Utjecaj predobrade ulaznih datoteka na točnost optičkog prepoznavanja znakova

Majnarić, Mirela

Master's thesis / Diplomski rad

2019

Degree Grantor / Ustanova koja je dodijelila akademski / stručni stupanj: **University of Zagreb, University of Zagreb, Faculty of Humanities and Social Sciences / Sveučilište u Zagrebu, Filozofski fakultet**

Permanent link / Trajna poveznica: https://urn.nsk.hr/urn:nbn:hr:131:614062

Rights / Prava: In copyright / Zaštićeno autorskim pravom.

Download date / Datum preuzimanja: **2024-03-28**

Filozofski fakultet
Sveučilišta u Zagrebu

Repository / Repozitorij:

ODRAZ - open repository of the University of Zagreb
Faculty of Humanities and Social Sciences

zir.nsk.hr

dabar
DIGITALNI AKADEMSKI ARHIVI I REPOZITORIJI

SVEUČILIŠTE U ZAGREBU
FILOZOFSKI FAKULTET
ODSJEK ZA INFORMACIJSKE I KOMUNIKACIJSKE ZNANOSTI
Ak. god. 2018./2019.

Mirela Majnarić

Utjecaj predobrade ulaznih datoteka na točnost optičkog prepoznavanja znakova

Diplomski rad

Mentor: prof. dr. sc. Hrvoje Stančić

Zagreb, rujan 2019.

Izjava o akademskoj čestitosti

Izjavljujem i svojim potpisom potvrđujem da je ovaj rad rezultat mog vlastitog rada koji se temelji na istraživanjima te objavljenoj i citiranoj literaturi. Izjavljujem da nijedan dio rada nije napisan na nedozvoljen način, odnosno da je prepisan iz necitiranog rada, te da nijedan dio rada ne krši bilo čija autorska prava. Također izjavljujem da nijedan dio rada nije korišten za bilo koji drugi rad u bilo kojoj drugoj visokoškolskoj, znanstvenoj ili obrazovnoj ustanovi.

_____

(potpis)

Sadržaj

Sadržaj ii
1. Uvod 1
2. Princip rada OCR programa 3
2.1. Obrada bitmape teksta OCR programom 3
3. Vrste OCR programa 5
4. Povijesni pregled razvoja OCR programa 8
5. Faktori utjecaja na točnost OCR-a 11
5.1. Kvaliteta izvornika 11
5.2. Razlučivost skeniranja 12
5.3. Formati slikovnih datoteka 13
5.3.1. JPG (JPEG) format 14
5.3.2. TIFF format 14
5.3.3. PNG format 15
5.3.4. BMP format 15
5.4. Optimizacija slike 15
5.4.1. Sjajnost i kontrast 16
5.4.2. Bitna dubina boje 16
5.4.3. Binarizacija 17
5.5. Pogreške pri skeniranju 18
5.5.1. Zakrenutost stranice 18
5.5.2. Obrubi skeniranih stranica 19
6. Predobrada ulaznih datoteka tijekom skeniranja 21
7. Predobrada ulaznih datoteka GIMP softverom za obradu slika 23
7.1. Stupanj kompresije slike u JPG formatu 24
7.2. Sjajnost i kontrast 25

7.3. Binarizacija slike 25
7.4. Zakrenutost i obrubi stranice 27
8. Istraživanje 29
8.1. Uzorak 29
8.2. Korištena softverska rješenja 30
8.2.1. ABBYY FineReader 12 30
8.2.2. ISRI analitički alati 32
8.3. Utjecaj izbora slikovnog formata, bitne dubine boje i razlučivosti na točnost OCR-a 34
8.4. Utjecaj stupnja kompresije JPG formata na točnost OCR-a 36
8.5. Utjecaj naknadne binarizacije slikovnih datoteka na točnost OCR-a 37
8.6. Ostala zapažanja tijekom istraživanja 40
9. Zaključak 41
10. Literatura 44
Popis oznaka i kratica 48
Popis slika 49
Popis tablica 50
Prilozi 51
Prilog 1 – testirani uzorak, TIFF format u boji skeniran razlučivošću od 600 DPI 51
Sažetak 52
Summary 53

1. Uvod

Informacijsko društvo u kojem živimo karakterizira neprestana proizvodnja velikih količina informacijskih sadržaja koje informacijske ustanove kao što su arhivi, knjižnice i muzeji nastoje prikupiti, organizirati i učiniti dostupnima svojim korisnicima. Ti su sadržaji danas u sve većoj mjeri izvorno proizvedeni u digitalnom okruženju te se nazivaju digitalno rođenom (engl. digitally born) građom koja korisnicima tih ustanova postaje lako dostupna putem Interneta. To znači da korisnici više ne moraju osobno dolaziti u prostor ustanove kako bi pristupili pretraživanju i pregledavanju željene građe, već to mogu učiniti iz komfora vlastitog doma.

Osim izvorno digitalne građe, velika količina građe dugo je postojala isključivo u analognom obliku te je kao takva zahtijevala fizičku prisutnost pojedinca koji joj želi pristupiti. Zahvaljujući konstantnom napretku informacijsko-komunikacijske tehnologije te primjeni različitih postupaka digitalizacije i ona današnjim korisnicima postaje sve dostupnija na daljinu. Digitalizirana građa lako je i brzo dostupna i pretraživa, što je iz perspektive današnjeg korisnika koji žive u svijetu gdje je „vrijeme novac“ od iznimnog značaja.

Sama građa predviđena za digitalizaciju može biti tekstualna, slikovna, zvučna, video i trodimenzionalna (Stančić, 2009, str. 33), a u ovom će radu fokus biti na digitalizaciji tekstualne građe uz uporabu softvera za optičko prepoznavanje znakova (engl. Optical Character Recognition, OCR). Optičko prepoznavanje znakova odnosi se na tehnologiju koja omogućuje konverziju različitih vrsta dokumenata, poput skeniranih papirnatih dokumenata, PDF datoteka ili slika dobivenih digitalnim fotoaparatom u pretražive podatke koje je moguće naknadno uređivati („What is OCR and OCR technology“, bez dat.). Iako se tekstualna građa može digitalizirati i prepisivanjem, taj postupak nije prikladan za opsežne količine građe, ali je nužan kod rukopisne građe kod koje OCR nije učinkovit. Stoga se češće primjenjuje metoda skeniranja ili snimanja digitalnim fotoaparatom u kombinaciji s primjenom OCR softvera. Bez uporabe takvog specijaliziranog softvera, digitalizacija tekstualne građe ne bi bila svrhovita, jer dobivena slikovna datoteka nije pogodna za daljnje pretraživanje i obradu teksta (Stančić, 2009, str. 55-56).

OCR tehnologija danas je prisutna u različitim domenama ljudskih djelatnosti među kojima se ističu postupci automatizacije poslovanja, primjerice u bankama, uredima ili poštama. Koristi se i u postupku verifikacije podataka poput potpisa, osobnih dokumenata, registarskih tablica vozila, ali i pri prijevodu tekstova u stvarnom vremenu korištenjem mobilnih aplikacija. No

bez obzira na sofisticiranost postojećeg OCR softvera, on još uvijek ne može konkurirati ljudskim sposobnostima čitanja na željenoj razini točnosti (Chaudhuri, Mandaviya, Badelia i Ghosh, 2017, str. 35-36).

Točnost dobivenih izlaznih podataka ovisi o mnogim čimbenicima, poput kvalitete izvornog predloška, vrste i veličine fonta, kontrasta između teksta i pozadine, ali i nečistoća na predlošku, kao i deformacije teksta. Ukoliko su znakovi na neki način oštećeni, primjerice potezi znakova nisu spojeni gdje bi trebali biti, to može utjecati na točnost njihovog prepoznavanja. Također, veća količina informacija o slici koja se obrađuje osigurava veći stupanj prepoznavanja, a to se postiže odabranom rezolucijom pri skeniranju predloška, dubinom boje te formatom slikovne datoteke koja se učitava u OCR softver (European Comission on Preservation and Access, 1997). Da bi se osigurali što bolji izlazni rezultati, moguće je primijeniti određene tehnike predobrade ulazne datoteke koje će osigurati optimalno funkcioniranje OCR programa. Neke od tehnika korisnik mora manualno provesti koristeći neki od alata za obradu slika i fotografija, dok su neke tehnike ugrađene u softverska rješenja OCR programa, ovisno o njihovoj sofisticiranosti.

U ovom radu bit će ispitane različite tehnike predobrade slikovnih datoteka i njihov utjecaj na konačni rezultat konverzije iz slikovne u tekstualnu datoteku po upotrebi OCR softvera. One će biti podijeljene na mogućnosti predobrade slika tijekom procesa skeniranja te dodatne mogućnosti predobrade u GIMP softveru za obradu slika. Potom će dobivene slikovne datoteke biti obrađene odabranim OCR alatom te će dobiveni rezultati biti uspoređeni kako bi se ustanovilo pri kojim uvjetima se dobiva najveća točnost prepoznavanja znakova pomoću OCR alata. Točnost prepoznavanja bit će ispitana ISRI analitičkim alatima za evaluaciju OCR-a.

2. Princip rada OCR programa

Čovjek razne oblike, pa tako i tekst, najčešće prepoznaje automatski, bez previše analize i razmišljanja, zahvaljujući vježbi i iskustvu stečenima tijekom života. Takav pristup kod računala naravno nije moguć, jer pri strojnom prepoznavanju znakova moraju postojati jasno definirane faze prepoznavanja opisane programskim kodom. Prema Radoševiću (1996, str. 18-19), sam postupak prepoznavanja teksta može se podijeliti i opisati u tri faze.

Prvu fazu čini dobivanje bitmape teksta, uglavnom korištenjem skenera. Bitmapu teksta čini pravilna pravokutna mreža piksela (engl. pixel, picture element), odnosno slikovnih elementa, a oni su osnovni elementi računalne rasterske slike. Gušća mreža i sitniji elementi rezultiraju kvalitetnijom slikom i većom razlučivošću. Preporučena razlučivost skeniranja ovisi o vrsti predloška koji se skenira, a izražava se u točkama po inču (engl. DPI - dot per inch), uobičajenoj mjernoj jedinici za iskazivanje kvalitete slike digitalizirane skenerom („Slikovni element“, bez dat.). Također, preporuča se i jednobojno skeniranje čime se dobiva crno-bijela slika, a samim time i bolji kontrast između piksela koji predstavljaju tekst, odnosno pozadinu.

Druga faza je obrada bitmape teksta pomoću OCR programa čime se dobiva tekst u ASCII formatu (engl. American Standard Code for Information Interchange). Ovaj akronim označava američki normirani kod za razmjenu informacija („ASCII“, bez dat.), a ekstenzija tako dobivene tekstovne datoteke je *.txt*. Rezultat prepoznavanja teksta obično može biti pohranjen i u nekom drugom općeprihvaćenom formatu tekstovnih datoteka, kao što su *.doc* ili *.docx* (ekstenzije Microsoftovog programa Word za obradu teksta), *.rtf* (engl. rich text format), pretraživi *.pdf* („ABBYY Finereader 12“, bez dat.). Osim očekivane visoke točnosti prepoznavanja znakova, i brzina strojnog prepoznavanja znakova mora premašivati onu ljudske sposobnosti čitanja.

Treća faza je obrada teksta dobivenog u drugoj fazi za što se koriste programi za obradu teksta poput Microsoft Worda. Upotrebom alata za provjeru gramatičke ispravnosti teksta moguće je prepoznati i ispraviti pogreške nastale pri prepoznavanju teksta te se može pristupiti daljnjem uređivanju teksta do željene konačne verzije.

2.1. Obrada bitmape teksta OCR programom

Druga faza, obrada bitmape teksta OCR programom, bit će podrobnije objašnjena u nastavku, jer se u njoj mogu uočiti i izdvojiti principi rada OCR programa. U prvom koraku obrade

bitmape teksta stranica se dijeli u blokove teksta te se izdvajaju različiti elementi na stranici, kao što su tekst, slika, tablice i sl., uz mogućnost korisničke intervencije i unosa izmjena. Slijedi prepoznavanje teksta unutar izdvojenih blokova, pri čemu se izdvaja znak po znak, pridružuje mu se pripadajući kod i konačno se sastavlja tekst. Pretpostavka od koje kreće izdvajanje znakova jest da međusobno povezani crni pikseli predstavljaju jedan znak koje računalo valja prepoznati, dok bijeli pikseli koji ga okružuju čine pozadinu (Radošević, 1996, str. 19-20).

Obradu bitmape teksta Radošević ujedno prepoznaje i kao najosjetljiviju fazu prepoznavanja teksta (1996, str. 20-21). Svoju tvrdnju potkrepljuje činjenicom kako se mnogi znakovi sastoje od više dijelova (npr. točke i kvačice na slovima) koje treba ispravno izdvojiti, prepoznati i zatim ih sastaviti i pridružiti im jedinstveni kod. Zbog lošije kvalitete predloška ili slabije kvalitete skeniranja može se dogoditi spajanje dvaju susjednih znakova ili izlomljenost jednog znaka koji bi trebao biti cjelovit. Ovisno o sofisticiranosti programa za prepoznavanje znakova, navedeni problemi mogu bitno utjecati na točnost prepoznavanja znakova.

Po izdvajanju dijelova znakova, slijedi njihovo prepoznavanje, pri čemu se kombinirano koriste dvije metode, prepoznavanje na temelju predložaka i prepoznavanje na temelju svojstava oblika (Radošević, 1996, str. 22-24).

Prepoznavanje na temelju predložaka temelji se na usporedbi izdvojenih oblika s gotovim predlošcima pohranjenima u bazi podataka OCR programa. Prethodno se provodi normiranje veličine oblika kako bi oni bili usporedivi s pohranjenim predlošcima, a zatim se usporedbom pronalazi najsličnijih predložak i njegov pripadajući kod.

Prepoznavanje na temelju svojstava oblika bazirano je na utvrđivanju osnovnih značajki oblika, odnosno uzoraka. Cilj je utvrđivanje što manjeg broja osnovnih značajki uzorka koje istovremeno nose dovoljno informacija za prepoznavanje oblika. Izdvajanjem najreprezentativnijih informacija iz sirovih podataka minimiziraju se varijacije uzoraka unutar klase, dok se istovremeno ističu razlike uzoraka između klasa. Nakon prepoznavanja slijedi postupak sastavljanja znakova ukoliko ih sačinjava više zasebnih dijelova, te sastavljanje teksta kojega čine riječi odvojene razmacima i redci teksta.

3. Vrste OCR programa

Programe za optičko prepoznavanje znakova moguće je podijeliti po nekoliko kriterija.

Prvi kriterij podjele koji nudi Walls (2008) temelji se na tipu teksta (strojni ili rukopisni) i tipu prepoznavanja (na razini znaka ili riječi) pa se razlikuju sljedeća softverska rješenja:

1. OCR (engl. Optical Character Recognition) programi prepoznaju pojedinačne znakove kod strojno otisnutog teksta.
2. OMR (engl. Optical Mark Recognition) programi prepoznaju prekrižena, zacrnjena ili na neki drugi način označena polja. Primjena ove tehnologije uobičajena je pri obradi rezultata standardiziranih ispita i anketa.
3. ICR (engl. Intelligent Character Recognition) programi su naprednija verzija OCR-a koji prepoznaju rukom pisana tiskana slova, a ona su pišu zasebno u za to predviđenim poljima, kao primjerice kod uplatnica.
4. IWR (engl. Intelligent Word Recognition) programi prepoznaju rukopis, i to pisana slova bez ikakvih ograničenja u stilu pisanja, a prepoznavanje se vrši na razini riječi ili fraze.

Slika 1. Optičko prepoznavanje znakova ili riječi prema složenosti postupka (preuzeto 23.04.2019. s https://slideplayer.com/slide/10330560/)

Drugi kriterij podjele OCR programa jest prema području očitavanja koje program obuhvaća („Simple Software - OCR Software Guide“, bez dat.; „Zonal OCR - What's it good for?“, 2015) pa je moguće izdvojiti:

1. Cjelovite OCR programe koji očitavaju čitav dokument i pretvaraju ga u jedan od navedenih formata:
    a. običan tekst kod kojega su zadržane samo osnovne tekstualne informacije u uzastopnom redoslijedu,
    b. formatirani tekst kod kojega je očitani tekst podijeljen u odlomke te su očuvane veličina i stil fonta, a omogućuje i očuvanje tablica te podataka unesenih u njih,

c. identična kopija kod koje su sačuvane sve informacije iz izvornika, uključujući grafičke elemente, kao i izvorni raspored elemenata na stranici,
d. pretraživi dokument kod kojeg su tekstualne informacije pohranjene na zasebnom, skrivenom sloju iza skeniranog izvornika, što omogućuje pretraživanje dokumenta iz istovremeno zadržavanje izgleda izvornika.

2. Zonske OCR programe koji očitavaju tekst u označenim područjima, odnosno zonama dokumenta, a pomažu pri indeksiranju i upravljanju dokumentima. Informacije očitane u unaprijed određenim zonama mogu služiti za imenovanje dokumenata, njihovu pohranu na željenu lokaciju ili arhiviranje pojedinih informacija u baze podataka.

Sljedeća podjela OCR programa jest na desktop i serverske OCR programe. Za korištenje desktop verzija OCR programa nužno ih je preuzeti i instalirati na osobno računalo, pri čemu valja voditi računa o tome je li odabrani OCR softver kompatibilan s operativnim sustavom računala, kao i o jezicima za koje je namijenjen. Pri pokretanju programa koriste se resursi računala na kojem je instaliran. *OCRopus* (https://github.com/tmbdev/ocropy) i *SimpleOCR* (https://www.simpleocr.com/) su primjeri OCR programa koji za sada postoje isključivo u desktop verzijama. Ukoliko je potrebno provesti prepoznavanje teksta na velikom broju dokumenata u kratkom vremenskom roku, jedno računalo tada možda neće imati dovoljno procesorske snage za uspješno izvršavanje zadatka. Zahvaljujući globalnom razvoju informacijske tehnologije, nastali su serverski OCR programi kod kojih se sam proces prepoznavanja teksta preusmjerava na server na kojemu se pokreće željeni softver, u ovom slučaju OCR softver. Na taj se način rasterećuje resurse osobnog računala, a serverski OCR može inteligentno rasporediti izvršavanje zadataka između više procesora i time bitno ubrzati cjelokupni proces pretvorbe slike u tekst („Simple Software - OCR Servers“, bez dat.). Primjeri poznatih serverskih OCR programa su *IRISPowerscan 10 Server* (https://www.irislink.com/EN-HR/c1839/IRISPowerscan-10-Server---Central-Management.aspx) i *ABBYY FineReader Server* (https://www.abbyy.com/en-us/finereader-server/). Valja istaknuti da su pojedini OCR programi poput *Tesseracta* dostupni u desktop (https://github.com/tesseract-ocr/tesseract) i serverskim (https://tesseract.projectnaptha.com/ inačicama.

Konačno, poput mnogih drugih softverskih rješenja, i OCR programi postoje u komercijalnim i besplatnim varijantama. Za komercijalne OCR programe uobičajena je upotreba vlasničke licence (engl. proprietary licence). Njome proizvođač uz novčanu naknadu daje pravo korištenja softvera pod određenim uvjetima, no zadržava pravo intelektualnog vlasništva, što

znači da korisnici ne mogu vidjeti ni mijenjati izvorni softverski kod („Proprietary software“, bez dat.). U ovu kategoriju spadaju *ABBYY FineReader 14* (https://www.abbyy.com/en-us/finereader/) i *Nuance OmniPage Ultimate* (https://www.nuance.com/print-capture-and-pdf-solutions/optical-character-recognition/omnipage/omnipage-ultimate.html). Suprotno tome, postoje i OCR programi kao besplatna softverska podrška (engl. freeware) i softver otvorenog koda (engl. open source software). Besplatna softverska podrška ili slobodni softver odnosi se na računalni program čija je uporaba besplatna, ali izmjene ili kopiranje programskoga koda nisu dozvoljene, kao ni njegovo raspačavanje i prodaja bez znanja i dozvole autora („Freeware“, bez dat.). Neki od freeware OCR programa su *SimpleOCR*, *MeOCR* (http://www.meocr.com/), *Online OCR* (https://www.onlineocr.net/), a podržani su i u Googleovom servisu *Google Drive* (https://www.google.com/drive/). Softver otvorenog koda (engl. open source software) označava besplatne i slobodne programe, na način da autor odabire licencu kojom korisnicima daje pristup izvornom kodu, a korisnici time dobivaju pravo njegovog proučavanja i izmjene bez ograničenja, kao i njegovu daljnju distribuciju pod istim uvjetima određenima licencom izvornog softvera („The Open Source Definition“, 2007). Kao najčešće korištena slobodna licenca navodi se GNU licenca (engl. GNU General Public Licence - GPL) („Open source licence“, bez dat.). U ovu kategoriju spadaju već prethodno spomenuti *Tesseract*, *OCRopus*, te *Ocrad* (https://www.gnu.org/software/ocrad/).

4. Povijesni pregled razvoja OCR programa

Iako se optičko prepoznavanje znakova kao ideja rađa još početkom 20. stoljeća, intenzivnije istraživanje OCR-a započinje zajedno s razvojem digitalnih računala. Chaudhuri i sur. (2017, str. 13) navode kako su 1950-ih godina izrađeni mehanički i optički uređaji koji su mogli digitalizirati redak po redak teksta, a već 1954. godine se u američkom časopisu *Reader's Digest* u primjeni našao OCR uređaj kojim su se strojopisna izvješća o prodaji pretvarala u bušene kartice za pohranu tih podataka u računalu.

Komercijalne OCR sustave nastale između ranih 1960-ih i ranih 1970-ih godina karakterizira mogućnost prepoznavanja jednog fonta, a taj font je obično posebno osmišljen za strojno čitanje, za postizanje što veće točnosti pri optičkom prepoznavanju znakova. Tako je nastao američki standardni font OCR-A, pojednostavljen i osmišljen za lakše optičko prepoznavanje, no istovremeno razumljiv ljudima. Nedugo nakon njega razvijen je i europski standardni font pod nazivom OCR-B, koji je svojim izgledom od američkog standarda bio mnogo ugodniji ljudskom oku te manje tehničkog izgleda. Usporedba izgleda ova dva fonta prikazana je na slikama 2 i 3.

Slika 2. Font OCR-A (preuzeto 13.04.2019. s https://en.wikipedia.org/wiki/OCR-A)

Slika 3. Font OCR-B (preuzeto 13.04.2019. s https://en.wikipedia.org/wiki/OCR-B)

U tom su periodu u SAD-u nastali prvi OCR uređaji za razvrstavanje pošte i prvi skener koji je uspješno prepoznavao rukom pisane brojeve („Timeline of optical character recognition“, bez dat.). Naravno, rani OCR sustavi su imali svoja ograničenja po tome što su bili sposobni prepoznati samo savršen tekst, a prepoznavanje se vršilo usporedbom svakog pojedinog očitanog znaka s predloškom pohranjenim u bazi podataka kako bi se pronašlo podudaranje. S

obzirom na memorijska ograničenja, te baze podataka nisu mogle biti opširne i bez savršenog podudaranja softver nije uspijevao prepoznati znak („A brief history of OCR: the technology inside your scanmarker", 2019).

U razdoblju 1970-ih i 1980-ih godina cijena hardvera pada, brzina rada računalnih procesora sve je veća, a OCR sustavi postaju sve rašireniji u komercijalnoj upotrebi. Počinju se koristiti u očitavanju računa kreditnih kartica, putovnica, cijena na artiklima. Godine 1989. osnovana je i softverska tvrtka ABBYY, danas jedan od vodećih komercijalnih proizvođača OCR softverskih rješenja („Timeline of optical character recognition", bez dat.). Kvaliteta samih OCR sustava raste te oni postaju sposobni prepoznati razne fontove, a pri prepoznavanju više se ne traži savršeno podudaranje, već se znakovi prepoznaju po njihovom općenitom obliku, neovisno o razlikama u dizajnu fonta. (Chaudhuri i sur., 2017, str.14).

Od 1990-ih godina nadalje u OCR sustavima se primjenjuju novi razvojni alati i metodologije temeljene na umjetnoj inteligenciji te se razvijaju još sofisticiraniji algoritmi za prepoznavanje znakova. Neke od korištenih metodologija su umjetne neuronske mreže (engl. ANN - artificial neural networks), skriveni Markovljevi modeli (engl. HMM - hidden Markov models), teorija neizrazitih skupova (engl. fuzzy sets) i obrada prirodnog jezika (engl. NLP - natural language processing) (Chaudhuri i sur., 2017, str.15).

Od 2000. godine OCR tehnologija postaje dostupna kao usluga putem interneta (engl. WebOCR), što znači da instalacija softvera na vlastito osobno računalo više nije potrebna. Godine 2015. Google objavljuje OCR alate za besplatno skeniranje Google Drive dokumenata u preko 200 jezika. Razvijene su i mobilne aplikacije koje koriste OCR za prijevode znakova na stranom jeziku u stvarnom vremenu pomoću pametnih mobilnih uređaja. Dovoljno je snimiti fotografiju, učitati je u OCR aplikaciju koja prepoznaje jezik i riječi i korisniku vraća prijevod snimljenog teksta. Varijante OCR softvera danas se koriste i u prepoznavanju i očitavanju faktura i računa, putovnica, registarskih tablica vozila te dokumenata osiguranja („Timeline of optical character recognition", bez dat.).

Suvremena OCR tehnologija čini osnovu usluga brojnih internetskih stranica koje svojim korisnicima nude pristup knjigama u elektroničkom obliku. Tako *Projekt Gutenberg* zahvaljujući OCR tehnologiji korisnicima besplatno nudi točne i potpune elektroničke verzije za više od 58 tisuća knjiga u javnom vlasništvu (Project Gutenberg, 2019), dok se usluga stranice *Google Books* temelji na skeniranju knjiga i časopisa, njihovoj konverziji u pretraživi

tekstualni oblik upotrebom OCR tehnologije te pohrani u digitalnoj bazi podataka (Google Books, 2011).

5. Faktori utjecaja na točnost OCR-a

Točnost optičkog prepoznavanja znakova može se mjeriti na razini pouzdanosti prepoznavanja znaka ili riječi. Proizvođači OCR softvera točnost OCR tehnologije uglavnom izražavaju postotkom uspješno prepoznatih znakova, a taj postotak označava stopu pouzdanosti prepoznavanja nekog znaka. Primjerice, razina točnosti od 99% govori da je jedan od 100 prepoznatih znakova nije pouzdano prepoznat. Za povećanje točnosti na razini riječi, brojni OCR programi prepoznate riječi uspoređuju s rječnicima jezika korištenog u tekstu. Riječi koje sadrže nepouzdano prepoznat znak pretražuju se u rječniku. Ako se prepoznata riječ ne nalazi u rječniku, pretpostavlja se da se radi o pogrešci te se riječ mijenja u onu s najvećim stupnjem sličnosti („Improve OCR Accuracy With Advanced Image Preprocessing“, bez dat.).

Prema Holley (2009), pravu mjeru točnosti može odrediti isključivo čovjek. Ukoliko se ispituje točnost prepoznavanja na razini znaka, broje se točno prepoznati znakovi, a ako se ispituje točnost prepoznavanja na razini riječi, broje se ispravno prepoznate riječi. Autorica navodi i metode koje se pritom koriste, a to su korektura teksta ili ručni unos cijeloga teksta i njegova usporedba s rezultatom dobivenim OCR-om te ističe kako su one vremenski vrlo dugotrajne.

Holley (2009) izdvaja sljedeće faktore koji utječu na točnost optičkog prepoznavanja znakova:

- kvaliteta izvornika,
- razlučivost skeniranja,
- format slikovne datoteke,
- optimizacija slike, koja uključuje sjajnost i kontrast te bitnu dubinu boje,
- pogreške pri skeniranju, koji uključuju zakrenutost stranice i tamne obrube.

U nastavku će svaki od navedenih faktora ukratko biti predstavljen i objašnjen njegov utjecaj na postupak optičkog prepoznavanja znakova.

5.1. Kvaliteta izvornika

Kvaliteta izvornika izravno utječe na kvalitetu slikovne datoteka koja će se iz njega dobiti, a time i rezultat optičkog prepoznavanja znakova. Što je kvaliteta izvornika bolja, znakove je lakše razlikovati od pozadine i ostalih elemenata koje izvornik sadrži te će i točnost OCR-a biti viša. Kadgod je moguće, nužno je pronaći što kvalitetniji izvornik koji ne sadrži nikakva oštećenja ni deformacije, jer će u suprotnome njegova skenirana verzija zahtijevati dodatnu obradu u nekom od programa za obradu slike prije upotrebe samog OCR programa. Ukoliko

se želi digitalizirati starija građa, ovaj je uvjet teško ispuniti zbog ograničenog broja izvornika i fizičkog propadanja materijala na kojem je građa otisnuta. Izvornik može biti zgužvan, poderan ili na bilo koji drugi način oštećen, slova mogu izblijedjeti, a papir požutjeti ili se na njemu pojaviti mrlje i ostale promjene u boji. Tekst može biti ispisan tintom niskoga kontrasta, nestandardnim fontom, strojopisom ili rukom, a znakovi biti izlomljeni (slika 5) ili se više znakova može spojiti u jedan (slika 4). Navedene promjene u boji utječu na kontrast između znakova i pozadine pa softver teže raspoznaje bjeline između znakova, riječi i redaka. Isto vrijedi i za oštećenja papira i lošu kvalitetu otiska znakova koji mogu smanjiti točnost prepoznavanja znakova (Henson, 2018; „Improve OCR Accuracy With Advanced Image Preprocessing“, bez dat.).

Slika 4. Spojeni znakovi „k” i „r” (Radošević, 1996, str. 21)

Slika 5. Razlomljeni znakovi „a”, „o” i „m” (Radošević, 1996, str. 22)

5.2. Razlučivost skeniranja

Razlučivost ili rezolucija (engl. resolution) je jedna od bitnih značajki skenera. Ona označava broj piksela koje skener može očitati u području skeniranja, a obično se izražava u točkama po inču (engl. DPI - dot per inch). Viša razlučivost daje bolju kvalitetu skenirane slike, no ujedno usporava rad skenera (Stančić, 2009, str. 43). Kod stolnog skenera razlikuju se horizontalna i vertikalna razlučivost. Horizontalna razlučivost ovisi o broju točaka na fotoosjetljivom elementu skenera, dok je vertikalna razlučivost ovisna o preciznosti pomicanja glave skenera duž predloška (Strgar Kurečić, bez dat.).

Također, kod skenera se razlikuje optička i interpolirana razlučivost. Optička razlučivost je stvarna razlučivost skenera koja se ostvaruje korištenjem optike samog uređaja, dok se interpolirana razlučivost dobiva softverskim izračunom. Na temelju svojstava optički prepoznatih susjednih piksela izračunavaju se svojstva dodatnog piksela koji će između njih biti umetnut. Iako se time dobiva slika veće razlučivosti, njena je kvaliteta obično lošija, jer su dodani pikseli rezultat procjene, a ne stvarnog prepoznavanja s uzorka (Stančić, 2009, str. 43).

Iako izbor razlučivosti skeniranja uvelike ovisi o samoj vrsti građe i korištenoj veličini fonta, mnogi autori preporučaju 300 DPI kao optimalnu razlučivost skeniranja. Razlučivost od 200 DPI i niža daje nedovoljno jasne rezultate, dok razlučivost iznad 600 DPI nepotrebno povećava vrijeme skeniranja i veličinu izlazne datoteke bez značajnog povećanja njene kvalitete (Henson, 2018; „Improve OCR Accuracy With Advanced Image Preprocessing“, bez dat.; Stančić, 2009, str. 57).

Slika 6. Slova skenirana različitim razlučivostima (preuzeto 13.05.2019. s http://preservationtutorial.library.cornell.edu/conversion/conversion-04.html)

Povećanje prostorne razlučivosti pri skeniranju tekstualnog gradiva može čak unijeti i dodatni šum (engl. noise) u slikovnu datoteku. Pod šumom podrazumijevamo razne smetnje, odnosno nasumične varijacije u sjajnosti ili bojama na slikama, koje se mogu pojaviti u slikovnoj datoteci, a ne sadrže za nju relevantne informacije. Šum može proizvesti senzor i sklopovlje skenera ili digitalnog fotoaparata („Image noise“, bez dat.). Primjerice, povećanjem razlučivosti skener može registrirati i nepravilnosti i nečistoće papira ili pogreške pri otisku znakova, a takvi šumovi OCR programu otežavaju prepoznavanje znakova te se za njihovo uklanjanje u predobradi obično koriste filteri za zaglađivanje, izoštravanje, binarizaciju (engl. thresholding ili binarization, pretvaranje slike u boji, odnosno skali sive boje u crno-bijelu), uklanjanje pozadinske teksture ili boje i prilagodbu kontrasta. U istu svrhu koriste se i morfološke operacije koje ističu oblike relevantne za optičko prepoznavanje znaka, a one nebitne za njihovu detekciju prikrivaju. (Chaudhuri i sur., 2017, str. 18).

5.3. Formati slikovnih datoteka

Zbog uštede podatkovnog prostora slikovni se zapisi komprimiraju, za što se koriste dva načina kompresije – kompresija bez gubitaka (engl. lossless compression) i kompresija uz gubitke (engl. lossy compression). Primjenom kompresije bez gubitaka smanjuje se veličina slike, no njena izvorna kvaliteta je sačuvana u cijelosti. Kompresija uz gubitke složenim algoritmima izračunava koji se dijelovi informacija mogu izostaviti pri čemu dolazi do kontroliranog gubitka kvalitete. Odnos stupnja kompresije i kvalitete slike je obrnuto proporcionalan, što

znači da viši stupanj kompresije rezultira slikom niže kvalitete. Ipak, prednost kompresije uz gubitke je značajna mogućnost smanjenja veličine slikovne datoteke (Stančić, 2009, str. 75).

Za pohranu slikovnih datoteka koriste se različiti formati slikovnih datoteka, a izbor formata ovisi o načinu korištenja te datoteke, primjerice služe li kao priprema za tisak, za trajnu pohranu slikovnog gradiva ili distribuciju putem mreže i prikaz na ekranu. Svaki od formata razvijen je za specifičnu uporabu i kao takav ima svoje prednosti i nedostatke. U nastavku su opisana četiri formata slikovnih datoteka ponuđenih softverom za skeniranje uređaja *Epson Stylus Photo RX500* korištenog u ovom istraživanju, a to su JPG, TIFF, PNG i BMP.

5.3.1. JPG (JPEG) format

JPG ili JPEG format imenovan je po udruzi koja ga je razvila (engl. Joint Photographic Experts Group). Radi se o najčešće korištenom formatu za digitalne fotografije i prikaz slika na internetu. Kod ovog formata sadržaj slike se komprimira uz gubitak kvalitete uz mogućnost izbora željenog stupnja kompresije. Čak i uz relativno visok stupanj kompresije, kvaliteta slike neće biti znatno narušena. JPG format funkcionira na način da analizira sliku i pronalazi informacije koje je moguće izostaviti, jer su najmanje uočljive prostim okom. Informacije pohranjuje kao 24-bitni zapis boje. Zbog manje veličine datoteka, JPG datoteke se preporuča za distribuciju putem mreže te bolju iskorištenost podatkovnog prostora. Kao nedostatci se navode gubitak kvalitete pri kompresiji te mogući neželjeni gubitci kvalitete pri opetovanoj obradi JPG datoteke u programima za obradu slike (Matthews, bez dat.).

5.3.2. TIFF format

TIFF (engl. Tagged Image File Format) je vrlo fleksibilan format uobičajen u izdavačkoj industriji. Omogućuje pohranu slika različite bitne dubine boje i visoke razlučivosti. U praksi se najčešće koristi kao format bez primjene kompresije što rezultira velikim datotekama, iako postoji i opcija kompresije bez gubitka upotrebom LZW (Lempel-Ziv-Welch) algoritma. Suvremeni digitalni fotoaparati, uz JPG, često nude i TIFF format za najbolju kvalitetu pohrane slike. TIFF format je prikladan kao radni format pri obradi slika u odabranom softveru, jer se višestrukim uređivanjem i spremanjem ne narušava kvaliteta datoteke. Također nudi mogućnost ispisa najviše kvalitete i velikih formata. Nedostatak je što zbog svoje veličine zahtjeva znatno više podatkovnog prostora za pohranu te dulje vrijeme prijenosa i učitavanja (Matthews, bez dat.). Većina međunarodnih smjernica i autora koji se bave tematikom optičkog

prepoznavanja znakova preporučaju upravo TIFF format kao standardni format za pohranu slikovnih datoteka i pohranu master slike, pri čemu nije dozvoljeno korištenje nikakvog oblika komprimiranja (Holley, 2009; „Ministarstvo kulture“, 2007; Stančić, 2009).

5.3.3. PNG format

PNG (engl. Portable Network Graphics) osmišljen je kao svojevrsna nadogradnja GIF formata te se koristi za objavljivanje slika na internetu. Za razliku od GIF formata koji podržava prikaz 256 boja, PNG format omogućuje prikaz 16 milijuna boja te je pogodan je za slike s velikim površinama jednake boje, ali koje sadrže više od 256 boja. Kako ovaj format koristi kompresiju bez gubitaka, slike su bolje kvalitete, no ipak nisu toliko velike da bi usporavale učitavanje internetskih stranica na kojima su sadržane. PNG format je prikladan za pohranu slika visoke kvalitete zahvaljujući velikom rasponu boja te mogućnosti upotrebe alfa-kanala (engl. alpha-channel) za promjenjivu razinu prozirnosti (engl. transparency) slike (Matthews, bez dat.).

5.3.4. BMP format

BMP (engl. Bitmap) je format slikovnih datoteka bez kompresije koji je razvila tvrtka Microsoft, izvorno za upotrebu u operativnom sustavu Windows. Veličina BMP datoteka proizlazi iz činjenice što su podatci o boji pohranjeni u svakom pojedinačnom pikselu koji čini sliku, no bez kompresije. U BMP formatu pohranjuju se dvodimenzionalne digitalne slike, monokromatske ili u boji, različitih bitnih dubina boja, uz dodatnu mogućnost kompresije podataka, alfa kanala i različitih formata boja. Iako je njegova upotreba u mrežnom okruženju podržana, drugi formati su za tu svrhu primjereniji te ga je stoga bolje ne koristiti. Za razliku od toga, visoka kvaliteta slike čini ga prikladnim za ispis („BMP file format“, bez dat.; Dadfar, bez dat.).

5.4. Optimizacija slike

Pod optimizacijom slike podrazumijeva se stvaranje dobrog kontrasta između teksta i pozadine dokumenta. Uključuje podešenja postavki sjajnosti i kontrasta, kao i izbor bitne dubine boje. Provodi se u procesu predobrade datoteke, prije optičkog prepoznavanja znakova, pomoću postavki dostupnih u softveru za skeniranje ili nekom od programa za obradu slike (Holley, 2009).

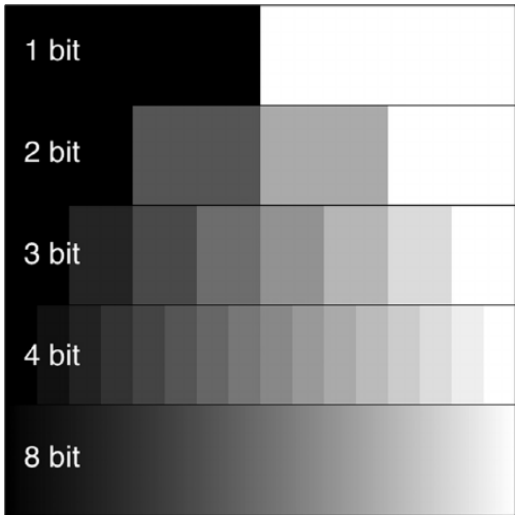
5.4.1. Sjajnost i kontrast

Sjajnost slike (engl. brightness) se odnosi na cjelokupnu svjetlinu ili zatamnjenost slike, dok je kontrast (engl. contrast) razlika u svjetlini između objekata ili regija neke slike. Dakle, podešenja sjajnosti i kontrasta obično su potrebna kada je slika pretamna ili presvjetla, ili kada se želi dodatno istaknuti razlika svjetlijih i tamnijih dijelova slike. Povećanjem sjajnosti svaki pojedini piksel slike se posvjetljuje, dok povećanjem kontrasta svjetla područja postaju svjetlijima, a tamna još tamnijima (Smith, 1997, str. 387). Povećanjem kontrasta znak se bolje ističe u odnosu na pozadinu što znači da će ga OCR softver lakše izdvojiti i prepoznati.

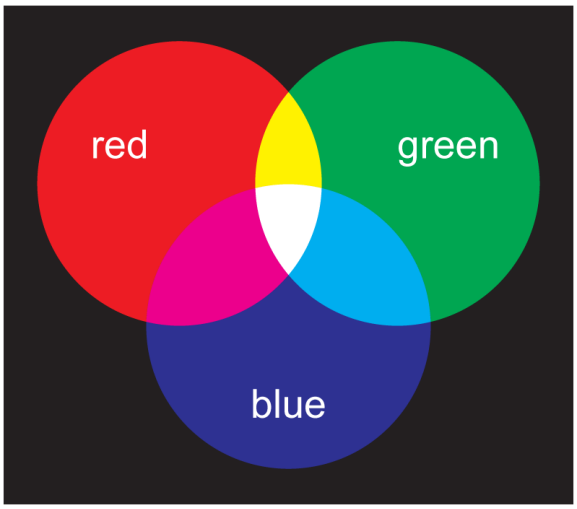
5.4.2. Bitna dubina boje

Digitalna slika je sačinjena od polja točaka, a svaka od tih točaka opisana je određenim brojem bitova, što se naziva bitnom dubinom točke. Bitnom dubinom boje u stvari izražavamo mjeru osjetljivosti skenera. Što je bitna dubina veća, to se nijansa boje svake pojedine točke može preciznije opisati i time je i konačni rezultat skeniranja kvalitetniji. Primjerice, jednim se bitom može opisati samo dvije mogućnosti – 1 ili 0, gdje 1 predstavlja crnu boju, a 0 bijelu. Sa dva bita mogu se izraziti četiri vrijednosti – crno, bijelo i dvije nijanse sive, a sa osam bita 256 mogućih kombinacija – crno, bijelo i 254 nijansi sive (slika 7). Za prikaz digitalne slike u boji koristi se nekoliko sustava prikaza boje, a najčešći su RGB (engl. RGB – Red, Green, Blue), CMYK (engl. CMYK – Cyan, Magenta, Yellow, blacK) i CIELAB sustav. Kod svih navedenih sustava boja se prikazuje miješanjem nekoliko kanala osnovnih boja. Tako se pri skeniranju u boji koristi RGB sustav sa tri kanala - crvenim, zelenim i plavim kanalom. Ako se radi o 24-bitnom skeneru, to znači da može razlikovati 2^8, odnosno 256 nijansi po svakom kanalu, što sveukupno čini 16.777.216 nijansi. (Stančić, 2009, str. 60-62).

Za materijal nad kojim se namjerava provesti optičko prepoznavanje znakova preporuča se korištenje 8-bitne sive skale, dok za kvalitetni tisak i 1-bitna crno-bijela slika daje dobre rezultate (Holley, 2009; „Ministarstvo kulture“, 2007; Stančić, 2009). Skeniranjem slike u sivoj skali uz potrebna podešenja sjajnosti i kontrasta manje se gube detalji no izravnim stvaranjem crno-bijele slike kada softver automatski određuje prag binarizacije („B Is for Binarize“, bez dat.). Postupak binarizacije detaljnije je objašnjen u narednom odjeljku teksta.



Slika 7. Pet različitih bitnih dubina točke i nijanse sive skale koje se njima dobivaju (Lambert, Waters, 2014, str. 41)



Slika 8. RGB sustav boja (preuzeto 29.04.2019. s http://www.supertisak.hr/boje/cmyk-rgb-spot-boje-o-cemu-se-tu-radi)

5.4.3. Binarizacija

Tiskane dokumente obično čini crni tisak na bijeloj pozadini pa je kod optičkog prepoznavanja znakova uobičajena praksa konverzija višebojne slike u binarnu, odnosno crno-bijelu sliku. Taj se proces naziva binarizacija (engl. binarization, thresholding) i često se provodi tijekom samog procesa skeniranja kako bi se uštedili podatkovni prostor i procesorsko vrijeme.

Proces binarizacije iznimno je važan zbog toga što o njemu izravno ovise rezultati optičkog prepoznavanja znakova. Binarizacija koja se izvodi prilikom skeniranja obično je vrlo jednostavna. Koristi se globalni prag, kod kojega se sve razine sive ispod praga fiksno određenoga za cijelu sliku prepoznaju kao znak u crnoj boji, a iznad praga kao pozadina bijele boje. Za dokumente visokog kontrasta s jednoličnom pozadinom i takav globalni prag može biti zadovoljavajući. No u praksi kontrast teksta i pozadine dokumenta može jako varirati te su u tom slučaju potrebne sofisticiranije, adaptivne metode binarizacije kako bi se dobili dobri rezultati.

Osim metoda koje koriste globalni prag, postoje i adaptivne metode binarizacije koje koriste lokalni prag. To znači da se prag izračunava i mijenja ovisno o svojstvima svakog pojedinačnog piksela i njegovog susjedstva. Takve se metode oslanjanju na skeniranje veće bitne dubine boje što zahtijeva više podatkovnog prostora i procesorskog vremena. (Chaudhuri i sur., 2017, str. 16-17). Adaptivne metode binarizacije nužne su za uspješno prepoznavanje tamnog teksta na tamnoj pozadini, svijetlog teksta na bijeloj pozadini, teksta na pozadini različitih boja (npr. gradijenta), ili pozadina kod kojih se nazire tekst s poleđine stranice. Tada se preporuča

skeniranje u sivoj skali ili u boji jer se njime dobiva razlika u tonovima koja nedostaje pri izravnom crno-bijelom skeniranju te se vrši naknadna konverzija u crno-bijelu sliku. No ujedno valja istaknuti da ne koriste svi OCR programi jednako sofisticirane metode binarizacije uz mogućnosti dodatnih podešavanja od strane korisnika, što znači da rezultati prepoznavanja teksta na slikama u boji uvelike ovise i o korištenom softverskom rješenju („B Is for Binarize“, bez dat.).

Slika 9. Primjer dobre binarizacije na uzorku loše kvalitete (Sauvola, Pietikäinen, 2000, str. 227)

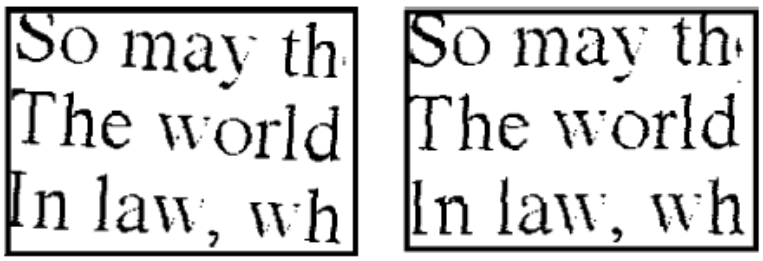
Uz binarizaciju, često se koriste i filteri za zaglađivanje slika (engl. image smoothing) kako bi se uklonili sitni detalji sa slike koji otežavaju optičko prepoznavanje znakova. Ovaj se postupak koristi za reduciranje šumova i dobivanje manje pikselizirane slike, kao i popunjavanje rubova znakova. Boja piksela se zamjenjuje prosječnom vrijednošću njegovih susjednih piksela čime se povećava čitljivost slova na neujednačenoj pozadini (Chaudhuri i sur., 2017, str. 17-18; Cheriet, Kharma, Liu i Suen, 2007, str. 30).

5.5. Pogreške pri skeniranju

5.5.1. Zakrenutost stranice

Linije teksta dokumenta su horizontalne ili vertikalne, pri čemu tekst može teći slijeva ili zdesna, ovisno o jeziku kojim je dokument pisan. Tijekom procesa skeniranja dokumenta, može doći do pogreške kod koje je cijeli dokument ili jedan njegov dio zakrenut u odnosu na x-os stranice (engl. skew). To se primjerice može dogoditi kada mehanizam za uvlačenje stranica dokumenata ne uvuče dobro individualni dokument, primjerice umjesto jednog lista uvuče dva, ili se listovi u procesu skeniranja zgužvaju, što može rezultirati ukošeno skeniranim cijelim dokumentom, pa i nedostatkom dijela teksta (slika 10). Kada se takva pogreška pojavi, treba biti uklonjena zbog toga što smanjuje točnost optičkog prepoznavanja znakova. Kako su algoritmi za analizu rasporeda elemenata na stranici i prepoznavanje znakova općenito vrlo osjetljivi na zakrenutost teksta, prepoznavanje i uklanjanje zakrenutosti su nužni koraci prije analize rasporeda elemenata stranice (Cheriet i sur., 2007, str. 32-34). Ukoliko se radi o

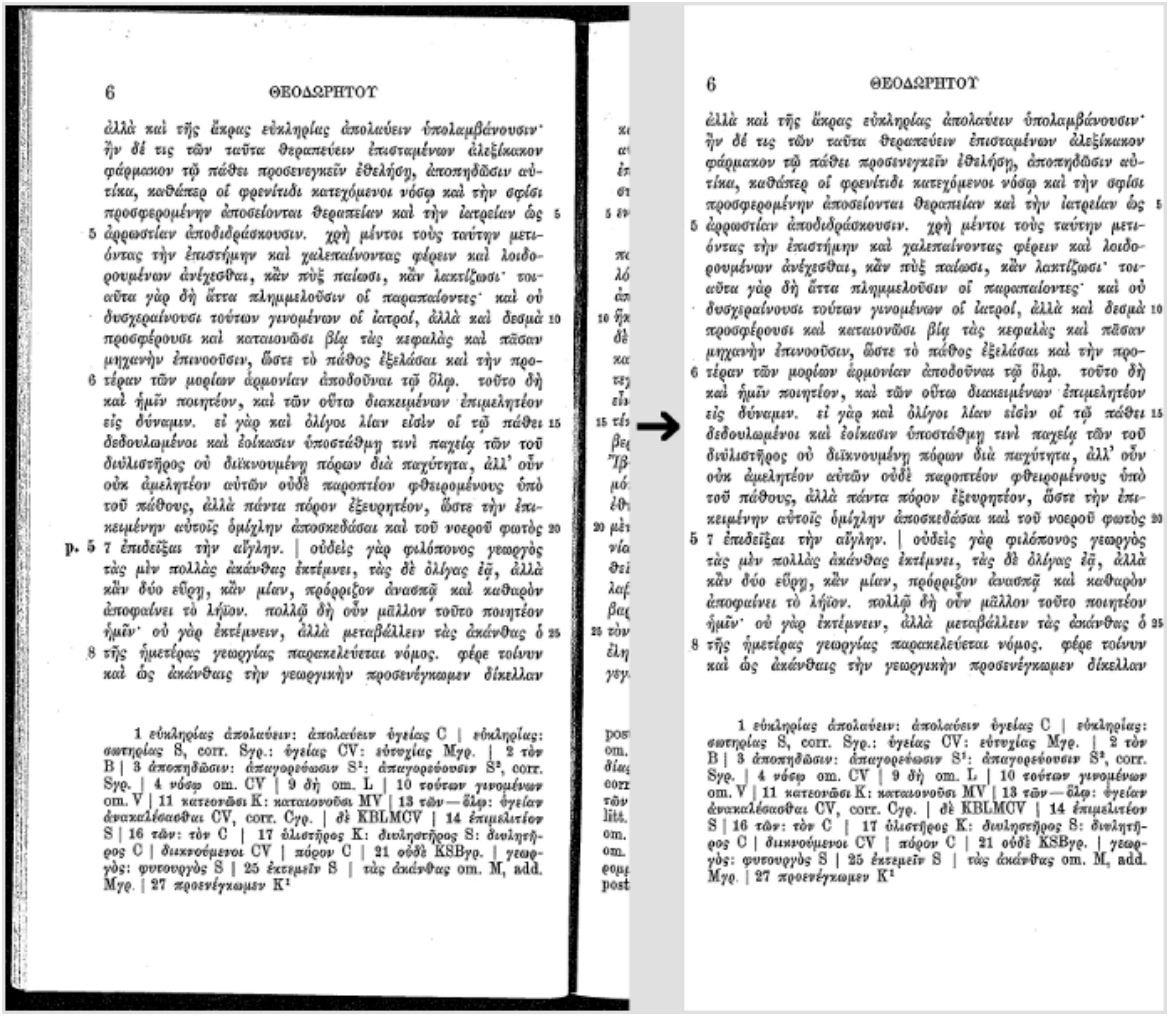
odstupanju zakrenutosti većem od 3 stupnja, preporuča se njegovo ispravljanje (Stančić, 2009, str. 57).



Slika 10. Primjer odstupanja zakrenutosti te korekcija odstupanja rotacijom (Cheriet i sur., 2007, str. 34)

5.5.2. Obrubi skeniranih stranica

Druga česta pogreška pri skeniraju odnosi se na obrube skeniranih stranica. Prilikom skeniranja dokumenata manje površine u odnosu na površinu skenera, na dobivenoj slikovnoj datoteci mogu se pojaviti tamni obrubi. Osim toga, zbog nedovoljno precizno pozicioniranog dokumenta na površini skenera, skeniranoj slici teksta može nedostajati ili biti znatno sužen jedan ili više obruba (margina) (slika 11). Obje spomenute pogreške otežavaju postupak optičkog prepoznavanja znakova. OCR softver tamne rubove može pogrešno prepoznati kao dodatne znakove, posebno ako kod njih postoji varijacija u obliku ili nijansi boje. Stoga se u oba slučaja preporuča prilagodba obruba pomoću nekog od postojećih programa za obradu slika. U prvom slučaju tamne obrube je potrebno izrezati (engl. crop), dok je u drugom slučaju potrebno dodati mali obrub („Improving the quality of the output“, bez dat.).



Slika 11. Vidljivi tamni obrubi i dio susjedne stranice nakon skeniranja i izgled stranice nakon njihovog uklanjanja (preuzeto 26.05.2109. s https://github.com/tesseract-ocr/tesseract/wiki/ImproveQuality)

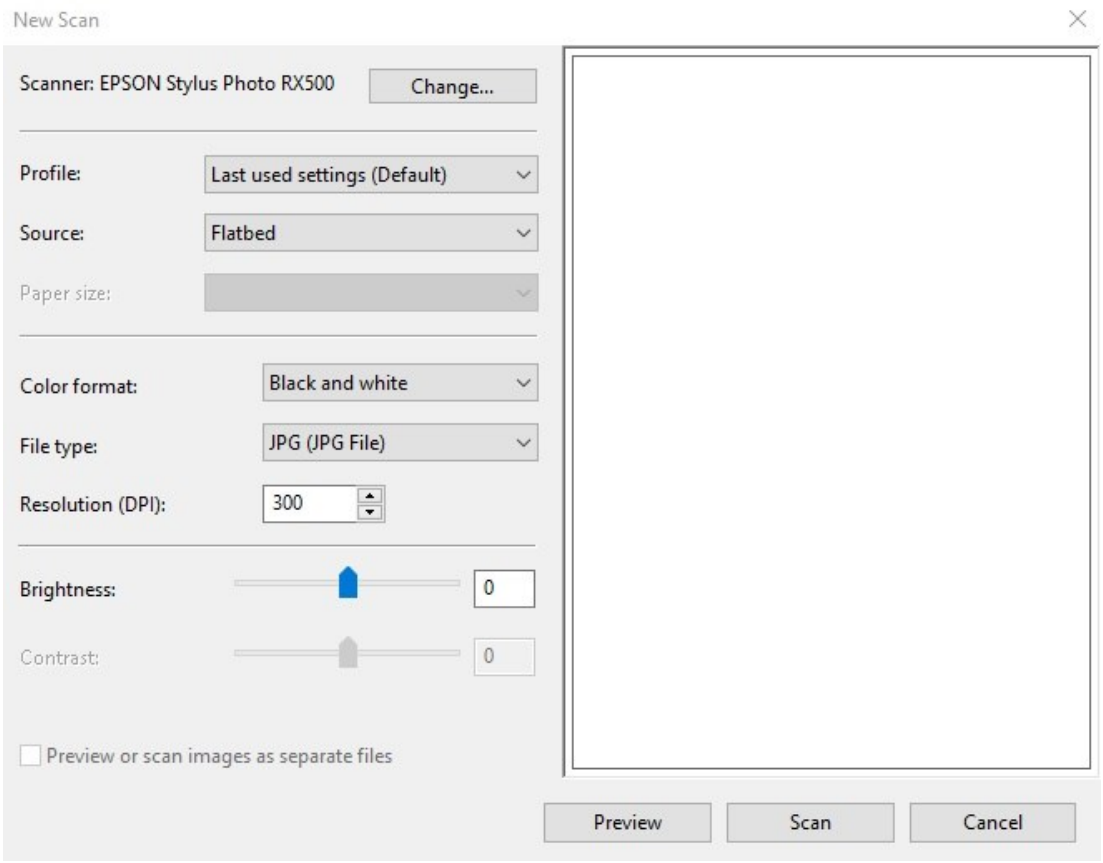
6. Predobrada ulaznih datoteka tijekom skeniranja

Nakon što je odabrano tekstualno gradivo za digitalizaciju, pristupa se samom postupku digitalizacije, za što se koriste ili skeneri ili digitalni fotoaparati. Time nastaje digitalna slika teksta kojeg se dodatno obrađuje OCR programom s ciljem dobivanja obradivog teksta (Stančić, 2009, str. 56). U ovom je radu za skeniranje odabranog predloška korišten *Epson Stylus Photo RX500*, višefunkcijski uređaj namijenjen kućnoj upotrebi koji objedinjuje funkcije pisača, skenera i fotokopirnog uređaja (slika 12). Od tehničkih specifikacija uređaja valja spomenuti 48-bitnu dubinu boje skenera, optičku razlučivost od 2400 x 4800 DPI i interpoliranu razlučivost od 9600 x 9600 DPI (Epson, 2003).

Slika 12. Epson Stylus Photo RX 500 (preuzeto 28.04.2019. s https://www.epson.de/en/products/printers/inkjet-printers/for-home/epson-stylus-photo-rx500)

Njegov softver za skeniranje nudi sljedeće postavke (prikazane na slici 13):

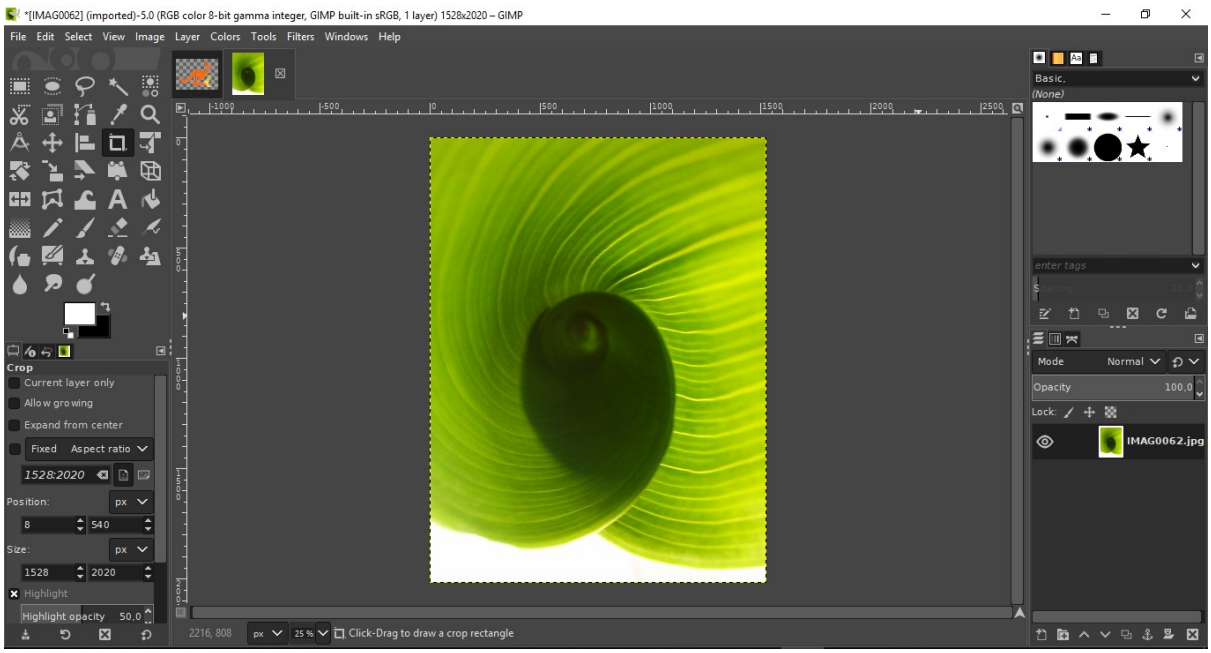
- bitna dubina boje: crno-bijela, skala sive boje, u boji
- format slikovne datoteke: JPG, TIFF, PNG, BMP
- rezolucija (razlučivost) u točkama po inču (engl. DPI – dot per inch): unaprijed zadana vrijednost jest 300 DPI
- podešavanje razine sjajnosti (engl. brightness)
- podešavanje razine kontrasta (engl. contrast).



Slika 13. Slika zaslona - program za skeniranje uređaja Epson Stylus Photo RX500

7. Predobrada ulaznih datoteka GIMP softverom za obradu slika

GIMP (skraćeno od GNU Image Manipulation Program) je softver otvorenog koda koji služi za stvaranje i obradu rasterske grafike (slika 14). Neke od funkcionalnosti koje nudi su retuširanje i uređivanje slika, slobodno crtanje, pretvaranje slika u različite formate. Osim osnovnih radnji, dostupne su i kompleksnije mogućnosti za naprednije korisnike zahvaljujući opciji nadogradnje softvera željenim proširenjima (engl. extension) i dodacima (engl. plug-in, vrsta softverskog dodatka izrađenih od drugih proizvođača koje glavni GIMP program izvodi i kontrolira i time mu omogućuje nove funkcije). Prvu inačicu GIMP-a objavili su Spencer Kimball i Peter Mattis 1996. godine, dok je zadnja stabilna inačica softvera GIMP 2.10.10 objavljena u travnju 2019. godine i ona je korištena u ovom radu. Danas su verzije ovog softvera dostupne za operativne sustave Linux, Mac OS X i Microsoft Windows, a zaštićen je licencom GPLv3+ („About GIMP“, bez dat.; „GIMP“, bez dat.).



Slika 14. Slika zaslona - softver za obradu slika GIMP 2.10.10

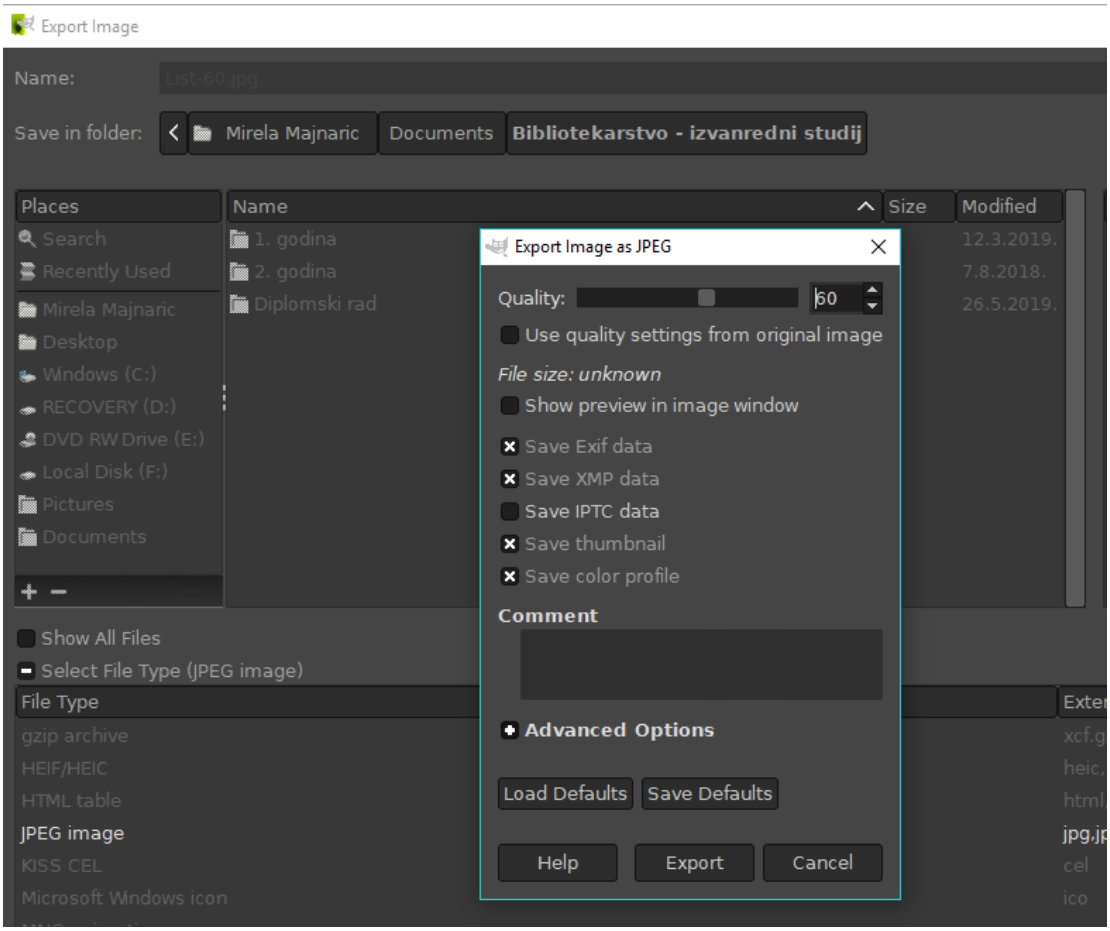
GIMP softverom moguće je izvršiti potrebne korekcije za uspješnije optičko prepoznavanje znakova, a pritom je uglavnom dostatno korištenje njegovih osnovnih mogućnosti. Tako je u GIMP-u moguće izvršiti prilagodbe sljedećih elemenata slikovne datoteke koje mogu utjecati na točnost optičkog prepoznavanja znakova odabranim OCR programom:

- stupanj kompresije slike u JPG formatu,
- sjajnost i kontrast,

- binarizacija slike,
- zakrenutost stranice,
- obrubi skeniranih stranica.

7.1. Stupanj kompresije slike u JPG formatu

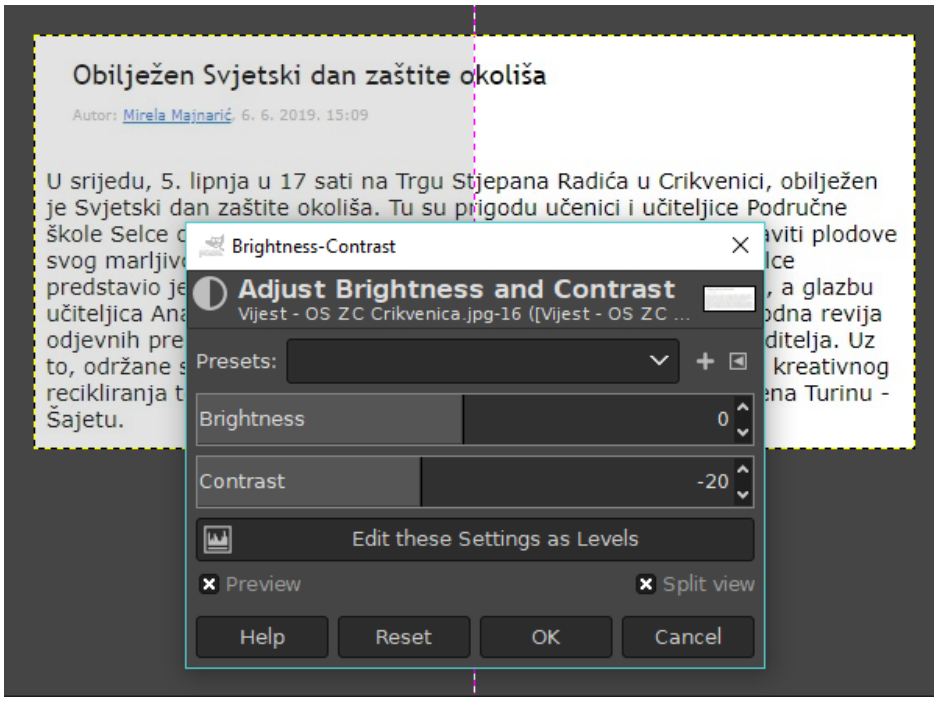
Kao što je u petom poglavlju bilo spomenuto pri opisu formata slikovnih datoteka, upotrebom JPG formata sadržaj slike se komprimira uz gubitak kvalitete pri čemu je moguće izabrati željeni stupanj kompresije. Pritom i kod visokog stupnja kompresije kvaliteta slike može biti sasvim zadovoljavajuća. Za promjenu stupnja kompresije slike u JPG formatu pomoću GIMP-a, potrebno ju je otvoriti u programu i iz padajućeg izbornika *File* (datoteka) odabrati opciju *Export As...* (Izvezi kao...). U dijaloškom okviru *Export Image* (izvezi sliku) moguće je preimenovati sliku (*Name*), odabrati mapu u kojoj će se ona pohraniti (*Save in folder*) te izabrati vrstu datoteke (*Select File Type*). Nakon izbora JPG formata, otvara se dijaloški okvir *Export Image as JPEG* (Izvezi sliku kao JPEG) u kojem se opcijom *Quality* (kvaliteta) podešava željeni stupanj kompresije (slika 15).



Slika 15. Slika zaslona – podešavanje stupnja kompresije JPG formata u programu GIMP

7.2. Sjajnost i kontrast

Izmjene postavki sjajnosti i kontrasta u programu GIMP vrše se otvaranjem padajućeg izbornika *Colors* (Boje) u kojem se nalazi opcija *Brightness-Contrast* (sjajnost-kontrast). Korisnik samostalno određuje vrijednosti svjetline i kontrasta koje se mogu pregledati prije prihvaćanja izmjena uključivanjem opcije *Preview* (pregled). Osim toga, ponuđena je i opcija *Split view* (podijeljeni prikaz) kojom se uređivana slika dijeli na dva dijela te se na jednome prikazuju izmjene, dok drugi ostaje u izvornom obliku (slika 16).

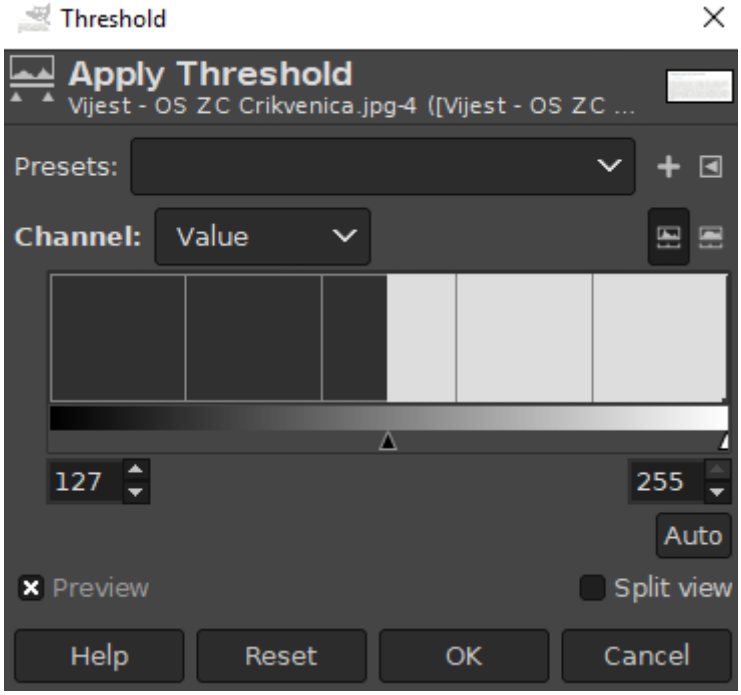


Slika 16. Slika zaslona – mijenjanje postavki sjajnosti i kontrasta u programu GIMP

7.3. Binarizacija slike

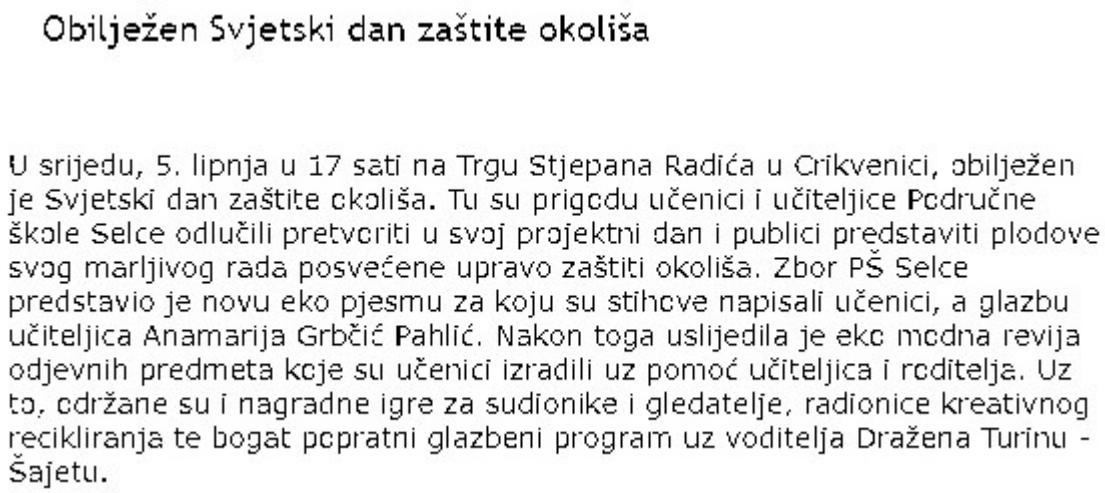
Konverzija višebojne slike u binarnu, odnosno crno-bijelu sliku naziva se binarizacija (engl. binarization, thresholding), a može se provesti ili tijekom procesa skeniranja, ili naknadno, u odabranom programu za obradu slike. Tako i program za obradu slike GIMP nudi mogućnost binarizacije slike u padajućem izborniku *Colors* (boje) pod nazivom *Threshold* (vrijednost praga, granična vrijednost). Program automatski određuje vrijednost praga te ga primjenjuje na odabranu sliku. Kao i kod postavki sjajnosti i kontrasta, rezultati se mogu pregledati prije prihvaćanja izmjena upotrebom opcija *Preview* (pregled) ili *Split view* (podijeljeni prikaz). Ukoliko korisnik nije zadovoljan rezultatima koje predlaže softver upotrebom opcije *Auto*

kojom se automatski određuje optimalni prag binarizacije, sam može podesiti postavke pomicanjem klizača za vrijednosti crne i bijele boje na linearnom histogramu (slika 17).

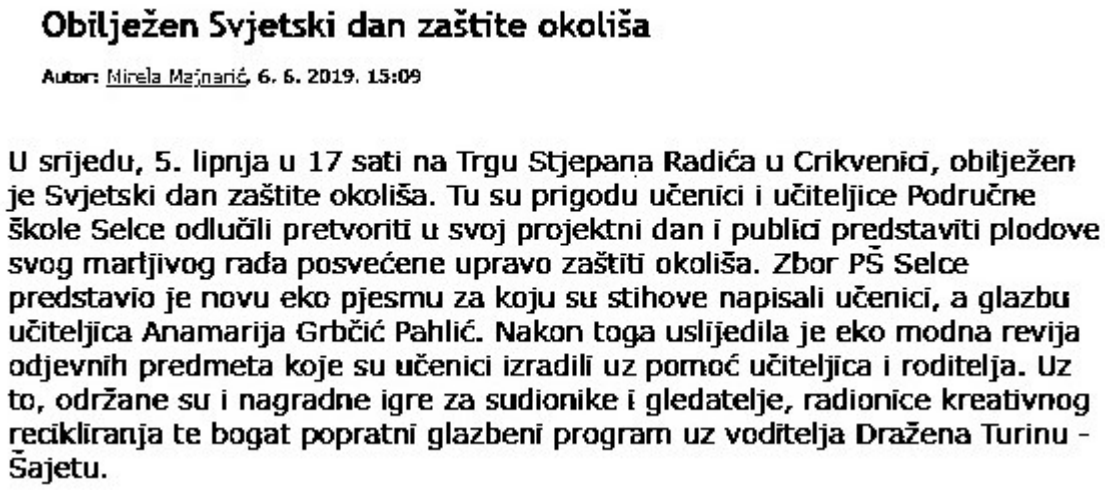


Slika 17. Slika zaslona – određivanje vrijednosti praga u programu GIMP

Na primjeru koji slijedi prikazana je razlika u rezultatima automatski određene vrijednosti praga (slika 18) te korisnički postavljene vrijednosti praga (slika 19). Glavnina teksta članka odabranog za primjer napisana je crnom bojom koja se dobro ističe u odnosu na bijelu pozadinu, no ime autora i vrijeme objave napisani su fontom manje veličine te sivom, odnosno plavom bojom zbog čega su slabije uočljivi na svijetloj pozadini. Ukoliko se korisnik opredijeli za automatski određenu vrijednost praga (globalni prag), na binariziranoj slici ti će podatci nestati zbog nedovoljnog kontrasta teksta i pozadine, dok će naslov i sadržaj članka biti zadržani. Istovremeno se na tekstu sadržaja javljaju neke druge promjene koje bi mogle otežati proces optičkog prepoznavanja znakova. Radi se o zadebljanju pojedinih slova, primjerice velikog slova P, dok su zakrivljene linije nekih malih slova razlomljene, primjerice slovo „o“ nalikuje slovu „c“. Samostalnim određivanjem vrijednosti praga binarizacije moguće je vrijednosti podesiti da i svjetlija slova postanu vidljiva i čitljiva na bijeloj pozadini. Problem je što se istovremeno i sva ostala slova podebljavaju pa se dva ili više susjednih slova mogu spojiti, a nakon određene vrijednosti pojavljuje se i šum na pozadini u obliku nasumično raspoređenih crnih točkica. Oba spomenuta problema mogla bi nepovoljno utjecati na uspješno optičko prepoznavanje znakova.



Slika 18. Slika binarizirana primjenom globalnog praga u programu GIMP

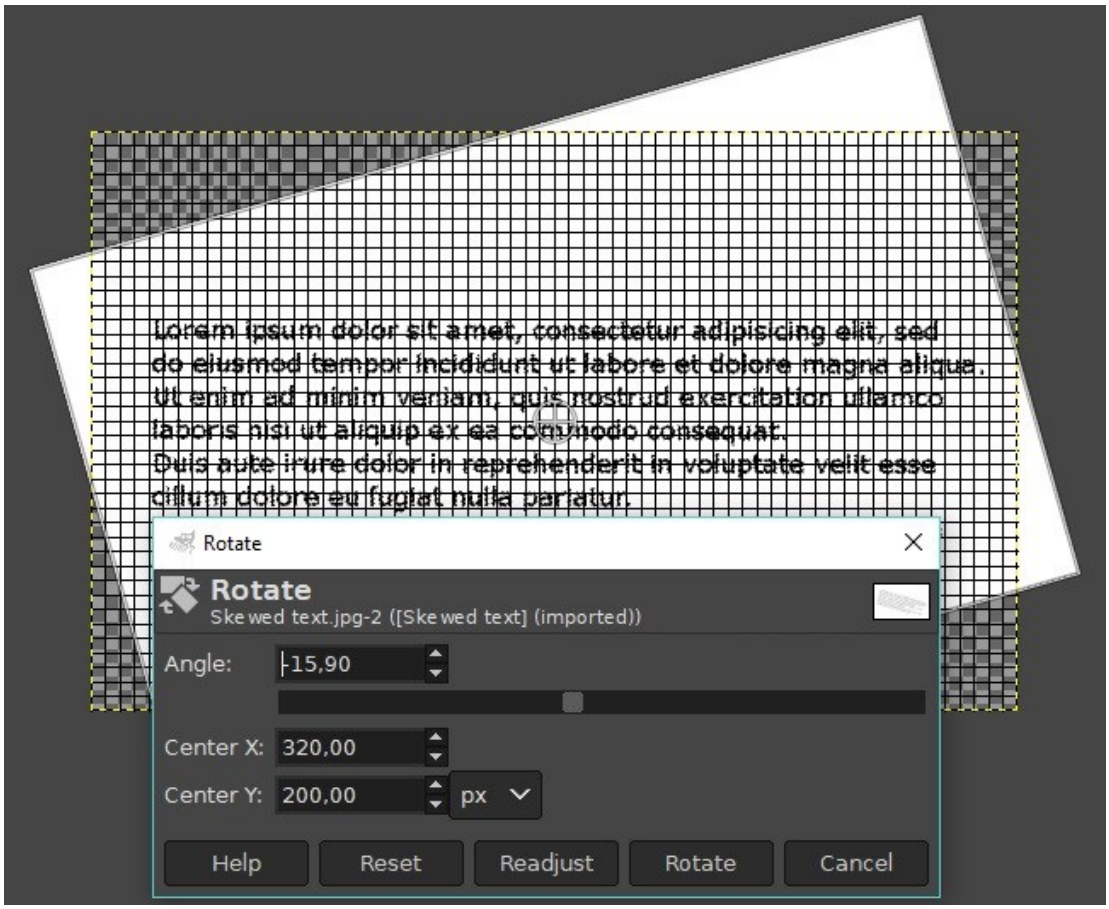


Slika 19. Slika binarizirana samostalnim određivanjem vrijednosti praga u programu GIMP

7.4. Zakrenutost i obrubi stranice

U petom poglavlju istaknuta je važnost horizontalne poravnatosti linija teksta s x-osi stranice radi uspješnijeg optičkog prepoznavanje znakova. Ponekad se pri postupku skeniranja dogodi da linije teksta nisu savršeno poravnate s x-osi, a što je to odstupanje veće, to je i vjerojatnost pogreške pri optičkom prepoznavanju znakova veća. Stoga je jedan od preporučenih koraka predobrade slikovne datoteke upravo ispravljanje zakrenutosti teksta u odnosu na x-os, posebno ako je odstupanje veće od tri stupnja. Ta se korekcija u programu za obradu slike GIMP vrši izborom padajućeg izbornika *Tools* (alati), te se u daljnjem podizborniku izabire opcija *Transform Tools* (alati za transformaciju), *Rotate* (rotiraj, zakreni). Otvara se dijaloški okvir *Rotate* u kojem se odabire željeni kut rotacije kojega je moguće vrlo precizno podesiti,

na stotinke kuta. Nudi se i mogućnost izmjene središta rotacije slike, no to za svrhu ovoga rada nije potrebno. Uključivanje opcije *Show Grid* (prikaži rešetku) iz padajućeg izbornika *View* (pregled) pomaže pri određivanju i provjeri poravnatosti teksta s x-osi stranice. Potvrdom željenog kuta rotacije slika se rotira, nakon čega se obično koristi alat *Crop* (izreži) za izrezivanje suvišnog prostora na slikama. Sljedeći primjer (slika 20) ilustrira odlomak teksta zakrenut za više od 3 stupnja u odnosu na x-os stranice s prikazanom rešetkom radi lakšeg određivanja kuta poravnanja te dijaloški okvir *Rotate* pomoću kojega se ispravlja navedeni nedostatak.



Slika 20. Slika zaslona - ispravljanje zakrenutosti teksta alatom *Rotate* u programu GIMP

Nakon poravnavanja ostaju suvišni obrubi koje se obično uklanja pomoću ikone *Crop Tool* ponuđene u okviru s alatima (engl. *Toolbox*) s lijeve strane prozora ili označavanjem dijela slike koji želimo izrezati te izborom opcije *Crop to Selection* (izreži označeno) iz padajućeg izbornika *Image* (slika).

8. Istraživanje

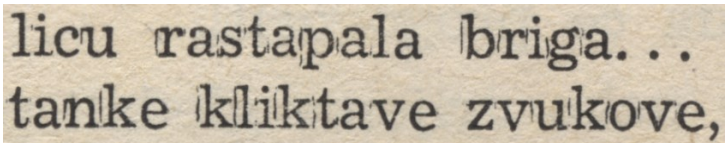
Praktični dio ovog diplomskog rada odnosi se na ispitivanje utjecaja predobrade ulaznih datoteka na točnost optičkog prepoznavanja znakova. U prvom dijelu istraživanja bit će ispitan utjecaj JPG i TIFF formata slikovne datoteke, bitne dubine boje i kvalitete slike iskazane u točkama po inču na točnost optičkog prepoznavanja znakova. Drugi dio istraživanja usredotočit će se na JPG format različitih stupnjeva kompresije i koliko kompresija utječe na točnost optičkog prepoznavanja znakova. Treći dio istraživanja odnosi se na binarizaciju slika u sivoj skali i u boji upotrebom programa GIMP, kao korak koji prethodi optičkom prepoznavanju znakova. Točnost dobivenih rezultata usporedit će se s rezultatima bez provedenog međukoraka binarizacije.

8.1. Uzorak

Kao što je spomenuto u petom poglavlju, kvaliteta izvornika koji se digitalizira jedan je od čimbenika koji znatno utječu na točnost optičkog prepoznavanja znakova. Visok postotak točnosti optičkog prepoznavanja znakova postiže se uz jasno otisnut tekst uzorka i dobar kontrast tamnog teksta na svijetloj pozadini. Takve je rezultate moguće očekivati kod novije otisnutih tekstova zahvaljujući inovacijama na polju tiska i kvaliteti korištenog papira te takvi uzorci ne bi trebali stvarati problem OCR softveru. No što je građa koja se digitalizira starija, to je veći izazov postići višu razinu točnosti, jer su na takvoj građi nerijetka oštećenja podloge ili otiska. Stoga je za ovo istraživanje odabran uzorak koji sadrži elemente koji bi mogli otežati optičko prepoznavanje znakova kako bi se utvrdilo u kojoj mjeri predobrada takvog uzorka utječe na točnost dobivenih rezultata.

Za uzorak je odabrana prva stranica priče „*Gost*“ iz knjige „*Pripovijetke*“ Vjekoslava Kaleba u izdanju Matice hrvatske iz 1963. godine (slika 21). Ta se stranica skenirana u u TIFF formatu u boji razlučivošću od 600 DPI nalazi u prilogu 1 ovome radu. Što se same knjige tiče, radi se o meko ukoričenoj knjizi lijepljenog uveza kod koje je vidljiva deterioracija podloge, posebno što se tiče promjene boje papira, što je razumljivo obzirom na godinu izdavanja. Osim požutjelog papira, i kod samog otiska teksta uočljivi su određeni nedostatci koji bi mogli dovesti do neispravnog optičkog prepoznavanja znakova. Zbog tehnike tiska korištene pri izradi ove knjige između mnogih znakova iste riječi vidljive su otisnute linije koje ne bi smjele biti prisutne, već su nastale kao rezultat slabe kvalitete tiska. Na primjeru prikazanom na slici 22 izdvojen je dio teksta skeniran u boji rezolucijom od 600 DPI na kojemu su jasno vidljive manjkavosti tiska kod svake od prikazanih riječi, a posebno su uočljive kod riječi „kliktave“.

Slika 21. Prednja korica knjige „*Pripovijetke*“ Vjekoslava Kaleba



Slika 22. Nedostaci tiska vidljivi na odabranom uzorku

8.2. Korištena softverska rješenja

Za optičko prepoznavanje znakova u istraživanju je korišten softverski alat *ABBYY FineReader 12*, dok je za analizu točnosti prepoznatih znakova u uzorku primijenjeni *ISRI* analitički alati. U nastavku će oba alata biti ukratko opisana, kao i način na koji su korišteni.

8.2.1. ABBYY FineReader 12

ABBYY FineReader je softver za optičko prepoznavanje znakova koji omogućuje pretvorbu slikovnih datoteka u brojne druge formate uz mogućnost njihovog uređivanja i pretraživanja. Prva je inačica softvera objavljena 1993. godine, dok je posljednja stabilna inačica *ABBYY FineReader 14* izašla u siječnju 2017. godine. Osim optičkog prepoznavanja znakova,

najnovija inačica softvera nudi i razne funkcije za rad s dokumentima u PDF formatu, kao i mogućnost usporedbe dokumenata u različitim formatima s ciljem uočavanja izmjena sadržaja. Softver je na tržištu dostupan u Standard verziji namijenjenoj privatnim korisnicima i manjim poduzećima te Corporate verziji namijenjenoj srednjim i većim poduzećima i organizacijama. Do sada su razvijene verzije softvera za Windows, Mac i Linux operativne sustave („ABBYY FineReader celebrates 25 years on the market“, 2018.).

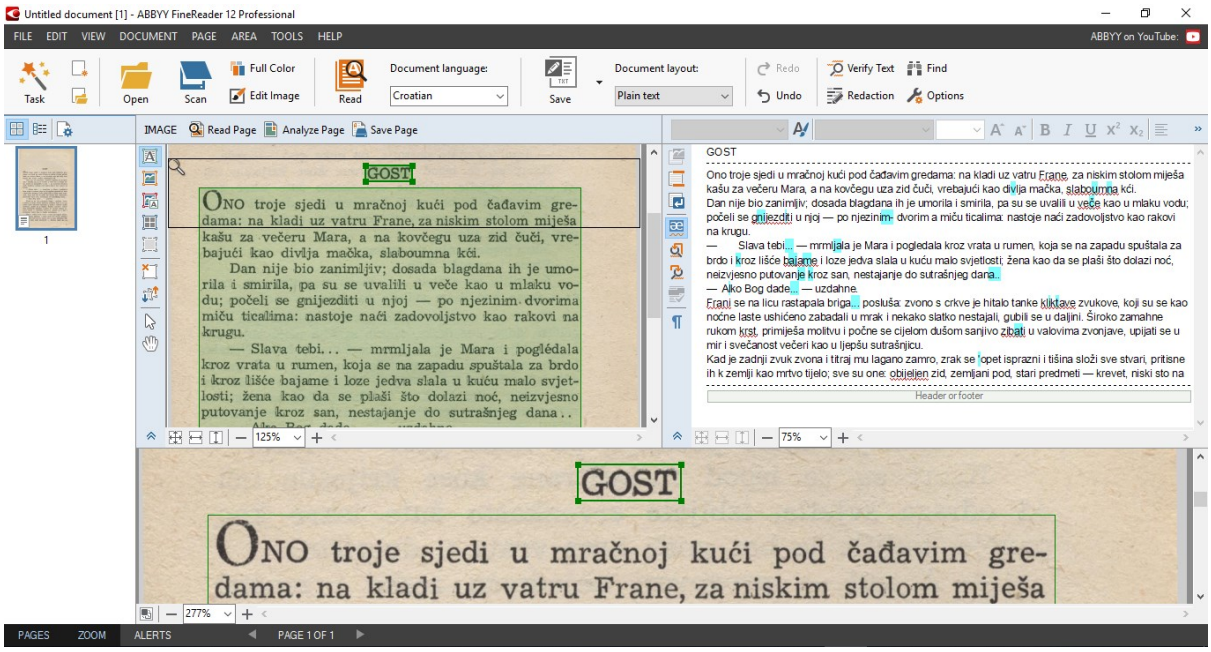
U istraživanju je korištena inačica *ABBYY FineReader 12* koja podržava prepoznavanje do 190 jezika, od čega je za 48 jezika dostupan i rječnik, uključujući i rječnik hrvatskog jezika (Dvornikova, bez dat.). U odnosu na prethodne inačice, FineReader 12 se izdvaja time što korisnicima omogućuje rad na dokumentu dok se istovremeno u pozadini provodi njegova OCR obrada. Time se ostvaruje znatna ušteda vremena pri optičkom prepoznavanju opsežnijih dokumenata („Novi ABBYY FineReader 12 uvelike ubrzava raspoznavanje i citiranje teksta iz skenova i fotografija“, 2014.).

Kako je svrha istraživanja ispitivanje utjecaja predobrade ulaznih datoteka na točnost optičkog prepoznavanja znakova, korištene su samo temeljne funkcije odabranog OCR softvera. U nastavku će biti opisane opcije iz alatne trake koje su bile korištene prilikom OCR-a datoteka te način prikaza i označavanja dijelova teksta tijekom čitanja i analize stranice (slika 23).

Pomoću opcije *Open* (otvori) u program se učitavaju slikovne datoteke, odnosno skenirani dokumenti nad kojima se želi provesti optičko prepoznavanje znakova, dok se opcijom *Read* (pročitaj) pokreće proces analize i optičkog prepoznavanja znakova učitane datoteke. *Document language* (jezik dokumenta) omogućuje izbor jezika na kojem je dokument pisan kako bi se osiguralo ispravno prepoznavanje znakova karakterističnih za taj jezik, kao i korištenje rječnika odabranog jezika, ukoliko je ta opcija dostupna. U ovom je slučaju kao jezik dokumenta odabran hrvatski jezik, kako bi prilikom optičkog prepoznavanja znakova bili prepoznati dijakritički znakovi korišteni u hrvatskom jeziku. Opcija *Save* (spremi) daje na izbor formate za pohranu pročitanog teksta, ovisno o potrebama korisnika i daljnjoj namjeni pročitanog dokumenta. Uzorci korišteni u istraživanju su nakon optičkog prepoznavanja znakova spremljeni u TXT formatu kako bi se zasebnim alatom mogla provjeriti točnost prepoznatih znakova svakoga od uzoraka.

U krajnje lijevom dijelu ekrana ispod alatne trake prikazana je umanjena slika (engl. thumbnail) uzorka nad kojim se provodi optičko prepoznavanje teksta. S desne strane je uvećani prikaz istog tog uzorka uz ponuđene opcije označavanja prepoznavanja teksta na čitavoj ili dijelu

stranice, kao i označavanja slike, pozadinske slike ili tablice prikazane na stranici. Nakon izvršenog optičkog prepoznavanja znakova zelenom su bojom označeni blokovi teksta koje je softver prilikom čitanja prepoznao i izdvojio. Na samom dnu stranice prikazan je uzorak uz mogućnost dodatnog povećavanja (engl. zoom) radi detaljnijeg uvida u sadržaj stranice. Krajnje desno se prikazuje prepoznati tekst nakon postupka optičkog prepoznavanja gdje korisnik bira želi li zadržati slike, zaglavlja i podnožja dokumenta ili samo tekst. Moguće je i uključiti opciju kojom se ističu znakovi s niskim stupnjem pouzdanosti prepoznavanja te podcrtavaju riječi koje nisu sadržane u rječniku toga jezika. Zahvaljujući toj mogućnosti korisnik prije same pohrane dokumenta može brže i lakše samostalno izvršiti korekciju pogrešaka bez potrebe čitanja cjelokupnog teksta. U sklopu ovog istraživanja nije zadržano podnožje dokumenta, odnosno broj stranice, dok su svi ostali prepoznati elementi ostavljeni bez izmjene zbog provjere točnosti prepoznavanja znakova.



Slika 23. Slika zaslona – ABBYY FineReader 12 Professional

8.2.2. ISRI analitički alati

ISRI analitički alati služe za provjeru točnosti teksta prepoznatog nekim OCR programom. Razvijeni su na Institutu za informacijske znanosti u Las Vegasu (engl. Information Science Research Institute, skraćeno ISRI) Sveučilišta u Nevadi gdje su tijekom 1990-ih godina provođena godišnja testiranja točnosti izvedbe OCR programa. U tu je svrhu nastao softverski paket *ISRI OCR Experimental Environment,* čiji su sastavni dio bili i *ISRI* analitički alati za procjenu točnosti pročitanih tekstova. Bili su napisani u programskom jeziku C, a namijenjeni

za upotrebu u operativnom sustavu Unix te su činili dio doktorskog rada Stephena V. Ricea. Iako od 1996. godine kreatori alata nisu izdali novije inačice, ovaj je alat i dalje iznimno koristan za procjenu novih OCR programa, pri čemu su najčešće u upotrebi alati „accuracy“ za procjenu točnosti na razini znaka i „wordacc“ za procjenu točnosti na razini riječi. Valja istaknuti da su ISRI analitički alati nedavno nadograđeni i unaprijeđeni za potrebe istraživanja ugroženih jezika pod nazivom *ocreval* i čija je najznačajnija osobina potpora za aktualni Unicode standard (Santos, 2019, str. 23-24). Unicode je univerzalni standard za kodiranje znakova korišten u tekstualnim datotekama, internetskim stranicama i drugim vrstama dokumenata, a nastao je da bi se njime opisali znakovi korišteni u svim svjetskim jezicima. Iako postoji više načina Unicode kodiranja, trenutno su u upotrebi najzastupljeniji UTF-8 i UTF-16 (Christensson, 2012). Zahvaljujući tom unapređenju, postupak analize može obuhvatiti više znakova, a samim time i više jezika. Ovaj je alat danas dostupan za preuzimanje na GitHub platformi kao softver otvorenog koda zaštićen licencom Apache 2.0 (https://github.com/eddieantonio/ocreval).

ISRI analitičke alate za procjenu OCR-a moguće je podijeliti u četiri kategorije koje sadrže sveukupno 19 programa. Te četiri skupine programa odnose se na ispitivanje točnosti prepoznavanja znakova, prepoznavanja riječi, ispravnosti prepoznavanja zona teksta te OCR stranih jezika (Nartker, Rice i Lumos, 2005). Kako je istraživanje u sklopu ovog rada usredotočeno na točnost prepoznavanja znakova, ta će kategorija programa u nastavku biti pobliže opisana. Prije upotrebe bilo kojega od ISRI analitičkih alata korisnik mora stvoriti tekstualnu datoteku koja sadrži u potpunosti točan tekst (engl. *ground truth*) koji se zatim uspoređuje s tekstom kojeg je prepoznao OCR program. Tako se za provjeru točnosti prepoznavanja znakova koristi program pod nazivom *accuracy* (točnost) koji se pokreće sljedećom naredbom:

accurracy *correctfile generatedfile* [*accuracy_report*].

Correctfile predstavlja naziv datoteke koja sadrži točan tekst, *generatedfile* naziv datoteke s tekstom kojeg je prepoznao OCR softver, a *accuracy_report* jest naziv datoteke u koju će se pohraniti izvještaj s informacijama o točnosti prepoznavanja znakova (Rice i Nartker, 1996, str. 3).

8.3. Utjecaj izbora slikovnog formata, bitne dubine boje i razlučivosti na točnost OCR-a

Uzorak opisan u uvodnom dijelu ovoga poglavlja skeniran je u komprimiranim slikovnim formatima JPG i TIFF u crno-bijeloj, sivoj skali i u boji. Korištene su razlučivosti od 100, 200, 300, 400, 500 i 600 DPI. Pomoću internetske stranice http://fotoforensics.com/ za provjeru metapodataka JPG i PNG datoteka utvrđeno je da je stupanj kompresije JPG datoteka prilikom skeniranja automatski postavljen na 75% za što su korištene diskretna kosinusna transformacija (engl. DCT – discrete cosine transform) i Huffmanovo kodiranje. TIFF datoteke komprimirane su bez gubitka LZW (Lempel-Ziv-Welch) algoritmom, što je provjereno upotrebom internetske stranice https://www.get-metadata.com/ za utvrđivanje metapodataka TIFF formata, ali i brojnih drugih tekstualnih, slikovnih i video formata. Točan tekst uzorka prepisan je i pohranjen kao TXT datoteka, a kao način kodiranja odabran je UTF-8. Ta je datoteka korištena u *ISRI* analitičkim alatima za usporedbu s tekstovima prepoznatima pomoću *ABBYY FineReadera* kako bi se dobili izvještaji s detaljnim informacijama o točnosti prepoznavanja znakova za svaku od datoteka pročitanih OCR programom.

Tablice 1 i 2 prikazuju dobivene rezultate, odnosno postotke točno prepoznatih znakova za TIFF i JPG formate, ovisno o bitnoj dubini boje i razlučivosti. Osim toga, navedene su i veličine izvornih slikovnih datoteka, radi usporedbe isplativosti skeniranja pri većoj razlučivosti i postotka točno prepoznatih znakova.

Tablica 1. Točnost OCR-a za TIFF datoteke uz različite bitne dubine boje i razlučivosti

Bitna dubina boje	Razlučivost	Veličina slikovne datoteke (MB)	Točnost OCR-a	Prosječna točnost OCR-a
Crno-bijela	100	0,01	70,62%	93,86%
	200	0,03	97,98%	
	300	0,05	98,53%	
	400	0,08	98,60%	
	500	0,11	98,99%	
	600	0,15	98,45%	
Siva skala	100	0,22	98,68%	98,73%
	200	0,83	98,76%	
	300	1,92	98,68%	
	400	3,22	98,84%	
	500	5,07	98,68%	
	600	6,85	98,76%	
U boji	100	0,60	98,68%	98,73%
	200	**2,12**	**99,07%**	
	300	5,32	98,60%	
	400	8,34	98,68%	
	500	13,20	98,68%	
	600	17,70	98,68%	

Tablica 2. Točnost OCR-a za JPG datoteke uz različite bitne dubine boje i razlučivosti

Bitna dubina boje	Razlučivost	Veličina slikovne datoteke (MB)	Točnost OCR-a	Prosječna točnost OCR-a
Crno-bijela	100	0,06	69,69%	93,77%
	200	0,20	97,91%	
	300	0,35	98,84%	
	400	0,56	98,84%	
	500	0,78	98,68%	
	600	0,99	98,76%	
Siva skala	100	0,06	98,84%	98,71%
	200	0,18	98,68%	
	300	0,42	98,60%	
	400	0,70	98,68%	
	500	1,06	98,60%	
	600	1,45	98,84%	
U boji	100	0,06	98,22%	98,58%
	200	0,19	98,60%	
	300	0,43	98,60%	
	400	0,72	98,84%	
	500	1,09	98,84%	
	600	1,48	98,37%	

Crno-bijele datoteke skenirane razlučivošću od 100 DPI imaju relativno nisku točnost OCR-a s oko 70% ispravno prepoznatih znakova, što je i razumljivo zbog niske razlučivosti, odnosno lošije kvalitete predloška. Pritom i sam OCR softver korisnika upozorava porukom kako bi za bolje rezultate bilo preporučljivo sliku skenirati većom razlučivošću. Iz ostalih dobivenih rezultata vidljivo je da su dobiveni postotci točnosti optičkog prepoznavanja znakova vrlo ujednačeni, neovisno o slikovnom formatu, izabranoj bitnoj dubini boje i razlučivosti. U tablicama su ružičastom bojom pozadine označeni najniži dobiveni postotci ispravno prepoznatih znakova, dok su zelenom bojom označeni najviši dobiveni postotci prema kategorijama bitne dubine boje. Podebljanim slovima istaknut je najviši dobiveni postotak točnosti OCR-a koji iznosi 99,07%, a dobiven je za TIFF datoteku u boji skeniranu razlučivošću od 200 DPI. Prema dobivenim rezultatima, može se izvesti zaključak da izravno skeniranje predloška u crno-bijele tonove nižom razlučivošću u pravilu daje nešto niže rezultate točnosti OCR-a. Nasuprot tome, uočeno je da povećanje razlučivosti skeniranja iznad 400 DPI nužno ne rezultira većom točnošću OCR-a. Izračunate prosječne vrijednosti točnosti OCR-a ukazuju i na to da TIFF datoteke u sivoj skali i u boji daju najbolju točnost OCR-a (98,73%), a JPG datoteke u sivoj skali za njima neznatno zaostaju (98,71%).

8.4. Utjecaj stupnja kompresije JPG formata na točnost OCR-a

Prateći postupak opisan u poglavlju 7.1., TIFF datoteke različitih bitnih dubina boje i vrijednosti razlučivosti 200, 400 i 600 DPI su pomoću programa GIMP pretvorene u JPG format i pohranjene u različitim stupnjevima kompresije. Kako su skenirane JPG datoteke komprimirane na 75%, iz TIFF datoteka izvedene su ostale JPG datoteke komprimirane na 100% (bez kompresije), 50% i 25%, nakon čega je upotrijebljen OCR program. Prepoznati tekstovi su uspoređeni s točnim tekstom, a rezultati točnosti OCR-a su izloženi u tablici 3.

Tablica 3. Točnost OCR-a za JPG datoteke uz različite stupnjeve kompresije

Stupanj kompresije	Bitna dubina boje	Razlučivost	Veličina slikovne datoteke (MB)	Točnost OCR-a	Prosječna točnost OCR-a
100%	Crno-bijela	200	0,36	98,45%	98,75%
		400	**0,98**	**98,91%**	
		600	1,73	98,68%	
	Siva skala	200	0,72	98,68%	
		400	2,73	98,84%	
		600	5,63	98,76%	
	U boji	**200**	**1,33**	**98,91%**	
		400	5,22	98,68%	
		600	11,00	98,84%	
75%	Crno-bijela	200	0,20	97,91%	98,61%
		400	0,56	98,84%	
		600	0,99	98,76%	
	Siva skala	200	0,18	98,68%	
		400	0,70	98,68%	
		600	1,45	98,84%	
	U boji	200	0,19	98,60%	
		400	0,72	98,84%	
		600	1,48	98,37%	
50%	Crno-bijela	200	0,15	97,98%	98,58%
		400	0,40	98,60%	
		600	0,72	98,37%	
	Siva skala	200	0,13	98,68%	
		400	0,45	98,60%	
		600	**0,97**	**98,91%**	
	U boji	200	0,13	98,68%	
		400	0,49	98,68%	
		600	1,01	98,76%	
25%	Crno-bijela	200	0,11	97,91%	98,59%
		400	0,29	98,84%	
		600	0,54	98,53%	
	Siva skala	200	0,09	98,37%	
		400	**0,28**	**98,91%**	
		600	0,59	98,76%	
	U boji	200	0,09	98,84%	
		400	0,29	98,60%	
		600	0,61	98,53%	

U svakoj su kategoriji prema stupnju kompresije i bitnoj dubini boje zelenom pozadinskom bojom označene najviše dobivene vrijednosti točnosti OCR-a. Iako niti jedna od njih ne prelazi 99%, valja izdvojiti i da su svega tri vrijednosti ispod 98%, a radi se o crno-bijelim datotekama skeniranima razlučivošću od 200 DPI i stupnjevima kompresije od 25 do 75%. U zadnjem su stupcu navedene izračunate prosječne vrijednosti točnosti OCR-a datoteka testiranih u ovom stadiju istraživanja. Najbolje su prosječne rezultate točnosti OCR-a očekivano dale najmanje komprimirane JPG datoteke, no relativno visok postotak prosječne točnosti OCR-a datoteka komprimiranih na 25% jest ponešto iznenađujući. Moguće obrazloženje tog rezultata jest činjenica da u ovu fazu testiranja nisu bile uključene datoteke skenirane na 100 DPI koje bi, kako je pokazala prethodna faza ispitivanja, vjerojatno imale niži postotak točnosti OCR-a te bi u konačnici snizile sve prosječne vrijednosti točnosti OCR-a. Također, postoji mogućnost da veći stupanj kompresije slikovnih datoteka koje sadrže informacije u tekstualnom obliku manje narušava kvalitetu i prepoznatljivost znakova no što bi to vjerojatno bio slučaj za fotografije ili crteže. Uz to, kvalitetan OCR softver uspješno će kompenzirati mnoge potencijalne nedostatke kvalitete slikovne datoteke nad kojom se provodi optičko prepoznavanje znakova što će rezultirati uspješno prepoznatim tekstovima s vrlo malo pogrešaka.

8.5. Utjecaj naknadne binarizacije slikovnih datoteka na točnost OCR-a

Za ovaj segment testiranja točnosti optičkog prepoznavanja znakova odabrane su slikovne datoteke u sivoj skali i u boji s dobivenim najvišim postotcima točnosti OCR-a u prethodnim fazama testiranja. Cilj je bio ispitati hoće li naknadna binarizacija slikovnih datoteka poboljšati točnost OCR-a. Kao granična vrijednost točnosti OCR-a odabran je postotak od 98,68%. Te su datoteke pomoću postupka binarizacije slike opisanog u poglavlju 7.3. naknadno binarizirane programom GIMP uz korištenje opcije automatskog određivanja optimalnog praga binarizacije. Za TIFF datoteke zadržana je LZW kompresija, dok je za JPG datoteke kod njihovog izvoza i spremanja uključena opcija zadržavanja postavki kvalitete izvorne slike. Binarizaciji su podvrgnuti svi ispitani stupnjevi kompresije JPG datoteka, uključujući i 25% i 50%, zbog toga što je u prethodnom koraku ispitivanja unatoč visokom stupnju kompresije točnost OCR-a pojedinih datoteka bila iznad odabrane granične vrijednosti. Binarizirane slike dobivene na ovaj način učitane su u ABBYY FineReader, a točnost prepoznatih tekstova uspoređena je s točnošću prepoznatih tekstova izvornih slikovnih datoteka. Usporedba rezultata za TIFF i JPG formate prikazana je u tablicama 4 i 5.

Tablica 4. Usporedba točnosti OCR-a za izvorne i naknadno binarizirane TIFF datoteke

Izvorne slikovne datoteke				Izvedene binarizirane datoteke			
Format	**Boja i razlučivost**	**Veličina slikovne datoteke (MB)**	**Točnost OCR-a**	**Format**	**Boja i razlučivost**	**Veličina slikovne datoteke (MB)**	**Točnost OCR-a**
TIFF	Siva skala 100	0,22	98,68%	TIFF	Crno-bijela 100	0,15	96,51%
TIFF	Siva skala 200	0,83	98,76%	TIFF	Crno-bijela 200	0,17	97,83%
TIFF	Siva skala 300	1,92	98,68%	TIFF	Crno-bijela 300	0,20	98,68%
TIFF	Siva skala 400	3,22	98,84%	TIFF	Crno-bijela 400	0,24	98,60%
TIFF	Siva skala 500	5,07	98,68%	TIFF	Crno-bijela 500	0,28	98,60%
TIFF	Siva skala 600	6,85	98,76%	TIFF	Crno-bijela 600	0,32	98,37%
TIFF	U boji 100	0,60	98,68%	TIFF	Crno-bijela 100	0,16	96,59%
TIFF	U boji 200	2,12	99,07%	TIFF	Crno-bijela 200	0,20	97,36%
TIFF	U boji 400	8,34	98,68%	TIFF	Crno-bijela 400	0,32	98,68%
TIFF	U boji 500	13,20	98,68%	TIFF	Crno-bijela 500	0,40	98,60%
TIFF	U boji 600	17,70	98,68%	TIFF	Crno-bijela 600	0,47	**98,84%**

Tablica 5. Usporedba točnosti OCR-a za izvorne i naknadno binarizirane JPG datoteke

Izvorne slikovne datoteke				Izvedene binarizirane datoteke			
Format i kompresija	**Boja i razlučivost**	**Veličina slikovne datoteke (MB)**	**Točnost OCR-a**	**Format i kompresija**	**Boja i razlučivost**	**Veličina slikovne datoteke (MB)**	**Točnost OCR-a**
JPG, 100%	Siva skala 200	0,72	98,68%	JPG, 100%	Crno-bijela 200	0,36	97,75%
JPG, 100%	Siva skala 400	2,73	98,84%	JPG, 100%	Crno-bijela 400	1,00	98,68%
JPG, 100%	Siva skala 600	5,63	98,76%	JPG, 100%	Crno-bijela 600	1,77	98,76%
JPG, 100%	U boji 200	1,33	98,91%	JPG, 100%	Crno-bijela 200	0,36	97,60%
JPG, 100%	U boji 400	5,22	98,68%	JPG, 100%	Crno-bijela 400	1,00	**98,84%**
JPG, 100%	U boji 600	11,00	98,84%	JPG, 100%	Crno-bijela 600	1,77	98,76%
JPG, 75%	Siva skala 100	0,06	98,84%	JPG, 75%	Crno-bijela 100	0,08	95,50%
JPG, 75%	Siva skala 200	0,18	98,68%	JPG, 75%	Crno-bijela 200	0,18	98,45%

JPG, 75%	Siva skala 400	0,70	98,68%	JPG, 75%	Crno-bijela 400	0,49	98,53%
JPG, 75%	Siva skala 600	1,45	98,84%	JPG, 75%	Crno-bijela 600	0,87	98,68%
JPG, 75%	U boji 400	0,72	98,84%	JPG, 75%	Crno-bijela 400	0,48	98,68%
JPG, 75%	U boji 500	1,09	98,84%	JPG, 75%	Crno-bijela 500	0,66	98,76%
JPG, 50%	Siva skala 200	0,13	98,68%	JPG, 50%	Crno-bijela 200	0,15	98,14%
JPG, 50%	Siva skala 600	0,97	98,91%	JPG, 50%	Crno-bijela 600	0,72	98,68%
JPG, 50%	U boji 200	0,13	98,68%	JPG, 50%	Crno-bijela 200	0,15	97,91%
JPG, 50%	U boji 400	0,49	98,68%	JPG, 50%	Crno-bijela 400	0,39	98,68%
JPG, 50%	U boji 600	1,01	98,76%	JPG, 50%	Crno-bijela 600	0,72	**98,91%**
JPG, 25%	Siva skala 400	0,28	98,91%	JPG, 25%	Crno-bijela 400	0,29	98,53%
JPG, 25%	Siva skala 600	0,59	98,76%	JPG, 25%	Crno-bijela 600	0,53	**98,91%**
JPG, 25%	U boji 200	0,09	98,84%	JPG, 25%	Crno-bijela 200	0,11	97,36%

Na trideset jednom uzorku provedena je naknadna binarizacija te je ponovljeno optičko prepoznavanje znakova i testiranje točnosti OCR-a. Uvidom u rezultate testiranja OCR-a uočeno je da su samo četiri datoteke imale viši postotak točnosti OCR-a po provedenom postupku. Ti su primjeri u tablicama označeni zelenom bojom pozadine, dok su dobiveni viši postotci točnosti istaknuti podebljanim slovima. Kao jedinu poveznicu između dobivenih rezultata moguće je navesti razlučivost skeniranja od 400 DPI ili višu, no iz rezultata je istovremeno vidljivo da viša razlučivost skeniranja za većinu preostalih testiranih datoteka nije doprinijela poboljšanju točnosti OCR-a. Osim toga, četiri binarizirane datoteke dale su istovjetne postotke točnosti OCR-a u odnosu na točnost OCR-a datoteka iz kojih su one izvedene, što je u tablici označeno sivom bojom pozadine. Kako su preostala dvadeset tri testirana uzorka imala niži postotak točnosti OCR-a od prvotnoga, može se zaključiti da binarizacija slikovnih datoteka programima za obradu slika prije OCR-a uglavnom neće imati pozitivan učinak na poboljšanje točnosti OCR-a te je stoga taj korak bolje izostaviti i pri optičkom prepoznavanju znakova koristiti datoteke u sivoj skali ili u boji.

8.6. Ostala zapažanja tijekom istraživanja

Iako nisu izravno vezana za cilj ovoga istraživanja, prilikom njegove provedbe uočeni su neki problematični segmenti optičkog prepoznavanja znakova ispitivanih uzoraka koje bi valjalo spomenuti te ispitati u nekom od narednih istraživanja.

1) U korištenom OCR programu izabran je hrvatski jezik kao jezik teksta, no niti u jednom uzorku slovo s dugouzlaznim naglaskom „é“ u riječi „poglédala“ nije prepoznato kao takvo, već kao obično slovo „e“. Iz toga se može pretpostaviti da slova s naglascima nisu uključena u rječnik hrvatskog jezika korištenog OCR softvera, iako se ona povremeno koriste u tekstovima pisanima hrvatskim jezikom.
2) Nedostatci pri tiskanju izvornika (prikazani na slici 22) utječu na neispravno prepoznavanje znakova, jer OCR softver vertikalne linije između slova, kao i sitne mrlje tinte na otisku prepoznaje kao znakove kojih u stvari na izvorniku nema. Primjerice, riječ „ako“ često je prepoznata kao „alko“ ili „aJko“, a ime „Mara“ netočno je prepoznato kao „Maira“. OCR softver pogrešno prepoznaje i dodaje nepostojeća slova, a takvo je pogrešno prepoznavanje učestalije kod uzoraka više razlučivosti, jer su na njima ti nedostaci otiska jasniji i lakše ih je pogrešno prepoznati kao slovo, odnosno znak.
3) Pojedina tipografska rješenja korištena u uzorku pokazala su se problematičnima u postupku optičkog prepoznavanja znakova. Na primjer, početna riječ prvog odlomka teksta napisana je velikim tiskanim slovima, a početno slovo je dodatno istaknuto većom veličinom fonta. OCR softver tu riječ u ponekad izdvaja u zaseban blok teksta i nakon provedenog optičkog prepoznavanja znakova stavlja u zaseban redak, a riječ nije napisana velikim tiskanim slovima kao u izvorniku. U rijetkim slučajevima OCR softver navedenu riječ čak ni ne prepoznaje kao tekst, već ju označava kao slikovni element i izostavlja iz pročitanog teksta.

9. Zaključak

Tekstualna građa se digitalizira kako bi se povećala njena iskoristivost i dostupnost, a da bi ona postala prikladna za pregledavanje, pretraživanje i uređivanje, koriste se OCR programi. Iako se željeni rezultat može postići i ručnim prepisivanjem tekstova, taj bi postupak za opsežnu količinu građe bio dugotrajan i mukotrpan te bi se tijekom prepisivanja mogle potkrasti brojne ljudske pogreške. Stoga su razvijena specijalizirana softverska rješenja za optičko prepoznavanje znakova koja zahvaljujući sofisticiranim algoritmima ubrzavaju i olakšavaju postupak prepoznavanja i konverzije teksta iz analognog u digitalni oblik. Osim prepoznavanja tekstualnih elemenata, neki od OCR programa omogućuju i zadržavanje izvornog izgleda dokumenta, uključujući tablice i ostale grafičke elemente, čime se doprinosi preglednosti digitaliziranog dokumenta. Također, proizvođači OCR softvera nerijetko navode visoke postotke točnosti optičkog prepoznavanja znakova koji se mogu postići korištenjem njihovog softvera, a taj postotak u stvari označava stopu pouzdanosti prepoznavanja nekog znaka. Kako je upravo visoka razina točnosti OCR-a ono što krajnji korisnici očekuju, valja znati o čemu sve ona ovisi te kako ju se može postići i unaprijediti.

Točnost optičkog prepoznavanja znakova ovisi o mnogim faktorima, a u njih, osim performansi korištenog OCR programa, spadaju kvaliteta izvornika, razlučivost skeniranja, format slikovne datoteke, sjajnost i kontrast, bitna dubina boje te moguće pogreške pri skeniranju, poput zakrenutosti stranice ili tamnih obruba. Dok se na kvalitetu korištenog izvornika ne može značajnije utjecati, svi ostali navedeni faktori su varijabilni te ih provoditelj digitalizacije može mijenjati s ciljem ostvarivanja što veće točnosti OCR-a, a pritom poduzeti postupci nazivaju se predobradom datoteka.

Pri odabiru građe za digitalizaciju, trebalo bi, kadgod je to moguće, pronaći najočuvaniji dostupni primjerak građe. Naravno, ukoliko se radi o projektu digitalizacije starije i rijetke građe s vidljivim znakovima fizičkog propadanja koju se tim postupkom želi zaštititi od daljnjeg propadanja, pronalaženje bolje očuvanih izvornika može biti teško i financijski zahtjevno, posebno ako oni nisu dio fonda ustanove koja provodi projekt. Stoga prilikom odabira građe valja postaviti jasne kriterije kojima će se odrediti što je najisplativije u odnosu na učestalost korištenja i financijske zahtjeve projekta digitalizacije.

Predobrada prilikom skeniranja građe obično se odnosi na odabir željene razlučivosti, formata slikovne datoteke, bitne dubine boje, sjajnosti i kontrasta. Različiti izvori kao optimalnu razlučivost skeniranja preporučuju 300 DPI, a ovisno o osobinama same građe ponekad i višu

kako bi dobivena slikovna datoteka sadržala što više informacija za kvalitetnije optičko prepoznavanje znakova. Kao preporučeni format slikovnih datoteka navodi se nekomprimirani TIFF, jer na taj način ne dolazi do gubitka informacija, ali se veličina datoteke povećava. Za kvalitetno otisnut izvornik smatra se da se skeniranjem u crno-bijelim tonovima postiže dovoljno dobar kontrast slova i pozadine, dok se za ostalu građu koristi 8-bitna siva skala. Ovdje valja istaknuti da je postizanje što boljeg kontrasta teksta i pozadine iznimno bitno za točno optičko prepoznavanje znakova.

Osim u postupku skeniranja, predobrada ulaznih datoteka može se izvršiti i u nekom od programa za obradu slike poput GIMP-a korištenog u ovom radu. Programi za obradu slike nude brojne opcije optimizacije slikovnih datoteka prije OCR-a pa se u njima može promijeniti njena razlučivost, format, bitna dubina boje, sjajnost i kontrast, kao i ispraviti pogreške nastale prilikom skeniranja kao što su zakrenutost stranice i tamni obrubi.

Istraživanje provedeno u sklopu ovog rada bilo je usmjereno na utjecaj predobrade ulaznih datoteka na točnost OCR-a kako bi se utvrdilo koje postavke daju najbolje rezultate. Kao uzorak poslužila je stranica knjige „*Pripovijetke*“ Vjekoslava Kaleba izdana 1963. godine za koju su karakteristični nedostatci poput požutjele boje papira i pogrešaka nastalih pri tisku. Testiranje na većem uzorku, odnosno broju stranica, zasigurno bi dalo još preciznije rezultate, no i rezultati dobiveni na uzorku od jedne stranice daju okvirne vrijednosti po kojima je moguće odabrati optimalne postavke za dokumente sličnih osobina onome ispitanome. Uspoređeni su slikovni formati TIFF i JPG, različiti stupnjevi kompresije JPG formata i slikovne datoteke binarizirane programom GIMP prije samog OCR-a. Točnost tekstova dobivenih primjenom OCR softvera ABBYY FineReader 12 potom je ispitana ISRI analitičkim alatima te izražena postotkom ispravno prepoznatih znakova.

Dobiveni rezultati su pokazali da je prosječna točnost OCR-a TIFF datoteka nešto viša od prosječne točnosti OCR-a JPG datoteka komprimiranih na 75% izvorne veličine. Usporedbom točnosti OCR-a obzirom na odabranu razlučivost zaključeno je da je bolje skenirati dovoljnom nego iznimno visokom razlučivošću, jer povećanje razlučivosti nije garancija veće točnosti, a na taj se način dobivaju znatno veće datoteke za koje je potrebno dulje vrijeme skeniranja i OCR-a. Za građu ove kvalitete izravno skeniranje i OCR crno-bijelih dokumenata nije polučilo najviše rezultate točnosti OCR-a, već se kao bolja opcija pokazalo skeniranje i OCR dokumenata u sivoj skali ili u boji. Ukoliko se radi uštede podatkovnog prostora ili nekog drugog razloga umjesto TIFF formata koristi JPG, preporučljivo je datoteke ne komprimirati ili koristiti što manji stupanj kompresije, jer se na taj način postiže bolja točnost OCR-a.

Također, ispitivanje je pokazalo da binarizacija slikovnih datoteka programom za obradu slika prije OCR-a u većini slučajeva nije povećala točnost OCR-a te je stoga taj postupak bolje ne primijeniti.

Kako je razvoj OCR softvera područje koje se i dalje aktivno i intenzivno istražuje, napredak u postizanju još veće točnosti optičkog prepoznavanja znakova nesumnjivo neće izostati. Neovisno o tome hoće li u budućnosti razvoj polja biti usmjeren na daljnje usavršavanje optičkog prepoznavanja znakova ili će veći napori biti usmjereni ka strojnom prepoznavanju rukopisa, današnja istraživanja usmjerena na točnost OCR-a i faktore koji na nju utječu zasigurno će doprinijeti kvalitetnijem ostvarivanju budućih ciljeva.

10. Literatura

1. *A brief history of OCR: the technology inside your scanmarker* (2019). Preuzeto 13.04.2019. s https://scanmarker.com/a-brief-history-of-ocr-the-technology-inside-your-scanmarker/
2. *ABBYY FineReader celebrates 25 years on the market* (2018). Preuzeto 17.07.2019. s https://finereaderblog.abbyy.com/abbyy-finereader-celebrates-25-years-on-the-market/
3. *ABBYY Finereader 12* (bez dat.). Preuzeto 23.04.2019. s https://www.abbyy.com/en-gb/support/finereader/12/formats/
4. *About GIMP* (bez dat.). Preuzeto 04.06.2019. s https://www.gimp.org/about/introduction.html
5. ASCII (bez dat.). U Hrvatska enciklopedija (mrežno izdanje). Preuzeto 23.04.2019. s http://www.enciklopedija.hr/Natuknica.aspx?ID=4136
6. *B Is for Binarize* (bez dat.). Preuzeto 14.05.2019. s http://www.how-ocr-works.com/OCR/binarization.html
7. *BMP file format* (bez dat.). U Wikipedia. Preuzeto 05.05.2019. s https://en.wikipedia.org/wiki/BMP_file_format
8. Chaudhuri, A., Mandaviya, K., Badelia, P. i Ghosh, S.K. (2017). *Optical Character Recognition Systems for Different Languages with Soft Computing*. Cham, Switzerland: Springer International Publishing AG.
9. Cheriet, M., Kharma, N., Liu, C.-L., i Suen, C.Y. (2007). *Character Recognition Systems - A Guide for Students and Practitioners*. Hoboken, NJ, USA: John Wiley & Sons Inc.
10. Christensson, P. (2012). Unicode Definition. U *Tech Terms Computer Dictionary*. Preuzeto 24.07.2019. s https://techterms.com/definition/unicode
11. Dadfar, K. (bez dat.). *Understanding all the Different Image File Formats*. Preuzeto 05.05.2019. s https://digital-photography-school.com/understanding-all-the-different-image-file-formats/
12. Dvornikova, V. (bez dat.). *FineReader 12 specification*. Preuzeto 17.07.2019. s https://support.abbyy.com/hc/en-us/articles/360004047940-FineReader-12-specification
13. Epson (2003). *Epson Stylus Photo RX500*. Preuzeto 28.04.2019. s https://files.support.epson.com/pdf/rx500_/rx500_sl.pdf

14. European Comission on Preservation and Access (1997). *Digitization as a Means of Preservation?* Preuzeto 07.04.2019. s https://www.clir.org/pubs/reports/digpres/digpres3/

15. *Freeware* (bez dat.). U Hrvatska enciklopedija (mrežno izdanje). Preuzeto 25.04.2019. s http://www.enciklopedija.hr/Natuknica.aspx?ID=68092

16. GIMP (bez dat.). U Wikipedia. Preuzeto 04.06.2019. s https://en.wikipedia.org/wiki/GIMP

17. Google Books (2011). *O Google pretraživanju knjiga*. Preuzeto 14.04.2019. s https://books.google.com/intl/hr/googlebooks/about.html

18. Henson, J. (2018). *What are the factors that affect the accuracy of OCR?* Preuzeto 12.05.2019. s https://marinersoftware.deskpro.com/kb/articles/what-are-the-factors-that-affect-the-accuracy-of-ocr

19. Holley, R. (2009). How Good Can It Get? Analysing and Improving OCR Accuracy in Large Scale Historic Newspaper Digitisation Programs. *D-Lib Magazine*, Vol.15(3/4). doi:10.1045/march2009-holley. Preuzeto 12.05.2019. s http://www.dlib.org/dlib/march09/holley/03holley.html

20. *Image noise* (bez dat.). U Wikipedia. Preuzeto 13.05.2019. s https://en.wikipedia.org/wiki/Image_noise

21. *Improve OCR Accuracy With Advanced Image Preprocessing* (bez dat.). Preuzeto 12.05.2019. s https://docparser.com/blog/improve-ocr-accuracy/

22. *Improving the quality of the output* (bez dat.). Preuzeto 26.05.2019. s https://github.com/tesseract-ocr/tesseract/wiki/ImproveQuality

23. Lambert, T.J., i Waters, J.C. (2014) Assessing camera performance for quantitative microscopy. *Methods in cell biology*, 123C, 35-53. doi: 10.1016/B978-0-12-420138-5.00003-3

24. Matthews, R. (bez dat.). *Digital Image File Types Explained*. Preuzeto 05.05.2019. s http://users.wfu.edu/matthews/misc/graphics/formats/formats.html

25. Ministarstvo kulture (2007). *Formati datoteka za pohranu i korištenje : radna verzija*. Preuzeto 05.05.2019. s www.kultura.hr/content/download/597/7937

26. Nartker, T. A., Rice, S. V. i Lumos, S. E. (2005). Software Tools and Test Data for Research and Testing of Page-Reading OCR Systems. *Proceedings of SPIE, Document Recognition*

and Retrieval XII, Vol. 5676, 17 January 2005. Preuzeto 23.07.2019. s http://stephenvrice.com/images/Nartker2005a.pdf

27. *Novi ABBYY FineReader 12 uvelike ubrzava raspoznavanje i citiranje teksta iz skenova i fotografija* (2014). Preuzeto 17.07.2019. s https://pcchip.hr/softver/novi-abbyy-finereader-12-uvelike-ubrzava-raspoznavanje-i-citiranje-teksta-iz-skenova-i-fotografija/
28. *Open source licence* (bez dat.). Preuzeto 25.04.2019. s https://www.carnet.hr/tematski/opensource/licence.html
29. Project Gutenberg (2019). *Free eBooks – Project Gutenberg*. Preuzeto 14.04.2019. s http://www.gutenberg.org/wiki/Main_Page
30. *Proprietary software* (bez dat.). U Wikipedia. Preuzeto 24.04.2019. s https://en.wikipedia.org/wiki/Proprietary_software
31. Radošević, D. (1996). Postupci i problemi optičkog raspoznavanja uzoraka. *Zbornik radova: journal of information and organizational sciences*, 21(2), 17-31.
32. Rice, S. V. i Nartker, T. A. (1996). *The ISRI Analytic Tools for OCR Evaluation. Version 5.1.* Technical report TR 96-02. Las Vegas, NV, USA: Information Science Research Institute (ISRI). Preuzeto 23.07.2019. s http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.216.9427&rep=rep1&type=pdf
33. Santos, E. A. (2019) OCR Evaluation Tools for the 21st Century. *Proceedings of the Workshop on Computational Methods for Endangered Languages*, Volume 1, Article 4, 23-27.
34. Savuola, J., i Pietikäinen, M. (2000) *Adaptive document image binarization*. Pattern Recognition, 33, 225-236. Preuzeto 14.05.2019. s http://www.ee.oulu.fi/mvg/files/pdf/pdf_24.pdf
35. *Simple Software - OCR Servers* (bez dat.). Preuzeto 24.04.2019. s https://www.simpleocr.com/OCR-Servers
36. *Simple Software - OCR Software Guide* (bez dat.). Preuzeto 24.04.2019. s https://www.simpleocr.com/OCR_Software_Guide
37. *Slikovni element* (bez dat.). U Hrvatska enciklopedija (mrežno izdanje). Preuzeto 23.04.2019. s http://www.enciklopedija.hr/Natuknica.aspx?ID=56642

38. Smith, S. W. (1997). *The Scientist and Engineer's Guide to Digital Signal Processing*. San Diego, CA, USA: California Technical Publishing.

39. Stančić, H. (2009). *Digitalizacija*. Zagreb: Zavod za informacijske studije Odsjeka za informacijske znanosti Filozofskog fakulteta Sveučilišta u Zagrebu.

40. Strgar Kurečić, M. (bez dat.) *Reprodukcijski procesi*. Reprodukcijska fotografija [predavanje - pdf]. Sveučilište u Zagrebu, Grafički fakultet, Zagreb. Preuzeto 05.05.2019. s http://repro.grf.unizg.hr/media/download_gallery/Reprodukcijski%20procesi.pdf

41. *The Open Source Definition* (2007). Preuzeto 25.04.2019. s https://opensource.org/osd

42. *Timeline of optical character recognition* (bez dat.). U Wikipedia. Preuzeto 14.04.2019. s https://en.wikipedia.org/wiki/Timeline_of_optical_character_recognition

43. Walls, J. (2008). *OCR and Content Management with SAP and Imaging* [Slideshare prezentacija]. Preuzeto 23.04.2019. s https://www.slideshare.net/verbella/ocr-and-content-management-with-sap-and-imaging

44. *What is OCR and OCR technology* (bez dat.). Preuzeto 07.04.2019. s https://www.abbyy.com/en-us/finereader/what-is-ocr/

45. *Zonal OCR - What's it good for?* (2015) [Blog objava]. Preuzeto 24.04.2019. s https://www.filehold.com/blog/15/08/zonal-ocr-whats-it-good

Popis oznaka i kratica

ANN (engl. Artificial Neural Networks) umjetne neuronske mreže

ASCII (engl. American Standard Code for Information Interchange) američki normirani kod za razmjenu informacija

BMP (engl. Bitmap) vrsta formata slikovnih datoteka

CMYK (engl. Cyan, Magenta, Yellow, blacK) sustav prikaza boje uobičajen za ispis slika pisačima u boji

DCT (engl. Discrete Cosine Transform) diskretna kosinusna transformacija

DPI (engl. Dots Per Inch) točke po inču

GIF (engl. Graphic Interchange Format) vrsta formata slikovnih datoteka

GIMP (engl. GNU Image Manipulation Program) softver otvorenog koda za stvaranje i obradu rasterske grafike

GNU GPL (engl. GNU General Public Licence) vrsta slobodne licence

HMM (engl. Hidden Markov Models) skriveni Markovljevi modeli

ICR (engl. Intelligent Character Recognition) inteligentno prepoznavanje znakova

ISRI Analytic Tools (engl. Information Science Research Institute) analitički alati za provjeru točnosti teksta prepoznatog OCR programom Instituta za informacijske znanosti Sveučilišta u Nevadi

IWR (engl. Intelligent Word Recognition) inteligentno prepoznavanje riječi, odnosno rukopisa

JPG / JPEG (engl. Joint Photographic Experts Group) vrsta formata slikovnih datoteka

LZW (engl. Lempel-Ziv-Welch) vrsta algoritma korištena za komprimiranje podataka bez gubitka

NLP (engl. Natural Language Processing) obrada prirodnog jezika

OCR (engl. Optical Character Recognition) optičko prepoznavanje znakova

OMR (engl. Optical Mark Recognition) optičko prepoznavanje oznaka

PNG (engl. Portable Network Graphics) vrsta formata slikovnih datoteka

RGB (engl. Red, Green, Blue) sustav prikaza boje uobičajen za prikaz boje na zaslonima računala

TIF / TIFF (engl. Tagged Image File Format) vrsta formata slikovnih datoteka

UTF (engl. Unicode Transformation Format) format kodiranja znakova u Unicode standardu

Popis slika

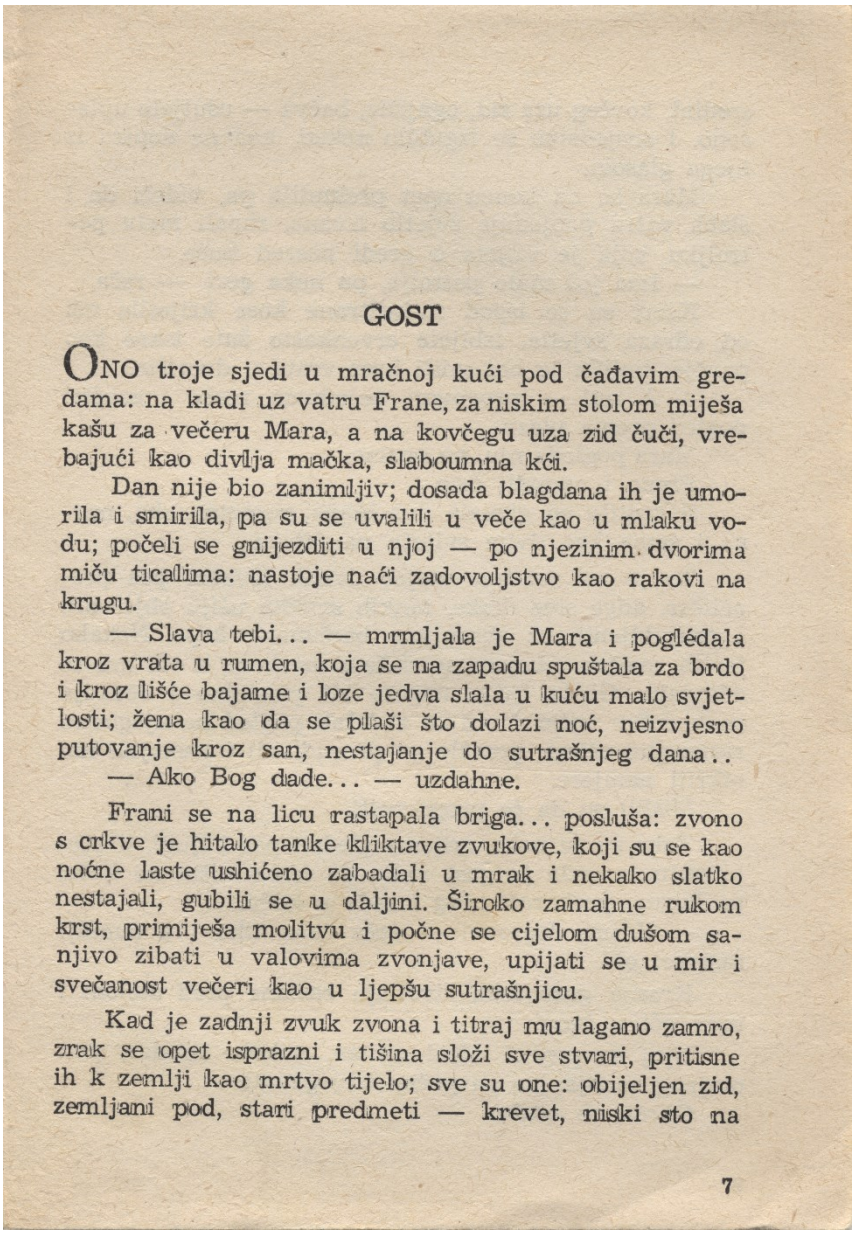
Slika 1. Optičko prepoznavanje znakova ili riječi prema složenosti postupka 5
Slika 2. Font OCR-A 8
Slika 3. Font OCR-B.....8
Slika 4. Spojeni znakovi „k" i „r".....12
Slika 5. Razlomljeni znakovi „a", „o" i „m" 12
Slika 6. Slova skenirana različitim razlučivostima 13
Slika 7. Pet različitih bitnih dubina točke i nijanse sive skale koje se njima dobivaju 17
Slika 8. RGB sustav boja 17
Slika 9. Primjer dobre binarizacije na uzorku loše kvalitete 18
Slika 10. Primjer odstupanja zakrenutosti te korekcija odstupanja rotacijom.....19
Slika 11. Vidljivi tamni obrubi i dio susjedne stranice nakon skeniranja i izgled stranice nakon njihovog uklanjanja 20
Slika 12. Epson Stylus Photo RX 500 21
Slika 13. Slika zaslona - program za skeniranje uređaja Epson Stylus Photo RX500.....22
Slika 14. Slika zaslona - softver za obradu slika GIMP 2.10.10 23
Slika 15. Slika zaslona – podešavanje stupnja kompresije JPG formata u programu GIMP ..24
Slika 16. Slika zaslona – mijenjanje postavki sjajnosti i kontrasta u programu GIMP 25
Slika 17. Slika zaslona – određivanje vrijednosti praga u programu GIMP.....26
Slika 18. Slika binarizirana primjenom globalnog praga u programu GIMP 27
Slika 19. Slika binarizirana samostalnim određivanjem vrijednosti praga u programu GIMP27
Slika 20. Slika zaslona - ispravljanje zakrenutosti teksta alatom *Rotate* u programu GIMP ..28
Slika 21. Prednja korica knjige „*Pripovijetke*" Vjekoslava Kaleba.....30
Slika 22. Nedostaci tiska vidljivi na odabranom uzorku 30
Slika 23. Slika zaslona – ABBYY FineReader 12 Professional.....32

Popis tablica

Tablica 1. Točnost OCR-a za TIFF datoteke uz različite bitne dubine boje i razlučivosti.....34
Tablica 2. Točnost OCR-a za JPG datoteke uz različite bitne dubine boje i razlučivosti 35
Tablica 3. Točnost OCR-a za JPG datoteke uz različite stupnjeve kompresije.....36
Tablica 4. Usporedba točnosti OCR-a za izvorne i naknadno binarizirane TIFF datoteke 38
Tablica 5. Usporedba točnosti OCR-a za izvorne i naknadno binarizirane JPG datoteke.....38

Prilozi

Prilog 1 – testirani uzorak, TIFF format u boji skeniran razlučivošću od 600 DPI



GOST

ONO troje sjedi u mračnoj kući pod čađavim gredama: na kladi uz vatru Frane, za niskim stolom miješa kašu za večeru Mara, a na kovčegu uza zid čuči, vrebajući kao divlja mačka, slaboumna kći.

Dan nije bio zanimljiv; dosada blagdana ih je umorila i smirila, pa su se uvalili u veče kao u mlaku vodu; počeli se gnijezditi u njoj — po njezinim dvorima miču ticalima: nastoje naći zadovoljstvo kao rakovi na krugu.

— Slava tebi... — mrmljala je Mara i poglédala kroz vrata u rumen, koja se na zapadu spuštala za brdo i kroz lišće bajame i loze jedva slala u kuću malo svjetlosti; žena kao da se plaši što dolazi noć, neizvjesno putovanje kroz san, nestajanje do sutrašnjeg dana..

— Ako Bog dade... — uzdahne.

Frani se na licu rastapala briga... posluša: zvono s crkve je hitalo tanke kliktave zvukove, koji su se kao noćne laste ushićeno zabadali u mrak i nekako slatko nestajali, gubili se u daljini. Široko zamahne rukom krst, primiješa molitvu i počne se cijelom dušom sanjivo zibati u valovima zvonjave, upijati se u mir i svečanost večeri kao u ljepšu sutrašnjicu.

Kad je zadnji zvuk zvona i titraj mu lagano zamro, zrak se opet isprazni i tišina složi sve stvari, pritisne ih k zemlji kao mrtvo tijelo; sve su one: obijeljen zid, zemljani pod, stari predmeti — krevet, niski sto na

7

Utjecaj predobrade ulaznih datoteka na točnost optičkog prepoznavanja znakova

Sažetak

Digitalizacija tekstualne građe danas je široko zastupljena u različitim domenama ljudskih djelatnosti. Najčešće se provodi skeniranjem ili fotografiranjem građe te upotrebom specijaliziranih programa za optičko prepoznavanje znakova. Na taj se način dobiva elektronička građa koju je moguće pregledavati, pretraživati i uređivati. Točnost dobivenih izlaznih podataka ovisi brojnim faktorima, a neki od njih su kvaliteta izvornika, razlučivost skeniranja, odabrani slikovni format, bitna dubina boje, ali i korišteni OCR softver. I postupak binarizacije, odnosno konverzije višebojne slike u crno-bijelu, utječe na rezultate optičkog prepoznavanja znakova. Binarizacija se obično provodi prilikom skeniranja ili kao jedan od koraka pri optičkom prepoznavanju znakova, a moguće ju je provesti i u nekom od programa za obradu slike prije korištenja OCR softvera. Cilj ovog rada jest istražiti utjecaj predobrade ulaznih datoteka na točnost optičkog prepoznavanja znakova s obzirom na navedene faktore. Točnost tekstova dobivenih primjenom OCR softvera ispitana je ISRI analitičkim alatima te je izražena postotkom ispravno prepoznatih znakova.

Ključne riječi: optičko prepoznavanje znakova, OCR, točnost OCR-a, predobrada, razlučivost skeniranja, DPI, slikovni format, TIFF, JPG, kompresija slike, bitna dubina boje, binarizacija, ISRI analitički alati.

The effects of the image pre-processing on the OCR accuracy

Summary

Digitization of textual materials is nowadays widely used in different domains of human activities. It is most commonly done by scanning or taking photos of the original source followed by using specialized software for optical character recognition. The result of this procedure is an electronic text which can be browsed, searched and edited. The OCR accuracy of the recognized text depends on numerous factors, including quality of the source materials, scanning resolution, chosen image file format, image bit depth as well as the OCR software used in the text recognition process. Image binarization or thresholding, a procedure in which a greyscale or colour image is converted into its bitonal variant, can also influence the OCR results. Binarization usually takes place during scanning or as a step in the OCR procedure, but it can also be accomplished by using a raster graphics editor, before the usage of the OCR software. The aim of this diploma thesis is to examine the effects of image pre-processing on OCR accuracy taking into account the above-mentioned factors. The OCR accuracy of the output texts was checked by ISRI Analytic Tools and is expressed as a percentage of correctly recognized characters.

Key words: optical character recognition, OCR, OCR accuracy, pre-processing, scanning resolution, DPI, image file format, TIFF, JPG, image compression, bit depth, binarization, thresholding, ISRI Analytic Tools.