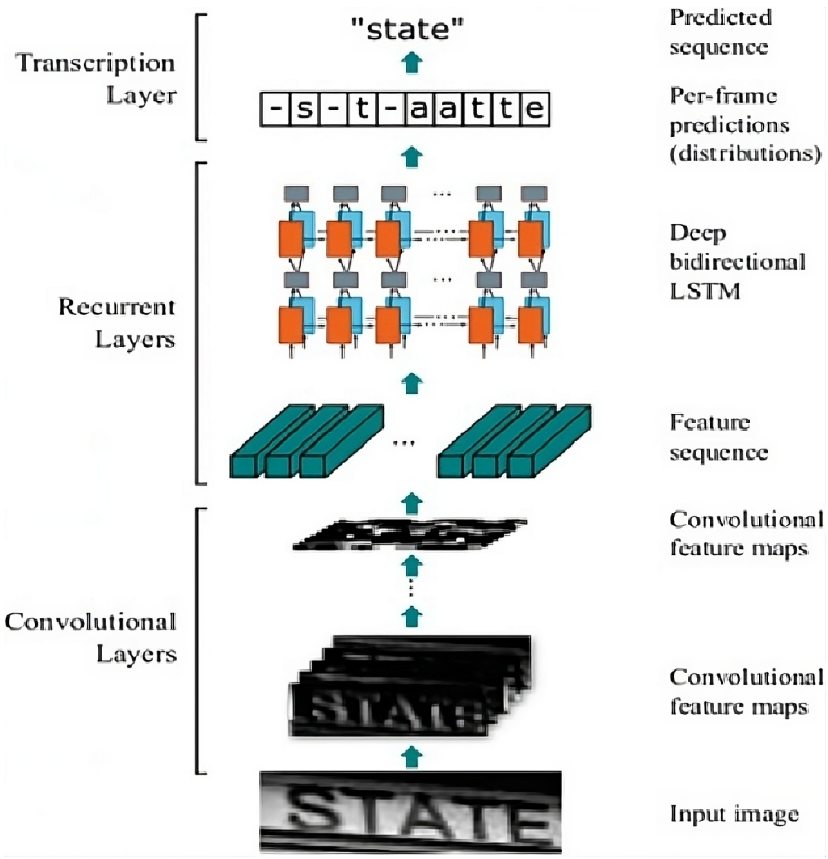
U procesu klasifikacije ključnu ulogu igra jezični model u kojem su spremljeni statistički uzorci jezika koje LSTM mreža koristi nakon inicijalnog raspoznavanja za upućivanje budućih pokušaja.

Tesseract također sprema prilagodljive predloške koje prilagođava varijacijama fonta, kurzivu, masnopisu, podcrtavanju i sl. te razmaku među znakovima.

4.3. Naknadna obrada

Tesseract se ne koristi jezičnim modelom samo u zadnjem koraku, nego zapravo prožima čitav proces klasifikacije. Na temelju rječnika daje se prednost poznatim znakovnim nizovima, a upućen međuovisnostima jezičnog modela ispravlja riječi u skladu s kontekstom.

Konačno, Tesseract prilagođava interpunkcijske znakove i razmake pravilima jezika.



Slika 4.1. CRNN arhitektura [30]

5. Ocular

Ocular [4] je sustav za optičko raspoznavanje teksta razvijen specifično za nenadziranu transkripciju povijesnih dokumenata, i koji je, kada je izdan i svojevremeno unaprijeđen [19], bio vrhunac tehnologije za to područje (eng. *state-of-the-art*).

Njegove glavne značajke su: [31]

- Nenadzirano učenje nepoznatih znakovlja rabeći slike ulaznog dokumenta i korpus teksta na ciljnom jeziku.
- Prilagođenost radu sa šumovitim dokumentima.
- Podrška za višejezične dokumente.
- Nenadzirano učenje ortografskih varijacija uslijed arhaičnog pravopisa.
- Istovremen ispis doslovnog teksta i normaliziranog oblika (prilagođenog standardnom jezku).

U 2. poglavlju spomenuta je klasifikacijska metoda uspoređivanja predložaka čiji je glavni nedostatak neprilagodljivost na različita znakovlja.

Ocular nadilazi tu poteškoću gradeći model znakovlja dinamički, tj. po potrebi za svaki dokument, te ne uspoređuje ulazni znak izravno s predloškom, već na temelju naučenog modela, uzimajući u obzir kontekst, nagnuće teksta, količinu tinte i šum, generira znak koji potom uspoređuje s pikselima ulaznog znaka.

To postiže četirima generativnim probabilističkim modelima koji predstavljaju aspekte procesa tiskanja: jezični model, slovoslagarski ili tipografski model, koji uključuje model tinte, te model šuma, koji združeni tvore skriveni polu-Markovljev model (HSMM).

Međuovisnosti ulaza i modela određuje sljedeća formula, [4] gdje E predstavlja tekst,

X piksele ulazne slike, T raspored znakova na slici, a R aspekte otiskivanja tinte:

$$\begin{aligned} P(E,T,R,X) = P(E) & \quad \text{[Jezični model]} \\ \cdot P(T|E) & \quad \text{[Slovoslagarski model]} \\ \cdot P(R) & \quad \text{[Model tinte]} \\ \cdot P(X|E,T,R) & \quad \text{[Model šuma]} \end{aligned}$$

5.1. Jezični model

Jezični model jest Kneser-Neyev uglađeni znakovni n-gramski model [32] koji se uči na korpusu teksta u ciljnom jeziku. Specifičnost ovog modela, koja ga razlikuje od uobičajenih NLP modela, je da nema zaustavni znak, nego tretira cijeli redak kao jednu cjelinu, sadržavala ona jednu rečenicu ili pak više njih.

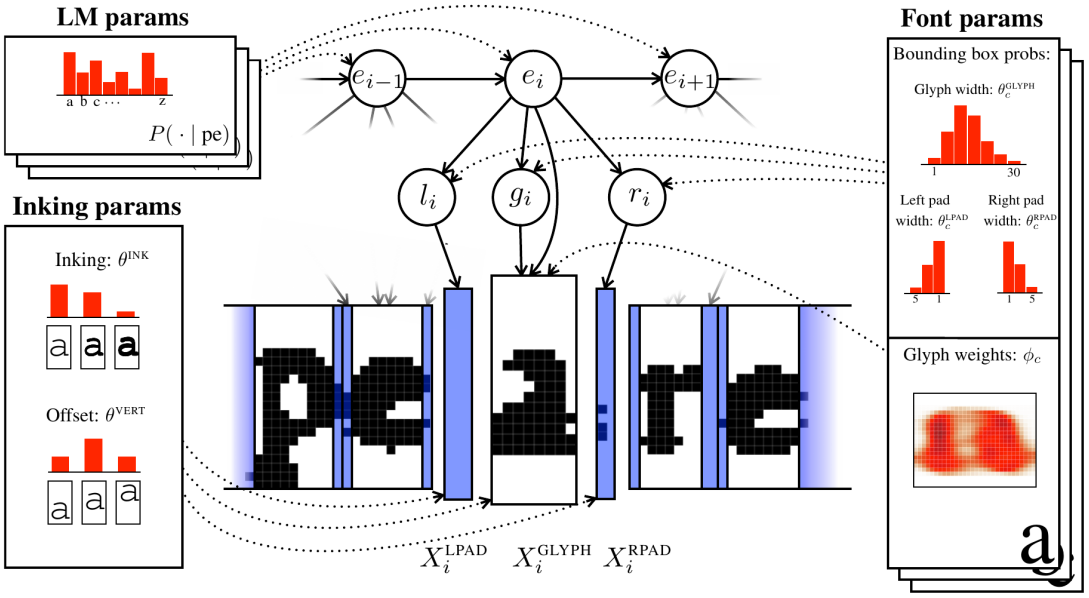
Blisko povezan s jezičnim modelom pojam je veličine snopa (eng. *beam size*) pretraživanja skupa stanja. Naime, prilikom treniranja modela znakovlja, a i same transkripcije, na temelju jezičnog modela generiraju se najvjerojatniji slijedovi znakova, koji se potom uspoređuju s ulazom i od kojih se odabire najsličniji.

Veličina snopa označava jednostavno broj tih nizova koji će se uzeti u obzir. [19] Veći snop će produljiti vrijeme izvršavanja, a i prevelik snop može dovesti do prenaučenosti.

5.2. Slovoslagarski model

Slovoslagarstvo (eng. *typesetting*) u kontekstu mehaničkog tiskanja podrazumijeva postavljanje glifova znakova prikladno razmaknute na traku za tiskanje.

Ocularov generativni model radi vrlo sličnu stvar. 5.1. Prvo generira širinu glifa g_i, zatim lijevi razmak l_i te desni razmak r_i, koji naravno ovise o prepoznatom znaku e_i.



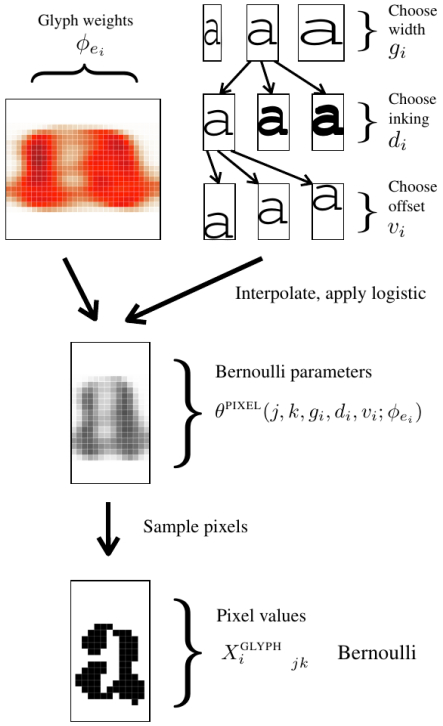
Slika 5.1. Stohastički proces generiranja glifa znaka [4]

5.2.1. Model tinte

Kako bi generirani znak bio sličan ulaznom potrebno je odabrati prikladnu količinu tinte, ali i vertikalni pomak.

To postiže generiranjem količine tinte za raspodjelu po pikselima d_i te pomak od dna v_i.

Ove vrijednosti obično ne ovise o vrsti znaka te su stoga tako modelirane i u stohastičkim modelima.



Slika 5.2. Stohastički proces zacrnjivanja piksela glifa [4]

5.3. Model šuma

Model šuma simulira degradaciju slike koja nastaje tijekom procesa tiskanja dodavanjem šuma u obliku neovisnog bijeljenja piksela prema Gaussovoj distribuciji.

Na kraju zadržava najcrnje piksele kao što je učinjeno na ulazu tijekom binarizacije.

6. Metodologija

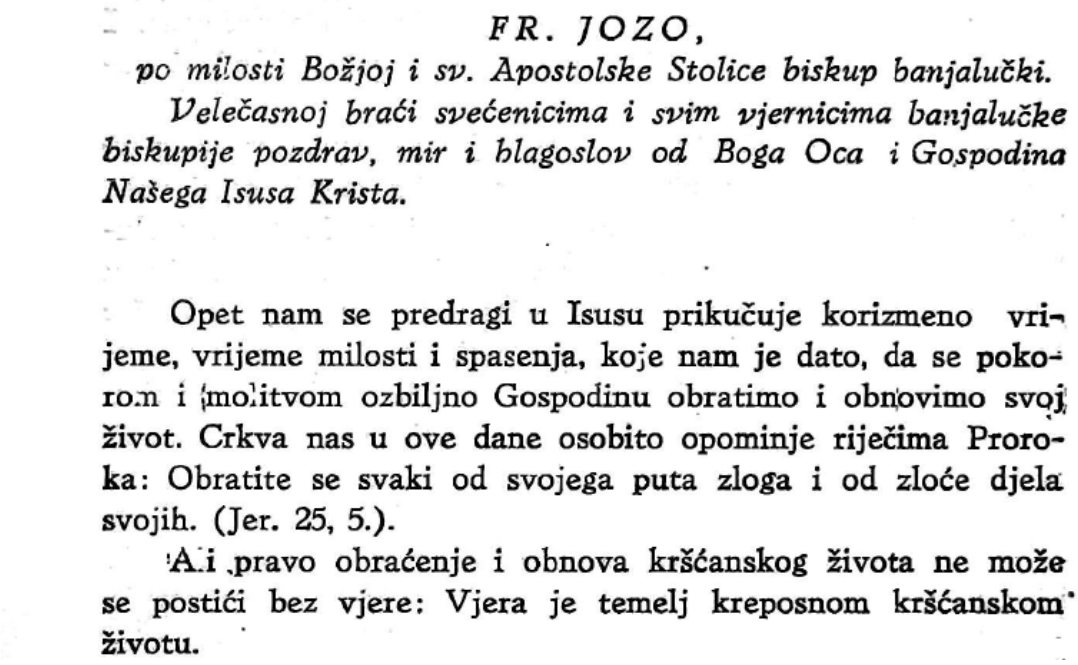
Ključan dio razvoja boljeg rješenja evaluacija je preciznosti Tesseracta i Oculara. U tu svrhu potreban je ispitni skup podataka prilagođen ograničenjima sustava.

Zatim treba prikupiti podatke za treniranje Oculara i namjestiti njegove hiperparametre da daju zadovoljavajuće rezultate, što će zapravo biti najznačajniji dio rada.

6.1. Ispitni skup podataka

Budući da Tesseract ima ugrađenu predobradu slika, radi pravednije usporedbe same klasifikacije teksta izabrani su već obrađeni dokumenti:

- Fra Jozo Garić, biskup – Korizmena okružnica (1932.) 6.1.
  - 3 stranice za trening znakovlja
  - 12 stranica s 4,623 riječi za transkripciju



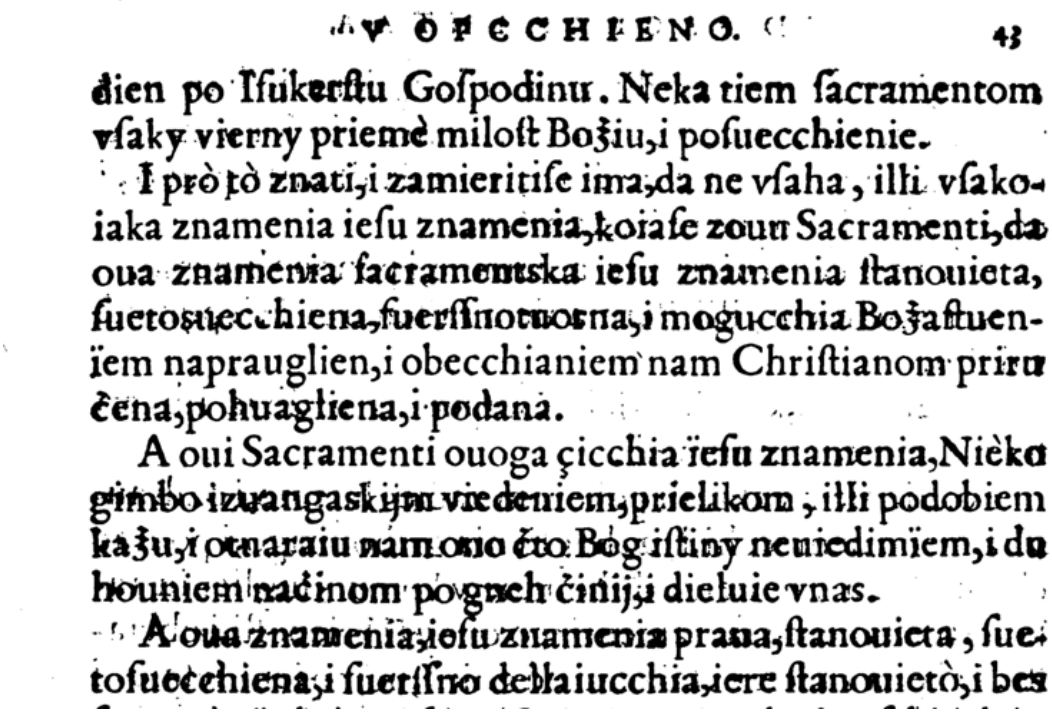
FR. JOZO,
po milosti Božjoj i sv. Apostolske Stolice biskup banjalučki.
Velečasnoj braći svećenicima i svim vjernicima banjalučke biskupije pozdrav, mir i blagoslov od Boga Oca i Gospodina Našega Isusa Krista.

Opet nam se predragi u Isusu prikučuje korizmeno vrijeme, vrijeme milosti i spasenja, koje nam je dato, da se pokorom i molitvom ozbiljno Gospodinu obratimo i obnovimo svoj život. Crkva nas u ove dane osobito opominje riječima Proroka: Obratite se svaki od svojega puta zloga i od zloće djela svojih. (Jer. 25, 5.).

A i pravo obraćenje i obnova kršćanskog života ne može se postići bez vjere: Vjera je temelj kreposnom kršćanskom životu.

Slika 6.1. Izvadak iz *Korizmene okružnice*

- Sv. Petar Kanizije – Summa nauka christianskoga (1583.) 6.2.
  - 6 stranica za trening modela znakovlja
  - 1 stranica s 241 riječi za transkripciju



V OBECCHIENO. 43

dien po Iſukerſtu Goſpodinu. Neka tiem ſacramentom vſaky vierny priemè miloſt Bożiu, i poſuecchienie.

I prò tò znati, i zamieritiſe ima, da ne vſaha, illi vſakoiaka znamenia ieſu znamenia, koiaſe zoun Sacramenti, da oua znamenia ſacramentska ieſu znamenia ſtanouieta, ſuetoſuecchiena, fuerſſnotuorna, i mogucchia Bożaſtuenïem naprauglien, i obecchianiem nam Chriſtianom prirečena, pohuagliena, i podana.

A oui Sacramenti ouoga çicchia ïeſu znamenia, Nièko gimbo izuangaskim viedeniem, prielikom, illi podobiem każu, i otnaraiu nam ono čto Bóg iſtiny neuiedimïem, i dohouniem načinom pogneh činij, i dieluie vnas.

A oua znamenia, ieſu znamenia prama, ſtanouieta, ſuetoſuecchiena, i fuerſſno debla iucchia, iere ſtanouietò, i bes

Slika 6.2. Izvadak iz *Summe nauka christianskoga*

Korizmena okružnica korištena je za sva ispitivanja osim za isprobavanje ortografskih mogućnosti gdje je izabrana *Summa* kao primjer zahtjevnog ulaza za sustav.

Oba sustava imaju određena ograničenja na ulaze:

Ocular radi jedino s PDF dokumentima zastarjele verzije 1.4 te ih je stoga bilo potrebno pretvoriti u taj format. Za to je korišten Ghostscript, [33] slobodno dostupan alat otvorenog koda.

Tesseract pak radi jedino na slikama te je zato bilo potrebno ekstrahirati ih iz PDF-a prije prepoznavanje teksta. Ovdje je zgodno napomenuti da treba paziti da se prilikom pretvorbe ne smanji DPI rezolucija jer to ima poguban utjecaj na preciznost.

6.2. Predobrada i naknadna obrada

Radi veće bliskosti realističnoj primjeni učinjene su jednostavne obrade teksta.

U Tesseractovom ispisu nizovi dva zareza ,, zamijenjeni su ravnim navodnicima " te

je za donje navodnike u ispitnom skupu podataka učinjeno isto.

Budući da su Ocular duge crte često bunile u skupu podataka za jezični model zamijenjene su crticama –. To ne predstavlja problem jer se može jednostavno naknadno utvrditi treba li biti gdje duga ili kratka crta.

6.3. Mjere uspješnosti

Za uspoređivanje znakovnih nizova najčešće korištena mjera uspješnosti je **Levenshteinova udaljenost** koja bilježi broj potrebnih zamjena, brisanja ili umetanja znakova da bi se iz jednog niza dobio drugi. [34]

Budući da ta mjera ovisi o duljini teksta, dijeljenjem Levenshteinove udaljenosti ukupnim brojem znakova dobivamo **stopu pogreške za znakove** (eng. *Character Error Rate*). Obično se dobrom vrijednošću smatra 1-2%.

Stopa pogreške za riječi (eng. *Word Error Rate*) dobiva se uzimanjem riječi za najmanju jedinicu zamjene pri računanju Levenshteinove udaljenosti, tj. ako su jedan ili više znakova u riječi pogrešni čitava riječ broji se kao pogrešna, te dijeljenjem te udaljenosti s ukupnim brojem riječi.

7. Optimizacija Oculara

Za razliku od Tesseracta, Ocular nije univerzalno primjenjiv za različite jezike i znakovlja već je potrebno naučiti jezični model, za koji je potreban korpus teksta na ciljnom jeziku, te model fonta, koji se gradi na temelju slika čiji će tekst kasnije prepoznavati. U ovom poglavlju, koje predstavlja i glavni dio rada, provest će se optimizacija Ocularovih hiperparametara.

7.1. Jezični model

Pri izgradnji skupa podataka za trening jezičnog modela najrelevantnije su dvije stavke: broj podataka i tematika teksta.

U izvornom radu pokazano je kako model malo precizniji (4 WER postotna boda) na dokumentima čija je tematika pokrivena u jezičnom modelu. Budući da je sustav namijenjen starijim knjigama, od kojih je dobar dio vjerske tematike, uključeno je Sveto Pismo i druge duhovne knjige pored novijeg i starijeg štiva koje doprinosi većoj raznolikosti izričaja i opsežnijem rječniku. [4]

U daljnjim eksperimentima korišten je jezični model treniran na tekstovima javno dostupnih knjiga, poput djela Augusta Šenoe, Marije Jurić-Zagorke, Charlesa Dickensa i sl. (7.5 milijuna riječi) uz Šarićev prijevod Svetog Pisma (670k riječi) i još 11 knjiga vjerske tematike (500k riječi).

Ukupan broj riječi odabranog skupa podataka od 8.7 milijuna usporediv je sa skupom podataka korištenim u izvornom radu koji ih ima 10 milijuna.

7.1.1. Veličina skupa podataka

Ispitani su i podskupi odabranog ali i veći skupovi podataka, temeljeni na prethodno navedenom uz dva tipa proširenja: stranom i domaćom beletristikom (11.6 milijuna riječi) te tekstovima vjerske tematike (6 milijuna riječi).

Nažalost, dodatak beletristike dovodio je do neizbježnog neuspjeha treninga te je stoga ispitano samo proširenje vjerskim štivom.

Tekstovi vjerskih knjiga nisu bili lektorirani nego nesavršeni proizvodi prepoznavanja teksta, a uključeni su svejedno kako bi se ispitalo doprinosi li kvantiteta potencijalno više od kvalitete ako greške nisu značajne.

Kako je vidljivo iz tablice 7.1., povećanje skupa podataka ipak je dovelo do smanjenja preciznosti. Ručnom provjerom lako se uviđa kako su greške iz jezičnog modela utjecale na ispis.

Primjerice, prvotni jezični model ispravno prepoznaje riječ `katekizam`, dok ju prošireni zamijenjuje nizom `katalo sam`, s tim da se niz `katalo` pojavljuje samo jednom u proširenom skupu podataka.

Očigledno, Ocularov jezični model nije robustan na greške u skupu podataka čak i kad se pojavljuju samo jednom, što je razumljivo, jer je poželjno da svaka viđena riječ postane dio vokabulara.

Ipak, pomalo je iznenađujuće da je `katalo` nadjačalo riječ koja se pojavljuje 80 puta u skupu podataka. Međutim, čini se kako je do toga došlo zbog kombinacije s nizom `sam` koji se pojavljuje preko 57 tisuća puta.

Jezični model	**CER**	**WER**
Izvorni	1.05	2.74
Izvorni + OCR vjerskih knjiga	1.59	3.64

Tablica 7.1. Uspješnost povećanja jezičnog modela

7.1.2. Veličina snopa

S obzirom na to da veći jezični model podrazumijeva više mogućih kombinacija riječi, povećanje snopa ispitivanih riječi u skrivenom Markovljevom modelu postaje potencijalno presudno kako bi točna riječ bila pronađena.

Ipak, kako je vidljivo iz donje tablice 7.2., povećanje snopa nije dovelo do poboljšanja niti pri treningu modela fonta niti prilikom transkripcije.

Rezultati impliciraju kako je bolje ne ispraviti prepoznati znakovni niz ako predloženi ispravak nije među prvih pedeset predloženih. To može biti do grešaka u skupu podataka za jezični model, gdje veći snop obuhvati i rijetke pogrešne, ali slične nizove

Veličina snopa		CER	WER
Trening	Transkripcija		
10	10	2.03	3.58
40	40	1.59	3.64
50	50	**1.05**	**2.74**
40	120	1.81	3.77
120	50	2.21	5.04
120	120	2.29	5.78

Tablica 7.2. Usporedba uspješnosti prema veličini snopa

ispravljanome.

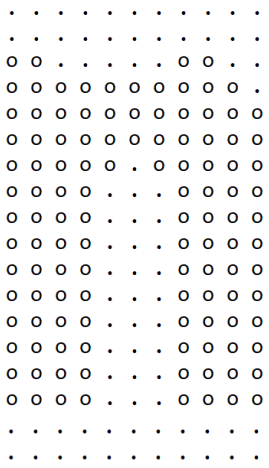
Općenitije gledano, povećanje snopa pretraživanja prostora stanja podržava "pamćenje" više informacija i stoga dovodi do prenaučenosti, kao i prekomjerno povećanje neuronske mreže. Drugim riječima, veća veličina snopa podložnija je utjecaju šuma.

Tomu u prilog ide, primjerice, zamjena rijetkog niza `vjere:`, koji se pojavljuje 4 puta u skupu podataka, riječju `vjetar`, koja se pojavljuje 897 puta.

7.2. Model znakovlja

Budući da se Ocularova klasifikacija zasniva na generiranju znakova na temelju modela znaka opisanog matricom vjerojatnosti, 7.1. ključno je te vjerojatnosti pomno odrediti.

Rezultati ispitivanja optimalnog broja faza treninga modela fonta vidljivi su u tablici 7.3. Suprotno očekivanjima, ponovno treniranje na prve dvije stranice nije povećalo preciznost.



Slika 7.1. Matrica vjerojatnosti za slovo `n`. Svaki kružić predstavlja vjerojatnost da je piksel zacrnjen.

Gledajući pobliže, uspješnost opada na jednoj od dvije stranice koje su prošle dodatne dvije iteracije treninga, a na drugoj se povećava. 7.4. Radi se o razlici od nekoliko znakova te se stoga ne može zaključiti poboljšava li se barem uspješnost na dotreniranim stranicama kao kod školskog primjera prenaučenosti.

Iteracije treninga	Veličina snopa		CER	WER
	Trening	Transkripcija		
3x3 stranice	10	10	2.03	3.58
	50	50	**1.05**	**2.74**
3x3 str. + 2x2 str.	50	50	1.19	2.9

Tablica 7.3. Uspješnost modela prema broju iteracija treninga

Ispitivana stranica	Model znakovlja	CER	WER
str. 5.	3x3	**0.73**	**1.49**
	3x3+2x2	1.20	3.08
str. 6.	3x3	1.17	3.16
	3x3+2x2	**0.95**	**1.46**

Tablica 7.4. Uspješnost modela prema broju iteracija treninga na stranicama koje su bile u dodatnim iteracijama

7.3. Ispitivanje ortografskih mogućnosti

Kako bi uspješno prepoznavali arhaične ortografske varijacije, u izvornom radu [21] pretvorbom suvremenih tekstova, na temelju pravila koja su definirali jezikoslovci, konstruiran je skup podataka s arhaičnim ortografskim varijacijama.

U ovom odjeljku ukratko se demonstrira kako je za uspješnu transkripciju vrlo starih dokumenata arhaičnog jezika potreban značajniji trud od jednostavnog treniranja modela na korpusu suvremenog teksta.

7.3.1. Jezični model

Moglo bi se očekivati da će veći rječnik više doprinijeti uspjehu na tekstovima s dotad manjom preciznošću prepoznavanja, kao što je *Summa nauka christianskoga*, međutim, povećanje skupa podataka nije dovelo do značajnog poboljšanja u preciznosti, ali je značajno usporilo trening i transkripciju.

Ispitivani dokument	Jezični model	CER	WER
Summa nauka christianskoga	Izvorni	17.67	66.79
	Izvorni + OCR vjerskih knjiga	16.52	64.64

Tablica 7.5. Usporedba uspješnosti jezičnih modela

Minimalno poboljšanje vidljivo u tablici 7.1. za *Summu* u granicama je slučajnosti,

osobito uzevši u obzir da je za tu knjigu ispitivana samo jedna stranica (iako je trening modela fonta bio na 6) naspram 12 za *Korizmenu okružnicu*.

8. Sinteza rješenja

Budući da je izuzev ispuštanja određenih redaka Ocular točniji od Tesseracta ovdje se predlaže jednostavan sustav glasanja kojim je Ocularova manjkavost otklonjena bez gubitka preciznosti.

Algoritam glasanja čita redak po redak Tesseractov ispis i traži odgovarajući redak Ocularovog ispisa na temelju sličnosti izračunate pomoću Levenshteinove udaljenosti. Ako pronađe dovoljno sličan redak odabire ga kao izlaz, inače preferira Tesseractov ispis.

```
      for t_line in tesseract_output:
        for c_line in ocular_output:
          distance = Levenshtein.distance(t_line, c_line)
          if distance < threshold * len(t_line):
            output.append(c_line)
          break
        output.append(t_line)
```

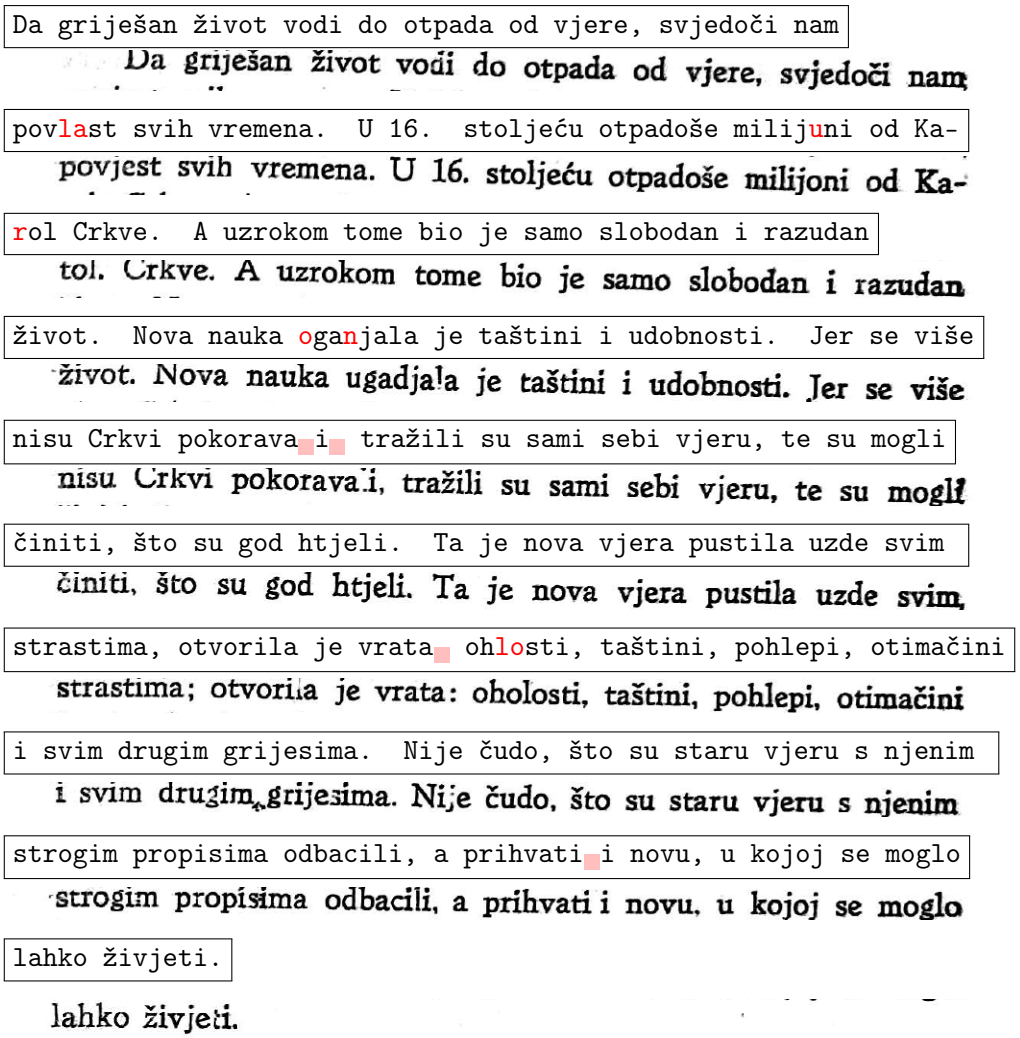
Slika 8.1. Algoritam glasanja predstavljen Python kodom

Suradnjom dvaju modela dobiva se 63.16% bolji CER i 65.78% bolji WER od samog Tesseracta kao što je vidljivo u tablici 8.1.

OCR sustav	Pojedinosti	CER	WER
Ocular		8.83	10.94
	Zanemareni retci s CER>20	1.05	2.74
Tesseract		1.52	3.77
Predložen sustav		**0.96**	**2.48**

Tablica 8.1. Uspješnosti sustava

Na slici 8.2. vidljiv je primjer transkripcije za najzahtjevniju stranicu ispitnog skupa.



Slika 8.2. Transkricpija isječka najlošije stranice.

9. Zaključak

Usprkos općenitom napretku OCR sustava, suvremena rješenja imaju određene manjkavosti praktične naravi. Poglavito to je potreba za skupovima označenih podataka, a često i velikim jezičnim korpusom kako bi se prilagodila novom jeziku.

Tesseract pak predstavlja jednostavno, gotovo rješenje no uz nezadovoljavajuća preciznost. Ocular, s druge strane, izbjegavajući zahtjeve nadziranog učenja te nadilazeći Tesseractovu preciznost na starim dokumentima, ipak pati od svojih problema.

Prvenstveno to je resursno zahtjevan i dugotrajan proces učenja i transkripcije (min. po str.), i to usprkos OpenCL paralelizaciji, koja je korištena umjesto bolje CUDA paralelizacije, jer se potonju, notornu za lošu kompatibilnost s Java-om, nije uspjelo omogućiti.

Osim toga, često ispušta kraće retke, što je za cjelovito rješenje neprihvatljivo te je stoga Ocularov ispis ispravljen Tesseractovim.

Dodatno poboljšanje preciznosti predloženog sustava moglo bi se postići složenijim sustavom glasanja temeljenom na detaljnijoj analizi grešaka pojedinih modela.

Primjerice, uzevši u obzir kako Ocularov jezični model djeluje nad retcima umjesto nad rečenicama, očekivalo bi se da čini više grešaka na prvoj riječi retka nego Tesseract, osobito uslijed hifenacije.

S druge strane, čini se kako Tesseract češće prepoznaje interpunkciju kada treba i kada ne treba uslijed većeg oslanjanja na predobradu naspram jezičnog modela te bi u tom slučaju preferiranje Ocularove interpunkcije poboljšalo uspješnost.

Na žalost za Ocular, ali na sreću optičkog raspoznavanja teksta, modernije arhitekture pružaju sve bolje rezultate, a i razvojem velikih jezičnih modela dolazi do sve veće konvergencija OCR-a i obrade prirodnog jezika.

Imajući taj trend u vidu, budućnost transkripcije povijesnih dokumenata nazire se u naknadnoj obradi teksta, koja bi pružala svevremensko rješenje neovisno o klasifikacijskoj arhitekturi.

Literatura

[1] M. Fujitake, "Dtrocr: Decoder-only transformer for optical character recognition", 8 2023. [Mrežno]. Adresa: http://arxiv.org/abs/2308.15996

[2] K. Olejniczak i M. Šulc, "Text detection forgot about document ocr", 10 2022.

[3] R. Smith, "An overview of the tesseract ocr engine", u *Ninth International Conference on Document Analysis and Recognition (ICDAR 2007) Vol 2*. IEEE, 9 2007., str. 629–633. https://doi.org/10.1109/ICDAR.2007.4376991

[4] T. Berg-Kirkpatrick, G. Durrett, i D. Klein, "Unsupervised transcription of historical documents", str. 207–217, 2013. [Mrežno]. Adresa: https://aclanthology.org/P13-1021/

[5] E. Borovikov, "A survey of modern optical character recognition techniques", 2014.

[6] P. Dhande i R. Kharat, "Recognition of cursive english handwritten characters", *Proceedings - International Conference on Trends in Electronics and Informatics, ICEI 2017*, sv. 2018-January, str. 199–203, 7 2017. https://doi.org/10.1109/ICOEI.2017.8300915

[7] Jyotsna, S. Chauhan, E. Sharma, i A. Doegar, "Binarization techniques for degraded document images-a review", *2016 5th International Conference on Reliability, Infocom Technologies and Optimization, ICRITO 2016: Trends and Future Directions*, str. 163–166, 12 2016. https://doi.org/10.1109/ICRITO.2016.7784945

[8] "Let's take things step by step, shall we? | how ocr works". [Mrežno]. Adresa: https://how-ocr-works.com/OCR/OCR.html

[9] N. Otsu, "Threshold selection method from gray-level histograms." *IEEE Trans Syst Man Cybern*, sv. SMC-9, str. 62–66, 1979. https://doi.org/10.1109/TSMC.1979.4310076

[10] B. Su, S. Lu, i C. L. Tan, "Robust document image binarization technique for degraded document images", *IEEE Transactions on Image Processing*, sv. 22, str. 1408–1417, 2013. https://doi.org/10.1109/TIP.2012.2231089

[11] A. S. Hassanein, S. Mohammad, M. Sameer, i M. E. Ragab, “A survey on hough transform, theory, techniques and applications”.

[12] Øivind Due Trier, A. K. Jain, i T. Taxt, “Feature extraction methods for character recognition-a survey”, *Pattern Recognition*, sv. 29, str. 641–662, 4 1996. https://doi.org/10.1016/0031-3203(95)00118-2

[13] K. M. G. Noaman, J. A. M. Saif, i I. A. A. Alqubati, “Optical character recognition based on genetic algorithms”, 2015.

[14] “Deep learning architectures - ibm developer”. [Mrežno]. Adresa: https://developer.ibm.com/articles/cc-machine-learning-deep-learning-architectures/

[15] T. Wang, D. J. Wu, A. Coates, i A. Y. Ng, “End-to-end text recognition with convolutional neural networks”, u *Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012)*, 2012., str. 3304–3308. [Mrežno]. Adresa: https://ieeexplore.ieee.org/document/6460871

[16] T. M. Breuel, A. Ul-Hasan, M. A. Al-Azawi, i F. Shafait, “High-performance ocr for printed english and fraktur using lstm networks”, *Proceedings of the International Conference on Document Analysis and Recognition, ICDAR*, str. 683–687, 2013. https://doi.org/10.1109/ICDAR.2013.140

[17] A. Vaswani, G. Brain, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Łukasz Kaiser, i I. Polosukhin, “Attention is all you need”, 2023.

[18] C.-A. Boiangiu, R. Ioanitescu, i R.-C. Dragomir, “Voting-based ocr system”, *Journal of Information Systems and Operations Management*, sv. 10, str. 470–486, 9 2016.

[19] T. Berg-Kirkpatrick i D. Klein, “Improved typesetting models for historical ocr”, str. 118–123, 2014. https://doi.org/10.3115/v1/P14-2020

[20] D. Garrette, H. Alpert-Abrams, T. Berg-Kirkpatrick, i D. Klein, “Unsupervised code-switching for multilingual historical document transcription”, str. 1036–1041, 2015.

[21] D. Garrette i H. Alpert-Abrams, “An unsupervised model of orthographic variation for historical document transcription”, 2016.

[22] U. Springmann, D. Najock, H. Morgenroth, H. Schmid, A. Gotscharek, i F. Fink, “Ocr of historical printings of latin texts: Problems, prospects, progress”, u *ACM International Conference Proceeding Series*. Association for Computing Machinery, 2014., str. 71–75. https://doi.org/10.1145/2595188.2595205

[23] U. Springmann i A. Lüdeling, “Ocr of historical printings with an application to building diachronic corpora: A case study using the ridges herbal corpus”, *Digital Humanities Quarterly*, sv. 11, 2017.

[24] M. Christy, A. Gupta, E. Grumbach, L. Mandell, R. Furuta, i R. Gutierrez-Osuna, “Mass digitization of early modern textswith optical character recognition”, u *Journal on Computing and Cultural Heritage*, sv. 11. Association for Computing Machinery, 12 2017. https://doi.org/10.1145/3075645

[25] C. Wick, C. Reul, i F. Puppe, “Improving ocr accuracy on early printed books using deep convolutional networks”, 2 2018. [Mrežno]. Adresa: http://arxiv.org/abs/1802.10033

[26] S. Rijhwani, G. Neubig, A. Black, T. Berg-Kirkpatrick, i D. Rosenblum, “Improving optical character recognition for endangered languages”, 2022.

[27] “Leptonica: Leptonica reference documentation”. [Mrežno]. Adresa: https://tpgit.github.io/Leptonica/index.html

[28] J. Sauvola, T. Seppanen, S. Haapakoski, i M. Pietikainen, “Adaptive document binarization”, *Proceedings of the International Conference on Document Analysis and Recognition, ICDAR*, sv. 1, str. 147–152, 1997. https://doi.org/10.1109/ICDAR.1997.619831

[29] “tesseract-ocr/tesseract: Tesseract open source ocr engine (main repository)”. [Mrežno]. Adresa: https://github.com/tesseract-ocr/tesseract/

[30] G. Dharmale, P. Shirsath, A. Shinde, V. Sawant, i A. Chougule, “Remi-care—medicine intake tracker and healthcare assistant”, *Lecture Notes in Networks and Systems*, sv. 540, str. 273–283, 2023. https://doi.org/10.1007/978-981-19-6088-8_25

[31] “tberg12/ocular: Ocular is a state-of-the-art historical ocr system.”. [Mrežno]. Adresa: https://github.com/tberg12/ocular/

[32] R. Kneser i H. Ney, “Improved backing-off for m-gram language modeling”, *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, sv. 1, str. 181–184, 1995. https://doi.org/10.1109/ICASSP.1995.479394

[33] “Ghostscript”. [Mrežno]. Adresa: https://www.ghostscript.com/

[34] V. I. Levenshtein, “Binary codes capable of correcting deletions, insertions, and reversals”, *Doklady Akademii Nauk SSSR*, sv. 163, str. 845–848, 1965.

Sažetak

SUSTAV ZA OPTIČKO RASPOZNAVANJE TEKSTA STARIJIH KNJIGA NA HRVATSKOME JEZIKU

Dominik Agejev

Cilj rada nadići je uspješnost gotovih sustava za optičko raspoznavanje teksta na starijim knjigama hrvatskoga jezika koristeći se nenadziranim metodama učenja, uz predobradu i naknadnu obradu. Razmatraju se najznačajniji slobodno dostupni OCR alati prikladni zadatku, Tesseract, OCR sustav opće namjene koji održava Google, te Ocular, razvijen specifično za primjenu na antikvarnim dokumentima. Nakon treniranja i optimiziranja hiperparametara Oculara, uspoređen je s Tesseractom gdje se pokazuje da usprkos starijoj arhitekturi u bitnome nadjačava Tesseract, ali uz određena ograničenja. Konačno, izveden je sustav glasanja kojim se postiže veća uspješnost od one samostalnih modela.

Ključne riječi: OCR; optičko raspoznavanje teksta; računalni vid; Ocular; Tesseract;

Abstract

Optical character recognition system for older books in Croatian

Dominik Agejev

The aim of the paper is to surpass the accuracy of out-of-the-box systems at Optical Character Recognition of historical documents in the Croatian language relying on unsupervised learning methods, preprocessing and postprocessing. The most appropriate freely available OCR tools are evaluated, namely Tesseract, a general-purpose OCR system maintained by Google, and Ocular, developed specifically for use on historical documents. After training and optimizing Ocular's hyperparameters it is compared to Tesseract where it is shown that despite its older architecture Ocular in the main still bests Tesseract, with certain caveats. Finally, a voting-based system is implemented which achieves greater success than each model alone.

Keywords: OCR; Optical Character Recogniton; Computer Vision; Ocular; Tesseract;