Usporedba aplikacija za optičko prepoznavanje znakova

Jelovčić, Sunčana

Sveučilište u Zadru

Odjel za informacijske znanosti
Diplomski sveučilišni studij informacijskih znanosti

Sunčana Jelovčić

Usporedba aplikacija za optičko prepoznavanje znakova

Diplomski rad

Zadar, 2021.

Sveučilište u Zadru

Odjel za informacijske znanosti
Izvanredni diplomski sveučilišni studij Informacijske znanosti

Usporedba aplikacija za optičko prepoznavanje znakova

Diplomski rad

Studentica:
Sunčana Jelovčić

Mentor:
doc. dr. sc. Mirko Duić

Zadar, 2021.

Izjava o akademskoj čestitosti

Ja, Sunčana Jelovčić, ovime izjavljujem da je moj diplomski rad pod naslovom **Usporedba aplikacija za optičko prepoznavanje znakova** rezultat mojega vlastitog rada, da se temelji na mojim istraživanjima te da se oslanja na izvore i radove navedene u bilješkama i popisu literature. Ni jedan dio mojega rada nije napisan na nedopušten način, odnosno nije prepisan iz necitiranih radova i ne krši bilo čija autorska prava.

Izjavljujem da ni jedan dio ovoga rada nije iskorišten u kojem drugom radu pri bilo kojoj drugoj visokoškolskoj, znanstvenoj, obrazovnoj ili inoj ustanovi.

Sadržaj mojega rada u potpunosti odgovara sadržaju obranjenoga i nakon obrane uređenoga rada.

Zadar, 25. veljače 2021.

Sažetak

Cilj ovog rada je predstavljanje procesa optičkog prepoznavanja znakova i njegove primjene u praksi. Kako je optičko prepoznavanje znakova pretvaranje analognog dokumenta u pretraživu digitalnu inačicu, u prvom teoretskom dijelu rada objašnjen je pojam digitalizacije te su predstavljeni pojedini projekti i inicijative digitalizacije. Proces optičkog prepoznavanja znakova isprepliće se sa nekoliko srodnih disciplina od kojih su u radu ukratko objašnjene sljedeće: prepoznavanje uzoraka, računalni vid te strojno učenje. Teoretski dio rada također opisuje osam glavnih komponenti procesa optičkog prepoznavanja znakova: optičko skeniranje, segmentiranje regija slike, pretprocesiranje, normalizacija, segmentacija, reprezentacija, ekstrahiranje značajki te poučavanje i prepoznavanje. Potom su predstavljene aplikacije za optičko prepoznavanje znakova te je napravljeno istraživanje usporedbe rada u aplikacijama ABBYY FineReader 15, Free OCR-u, Google Drive OCR, I2OCR-u i Convertio. U istraživanju je napravljena usporedba osobina i uspješnosti pet različitih aplikacija za optičko prepoznavanje znakova. Provedena je rasprava o dobivenim rezultatima te su na temelju uvida iz istraživanja i analizirane teorije predstavljena zaključna razmatranja.

Ključne riječi: optičko prepoznavanje znakova, digitalizacija, ABBYY FineReader 15, Tesseract Google Drive OCR, Free OCR, Convertio, I2OCR

Kazalo sadržaja

Kazalo slika

Kazalo tablica

1. Uvod

Okruženi smo brojnim tehnologijama koje uzimamo zdravo za gotovo a da često uopće ne znamo o kakvim se složenim procesima radi u pozadini. Televizor, stolno računalo, laptop, mobilni telefon, tablet, automobil i mnoge druge tehnologije koristimo svakodnevno a da o njima znamo vrlo malo ili gotovo ništa. Upravo iz ovog razloga, ono na čemu je poseban naglasak u ovom radu je sam proces optičkog prepoznavanja znakova, odnosno što se zapravo događa kada tekst u slikovnom formatu želimo pretvoriti u čitljiv, obradiv i pretraživ tekst. Iz tog razloga najveći dio teoretskog dijela rada bavi se analizom komponenti optičkog prepoznavanja znakova odnosno tehničkim dijelom procesa, a potom se u istraživačkom dijelu rada istražuju različiti programi odnosno aplikacije za optičko prepoznavanje znakova. Rad je pisan kako bi bio razumljiv svim članovima akademske zajednice odnosno analiza komponenti optičkog prepoznavanja znakova seže duboko ali je pisana vrlo jednostavnim jezikom. Kada malo dublje istražujete funkcioniranje mnogih procesa koje nikako ne možete shvatiti laičkim promatranjem često ćete se susresti sa svijetom logaritama odnosno skupom naredbi koje su definirane logičkim slijedom i služe rješavanju kompleksnih problema. Upravo do te granice, u dubljoj analizi, uranja i ovaj rad, dolazi do tih skrivenih dijelova, promatra ih i kad treba prikazuje na jednostavan, svima razumljiv način.

Kako je proces optičkog prepoznavanja znakova usko povezan sa digitalizacijom odnosno često se spominje kao dio procesa digitalizacije, u ovom radu date su neke osnovne definicije kao, na primjer, što je to digitalizacija (posebno digitalizacija teksta), koji su poznati projekti digitalizacije na globalnoj ali i nacionalnoj razini te koji su koraci projekta digitalizacije. Digitalizacija je od posebne važnosti za četiri baštinske djelatnosti: knjižničnu, muzejsku, arhivsku i konzervatorsko–restauratorsku. U digitalnom svijetu uzorak je sve, bilo da se radi o bitovima ili bajtovima ili nekoj drugoj vrsti informacije. Uzorak može biti vidljiv golim okom ili promatran matematički primjenjujući algoritme. Prepoznavanje uzoraka je proces u kojem se koriste algoritmi za strojno učenje. Prepoznavanje uzoraka može se također definirati kao klasifikacija određenih podataka baziranih na informacijama koje su prethodno statistički ekstrahirane iz uzoraka. Digitalizacija i optičko prepoznavanje znakova su „simbiotski organizam“ na način da se brojni, ali ne svi, projekti digitalizacije temelje na procesu optičkog prepoznavanja znakova. Optičko prepoznavanje znakova često se koristi i neovisno o digitalizaciji, na primjer kada se optički prepoznaju odnosno verificiraju računi, formulari, potpisi, i brojne druge stavke u sklopu neke vrste poslovanja.

Strojna replikacija ljudskih funkcija, poput čitanja, predstavlja veliki drevni san. Mnogi procesi, koji su nekad uključivali zaprimanje dokumenata, sortiranje, ručno procesiranje, ispunjavanje formulara te njegov povrat, danas su automatizirani. Optičko prepoznavanje znakova proces je koji se pojavio uslijed i zbog povećane automatizacije te je utjecao na brže procesiranje podataka i smanjeno poslovanje u papirnatom obliku. Velik broj informacija u svijetu sačuvan je u papirnatom obliku te optičko prepoznavanje znakova omogućuje prebacivanje ovih informacija u elektronički oblik. Ovaj čin omogućava brže pretraživanje dokumentacije prema predmetu interesa. Međutim, sustavi za optičko prepoznavanje znakova ne izvode konverziju bez greške te se elektroničke verzije najčešće ne podudaraju sa papirnatom verzijom. Jedan od najznačajnijih problema su greške kod optičkog prepoznavanja znakova prilikom konverzije analognog dokumenta u digitalni, binarni dokument. Greške u prepoznavanju znakova mogu nastati najčešće zbog nečistoće papira, razmazanih znakova i nejasnih poteza, ukošenosti linije teksta u odnosu na list papira, ukošenosti slova i drugo. Dakle, sustav za optičko prepoznavanje znakova koji radi previše grešaka i nije od neke koristi, stoga treba sagledati koji su to problemi s kojima se sustavi za optičko skeniranje susreću te na koji način ih rješavaju.

2. Digitalizacija

2.1. Uvodno o procesu digitalizacije

Digitalna konverzija knjižničnog materijala, odnosno digitalizacija, u posljednjih nekoliko godina napredovala je velikom brzinom. Digitalizacija je moguća za gotovo svaki format i medij koji knjižnica posjeduje, od mapa do rukopisa, pomičnih slika do notnih zapisa. Korištenje hardvera i softvera za snimanje predmeta digitalizacije i njegovu konverziju u bitove i bajtove, usporedno sa razvitkom različitih praksi za opis i nabavu digitalnih objekata, dovelo je do fenomena "knjižnice bez zidova".

U digitalnom svijetu, sve znanje je zapisano u nizovima binarnih znakova nula i jedinica koje čine genetički kod informacija. Na ovaj način ljudima je omogućeno da stvaraju, manipuliraju i dijele informacije na revolucionaran način. Često se kaže da digitalne informacije mijenjaju način na koji učimo, komuniciramo pa čak i način na koji mislimo. Ne samo da mijenjaju način na koji knjižnice rade već i sam rad koji obavljaju.

Donedavno, sve pohranjene informacije bile su analogne, odnosno nisu bile zapisane putem nula i jedinica kao kod digitalnih zapisa. Analogna informacija može imati opseg od suptilnih tonova i gradacija na crno bijeloj fotografiji do promjena u volumenu, tonu i visini glasa snimljenima na kazeti. Međutim, kada su takve informacije pohranjene na računalo, razbijene u nule i jedinice te posložene u binarni kod, njihov karakter je promijenjen.

Digitalna informacija ne predstavlja beskonačno promjenjivu prirodu informacije vjerno kao analogni zapis informacije. Nule i jedinice su pridružene brojčane vrijednosti koje su fiksne te je postignuta velika preciznost umjesto beskonačne gradacije zapisane analognim zapisom. Na primjer, kada je fotografija digitalizirana za prikaz na kompjutorskom zaslonu, originalni ton slike podijeljen je na točke, sa njima pridruženim vrijednostima, koje su mapirane u mreži. Ovakvi nizovi bitova koji tvore informaciju vrlo lako mogu biti rekombinirani kako bi se informacijom lakše manipuliralo te ju se spremalo za pohranu.

Jedna od najvažnijih kvaliteta informacije u digitalnom obliku je da po svojoj prirodi nije fiksna kao tekst na papiru. Digitalni tekst nije konačan, osim kada je isprintan. Fleksibilnost je jedna od glavnih prednosti digitalne informacije što se očituje kada tekstom manipuliramo u nekom od programa za obradu teksta. Tada se tekst lako može uređivati, može mu se promijeniti format, stvoriti brojne kopije identičnog digitalnog dokumenta te naposljetku ispisati pod odabranim postavkama.

2.2. Što je to digitalizacija?

U 21. stoljeću digitalizacija je pojam koji polako ali sigurno ulazi u svakodnevni vokabular. Koriste ga pripadnici različitih struka, počevši od knjižnica, muzeja, arhiva, konzervatorsko-restauratorskih ureda do privatnih tvrtki i pojedinaca za svoje privatne potrebe. Razlog njezine široke upotrebe je kompjuterizacija koja je postala sastavni dio gotovo svakog procesa u današnje doba kada se najveći dio poslovnih transakcija odvija mrežnim putem. Popularni naziv "društvo znanja" i "informacijsko doba" također su usko vezani sa počecima digitalizacije. Ipak, prije svega, potrebno je stručno opisati što se sve krije iza pojma digitalizacije te kako se ona odvija.

Pri procesu digitalizacije jasno je da se događa nekakva vrsta pretvorbe jednog medija u drugi, no sam proces je, kao i kod računala, određen nizom prethodno definiranih naredbi koje se odvijaju u samo nekoliko sekundi, skrivene ljudskom oku. Prema enciklopediji Leksikografskog zavoda Miroslava Krleže, digitalizacija (engl. digitalization, od digit: znamenka), u najširem smislu, je prevođenje analognoga signala u digitalni oblik.[1] U užem smislu to je pretvorba teksta, slike, zvuka, pokretnih slika (filmova i videa) ili trodimenzionalnog oblika nekog objekta u digitalni oblik, u pravilu binaran kod zapisan kao računalna datoteka sa sažimanjem podataka ili bez sažimanja podataka, koji se može obrađivati, pohranjivati ili prenositi računalima i računalnim sustavima.[2] Sažimanje podataka vrši se kompresijom u format koji zauzima manje memorijskog prostora nego original te je pogodniji za slanje elektronskim putem. Digitalizacija se vrši procesom analogno-digitalne pretvorbe, odnosno pretvaranjem informacije iz analognog u digitalni (brojčani) oblik.[3] Digitalna je informacija pri obradi mnogo manje podložna smetnjama, izobličavanju, oštećenju i drugim utjecajima, nego što je to analogna informacija.[4]

[1] Usp. Digitalizacija. Leksikografski zavod Miroslava Krleže. Encikopedija. URL: http://enciklopedija.hr/Natuknica.aspxID=68025 (2019-02-17)
[2] Ibid.
[3] Ibid.
[4] Ibid.

2.2.1. Digitalizacija teksta

U prethodnom dijelu rada navedeno je da se digitalizirati mogu fotografije, tekst, zvuk, pokretne slike te kako postoji i 3D digitalizacija. S obzirom da se u ovom radu analizira optičko prepoznavanje znakova (slova) nadalje će više biti riječ o tome što je digitalizacija teksta te kako se ona provodi.

Digitalizacija teksta, kao što je navedeno u definiciji digitalizacije, je pretvorba analogne tekstualne informacije u digitalnu. Prema enciklopediji leksikografskog zavoda Miroslava Krleže, digitalizacija teksta, odnosno dokumenata, pojedinih stranica ili cijelih knjiga i drugih tiskovina s tekstualnim sadržajem, provodi se istim postupkom kao i digitalizacija slike te se time dobivaju digitalne slike teksta koje se mogu prikazati na zaslonu računala, čitati i prelistavati, ali ne i pretraživati i obrađivati.[5] To se, prema istom izvoru, postiže naknadnim pretvaranjem teksta u računalno čitljiv oblik uz pomoć posebnoga programa za optičko prepoznavanje znakova (engl. Optical Character Recognition, OCR), kojim se svakom znaku u tekstu dodjeljuje odgovarajući UTF-8, ASCII ili drugi binarni kod.[6] Optičko prepoznavanje znakova važno je u procesu digitalizacije. U virtualnom svijetu česti su primjeri digitaliziranih dokumenata, na primjer u PDF formatu, koji nisu procesirani optičkim prepoznavanjem znakova, a stoga ih nije moguće naknadno obrađivati niti pretraživati pojmove koji bi mogli zanimati čitatelja. Postoji više načina na koji se tekst može digitalizirati što ovisi o potrebama projekta digitalizacije. Stančić navodi kako se digitalizacija teksta provodi prepisivanjem (starijeg rukopisnog gradiva), korištenjem koračnih (za uvezano gradivo) ili protočnih skenera opremljenih uvlakačima stranica te slikanjem digitalnim fotoaparatom.[7] Ipak, za opsežnije projekte digitalizacije, od svega navedenog, najbolji izbor bi bili protočni skeneri jer su automatizirani i mogu samostalno listati stranice bez ljudske intervencije.[8] Stančić navodi kako digitalizacija skeniranjem ili fotografiranjem rezultira slikom teksta te ako se želi dobiti obradiv i pretraživ tekst, sliku teksta potrebno je obraditi programom za optičko prepoznavanje znakova (engl. Optical Character Recognition – OCR) koji na temelju kontrasta između pozadine i otisnutih znakova prepoznaje znakove i zapisuje ih kao tekst dok neki profesionalniji OCR programi mogu prepoznati strukturu grafičkih

[5] Usp. Digitalizacija. Leksikografski zavod Miroslava Krleže. Enciklopedija. URL: http://enciklopedija.hr/Natuknica.aspxID=68025 (2019-02-27)
[6] Ibid.
[7] Usp. Stančić, Hrvoje; Zanier, Katharina. Heritage Live. Upravljanje baštinom uz pomoć informacijskih alata. Str. 12. – 13. URL: https://www.had-info.hr/dokumenti/publikacije/Heritage%20live%20-%20Upravljanje%20bastinom%20uz%20pomoc%20informacijskih%20alata.pdf (2019-02-28)
[8] Ibid.

elemenata u izgledu skenirane stranice (tekstualni okviri, stupci i sl.) te tekst koji su prepoznali vizualno složiti tako da izgleda poput skeniranoga.[9] Proces optičkog skeniranja složen je postupak koji se sastoji od brojnih potkomponenti, o čemu je više riječ u 5. poglavlju: Komponente OCR sustava.

2.3. Razlozi za digitalizaciju

Stančić navodi kako se digitalizacija provodi radi: 1. zaštite izvornika; 2. povećanja dostupnosti; 3. stvaranja nove ponude i usluga; 4. upotpunjavanja fonda; 5. zahtijeva korisnika.[10] Slično ovome, autori Škrabo i Vrana, navode kako se digitalizacija knjižnične građe provodi radi zaštite izvornika, povećanja dostupnosti i mogućnosti korištenja građe, radi stvaranja nove ponude, odnosno usluga korisnicima ili pak radi upotpunjavanja postojećega fonda.[11] Prema IFLA – inim smjernicama, projekti digitalizacije se provode kako bi se povećao pristup građi, poboljšale usluge korisnicima, smanjilo korištenje osjetljive i ugrožene građe, izgradila potrebna infrastruktura i kadar u sklopu određene institucije, uspostavila suradnja s drugim institucijama putem kreiranja virtualnih zbirki.[12]

2.4. Projekti i inicijative digitalizacije

2.4.1. Masovni projekti digitalizacije

Unatoč čestom dolasku u sukob s vlasnicima autorskih prava, tiskana građa se sve više digitalizira i o projektima masovne digitalizacije sve je više riječ u stručnoj literaturi. Ipak, ukoliko se, na primjer, pretražuje digitalizirana građa, točnije digitalizirane knjige na platformi Google Books, lako se može zaključiti da građa većim dijelom nije dana na uvid u svom cjelovitom obliku već selektivno ili nikako. Ono što ovaj i slični projekti nude korisnicima je mogućnost pretraživanja po ključnim riječima što znači da su dokumenti obrađeni uz pomoć optičkog prepoznavanja znakova. Autorica Šapro-Ficović navodi kako masovna digitalizacija označava projekte kojima se knjige digitaliziraju u industrijskim razmjerima, uz uporabu napredne i suvremene tehnologije te uz velika ekonomska ulaganja u te projekte.[13] U istom radu navodi se kako cilj masovne digitalizacije nije stvaranje posebne

[9] Ibid, str. 13.
[10] Ibid, str. 14.
[11] Usp. Škrabo, Katarina; Vrana, Radovan. Digitalne zbirke u narodnim knjižnicama u Hrvatskoj.// Vjesnik bibliotekara Hrvatke 60, 1(2017), str. 107.
[12] Usp. IFLA Guidelines for digitization projects. Str. 6-7.,
URL: https://www.ifla.org/files/assets/preservation-and-conservation/publications/digitization-projects- g uidelines.pdf (2019-03-01)
[13] Šapro-Ficović, Marica. Masovna digitalizacija knjiga: utjecaj na knjižnice. // Vjesnik bibliotekara Hrvatske 54, 1/2(2011), str. 217. URL: https://hrcak.srce.hr/80483 (2019-03-01)

zbirke (iako se u konačnici zapravo stvara jedna velika zbirka), nego digitalizacija skoro svega, u ovom slučaju skoro svake tiskane knjige.[14] Primjeri su: Projekt Gutenberg, Open Content Alliance, Amazon Search Inside Books, European digital library, Million Book, Google Books i drugi. [15]

2.4.2. Europe 2020 i Europeana

Aleksandra Horvat u svom radu Digitalizacija i knjižnice govori o poznatom dokumentu *Europe 2020.* usvojenom u ožujku 2010. godine koji, među ostalim, potiče razvitak digitalne ekonomije što je posebno navedeno u popratnom dokumentu Digitalni plan za Europu (*A Digital Agenda for Europe*) iz kojeg je jasno kako se želi omogućiti da se većina dosadašnjih poslovnih transakcija i radnih postupaka može obaviti koristeći Internet.[16] U istom radu autorica navodi kako digitalni plan za Europu najavljuje stvaranje europske digitalne knjižnice s kojom bi, putem portala Europeana, građani Unije dobili pristup za oko dvije i po milijarde svezaka knjiga i časopisa.[17] Iz prethodno navedenog jasno je koliko je digitalizacija postala općim prioritetom što optičko prepoznavanje znakova također stavlja u prvi plan. Ono što je oduvijek bio izazov u ovom i sličnim poduhvatima je pribavljanje dopuštenja nositelja autorskih prava što je vidljivo i u Europeani u sklopu koje je digitalizirana samo tzv. slobodna građa, odnosno prvenstveno građa nastala prije dvadesetog stoljeća i početkom istog stoljeća.[18]

2.4.3. Strategija digitalizacije kulturne baštine 2020.

U sklopu dokumenta Strategija e-Hrvatska 2020. donesena je Strategija digitalizacije kulturne baštine 2020. kojom će se definirati načini i pravila digitalizacije muzejske, arhivske, knjižnične i audiovizualne građe prema standardima europske digitalne knjižnice Europeane.[19] Strategija e-Hrvatska 2020. također navodi kako će se u području kulturne i nacionalne baštine provesti konsolidacija i jačanje infrastrukture za digitalizaciju, korištenje i očuvanje digitalne kulturne baštine čime bi se omogućila koordinacija i upravljanje e-digitalnom kulturnom baštinom te pristup e-uslugama kulturne i nacionalne baštine.[20]

---

[14] Ibid.
[15] Ibid.
[16] Horvat, Aleksandra. Digitalizacija i knjižnice. // Vjesnik bibliotekara Hrvatske 55, 2(2012), str. 18-19. URL:https://www.hkdrustvo.hr/vjesnik-bibliotekara-hrvatske/index.php/vbh/article/view/307 (2019-03-02)
[17] Horvat, Aleksandra. Digitalizacija i knjižnice. // Vjesnik bibliotekara Hrvatske 55, 2(2012), str. 19.
[18] Horvat, Aleksandra. Op. cit, str. 19.
[19] Strategija e-Hrvatska 2020. URL: https://uprava.gov.hr/UserDocsImages/Istaknute%20teme/e-Hrvatska/Strategija_e-Hrvatska_2020.pdf str. 16. (2019-03-07)
[20] Ibid., str. 64.

2.4.4. Projekti digitalizacije u narodnim knjižnicama

Što se tiče digitalizacije u knjižnicama, autorice Seiter-Šverko i Križaj navode kako je hrvatskom Strategijom kulturnog razvitka u 21. st. kao jedan od ciljeva postavljeno donošenje nacionalnog plana digitalizacije knjižnične građe te njegova koordinacija s digitalizacijom analogne kulturne baštine, posebno muzejske i arhivske, no ovaj nacionalni plan još nije izrađen.[21] Iz istog rada saznajemo kako većina hrvatskih narodnih knjižnica nije pristupila postupcima digitalizacije iz različitih razloga kao što su nedostatak opreme, financijskih sredstava te stručnih djelatnika za postupke digitalizacije.[22] Autori Škrabo i Vrana navode da kako bi digitalizacija mogla biti uspješno provedena, knjižnica mora osigurati svoje raspoložive resurse i infrastrukturu koja podrazumijeva djelovanje politike, tehnologiju, financiranje, stručno znanje i dugoročnu obvezu ustanove te se ne smije improvizirati, pa je takvu aktivnosti potrebno planirati.[23]

2.4.5. Koraci projekta digitalizacije

Tekstualno se gradivo može digitalizirati na tri načina: ručnim prepisivanjem na računalo, skeniranjem i snimanjem digitalnim fotoaparatom, no prepisivanje je najdugotrajniji postupak te se malokad primjenjuje.[24] Koraci koji bi trebali biti primijenjeni u odvijanju bilo kojeg projekta digitalizacije su: odabir gradiva, digitalizacija gradiva, obrada i kontrola kvalitete, zaštita, pohrana i prijenos, pregled i korištenje te održavanje digitalnoga gradiva. Projekt digitalizacije, dakle, ukratko obuhvaća: 1) pripremu materijala za skeniranje; 2) skeniranje materijala, 3) automatsku obradu (OCR); 4) unos teksta u aplikaciju za obradu sekundarne dokumentacije te 5) računalno pohranjivanje u željenom formatu (npr. PDF) prema inventarnoj oznaci fonda sekundarne dokumentacije.[25]

[21] Seiter-Šverko, Dunja; Križaj Lana. Digitalizacija kulturne baštine u Republici Hrvatskoj: Od trenutne situacije prema nacionalnoj strategiji.// Vjesnik bibliotekara Hrvatske 55, 2(2012), str. 33.
[22] Ibid, str. 34.
[23] Škrabo, Katarina; Vrana, Radovan. Op. cit, str. 16.
[24] Ibid.
[25] Balog Vojak, Jelena; Šinkić, Zdenka. Projekt digitalizacije hemeroteke Hrvatskog povijesnog muzeja.//Informatica museologica 44, 1-4(2013), str. 178.
URL: https://hrcak.srce.hr/index.php?show=clanak&id_clanak_jezik=257178 (2019-03-02)

3. Optičko prepoznavanje znakova (OCR)

Optičko prepoznavanje znakova pojavilo se uslijed sve većeg razvoja digitalnih knjižnica, digitalnih poštanskih i bankovnih usluga te brojnih drugih zahtjeva koji su uključivali automatizaciju procesa. Prema autorima Karić i Krpić, optičko prepoznavanje se sastoji od slijedećih procesa: 1) dohvaćanja slike; 2) pretprocesiranja; 3) segmentacije; 4) ekstrahiranja značajki te 5) klasifikacije.[26] Isti autori navode kako su prva tri procesa optičkog prepoznavanja, posebno pretprocesiranje i segmentacija, vremenski najzahtjevniji te potom najviše utječu na kvalitetu optičkog prepoznavanja znakova. Kao što je za pretpostaviti, baze u koje su pohranjeni prethodno analizirani znakovi (slova, brojevi ili interpunkcijski znakovi) sadrže veliki broj uzoraka te su računala „istrenirana" kako da se snalaze u izrazito velikom broju situacija. Broj uzoraka u prethodno konstituiranim bazama ovisi od projekta do projekta. Koliko je neki proces optičkog prepoznavanja uspješniji od drugoga, ovisi o brzini optičkog prepoznavanja znakova, uspješnosti rezultata, razlici u broju prethodno procesiranih te potom pohranjenih znakova i drugim parametrima.[27] Iako se računala još uvijek ne mogu uspoređivati sa izvedbom čovjeka, optičko prepoznavanje znakova je proces pomoću kojeg se uspješno prevode rukom ispisani znakovi u računalno čitljive znakove. Počeci prepoznavanja znakova sežu u 19. stoljeće, kada je tijekom prve dekade napravljeno nekoliko pokušaja razvoja uređaja za slijepe osobe, s tadašnjom OCR tehnologijom.[28] S druge strane, autori Tafti et. al. navode kako je optičko prepoznavanje znakova klasično strojno učenje čija primjena je prepoznata u medicini, edukaciji, prilikom osiguranja i još mnogih aktivnosti u sklopu kojih se konvertiraju elektronički dokumenti u pretraživi tekst.[29] Sverastuća pojava digitalnog sadržaja postavila je optičko prepoznavanje znakova kao temelj i imperativ za svaku daljnju analizu teksta. Veliki broj podataka bio je pohranjen u papirnatom obliku sve do pojave skenera te potom i optičkog prepoznavanja znakova.

[26] Karić, Miran; Krpić, Zdravko. Optičko prepoznavanje znakova na grid i višejezgrenim platformama. // Tehnički vjesnik 20, 4(2013.), str. 647.
[27] Ibid.
[28] Ibid.
[29] Tafti, Ahmad P.; Baghaie, Ahmadreza; Assefi, Mehdi; Arabnia, Hamid R.; Yu, Zeyun; Peissig, Peggy. OCR as a Service: An Experimental Evaluation of Google Docs OCR, Tesseract; ABBYY FineReader and Transym.//ISVC (2014), str. 736. URL: https://www.researchgate.net/publication/310645810_OCR_as_a_Service_An_Experimental_Evaluation_of_Google_Docs_OCR_Tesseract_ABBYY_FineReader_and_Transym (2019-03-20)

3.1. Što je OCR?

Prema mrežnoj Britannica enciklopediji, OCR , odnosno Optičko Prepoznavanje Znakova (Optical Character Recognition), je tehnika skeniranja i usporedbe, čiji cilj je identifikacija tiskanog teksta ili brojčanih podataka.[30] Prema istom izvoru, aplikacije za optičko prepoznavanje znakova nastoje identificirati znakove i riječi uspoređujući oblike sa onima pohranjenima u programskoj biblioteci te s obzirom na blizinu znakova.[31] Prema Balog Vojak i Šinkić, optičko prepoznavanje znakova računalna je tehnologija koja omogućuje pretvorbu skeniranih papirnatih dokumenata, PDF dokumenata i slikovnih dokumenata snimljenih digitalnim fotoaparatom ili skeniranih, u formate koji se mogu uređivati.[32] Isti autori navode kako sustav OCR-a čine oprema (skener ili fotoaparat) i sofisticirani program.[33] Hairuman i Foong navode kako se radi o procesu u kojem se papirnati dokument optički skenira te potom konvertira u elektronički format na način da se svaki pojedinačni znak u dokumentu prvo "prepoznaje" te mu se zatim dodjeljuje simbolični identitet.[34] Prema Wallsu, optičko prepoznavanje znakova bavi se analizom znakova kojima pripisuje određenu vrijednost uspoređujući ih sa sustavom pohranjenih predložaka znakova te prema matematičkoj analizi značajki ili ekstrakcijom značajki.[35] Hmoumen Marouane optičko prepoznavanje znakova naziva sofisticiranim problemom zbog brojnih varijabli koje mogu utjecati na njega, kao što su različitost jezika, stilova i fontova teksta kao i zbog osvjetljenja koje je teško kontrolirati.[36] S obzirom na sve navedeno, Islam et. al. navode kako je potrebno znanje različitih disciplina računalne znanosti, kao što su procesiranje ili obrada slike, klasifikacija uzoraka, procesiranje prirodnog jezika i drugo, kako bi se riješili brojni izazovi. Vynckier navodi kako je optičko prepoznavanje znakova podgrana polja prepoznavanja uzoraka (pattern recognition) i temelji se na biometrijskim tehnologijama koje također omogućuju računalima prepoznavanja ljudskog glasa, lica, otiska prsta, očne mrežnice, ali se bitno razlikuju.[37]

---

[30] Optical character recognition. Britannica. URL: https://www.britannica.com/technology/OCR (2019-03-20)
[31] Ibid.
[32] Balog Vojak, Jelena; Šinkić, Zdenka. Op. cit, str. 178.
[33] Ibid.
[34] Intan, Fariza Bt Haruman; Foong, Oi-mean. OCR signage recognition with skew & slant correction for visually impaired people. // 11th International Conference on Hybrid Intelligent Systems, HIS 2011. URL: https://www.researchgate.net/publication/220981126_OCR_signage_recognition_with_skew_slant_correction_for_visually_impaired_people str. 306-310. (str. 306.)
[35] Walls, John. OCR and Content Management with SAP and Imaging (2008) URL: https://www.slideshare.net/verbella/ocr-and-content-management-with-sap-and-imaging (2019-05-25)
[36] Hmoumen, Marouane. A review of optical character recognition system. // Design of Machines and Structures 7, 2(2017), str. 5-12.
[37] Vynckier, Ivo. How OCR works. URL: http://www.how-ocr-works.com/ (2019-05-25)

3.2. Povijest OCR-a

Polje proučavanja optičkog prepoznavanja znakova prethodi i samoj pojavi računala. Prvi patenti povezani sa OCR tehnologijom datiraju već iz 1929. godine kada je austrijski inženjer Gustav Tauschek osmislio uređaj za optičko prepoznavanje, tada zvan "uređaj za čitanje" (Reading Machine).[38] Potom je slijedio patent Amerikanca Paula Handela 1933. godine, zvan "statistički uređaj" (Statistical Machine). Prema Taradiju, radilo se foto električnom mehanizmu u sklopu kojeg se optičko prepoznavanje znakova odvijalo uz pomoć svjetlosne zrake koja je prolazila kroz filtar s uzorcima znakova te ako bi dovoljno svjetlosti prošlo kroz njega te bi došlo do istovjetnosti sa znakom, informacija bi se formirala kao podudarnost uzorka i znaka.[39] Iz istog rada saznajemo kako jedna od prvih konkretnih primjena OCR programa datira iz 1956. godine kada je Pošta Sjedinjenih Američkih Država počela eksperimentirati na području skeniranja i optičkog prepoznavanja tiskanog teksta kako bi mogla skenirati i verificirati otisnuti poštanski broj i adresu. Eksperiment je funkcionirao na način da bi nakon skeniranja podataka računalo usporedilo skenirani poštanski broj i adresu s uzorcima koji su prethodno bili pohranjeni u memoriji.[40] Tek 1983. godine Pošta Sjedinjenih Američkih država je počela opremati svoje urede OCR sustavima koji su znatno ubrzali sortiranje pošiljaka a time se ubrzalo poslovanje te racionaliziralo troškove. Sustavi za prepoznavanje strojno tiskanog teksta datiraju iz kasnih 50-ih te doživljavaju široku primjenu na osobnim računalima početkom 90-ih. U početku su OCR sustavi radili mnogo pogrešaka, pa se pribjeglo stvaranju novih standardiziranih fontova. Prema Jurkoviću[41], novi fontovi OCRA i OCRB razvijeni su u 1970-ima od strane American National Standards Institute (ANSI) i the European Computer Manufactures Association (ECMA), a potom su prihvaćeni od međunarodne organizacije International Organization for Standardization (ISO) što je uzrokovalo njihovo veće korištenje u poslovnim sustavima i time omogućilo veću razinu raspoznavanja OCR sustava. Ipak najznačajniji napredak je direktno povezan sa elektroničkom erom. Obično, što je OCR tehnologija naprednija potrebnija je veća brzina

[38] History of computer. The Reading Machine (first OCR device) of Gustav Tauschek. URL: https://history-computer.com/ModernComputer/Basis/OCR.html (2019-05-25)
[39] Taradi, Ivan. Mogućnosti unapređenja programa za optičko prepoznavanje znakova (OCR programa). Završni rad. Str. 5. URL: http://darhiv.ffzg.unizg.hr/id/eprint/9125/1/03%20Taradi%20Ivan%20Mogucnosti%20poboljsanja%20OCR%20programa%20v2.pdf (2019-05-25)
[40] Taradi, Ivan. Mogućnosti unapređenja programa za optičko prepoznavanje znakova (OCR programa). Završni rad. Str. 5. URL: http://darhiv.ffzg.unizg.hr/id/eprint/9125/1/03%20Taradi%20Ivan%20Mogucnosti%20poboljsanja%20OCR%20programa%20v2.pdf (2019-05-25)
[41] Jurković, Mladen. Programski sustav za raspoznavanje tiskanog teksta. **(**Zav. Rad, Fakultet elektrotehnike i računarstva). Lipanj, 2009. URL: http://www.zemris.fer.hr/~kalfa/ZR/Jurkovic_ZR_2009.ppt (2019-05-25)

računalnog procesora. OCR i njoj srodna tehnologija ICR (Intelligent Character Recognition) odnosno pametno prepoznavanje znakova, mijenjaju način industrijske obrade dokumenata.

3.3. Vrste prepoznavanja znakova

Prema Wallsu[42], vrste prepoznavanja znakova su:

1) Optičko prepoznavanje znakova ili OCR (Optical Character recognition) o kojem je riječ u ovom radu, koristi se za prepoznavanje teksta unutar slike.
2) Optičko prepoznavanje znakova iz kvadratića ili OMR (Optical Mark Recognition) vrsta je optičkog prepoznavanja znakova za identifikaciju teksta koji se najčešće nalazi u posebno označenim kvadratićima (eng. Checkboxes) na poslovnim formularima.
3) Optičko prepoznavanje barkodova ili OBR (Optical Barcode Recognition) služi, kao što sam naziv kaže za identifikaciju barkodova a najjednostavniji primjer je upotreba u trgovinama.
4) Inteligentno prepoznavanje znakova ili ICR (Intelligent Character Recognition) služi za prepoznavanje rukom pisanog teksta ili ručno ispisanih dijelova na ispisanom (printanom) dokumentu. Stručno je definirano kao računalno prevođenje ručno unesenog teksta u računalno čitljive znakove.
5) Inteligentno prepoznavanje riječi ili IWR (Intelligent Word Recognition) koristi se za prepoznavanje riječi pisanih u kurzivu kao što su na primjer liječničke uputnice.

Lesić i Oblučar navode posebnu vrstu optičkog prepoznavanja glazbenih nota (eng. Optical music recognition), odnosno prevođenje notnog zapisa zapisanog na papiru u elektronički oblik i njegovo smještanje u korisniku razumljiv kontekst.[43]

[42] Walls, John. OCR and Content Management with SAP and Imaging. 2008. URL: https://www.slideshare.net/verbella/ocr-and-content-management-with-sap-and-imaging 8 (2019-05-25)
[43] Lesić, Dragan; Oblučar, Bojan. Optičko prepoznavanje glazbenog crtovlja. Fakultet elektrotenike i računarstva. Repozitorij. URL: https://www.fer.unizg.hr/_download/repository/Opticko_prepoznavanje_glazbenog_crtovlja.pdf (2019-06-01)

3.4. Upotreba OCR sustava

Optičko prepoznavanje znakova se ispočetka koristilo za sortiranje elektronske pošte a potom i za interpretiranje bankovnih računa i čekova, verifikaciju potpisa, procesiranje formulara i računa za plaćanje, provjeru valjanosti putovnica te za tablete sa digitalnim olovkama i pomaganje slijepim i slabovidnim osobama prilikom čitanja teksta.[44] Autori Chaudhuri et. al. prethodno navedenu upotrebu optičkog prepoznavanja znakova dijele na četiri posebno naglašena područja[45]:

1) Unošenje podataka (data entry area) – pokriva tehnologije za unos točno određenih podataka, koje su se prvotno koristile za bankarske aplikacije. Ove tehnologije dizajnirane su da optički prepoznaju korisničke brojeve, identifikacijske podatke, brojeve članaka, količinu novca i drugo. Identificira se samo ograničeni broj znakova te nekoliko specijalnih simbola. Procesira se oko 150 dokumenata po satu a postotak grešaka je od 0.0001 do 0.01 %.

2) Unošenje teksta (The text entry reading machines) – koriste se uređaji za optičko prepoznavanje tiskanih stranica teksta u sklopu uredske automatizacije. Ograničenja su format papira i široki raspon znakova te utječu na kvalitetu optičkog prepoznavanja znakova no unatoč svemu, točnost pojedinog znaka je od 0.01 do 0.1%.

3) Automatizirano procesiranje (Process Automation) – u ovom području glavna zadaća optičkog prepoznavanja znakova nije optičko prepoznavanje tiskanog dokumenta već upravljanje određenim procesom, odnosno sortiranjem pošte. Cilj je usmjeriti svako pismo u za njega prikladan sandučić pošte što ovisi koliko je informacija na pismu dostupno. Brzina sortiranja je obično oko 30 pisama na sat, velik broj pisama biva odbijen od strane sustava no prihvaćena pisma su u potpunosti točno sortirana.

4) Pomagala za slijepe (Aid for blind) – Uređaji za pomoć slijepim osobama pri čitanju pojavili su se netom prije ere računala, usporedno sa sustavima za sintezom govora. Pomoću OCR programa tiskani tekst se skenira, a potom pretvara u sintetički govor koji čita na glas.

---

[44] Islam, Norman; Islam, Zeeshan.; Noor, Nazia. A Survey on Optical Character Recognition System. // ITB Journal of Information and Communication Technology. 2015. URL: https://www.researchgate.net/publication/320442536_A_Survey_on_Optical_Character_Recognition_System (2019-06-01)

[45] Usp. Chaudhuri, Arindam…[et al.]. Optical Character Recognition Systems for Different Languages with Soft Computing.// Springer. Studies in Fuzziness and Soft Computing 352 (2017) str. 35. URL: https://www.researchgate.net/profile/Arindam_Chaudhuri2/publication/321518201_Optical_Character_Recognition_Systems_for_Different_Languages_with_Soft_Computing/links/5a5122e20f7e9bbc10543023/Optical-Character-Recognition-Systems-for-Different-Languages-with-Soft-Computing.pdf (2019-06-02)

Optičko prepoznavanje znakova također se koristi i za automatsko raspoznavanje registarskih tablica (ANPR – Automatic Number Plate Recognition). Martinović navodi kako je prvi takav sustav izumljen 1976. godine u Ujedinjenom Kraljevstvu.[46] Danas ANPR tehnologija omogućava stope prepoznavanja i do 98%. Kako bi se došlo do točne segmentacije i prepoznavanja znakova na registarskim tablicama, koriste se za to predviđeni algoritmi koji bi trebali ispravno izdvojiti znakove, odbacujući neispravne i prosljeđujući pronađene segmente za daljnju obradu. Prije klasifikacije ekstrahiraju se značajke pojedinih segmenata te se naposljetku tablice analiziraju prema skupu sintaktičkih pravila koja su različita za svaku državu pojedinačno.[47] U istom radu, redom je navedena procedura prepoznavanja registarskih tablica: 1) otkrivanje horizontalnih i vertikalnih rubova; 2) horizontalna i vertikalna projekcija; 3) lokalizacija tablice; 4) segmentacija uz pomoć binarizacije slike, horizontalne projekcije te izdvajanje znakova iz segmenata; 5) analiza znakova koja uključuje normalizaciju veličine te ekstrakciju značajki; 6) klasifikaciju znakova putem neuronskih mreža te 7) sintaksnu analizu prema različitim sintaksama različitih država.

3.5. Učinkovitost i točnost OCR sustava

Prema Balog Vojak i Šinkić, razne studije govore kako učinkovitost optičkog prepoznavanja znakova varira između 95 i 99 %, što nije dovoljno za apsolutno vjernu kopiju nekog dokumenta te nakon optičkog prepoznavanja znakova ručnu korekturu mora obaviti čovjek.[48] Autori Rice, Nagy i Natker navode kako je OCR tehnologija napredovala do točke kad se točnost od 99% i više rutinski postiže, no kao i kod svake tehnologije potrebna su dodatna poboljšanja i moramo imati u vidu da točnost i od 99% znači 30 pogrešaka na stranici od 3000 znakova.[49] Prema istraživanju istih autora, neki izvori grešaka češće su se pojavljivali kod određenog tipa dokumenta nego drugog pa tako mnogi tehnički izvještaji korišteni u istraživanju koji su fotokopije sadrže više defekata nego drugi dokumenti, dok su tamne

[46] Martinović, Anđelo. Raspoznavanje znakova na registarskim tablicama. (Zav. rad, Sveučilište u Zagrebu. Fakultet elektrotehnike i računarstva, 2008.) URL: http://www.zemris.fer.hr/~kalfa/ZR/Martinovic_ZR_2008.pdf (2019-05-27)
[47] Ibid.
[48] Balog Vojak, Jelena; Šinkić, Zdenka. Projekt digitalizacije hemeroteke hrvatskog povijesnog muzeja. // Iz muzejske teorije i prakse 44, 1-4(2013), str 179.
[49] Rice, Stephen V.; Nagy, George; Nartker, Thomas A. Optical Character Recognition: An Illustrated Guide to the Frontier.// The Springer International Series in Engineering and Computer Science. Springer Science & Business Media, 1999., str.58. URL: https://www.researchgate.net/publication/250171630_Optical_character_recognition_an_illustrated_guide_to_the_frontier (2019-05-28)

pozadine bile najčešći problem u člancima časopisa.[50] Prema Stipanović, točnost i pouzdanost optičkog prepoznavanja znakova ovisi o nizu čimbenika kao što su dimenzija i starost originala, kvaliteta papira, različite varijacije teksta (fontovi), lingvistička složenost pojedinih jezika, kvaliteta i rezolucija skenera, brzina prepoznavanja i drugo.[51] U istom radu, autor Stipanović navodi kako se kaže da je točnost i pouzdanost prihvatljiva pri onoj razini kod koje je rezultat optičkog prepoznavanja znakova prihvatljiviji od ručnog utipkavanja teksta, obično 98% i više. Prema Radoševiću, slijedeća važna komponenta mjerljivosti kvalitete optičkog prepoznavanja znakova je brzina prepoznavanja.[52] Da bi program za optičko prepoznavanje znakova bio vjerodostojan i upotrebljiv, tekst koji se želi analizirati mora se prepoznati barem onom brzinom kojom bi čovjek taj tekst čitao, što bi značilo nekoliko stotina znakova u minuti. Na današnjim računalima uglavnom se postiže prepoznavanje od najmanje tisuću znakova u minuti.

[50] Ibid. Str. 5.
[51] Stipanović, Zoran. Primjena OCR-a u vektorizaciji katastarskih planova. (dip. rad, Geodetski fakultet Zagreb, 2004) str. 31. URL: https://www.bib.irb.hr/147904 (2019-06-01)
[52] Radošević, Danijel. Postupci i problemi optičkog prepoznavanja teksta. // Zbornik radova 21(1996). Str. 18. URL: https://hrcak.srce.hr/index.php?show=clanak&id_clanak_jezik=117350 (2019-06-01)

4. Komponente OCR sustava

Postupak optičkog prepoznavanja znakova, prema Balog Vojak i Šinkić može se podijeliti na četiri glavne faze: 1) digitalizaciju (skeniranje ili digitalno fotografiranje) materijala; 2) računalnu obradu teksta; 3) korekturu obrađenog teksta te 4) spremanje teksta u željeni računalni format.[53] Radošević navodi kako se postupak optičkog prepoznavanja može podijeliti u tri faze[54]:

1) Dobivanje bitmape teksta na način da se tekst koji se želi prepoznati stavi u optički čitač (skener), preko kojega se bitmapa teksta unosi u računalo. Potrebna rezolucija skeniranja obično iznosi 200-300 DPI te uglavnom ovisi o veličini znakova od kojih se sastoji tekst.

2) Obrada bitmape teksta pomoću programa za optičko prepoznavanje teksta je središnja faza kojoj je cilj dobivanje teksta u UTF-8, ASCII formatu ili u nekom od formata za standardnu obradu teksta. Od analiziranih programa očekuje se točnost od najmanje 99% što uvelike ovisi i o kvaliteti teksta te kvaliteti skeniranja.[55]

3) Obrada teksta dobivenog u fazi 2 u nekom od standardnih programa za obradu teksta. Prvotno se ispravljaju pogreške nastale prilikom prepoznavanja teksta uz pomoć programa za ispravljanje gramatičkih grešaka (Speller) koji su uglavnom pisani za engleski jezik. Potom se podešavaju margine na stranici, poravnava se tekst, font, stil i ostalo u svrhu da tekst dobije svoj konačni oblik.

Glavni koncept optičkog prepoznavanja znakova odnosno automatskog prepoznavanja uzoraka se temelji na poučavanju uređaja da razlikuje koji set uzoraka bi se mogao pojaviti i kako bi mogao izgledati. Kod optičkog prepoznavanja znakova, uzorci su slova, brojevi i posebni simboli kao što je točka, upitnik i drugi dijakritički znakovi. Autori Chaudhuri et. al. navode kako uređaj uči primjere znakova različitih razreda te na temelju ovih primjera, gradi prototip opisa svakog razreda znakova.[56] Prilikom optičkog prepoznavanja, nepoznati znakovi se uspoređuju sa prethodno stvorenim opisima te se potom dodjeljuju razredu s kojim se najviše podudaraju. Kod većine komercijalnih sustava za optičko prepoznavanje znakova, proces učenja se izvodi unaprijed. Neki sustavi, međutim, uključuju sklopove za učenje novih

[53] Balog Vojak, Jelena; Šinkić, Zdenka. Projekt digitalizacije hemeroteke hrvatskog povijesnog muzeja. // Iz muzejske teorije i prakse 44, 1-4(2013), str 178.
[54] Radošević, Danijel. Postupci i problemi optičkog prepoznavanja teksta.// Zbornik radova 21(1996). Str. 18-19. URL: https://hrcak.srce.hr/index.php?show=clanak&id_clanak_jezik=117350 (2019-06-01)
[55] Radošević, Danijel. Postupci i problemi optičkog prepoznavanja teksta.// Zbornik radova 21(1996). Str. 18-19. URL: https://hrcak.srce.hr/index.php?show=clanak&id_clanak_jezik=117350 (2019-06-01)
[56] Usp. Chaudhuri, Arindam…[et al.]. Op. cit, str. 15.

razreda znakova. Prema Chaudhuri[57] et al. uobičajeni sustav za optičko prepoznavanje znakova se sastoji od sedam komponenti: 1) optičko skeniranje (optical scanning); 2) segmentacija regija slike (location segmentation); 3) pretprocesiranje (pre-processing); 4) segmentacija (segmentation); 5) reprezentacija (representation) 6) ekstrakcija značajki (feature extraction); 7) treniranje i prepoznavanje (training and recognition) te 8) postprocesiranje (post-processing). Za razliku od prethodnih autora, Ong i Suhatorno[58] navode kako su glavni zadaci optičkog prepoznavanja znakova: 1) dohvaćanje slike (image acquisition); 2) pretprocesiranje (preprocessing); 3) segmentacija (segmentation); 4) ekstrakcija značajki (feature extraction); 5) klasifikacija (classification) te 6) postprocesiranje (post processing). Mehta i Doshi proces optičkog prepoznavanja znakova dijele na: 1) pretprocesiranje; 2) procesiranje i 3) postprocesiranje. Ovaj rad će nadalje predstaviti detaljnije opis procesa optičkog prepoznavanja prema Chaudhuri et. al. Dakle, prvi korak optičkog prepoznavanja znakova je digitalizacija analognog dokumenta koristeći optički skener. Kada su definirane regije dokumenta koje sadrže tekst, svaki znak se ekstrahira putem procesa određivanja praga segmentacije (thresholding). Ekstrahirani znakovi se pretpocesiraju, eliminiraju se nejasnoće kako bi se omogućila ekstrakcija značajki. Identitet svakog znaka se utvrđuje uspoređivanjem ekstrahiranih značajki sa prethodno stvorenim opisima razreda znakova. Finalno, informacije dobivene iz cjelokupnog konteksta se koriste za rekonstrukciju riječi i brojeva originalnog teksta.

4.1. Optičko skeniranje

Prva komponenta optičkog prepoznavanja znakova je optičko skeniranje. Svaki dokument koji želimo digitalizirati odnosno provesti nad njim optičko prepoznavanje znakova prvo moramo skenirati skenerom ili fotografirati fotoaparatom. Autori Ong i Suhatorno ovaj prvi zadatak optičkog prepoznavanja znakova nazivaju dohvaćanje slike hardverom odnosno skenerom (Image Acquisition).[59] Za skeniranje se mogu koristiti različiti skeneri, no za veće projekte digitalizacije najpodobniji su protočni skeneri koji mogu automatizirano skenirati veći broj stranica odjednom. Skeniranjem se dobiva digitalna inačica originalnog dokumenta. Prema Chaudhuri et al. optički skener, kojim se obavlja skeniranje, se sastoji od transportnog

[57] Usp. Chaudhuri, Arindam…[et al.]. Op. cit, str. 16.
[58] Ong, Veronica; Suhatorno, Derwin. Using k-nearest neighbor in optical character recognition. // ComTech 7, 1(2016), str. 55. URL: https://media.neliti.com/media/publications/165987-EN-using-k-nearest-neighbor-in-optical-char.pdf (2019-10-09)
[59] Usp. Ong, Veronica; Suhatorno, Derwin. Op. cit, str. 55.

mehanizma i senzora koji konvertira intenzitet svjetla u sive nijanse.[60] „Prije skeniranja, analogni dokument je u RGB formatu te, da bi se procesom optičkog prepoznavanja dobili što točniji rezultati, treba biti konvertiran u sive tonove (grayscale).“ Autori Lesić i Oblučar ovaj korak optičkog prepoznavanja nazivaju pretprocesiranjem ulaznog parametra obrade, odnosno slike.[61] Slika u sivim tonovima odnosno grayscale slika sada je binarizirana odnosno prikazana bitovima od kojih 1 uglavnom predstavlja dio znaka odnosno slova, a 0 pozadinu. O ovom procesu više je riječ u slijedećem poglavlju o određivanju praga segmentacije.

4.2. Segmentiranje regija slike

Ovim elementom optičkog prepoznavanja teksta utvrđuju se sastavnice slike odnosno razlikuje se ispisani tekst od slika i grafičkih prikaza. Dakle, kada se primjenjuje na tekst, segmentacija je izolacija znakova ili riječi.[62] Većina OCR algoritama segmentira riječi u pojedinačne znakove. Ova tehnika je jednostavna za primjenu ali se problemi javljaju ukoliko se znakovi dotiču ili se sastoje od više dijelova. Glavni problemi prilikom segmentacije su: 1) ekstrakcija znakova koji se dodiruju ili su višedijelni; 2) razlikovanje buke (šuma) od teksta; 3) zamjena grafičkih i geometrijskih prikaza s tekstom i obrnuto.[63]

4.2.1. Thresholding ili određivanje praga segmentacije

Prema Lesiću, tiskani dokument (analogni), kao što je prethodno navedeno, se uglavnom sastoji od crnog tiska na bijeloj pozadini te kada se provodi optičko prepoznavanje znakova, višerazinska slika se konvertira u dvorazinsku crno bijelu sliku.[64] Ovaj proces skenera vrši se određivanjem praga segmentacije (thresholding) i koristi se za uštedu memorije računala. U istom radu se navodi kako je proces određivanja praga vrlo bitan zato što rezultat optičkog prepoznavanja u potpunosti ovisi o kvaliteti dvorazinske slike.[65] Međutim, u praksi se rijetko nalaze dokumenti sa adekvatnim kontrastom stoga određivanje praga segmentacije uvelike ovisi o tome koliko je prilagođen kontrast i svjetlina dokumenta. Šekoranja objašnjava osnovni algoritam thresholdinga koji transformira ulaznu sliku u binariziranu (segmentiranu) sliku i to na način: T je prag thresholdinga odnosno segmentacije te ukoliko objektni pikseli

[60] Usp. Chaudhuri, Arindam…[et al.]. Op. cit, str. 16.
[61] Usp. Lesić, Dragan; Oblučar, Bojan. Optičko prepoznavanje glazbenog crtovlja. (Predavanje, Fakultet elektrotenike i računarstva) URL: https://www.fer.unizg.hr/_download/repository/Opticko_prepoznavanje_glazbenog_crtovlja.pdf (2019-06-03)
[62] Usp. Chaudhuri, Arindam…[et al.].Op. cit, str. 17.
[63] Ibid.
[64] Usp. Lesić, Dragan; Oblučar, Bojan. Optičko prepoznavanje glazbenog crtovlja. (Predavanj bzte, Fakultet elektrotenike i računarstva) URL: https://www.fer.unizg.hr/_download/repository/Opticko_prepoznavanje_glazbenog_crtovlja.pdf (2019-06-03)
[65] Ibid. str. 16.

g(i,j) prelaze prag segmentacije ili su isti pragu m, predstavljaju sliku ili tekst dok preostali objekti f(i,j) koji ne prelaze prag predstavljaju pozadinu.[66] Prema mrežnoj stranici Cvision, thresholding ili određivanje praga je najjednostavnija metoda grupiranja slike u regije.[67] Rezultat procesa segmentacije su dvije vrste piksela: 1) oni koji imaju vrijednost 1 predstavljaju tekst te 2) oni koji imaju vrijednost 0 i predstavljaju pozadinu. Autor Šekoranja navodi kako brojni objekti slike posjeduju karakterističnu konstantnu refleksiju ili apsorpciju svjetla na njihovim površinama što omogućuje određivanje konstantnog iznosa odnosno praga koji razdvaja objekte od pozadine.[68] Isti autor ističe kako je određivanje praga računski nezahtjevna i jednostavna metoda – to je najstarija metoda koja još ima široku primjenu za jednostavnije zadatke.[69]

Postoji nekoliko različitih metoda za odabir "statičnog" praga. Najjednostavnija metoda bi bila odabir srednje vrijednosti piksela slike i ona bi vrlo dobro funkcionirala za fotografije bez šuma (without noise), međutim to često nije slučaj. Sofisticiraniji pristup bio bi stvaranje histograma koji prikazuje intenzitet piksela slike te se uzima srednja vrijednost. Histogram prvog reda predstavlja relativnu frekvenciju svjetlina točaka u slici.[70] Pristup sa histogramom pretpostavlja da pozadina i tekst imaju neku srednju vrijednost u pikselima, kao i da stvarne vrijednosti piksela variraju oko utvrđene srednje vrijednosti. Ukoliko na slici postoji više različitih regija takvih da je svjetlina točaka unutar regije otprilike jednaka, primjenjuje se bimodalan ili multimodalan histogram.[71]

4.2.2. Amplitudna segmentacija

Najjednostavniji pristup segmentaciji slike sivih razina baziran je na korištenju histograma relativnih frekvencija svjetlina točaka na slici, odnosno jednostavnije histogramom intenziteta sivih razina. Histogram predstavlja standardni način zapisa statističke distribucije frekvencija nijansi sive koji se dobija diskretizacijom domene u konačan broj ćelija i prebrojavanjem piksela slike koji se nalaze u svakoj od čelija (na sličan način može se izračunati i histogram distribucije boja diskretizacijom 3D domene boja u 3D čelije). Histogram može biti bimodalan ili multimodalan. Kod bimodalnog histograma svjetlina točaka pada ili u svijetle ili

[66] Usp. Bojan, Šekoranja. Segmentacija slike. (Predavanje, Fakultet strojarstva i brodogradnje u Zagrebu) URL: http://www.sjever.fsb.hr/vizija/predavanja/Segmentacija%20slike-sekoranja.ppt (2019-06-04)
[67] Cvisiontech. OCR software primer. Thresholding with OCR. URL: http://www.cvisiontech.com/resources/ocr-primer/thresholding-within-ocr.html (2019-06-08)
[68] Bojan, Šekoranja. Op. Cit., URL: http://www.sjever.fsb.hr/vizija/predavanja/Segmentacija%20slike-sekoranja.ppt (2019-06-04)
[69] Ibid.
[70] Ibid.
[71] Lončarić, Sven. Op. Cit.

u tamne točke te se u obzir uzima minimum između dva vrha histograma. Takva je na primjer segmentacija kada se želi izdvojiti tamni tekst na svijetloj pozadini. Multimodalan histogram sadrži više izraženih vrhova svjetline točaka te se određuje pripadnost svakog piksela slike jednom od dominantnih vrhova, za što se najčešće koristi metoda k-sredina.[72]

Sven Lončarić u svom predavanju o segmentaciji slike objašnjava kako se pragovi prilikom amplitudne segmentacije određuju pomoću minimuma u histogramu na način da određeni dio točaka ima svjetlinu nižu od praga te takve točke nisu objekti od interesa.[73] U istom predavanju navedeno je kako je amplitudna segmentacija korisna kad amplitudne značajke dovoljno precizno definiraju regije scene.

U svom doktorskom radu Krstinić ističe kako je segmentacija proces dijeljenja slike na homogene segmente (cjeline ili regije) te je ekvivalentna uočavanju rubova, tj. pronalaženju granica između nepovezanih regija na slici.[74] Krstinić navodi kako bi segmentirana područja na slici trebala bi biti jednolika i homogena s obzirom na neku karakteristiku kao što je boja ili tekstura te bi unutrašnjost područja trebala biti jednostavna i bez mnogo malih rupa a susjedna područja na segmentaciji trebala bi se značajnije razlikovati s obzirom na neku karakteristiku dok bi granice trebale bi biti jednostavne, glatke i prostorno točne." [75] Segmentacija slike se bavi dekompozicijom scene u dijelove koje imaju sljedeće karakteristike: a) regije su uniformne s obzirom na neko svojstvo ; b) granice regija moraju biti jednostavne; c) regije ne smiju imati male otvore; d) susjedne regije se moraju značajno razlikovati.[76]

4.3. Pretprocesiranje

Jednostavno rečeno, pretprocesiranje je ispravljanje eventualnih grešaka nastalih prilikom snimanja digitalne slike. Važnost ovog koraka je što nakon skeniranja slika često sadrži određenu količinu nejasnoća ili buke (noise in text) te ju je potrebno "očistiti" a potom i izdvojiti zone interesa. Ovisno o rezoluciji skenera i postavljenom pragu (thresholding), znakovi mogu biti razmazani ili nečitki te neki od ovih defekata uzrokuju teže optičko prepoznavanje znakova no pretprocesiranjem bivaju eliminirani.[77]

[72] Ibid.
[73] Lončarić, Sven. Segmentacija slike. (Predavanje, Fakultet elektrotehnike i računarstva u Zagrebu) URL: https://www.fer.unizg.hr/_download/repository/09-OI-SegmentacijaSlike%5B2%5D.pdf (2019-07-09)
[74] Krstinić, Damir. Op. cit, str. 21.
[75] Ibid.
[76] Lončarić, Sven Op.cit.
[77] Usp. Chaudhuri, Arindam…[et al.].Op. cit, str. 18.

Cilj pretprocesiranja je dobivanje podataka pomoću kojih OCR sustavi mogu lako postići točnost. Prije izvedbe glavne analize podataka odnosno znakova, ovaj korak je posebno važan. Glavni koraci pretprocesiranja su dakle: a) smanjenje šuma (noise reduction), b) normalizacija podataka te c) kompresija podataka koji se žele zadržati.[78]

4.3.1. Smanjenje šuma

Fotografije nastale skeniranjem vrlo često imaju mnogo nedostataka te je najčešće potrebno smanjiti šum na fotografijama ili izoštriti rubove znakova, popuniti male rupe ili ukloniti dodatnu tintu na rubovima znakova i slično. Važno je napomenuti kako različiti autori navode različite izvore šuma koji se u glavnini podudaraju ali postoje varijacije. U članku "Nova metoda za smanjenje buke prilikom OCR procesa" navedeni su slijedeći izvori šuma[79]:

4.3.1.1. Šum nastao zbog linija na linijskom papiru

Ova vrsta šuma nastaje kada se linija presijeca sa tekstom, kada debljina linije varira pa stvara probleme algoritmima za smanjenje buke, kada je linija izlomljena i kada se linija slova poklapa sa linijom papira kao na primjer kod slova z (tada algoritmi eliminiraju slovo z jer ga ne prepoznaju). Metode za uklanjanje ovog tipa šuma podijeljene su u dvije velike grupe: a) metode bazirane na matematičkoj morfologiji ovise o prethodno stečenom znanju i b) metode koje koriste Houghovu transformaciju (Hough Transform) – služe za izdvajanje značajki i pronalazak linija u svim smjerovima.[80]

4.3.1.2. Šum nalik potezu kista

Autori Subduhi, Sahu i Mohapatra ističu kako je ovakva vrsta šuma slična dijakritičkim znakovima te njegova blizina tekstu ili tekstualnim komponentama može promijeniti značenje riječi.[81] Navedeni šum u tekstu nastaje najvećim dijelom zbog neuspjelog uklanjanja linija na linijskom papiru ili pozadine koja je bila veća od teksta. „To se dešava kada su linije na linijskom papiru razlomljene ili nejasne te se ne mogu primijeniti uobičajene tehnike kao Houghova transformacija ili projekcijski profili te je predložena metoda razlikovanja linije i teksta uz pomoć algoritma koji prvo ekstrahira istaknute značajke komponenti teksta uz pomoć klasifikacije pod nadzorom te potom koristi njihovu povezanost / kohezivnost i širinu

[78] Ibid.
[79] Subduhi, Rajesh Kumar; Sahu, Biphuprasad; Mohapatra, Pratyush Rn. A Novel Noise Reduction Method For OCR System. // IJCST 5, 2(2014), str. 83-84., URL: http://ijcst.com/vol5/spl2/ec1145.pdf (2019-07-04)
[80] Subduhi, Rajesh Kumar; Sahu, Biphuprasad; Mohapatra, Pratyush Rn. A Novel Noise Reduction Method For OCR System. // IJCST 5, 2(2014), str. 83-84., URL: http://ijcst.com/vol5/spl2/ec1145.pdf (2019-07-04)
[81] Usp. Subduhi, Rajesh Kumar; Sahu, Biphuprasad; Mohapatra, Pratyush Rn. Op. cit, str. 84.

poteza kako bi se izvukle manje komponente teksta koristeći nenadziranu klasifikacijsku tehniku.“[82]

4.3.1.3. Marginalan šum

Marginalan šum čine tamne sjene koje se pojavljuju na vertikalnim ili horizontalnim marginama slike a često su rezultat debljine dokumenta koji se skenira. Prema Subduhi, Sahu i Mohapatra, za uklanjanje ovakve vrste šuma autori navode kako se koristi Zheng Zhang metoda vertikalne projekcije na način da se identificira na kojoj se strani uzorak nalazi te potom na temelju ekstrahiranih značajki uzorka utvrdi razgraničenje između sjena i čistog područja te kao bolju metodu autori navode identifikaciju komponenti teksta jer je jednostavnije pronaći uzorke teksta nego značajke šuma.[83]

4.3.1.4. Šum “sol i papar” ili impulsni šum

Šum „sol i papar“ se najčešće pojavljuje zbog nečistoće na dokumentu ili papiru i može se sastojati od jednog ili više piksela ali je uglavnom manji nego objekti teksta odnosno slova i dijakritički znakovi te ukoliko je ovaj tip šuma izoliran lako ga se može ukloniti filterima kao što je medijan, no ukoliko je količina šuma veća koriste se algoritmi kao k-fill ili morfološki operatori.[84]“Sol i papar” šum izgleda kao da je nestalo tinte u pisaču prilikom ispisa te ukoliko je fragmentiranost visoka, smanjuje se segmentacija a potom i točnost optičkog prepoznavanja.

4.3.1.5. Pozadinski šum

Ovakva vrsta šuma najčešća je na povijesnim rukopisima i pojavljuje se u obliku degradacija kao što su nejednak kontrast, ispresijecane linije, pozadinske točke, vlaga na papiru ili neravnost papira. Metode za uklanjanje ove vrste šuma su[85]:

1) Metode koje koriste neizrazitu (fuzzy) logiku i temelje se na poboljšanju kvalitete slike operatorima neizrazite logike na način da se odrede ključne značajke proporcionalne nekim značajkama slike kao što su prosječni intenzitet ili pojačani kontrast.

2) Metode temeljene na histogramu odnosno grafičkom prikazu intenziteta na slici. Iscrtava se određeni broj piksela za svaku vrijednost intenziteta. Histogram za vrlo tamnu sliku će većinu podataka sadržavati na lijevoj strani i na centralnom dijelu grafičkog prikaza. S druge strane,

[82] Ibid.
[83] Ibid.
[84] Ibid.
[85] Usp. Subduhi, Rajesh Kumar; Sahu, Biphuprasad; Mohapatra, Pratyush Rn. Op. cit, str. 84.

histogram za vrlo svijetlu sliku sa malo zatamnjenih područja imat će većinu podataka na desnoj strani i centralnom dijelu grafičkog prikaza. 2001. godine predložena je metoda POSHE (Partially Overlapped Sub-Block Histogram Equalization) u kojoj je slika podijeljena u blokove te se za svaki blok zasebno radi histogram. Ova metoda je jedna od boljih za poboljšanje kontrasta jer koristi ekstrahiranje značajki i kada se kombinira sa neizrazitim (fuzzy) operatorima za poboljšanje pozadinske kvalitete dolazi se do točnijih rezultata.

Metode bazirane na matematičkoj morfologiji su vrlo učinkovite za poboljšanje nejednake pozadine. Morfološki operatori su vrlo moćan alat za procesiranje i analiziranje oblika koji imaju konture, površinu i drugo. Metodama ove tehnike uočavaju se uzorci šuma koji se pojavljuju kao sjene u pozadini strukturalnih elemenata odnosno znakova, potom morfološki operatori za istanjavanje i čišćenje uklanjaju sjene. Neki od algoritama ove metodologije počinju već sa fazom pretprocesiranja.

4.3.2. Normalizacija podataka

Usporedno sa izglađivanjem, pred-procesiranje uključuje i normalizaciju. Normalizacija znakova se smatra jednim od najbitnijih koraka prilikom obrade dokumenta. Cilj normalizacije znaka je smanjenje varijacije znaka te njegova preobrazba u standardni oblik, na način da se poboljša njegova orijentacija, veličina, nagib i debljina poteza, kako bi se olakšala ekstrakcija (izvlačenje) značajki te potom i točnost optičkog prepoznavanja znakova.[86] Neke od uobičajenih metoda normalizacije su 1) ispravljanje ukošenosti linije teksta te ekstrahiranje (dobijanje) osnovne linije (skew normalization), 2) ispravljanje nagiba samog znaka (slant normalization), 3) normalizacija veličine te 4) izglađivanje kontura.[87]

4.3.2.1. Ispravljanje ukošenosti linije teksta

Prilikom procesa skeniranja, kao i kod rukom pisanog teksta, često se događaju netočnosti i to uglavnom zbog blage ukošenosti ili zaobljenosti teksta u odnosu na čitavu sliku, odnosno u odnosu na x os. To može uvelike utjecati na učinkovitost algoritama koji se provode prilikom procesa optičkog prepoznavanja teksta te bi se netočnosti trebalo ustanoviti (detektirati) i potom ispraviti. Neki znakovi se optički prepoznaju tek prema poziciji u odnosu na osnovnu liniju, kao što su broj 9 ili slovo g. Neke od metoda ekstrahiranja (dobivanja) osnovne linije su korištenje projekcijskog profila slike, grupiranje prema algoritmu najbližih susjeda (K-

[86] Usp. Chaudhuri, Arindam…[et al.]. Op. cit, str. 18.
[87] Ibid.

nearest neighbor algorithm), Houghova transformacija, metoda kros korelacije i drugo. [88]

4.3.2.2. Metoda projekcijskog profila

Metode projekcijskog profila, uglavnom, su ograničene na procjenjivanje kuta nagiba sa odstupanjem ±10 stupnjeva.[89] Prema Brodicu, riječ je o vrlo jednostavnoj kompjutorski nezahtjevnoj metodi koja koristi kao osnovu binarnu sliku za kreiranje histograma intenziteta piksela horizontalno ili vertikalno a nagib teksta se procjenjuje prema najvišem vrhu histograma.[90] Isti autor navodi kako se ova metoda ne može koristiti za analizu teksta koji je podijeljen u više kolumni ili stupaca. Ova metoda nije osjetljiva na šum, što je vrlo pozitivno.

4.3.2.3. Algoritam k-najbližih susjeda

Algoritam najbližih susjeda je algoritam koji se koristi pri treniranju računala za prepoznavanje uzoraka. Smatra se jednim od deset najutjecajnijih algoritama za rudarenje podataka u znanstvenoj zajednici. Spada u neparametrijski tip algoritama, što znači da ne pretpostavlja svoje okruženje.[91] Jednostavno rečeno, algoritam k-najbližih susjeda djeluje na način da se prvotno odredi k odnosno broj s koliko susjednih znakova je „blizu“ i koji su znakovi „slični“. Problem ovog algoritma je postojanje velikog broja atributa sličnosti i teško je odrediti udaljenost svih primjera u memoriji i odlučiti se koji su najbliži susjedi.

Autori Lu i Lim Tam u svom radu o normalizaciji ukošenosti teksta metodom lanca najbližih susjeda (nearest neighbor chain NNC) govore kako ova metoda u odnosu na mnoge druge kao što su metode temeljene na projekcijskim profilima, Houghovoj transformaciji, kros korelaciji ili morfološkoj transformaciji, daje najtočniju procjenu kuta nagiba teksta iz razloga što sve druge spomenute metode daju loše rezultate ukoliko se radi o kompleksnijem tekstu koji sadrži na primjer različite vrste fontova, različita pisma, neuobičajeno postavljen tekst ili regije sa grafičkim prikazima i tablicama.[92] Analiziranjem više različitih izvora literature, dolazi se do zaključka kako ova metoda funkcionira na povezivanju susjednih znakova u vektore nakon čega se iz histograma utvrđuje njihov smjer a konačno i nagib.

---

[88] Ibid, str. 19.
[89] Panwar, Subhash; Nain, Neeta. A Novel Approach of Skew Normalization for Handwritten Text Lines and Words. (Eighth International Conference on Signal Image Technology and Internet Based Systems, 2012), str. 296. URL: https://www.researchgate.net/publication/235218361_A_Novel_Approach_of_Skew_Normalization_for_Handwritten_Text_Lines_and_Words/download (2019-07-04)
[90] Brodić, Darko…[et al.]. Op. Cit.
[91] Ong, Veronica; Suhatorno, Derwin. Op. cit, str.55.
[92] Lu, Yue; Lim Tan, Chew. A nearest-neighbor chain based approach to skew estimation in document images. // Pattern Recognition Letters 24 (2003), str. 2316. URL: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.121.3963&rep=rep1&type=pdf (2019-06-07)

4.3.2.4. Houghova transformacija

Objekti na slici, kao što su linije, elipse, kružnice i drugo, lokaliziraju se uz pomoć Houghove transformacije. Prema Vrhovski i Herčeki, ova metoda, prije same transformacije, podrazumijeva lokalizaciju rubova objekata koji se nalaze na slici.

Često korištena Houghova transformacija omogućuje lokalizaciju rubova objekata koji se nalaze na slici. Pretpostavka je da su prije transformacije detektirani rubovi. Kombiniranjem rubnih piksela detektiraju se linije, a njihovim kombiniranjem i složenije konture. Za potrebe daljnje analize koristit će se osnovna linijska Houghova transformacija. Nakon detekcije rubova Sobelovim gradijentnim filtrom dobiva se slika rubova koja predstavlja ulazni podatak za Houghovu transformaciju. Vrhovski navodi da je radni prostor transformacije ravnina s dvije vrste piksela: a) pikselima koji predstavljaju rub i b) pikselima koji predstavljaju pozadinu.[93] Zato se obično koriste monokromatske slike, iako to nije uvjet, a cilj Houghovog algoritma je pronaći pravac koji je predstavljen najvećim brojem piksela koji leže na tom pravcu.[94] Donju granicu broja piksela koji predstavljaju pravac moguće je mijenjati s obzirom na praćeni prostor. Prema Vrhovskom, prednost Houghove transformacije je u tome što ona može lokalizirati pravac, iako se on sastoji od više segmenata (razlomljeni pravac) te je lokalizirani pravac ulazni podatak za algoritam navođenja mobilnog robota u smjeru ravne linije. [95]

4.3.2.5. Ispravljanje ukošenosti te standardizacija znaka

Ovaj segment optičkog prepoznavanja znakova najčešće se koristi za tekst u kurzivu ili rukom pisani ukošeni tekst koji varira od osobe do osobe. Nagib koji se ispravlja je kut između horizontalne linije teksta ili niza znakova i njihove vertikalne osi. Znakovi teksta ili slova mogu biti nagnuta u lijevo ili desno. Prema Chaudhuri, normalizacija ukošenosti linije teksta se koristi za normalizaciju svih znakova u standardni oblik.[96] Najuobičajenija metoda za procjenu kuta nagiba znaka je izračun prosječnog kuta nagiba obližnjih vertikalnih elemenata. Koordinate početne i završne točke svakog linijskog elementa daju kut nagiba.

Ispravljanje veličine znaka je jedna od strategija normalizacije znaka koja se koristi za smanjenje varijacije znaka. Prema autoru Rajeshu, normalizacija znaka se smatra najvažnijim

---

[93] Vrhovski, Z. Lokalizacija ravne linije u slikovnoj sekvenci. // Tehnički glasnik 5, 2(2011), str. 7. URL: https://hrcak.srce.hr/index.php?show=clanak&id_clanak_jezik=124687 (2019-06-07)
[94] Ibid.
[95] Usp. Vrhovski, Z. Op. cit, str. 10.
[96] Usp. Chaudhuri, Arindam…[et al.]., str. 19.

postupkom pretprocesiranja. Autori Lei He et. al. napravili su istraživanje koliko različita veličina znaka, primjenjujući iste klasifikatore i značajke, utječe na rezultat te su došli do zaključaka kako prilikom optičkog prepoznavanja rukom pisanog teksta dolazi do kvalitetnijih rezultata ukoliko se radi o većem fontu znakova/slova.[97] Točnije, povećanje slike veličine 20x20 na 26x26 uvelike poboljšava uspjeh optičkog prepoznavanja znakova dok ga distorcije na slici umanjuju.[98] Iako normalizacija slika na veću veličinu čini optičko prepoznavanje kvalitetnijim, radi se također o većem financijskom trošku.

4.3.2.6. Izglađivanje kontura

Slika u originalnom obliku prvo je binarizirana uz pomoć različite metode (na primjer Otsu metoda) te potom slijedi izglađivanje kontura (smoothing process) kako bi se smanjila buka odnosno neravnine na linijama slova. Prema Eikvilu, izglađivanje uključuje i ispunjavanje (rupa) i istanjavanje.[99] Izglađivanjem se eliminiraju manja udubljenja ili rupe u znaku dok istanjavanje smanjuje širinu linije.

4.3.3. Kompresija

Optičko prepoznavanje znakova zahtijeva kompresijske tehnike prilikom čijeg korištenja se ne gube nikakve informacije odnosno gdje se zadržava prvotni oblik informacije. U digitalnom svijetu govori se sažimanju bez gubitaka (lossless compression) u odnosu na sažimanje sa gubicima (lossy compression). Dok bi se u klasičnom slučaju slika sačuvala u TIFF (Tagged Image File Format) stilu, prilikom optičkog prepoznavanja znakova, prema autoru Chaudhuri, dvije su tehnike sažimanja među najpopularnijima: 1) thresholding ili određivanje praga segmentacije gdje se, kako bi se smanjili podaci za pohranu ili povećala brzina procesiranja, slike konvertiraju iz sivog prikaza ili prikaza u boji u binarizirani prikaz te 2) istanjavanje.[100]

[97] Usp. Lei He, Chun et. Al. The Role of Size Normalization on the Recognition Rate of Handwritten Numerals. (Centre for Pattern Recognition and Machine Intelligence, Concordia University Montreal, Quebec, Canada., 2009.) URL: http://www.iapr-tc11.org/archive/icdar2009/papers/3725a451.pdf (2019-07-03)
[98] Ibid.
[99] Eikvil, Line. Optical Character recognition. URL: https://www.nr.no/~eikvil/OCR.pdf (2019-06-09)
[100] Usp. Chaudhuri, Arindam…[et al.]., str. 21.

4.3.3.1. Određivanje praga

Kao što je prethodno navedeno, thresholding se provodi kako bi se poboljšala kvaliteta procesa optičkog prepoznavanja znakova, ubrzao proces optičkog prepoznavanja znakova te smanjili podaci za pohranu. Autor Caudhuri navodi da se za određivanje praga segmentacije koriste globalne i lokalne ili adaptivne thresholding tehnike u sklopu kojih se dodjeljuju različite vrijednosti svakom pojedinačnom pikselu.[101] U istom dijelu navodi se kako se u sklopu globalnog određivanja praga segmentacije odabire jedan prag za cijeli slikovni prikaz znaka dok se za lokalno ili adaptivno određivanje praga segmentacije koriste različite vrijednosti za svaki piksel pojedinačno.[102] Autor Gross drugim riječima govori o odabiru statičnog praga koji se najčešće koristi kada se radi o crnom tekstu na bijeloj pozadini odnosno o slici bez šuma i varijabilnog praga segmentacije koji se koristi kada analizirana slika sadrži na primjer pozadinu u boji sa svijetlim tekstom u prvom planu ili obrnuto.

4.3.3.2. Istanjavanje

Istanjavanje je tehnika koja uvelike smanjuje količinu podataka te ekstrahira konačan oblik znaka koji se potom može kvalitetnije optički prepoznati te potom i klasificirati. Prema Chaudhuri, postoje dvije mogućnosti istanjavanja: 1) kada se znak istanji na širinu linije od jednog piksela te 2) kada se odredu centralne linije znaka bez preispitivanja pojedinačnih piksela.[103] Isti autor navodi neke od preostalih metoda za istanjavanje kao što je metoda temeljena na klasterima (kada se kostur znaka uzima kao osnova podjele), te sekvencijski algoritmi.

4.4. Segmentacija

„Pojam segmentacije ili dekompozicije u računarskoj znanosti predstavlja proces rastavljanja kompleksnog sustava ili problema na više manjih, međusobno nepreklapajućih cjelina koje je lakše pojmiti, shvatiti i analizirati.“[104] Prema autorima Kaur, Baghla i Kumar, segmentacija je faza nakon pretprocesiranja kada se analizirani dokument segmentira odnosno dijeli na manje potkomponente ili logičke cjeline kao što su odvajanje teksta od grafičkih prikaza,

[101] Ibid.
[102] Ibid.
[103] Usp. Gross, Ari. Understanding OCR Technology. Thresholding within OCR. URL: http://www.cvisiontech.com/resources/ocr-primer/thresholding-within-ocr.html (2019-07-04)
[104] Križaić, Damjan. Postupci segmentacije objekata zadanih poligonalnom mrežom. Str. 2. (dip.rad, Fakultet elektrotehnike i računarstva u Zagrebu) URL: https://bib.irb.hr/datoteka/714522.diplomski-0036442298.pdf (2019-07-03)

pojedinačnih linija od paragrafa, znakova od riječi i drugo.[105] Ovo je jedna od najznačajnijih faza optičkog prepoznavanja znakova jer i najmanja pogreška prilikom prepoznavanja znaka, može utjecati na cjelokupan rezultat procesa optičkog prepoznavanja znakova. Ova metoda ima široku primjenu, od prepoznavanja registarskih tablica do analize medicinskih slika te segmentacijskom analizom jednog od ovih primjera, na primjer medicinskih slika, prema istraživanju autora Martinović, Stoić i Kiš, može se jednostavno predočiti proces segmentacije u sklopu kojeg se u prvoj fazi izrađuju algoritmi te se potom iz podataka izdvajaju obilježja koja kasnije služe kao ulaz u neuronsku mrežu da bi se nakon toga označilo regije i iscrtalo rezultat.[106]

Prema Chaudhuri, segmentacija se može podijeliti u tri kategorije[107]:

1) eksplicitna segmentacija – dijelovi se identificiraju na temelju značajki znaka
2) implicitna segmentacija – na slici se traže komponente koje se podudaraju sa prethodno pohranjenim predlošcima
3) mješovita segmentacija – kombinira eksplicitnu i implicitnu segmentaciju

4.5. Reprezentacija

Prema Chaudhuri, peta komponenta optičkog prepoznavanja znakova je reprezentacija slike te ukoliko se želi optički prepoznati kompleksnije prikaze nego što su binarizirane ili slike u sivim tonovima te postići veća točnost algoritama, potrebno je korištenje kompaktnije i karakterističnije reprezentacije.[108] Zato što se radi o kompleksnijim i karakterističnijim prikazima odnosno reprezentacijama, ova komponenta prethodi ekstrahiranju značajki kako bi se što jednostavnije izlučile značajke. Autor kategorizira reprezentacijske metode u četiri grupe: a) globalne, b) statističke i c) geometrijske i d) topološke.[109]

[105] Usp. Kaur, Amendeep; Baghla, Seema; Kumar, Sunil. Study of various character segmentation techniques for handwritten off-line cursive words: a review. // International Journal of Advances in Science Engineering and Technology 3, 3(2015) , Str. 154. URL: http://www.iraj.in/journal/journal_file/journal_pdf/6-162-1440573382154-158.pdf (2019-07-04)
[106] Usp. Martinović, Marko; Stoić, Antun; Kiš, Darko. Segmentacija CT slike pomoću samo-organizirajućih neuronskih mreža. // Tehnički vjesnik 15, 4(2008.), str. 23. URL: file:///C:/Users/BTP/Downloads/tv_15_2008_4_023_028.pdf (2019-08-07)
[107] Usp. Martinović, Marko; Stoić, Antun; Kiš, Darko. Segmentacija CT slike pomoću samo-organizirajućih neuronskih mreža. // Tehnički vjesnik 15, 4(2008.), str. 23. URL: file:///C:/Users/BTP/Downloads/tv_15_2008_4_023_028.pdf (2019-08-07)
[108] Ibid, str. 25.
[109] Ibid.

4.6. Ekstrahiranje značajki

Nakon dohvaćanja slike, optičkog skeniranja, dobivanja slike u sivim tonovima, pretprocesiranja i izlučivanja pojedinačnih znakova slijedi ekstrahiranje značajki znakova te potom njihova klasifikacija. Prema Trier, Jain i Taxt, ekstrahiranje značajki je izlučivanje pojedinačnih informacija iz sirovih podataka te je najvažnije za proces klasifikacije, u smislu da se varijabilnost ili razlikovnost uzoraka ili znakova unutar razreda smanji, a poveća između pojedinačnih razreda.[110] Međutim, određena metoda ekstrahiranja značajki koja je bila uspješna za rješavanje jednog problema optičkog prepoznavanja znakova ne mora biti uspješna za neki slijedeći iz razloga što se može raditi o drugačijem fontu teksta, orijentaciji, nagibu, rukom pisanom tekstu, osvjetljenju, šumu u tekstu, degradiranim znakovima i drugo. Pronalaženje ključnih obilježja, bez točnog podudaranja sa zadanim predloškom, je sastavnica optičkog prepoznavanja, odnosno radi se o inteligentnom računalnom prepoznavanju znakova čija je osnova pretraživanje osnovnih oblika kao što su otvorene površine, zatvoreni oblici, dijagonalne linije i drugo.[111]

4.7. Poučavanje i prepoznavanje

Poučavanje i prepoznavanje je sedma komponenta optičkog prepoznavanja znakova. Cilj poučavanja i prepoznavanja je prvotno izučiti računalo da prepoznaje određene znakove koji se potom na temelju sličnih značajki smještaju u određeni razred. Sustavi za optičko prepoznavanje naveliko koriste metodologije prepoznavanja uzoraka te potom njihovog pripisivanja unaprijed definiranom razredu znakova. Kada se govori o poučavanju računala misli se na strojno učenje. "Strojno učenje je grana umjetne inteligencije, znanstvena disciplina koja se bavi razvojem i konstrukcijom algoritama koji kao ulaznu vrijednost uzimaju empirijske podatke, npr. podaci iz baze podataka ili trenutno očitanje senzora, a daju predikacijski mehanizam na temelju generiranih podataka."[112] Drugim riječima, algoritmi koji se koriste pri strojnom učenju temelje se na podacima nastalima putem istraživanja te se nakon toga uspostavlja predikacija odnosno postupak u logici kojim se o subjektu nešto izriče pridodajući mu neki drugi pojam. Prema istom autoru, glavni fokus strojnog učenja je dizajn algoritma koji može prepoznati kompleksne uzorke i donijeti inteligentnu odluku baziranu na

[110] Trier, Ovind Due; Jain, Anil K.; Taxt, Torfinn. Feature extraction methods for character recognition – A Survey.//Pattern Recognition 29, 4(1996), str.2., URL: http://www.ee.bgu.ac.il/~dinstein/stip2002/FeatureExtractionReviewTrierJainTaxt95.pdf (2019-07-05)
[111] Usp. Stanić Loknar, Nikolina. Optical Character Recognition ili OCR. str. 24. (Predavanje, Grafički fakultet Zagreb) URL: http://slog.grf.unizg.hr/media/Optical%20Character%20Recognition%20ili%20OCR.ppt, (2019-07-06)
[112] Usp. Chaudhuri, Arindam …[et al.]. Op. cit, str. 29.

ulaznom podatku.[113]

Prema Chaudhuri, postoje tri glavna načina prepoznavanja uzoraka[114]: 1) podudaranje predloška (template matching), 2) statističke tehnike (statistical techniques) i 3) umjetne neuroske mreže (ANN – artificial neuron net).

4.7.1. Podudaranje predloška

Optičko prepoznavanje znakova najvećim dijelom temelji se na uspoređivanju znaka odnosno slova koje se želi prepoznati sa unaprijed strojno naučenim razredima znakova. Svaki razred znakova određen je kao takav na temelju sličnih obilježja ili značajki. Prema Chaudhuri, tehnike za optičko prepoznavanje znakova variraju ovisno o odabranom skupu značajki koje mogu biti jednostavne kao gray-level okviri sa znakovima ili kompliciranе kao grafički prikazi najprimitivnijih oblika znakova.[115] Autor Stanić Loknar ovu metodu uspoređivanja što skener vidi kao slovni znak sa popisom slovnih matrica ili predložaka, naziva matrix matchingom.[116] Prema istom radu, kada skenirana slika odgovara jednoj od zadanih matrica unutar postavljenog stupnja sličnosti, računalo joj dodjeljuje kod jednog od ASCII znakova.[117] Najjednostavnija je tehnika i temelji se na podudaranju znaka koji se optički prepoznaje sa pohranjenim prototipovima znakova, gdje se određuje stupanj sličnosti između dva vektora kao što su grupa piksela, oblici ili zakrivljenosti.[118]

4.7.1.1. Direktno podudaranje

Kod direktnog podudaranja, analizirani znakovi se uspoređuju sa standardiziranim skupom pohranjenih prototipova znakova te se pritom koriste različite mjere sličnosti (euclidian, yule i dr.) prema kojima se analizirani znakovi klasterskom analizom grupiraju u određene klase.[119] Prema Stjepanović, izračunava se udaljenost između uzoraka te se potom dodjeljuje klasa koja najviše odgovara ulaznom znaku.[120] Ova tehnika podudaranja može biti jednostavna kao što je pojedinačno uspoređivanje uzorka sa uzorkom ili vrlo kompleksna kada se koriste analize

[113] Ibid.
[114] Ibid.
[115] Ibid.
[116] Usp. Stanić Loknar, Nikolina. Optical Character Recognition ili OCR. URL: http://slog.grf.unizg.hr/media/Optical%20Character%20Recognition%20ili%20OCR.ppt
[117] Usp. Stanić Loknar. Op. cit.
[118] Usp. Chaudhuri, Arindam …[et al.]. Op. cit, str. 29.
[119] Ibid.
[120] Usp. Stjepanović, Nikolina. OCR tehnologije za digitalizaciju sadržaja. Str. 27.(Zav. rad, Odjel za informacijsko - komunikacijske tehnologije u Puli). URL: https://repozitorij.unipu.hr/islandora/object/unipu:2342/preview (2019-09-08)

putem stabla odlučivanja.[121] Međutim, autori Due Trier, Jain i Taxt navode kako se direktnim podudaranjem nikada ne može postići apsolutna točnost s obzirom da se pikseli analiziranog znaka rijetko kada točno poklapaju sa onima kod pohranjenih predložaka.

4.7.1.2. Deformirani predlošci i elastično podudaranje

Alternativna vrsta podudaranja je kada se deformirani predložak uspoređuje sa bazom prethodno pohranjenih uzoraka te se dobiva mjera nesličnosti proizašla iz količine deformacija te iz informacija koliko se poklapaju linije deformiranog predloška i onoga koji je prethodno pohranjen u bazi podataka. Osnovna ideja elastičnog podudaranja je optimalno podudaranje znaka koji se želi analizirati sa svim mogućim varijantama pohranjenog znaka.

4.7.2. Statističke tehnike

4.7.2.1. Parametrijsko i neparametrijsko prepoznavanje

Chaudhuri navodi kako se neparametrijsko prepoznavanje znakova koristi kada nije dostupno a priori znanje o znakovima koji su prethodno analizirani i pohranjeni, a najpoznatija metoda neparametrijske klasifikacije je Neuronska mreža (ANN – Artificial Neural Network).[122] Ovom tehnikom razdvajaju se različiti razredi klasa u hiperprostoru. Isti autor ističe kako parametrijsko prepoznavanje znakova, s druge strane, podrazumijeva dostupnost informacija o znakovima koji su prethodno analizirani i pohranjeni te se na temelju ovih informacija ususatvlja parametrijski model za svaki znak.[123]

4.7.2.2. Klasterska analiza

Sustavi za optičko prepoznavanje vrlo su učinkoviti prilikom prepoznavanja datog im teksta iz razloga što su za to „istrenirani“ ili strojno naučeni. Ono što predstavlja izazov je kako smanjiti troškove i gubitak vremena utrošenog na strojno učenje. Nameće se ideja grupiranja značajki znakova prema nekim svojstvima u različite razrede. Prema Jajuga, Sokolowski i Bock, klasterske tehnike mogu biti kategorizirane u dvije kategorije: hijerarhijske i nehijerarhijske odnosno partitivne.[124] Šmuc u sklopu svog predavanja ističe kako se radi o tehnikama strojnog učenja bez nadzora te kako se pod hijerarhijske tehnike podrazumijevaju hijerarhijski algoritmi (aglomerativni i divizivni) te se primjeri odnosno znakovi definiraju u grupe i podgrupe, dok partitivne tehnike obuhvaćaju partitivne algoritme gdje se primjeri

[121] Ibid.
[122] Usp. Chaudhuri, Arindam …[et al.]. Op. cit, str. 31.
[123] Ibid.
[124] Jajuga, Krzystof; Sokolowski, Andrzej; Bock, Hans-Hermann. Classification, Clustering and Dana Analysis: Recent Advances and Applications. Springer, 2002., str. 81.

odnosno znakovi grupiraju u distinktivne grupe bez podgrupa. Znakovi pojedinačnih klastera trebaju imati slične značajke koje se trebaju bitno razlikovati od značajki znakova neke druge skupine.

4.7.2.3. Skriveni Markovljevi modeli

Prema Chaudhuri, skriveni Markovljevi lanci su vrlo upotrebljavana i uspješna tehnika za optičko prepoznavanje rukom pisanog teksta.[125] Autorica Blažević navodi da je Markovljev model naziv za model u teoriji vjerojatnosti kojim se modeliraju sustavi koji imaju slučajne promjene stanja te se u ovom modelu pretpostavlja kako naredno stanje ovisi o određenom broju prethodnih stanja.[126] Sažeto rečeno, na temelju poznatog parametra, potrebno je predvidjeti parametar koji nas zanima, u slučaju optičkog prepoznavanja znakova, prilikom razrješavanja enigme o kojem se znaku radi, sustav prelazi iz jedne mogućnosti u drugu dok cijeli znak nije prepoznat u odnosu na već prepoznate znakove. Radi se o sustavu gdje se analiziraju prepoznati znakovi (vidljiva stanja), potom uzimaju u obzir brojne mogućnosti (skrivena stanja). Matematičkim vokabularom, ovakva uvjetna vjerojatnost može se zapisati prema Bayesovom pravilu. Ovaj model ima primjenu u prepoznavanju govora, pisanja, bioinformatici i drugim znanostima.

4.7.2.4. Neizrazit skup

Lotfi Zadeh je 1965. razvio fuzzy teoriju skupova, čija glavna ideja je da element neizrazitog odnosno fuzzy skupa ima odgovarajući stupanj pripadnosti.[127] Fuzzy logiku možemo lako predočiti ako se zapitamo koji dani su tjednu su dio vikenda: subota – u potpunosti točno (vrijednost 1); nedjelja – u potpunosti točno (vrijednost 1); utorak – u potpunosti netočno (vrijednost 0); petak – većim dijelom da, ali ne u potpunosti (vrijednost 0.8).[128] Elementi fuzzy skupa služe da opišu sličnosti značajki znakova na način da se znakovi promatraju kao skup točaka koji se uspoređuje sa referentnim uzorcima prema mjerama fuzzy skupa.[129] U sklopu optičkog prepoznavanja znakova, svaki novi znak se uspoređuje sa svim referentnim znakovima pohranjenima u razredima znakova te se pripisuje onom razredu znakova odnosno točno onom znaku s kojim ima najviše sličnosti. Ovaj tip razlučivanja koji znak predstavlja

[125] Usp. Chaudhuri, Arindam et. al. Op. cit, str. 32.
[126] Usp. Blažević, Antonela. Primjena skrivenih Markovljevih modela za modeliranje i predviđanje meteoroloških pojava. (dip. rad, Fakultet elektrotehnike i računarstvu u Zagrebu, 2017.), str. 1.
[127] Petrić Maretić, Grgur. Fuzzy logika. (Diplomski rad. Prirodoslovno matematički fakultet., 2010.) URL: https://www.math.pmf.unizg.hr/sites/default/files/pictures/petric-maretic-fuzzy-logika.pdf (2019-07-08)
[128] Umjetna inteligencija – Neizrazita logika – Skupovi. Predavanje. Fakultet prometnih znanosti. URL: http://www.fpz.unizg.hr/hgold/umin20092010/predavanja/47895-47816-uminteli-200809-03-fl01-skupovi.pdf (2019-07-08)
[129] Usp. Chaudhuri, Arindam…[et al.]. Op. cit, str. 32.

koje slovo se najčešće koristi za rukom pisani tekst gdje se svaki znak predstavlja grafom, a iz dobivenog algoritma se zaključuje o kojem znaku je riječ.[130]

4.7.3. Umjetne neuronske mreže

Umjetne neuronske mreže (Artificial Neuron Networks – ANN) danas se najčešće koriste za rješavanje problema za optičko prepoznavanje znakova, no kao i kod prethodnih metoda, sve ovisi o odabiru kvalitetnih klasifikatora. Prema autoru Gross, neuronske mreže su kolekcije matematičkih modela koje imitiraju značajke biološkog živčanog sustava kao i njegovu sposobnost prilagodljivog biološkog učenja.[131] Vynckier ističe kako su neuronske mreže korisne prilikom organizacije i klasterizacije podataka, rudarenja podataka te prepoznavanja znakova (teksta, medicinskih slika, glasova i drugo).[132]

Prema Ujević Andrijić, svaka neuronska mreža sastoji se od ulaznog, skrivenog i izlaznog sloja te se prilikom prolaska informacije kroz neuronsku mrežu generira izračunatu vrijednost koja se potom uspoređuje sa stvarnom vrijednošću te neuronska mreža na taj način uči predviđati.[133] U istom radu navedena je i topologija umjetnih neuronskih mreža[134]:

1) „umjetni neuron“ prima ulazne informacije koje su određene težinskim koeficijentima što je slično primanju ulaznog signala putem dendrita u sklopu bioloških neurona.
2) obrada informacija unutarnjim pragom (biasom) što je isto kao obrada signala u somi.
3) pretvaranje ulaza u izlaz (prijenosna funkcija) što je u biologiji funkcija aksona.
4)slanje informacija prema izlazu i sljedećim neuronima kao i slanje informacija putem sinapsi.

Hamad i Kaya također navode kako se radi o kompleksnoj vrsti arhitekture koja uključuje velik broj paralelno povezanih procesora (čvorova) gdje se izlazna informacija iz jednog čvora šalje u drugi čvor te konačan izbor informacije ovisi o složenoj suradnji svih čvorova.[135]

---

[130] Ibid.
[131] Gross, Ari. Understanding OCR Technology. Cvisiontech.com. URL: http://www.cvisiontech.com/resources/ocr-primer/ocr-neural-networks-and-other-machine-learning-techniques.html (2019-07-09)
[132] Vynckier, Ivo. A close look at optical character recognition. URL: http://www.how-ocr-works.com/index.html (2019-08-07)
[133] Ujević Andrijić, Željka. Umjetne neuronske mreže.// Kemija u industriji: Časopis kemičara i kemijskih inženjera Hrvatske 68, 5-6(2019.), str. 219.-220. URL: https://hrcak.srce.hr/220716 (2019-07-09)
[134] Ujević Andrijić, Željka. Umjetne neuronske mreže.// Kemija u industriji: Časopis kemičara i kemijskih inženjera Hrvatske 68, 5-6(2019.), str. 219.-220. URL: https://hrcak.srce.hr/220716 (2019-07-09)
[135] Hamad, Karez. Kaya, Mehmet. A Detailed Analysis of Optical Character Recognition Technology. // International Journal of Applied Mathematics, Electronics and Computers 4(2016), str. 247. URL:

4.8. Postprocesiranje

Naposljetku, kada je tekst optički prepoznat, potrebno je ustanoviti da li konačna inačica teksta i dalje sadrži određene greške. Chaudhuri navodi da je prva metoda postprocesiranja klasifikacija simbola koji su blizu u dokumentu ili imaju iste značajke (npr. udaljenost ili oblik slova), no problemi nastaju kada se pokušava prepoznati rukom pisani tekst čije su najčešće karakteristike ukošenost teksta te specifičnost oblikovanja znakova.[136] Prema Chaudhuri, ovaj proces se rješava uz pomoć osme komponente optičkog prepoznavanja znakova, postprocesiranja koje obuhvaća dvije potkomponente: 1) detekciju (pronalazak) te 2) korekciju otkrivenih grešaka.[137] Isti autori navode kako se greške u tekstu pronalaze pomoću dvije metode: 1) prepoznaju se individualni simboli u tekstu (interpunkcijski znakovi, slova i brojevi) zasebno ili u odnosu na druge znakove ili riječi te se grupiraju u nizove (strings) ovisno o lokaciji u dokumentu te se 2) riječi iz teksta uspoređuju sa riječima iz rječnika, što je vremenski vrlo zahtjevno.[138] Uspješnost ove komponente optičkog prepoznavanja znakova najviše ovisi o prethodne dvije komponente optičkog prepoznavanja znakova: ekstrakciji značajki te poučavanju računala za prepoznavanje znakova. Svaki od analiziranih znakova ili riječi uspoređuje se prethodno usustavljenim bazama znakova, te se traži najsličniji „kandidat“ dok ne dođe do podudaranja. Međutim, niti jedan sustav za optičko prepoznavanje znakova ne omogućuje apsolutno točnu detekciju te ispravljanje pronađenih grešaka, naročito ako se koristi samo jedna metoda za pronalazak te korekciju grešaka.

https://www.researchgate.net/publication/311851325_A_Detailed_Analysis_of_Optical_Character_Recognition_Technology (2019-08-07)

[136] Ibid.

[137] Usp. Chaudhuri, Arindam…[et al.]. Op. cit, str. 34.

[138] Ibid, str. 35.

5. Istraživački dio

5.1. Uvod u istraživanje aplikacija za optičko prepoznavanje znakova

Ovim istraživanjem doći će se do određenih zaključaka koliko je tehnologija ovog dijela znanosti napredovala analizirajući opcije te kolika je kvaliteta rezultata rada pet različitih aplikacija za optičko prepoznavanje znakova (ABBYY FineReader 15, Free OCR, Google Drive OCR, I2OCR i Convertio). U prethodnom teorijskom okviru bilo je više riječi o tome što se događa u pozadini procesa optičkog prepoznavanja znakova dok će se u istraživačkom dijelu rada analizirati praktičan rad u aplikacijama za optičko prepoznavanje znakova. Iza svakog postupka provedenog u nekom od ovih programa i dalje se u pozadini „skrivaju“ algoritmi i složeni matematički postupci što će se možda najbolje očitovati u situaciji kada program napravi grešku, što znači da neki od postupaka navedenih u prethodnom dijelu rada nije uspješno proveden. O nekima od ispitanih programa možemo pročitati na službenim mrežnim stranicama no to je neusporedivo sa konkretnim radom u programu te dobivanjem stvarne slike o stanju situacije.

5.2. Cilj i svrha istraživanja

Cilj ovog istraživanja je usporediti rezultate optičkog prepoznavanja znakova pet različitih aplikacija namijenjenih u tu svrhu te odgovoriti na isto toliko istraživačkih pitanja s ciljem dobivanja rezultata na osnovu kojih se mogu donijeti određeni zaključci o aplikacijama. Ovim istraživanjem nastojalo bi se saznati koji su nedostaci a koje prednosti pojedinačnih aplikacija za optičko prepoznavanje znakova. Posebno bi se pokušalo odgovoriti na slijedećih pet istraživačkih pitanja:

1. Da li je analizirana aplikacija otvorenog koda i kako ju se može koristiti?
2. Kakve mogućnosti pohrane nudi analizirana aplikacija?
3. Koje su dodane vrijednosti svake od aplikacija (faktor iznenađenja)?
4. Koliko jezika podržava analizirana aplikacija i da li je hrvatski jedan od njih?
5. S kolikom uspješnošću se mogu optički prepoznati slijedeći dokumenti:

A) tekst iz knjige na hrvatskom jeziku,

B) tekst iz knjige na engleskom jeziku,

C) zgužvani tekst,

D) tekst pisan rukom,

E) članak na hrvatskom jeziku,

F) članak na engleskom jeziku,

G) tekst pisan vrlo sitnim fontom,

F) tekst sa neuobičajenim fontom = strip,

H) svijetli tekst sa tamnom pozadinom

U istraživanju se polazi od pretpostavke da su neke aplikacije iznimno sofisticirane, kao što je na primjer ABBYY FineReader 15, u odnosu na druge jednostavnije, što ne mora utjecati na rezultate optičkog prepoznavanja znakova. Čak je za očekivati da će neke jednostavnije aplikacije pokazati bolje rezultate od nekih kompleksnijih. Aplikacije ABBYY FineReader, Convertio i Google Docs OCR izgrađene su od zasebne arhitekture dok se I2OCR, Free OCR i Simple OCR zasnivaju na Tesseract arhitekturi.

Svrha ovog istraživanja je usporediti značajke i uspješnost aplikacija za optičko prepoznavanje znakova te uspoređujući broj grešaka odnosno postotak uspješnosti optičkog prepoznavanja doći do određenih zaključaka.

5.3. Metodologija

U istraživanju se koristilo pet aplikacija za optičko prepoznavanje znakova (ABBYY Finereader 15, Free OCR, Google Drive OCR, I2OCR, Convertio) za slijedeće odabrane vrste dokumenta: tekst iz knjige, zgužvani tekst, tekst pisan rukom, članak iz novina, tekst pisan vrlo sitnim fontom, tekst sa neuobičajenim fontom odnosno strip, tekst koji je mnogo puta kopiran te svijetli tekst sa tamnom pozadinom. Od pet istraženih aplikacija tri su preuzete sa Interneta dok su preostale mrežne aplikacije s kojima se radilo mrežno kako bi se istražilo koje rezultate one daju u odnosu na aplikacije koje su zasebno instalirane na računalo. Tekstualni primjeri uzeti za analizu su iz različitih izvora (fizičkih i mrežnih) kako bi se pokazala uspješnost optičkog prepoznavanja znakova ne samo za prepoznavanje običnog teksta kao što je tekst u knjigama već i za primjere kao što su isječci iz stripa, rukom pisani tekst, zgužvani tekst. Samo neki od primjera popraćeni su slikovnim prikazima (screenshots) preuzetima iz programa za optičko prepoznavanje. U slučajevima gdje program omogućuje skeniranje, slike su skenirane te potom analizirane. U preostalim slučajevima primjeri su preuzeti sa Interneta. Koristeći skener, svaka stranica digitalizirana je u rezoluciji od 300 točaka po inču tj. 300 dpi (dots per inch). U sklopu istraživanja vezanih za opis osobina programa za optičko skeniranje odgovaralo se na istraživačka pitanja za svaki pojedini program kako bi se mogla provesti rasprava a potom izvesti zaključci, dok se za kvantitativni dio istraživanja istražio broj grešaka te uspoređivao između različitih programa.

5.4. Primjeri i analiza

5.4.1. ABBYY FineReader 15[139]

1. Da li je analizirana aplikacija otvorenog koda?

Ova aplikacija nije otvorenog koda, odnosno njezin dizajn ili arhitektura nisu javno dostupni i ne mogu se mijenjati od strane korisnika.

2. Kakve mogućnosti pohrane nudi analizirana aplikacija?

U sklopu aplikacije ABBYY Fine Reader procesirani dokumenti se mogu pohraniti u PDF, Word, Excell, EPUB, HTML, .txt, i još nekim formatima kao što su .djvu, .fb2 i .csv.

3. Koje su dodane vrijednosti ove aplikacije (faktor iznenađenja)?

Dodane vrijednosti aplikacije ABBYY FineReader 15 su što automatski odvaja slikovne prikaze, zaglavlja i podnožja od teksta (text i picture funkcija). Sve funkcije su pregledne i razumljive. (Slika 1.) Također, simultano se može raditi sa više OCR projekata odjednom. (Slika 2.)

Slika 1. Osnovne funkcije ABBYY FineReadera 1

Slika 2. Rad u ABBYY programu s tri dokumenta istovremeno

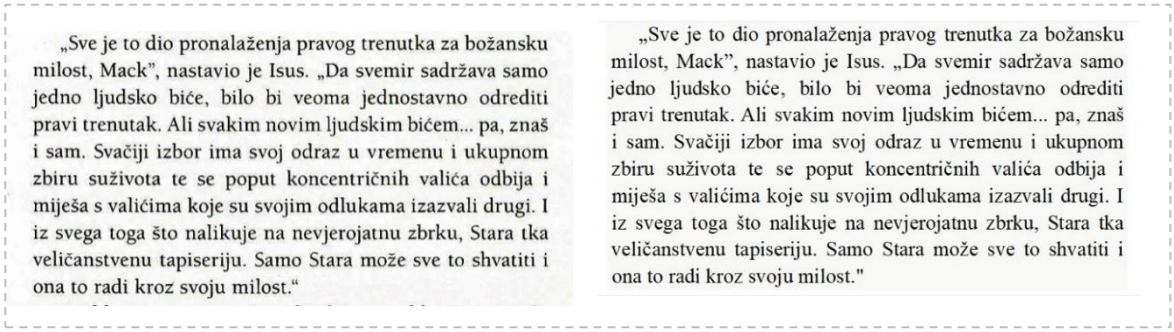
4. Koliko jezika podržava analizirana aplikacija i da li je hrvatski jedan od njih?

U sklopu aplikacije ABBYY FineReader 15 podržana su 192 jezika od kojih 42 jezika imaju gramatičku podršku u sklopu programa. Među ponuđenim jezicima je i hrvatski jezik.

[139] ABBYY FineRedaer 15. URL: https://pdf.abbyy.com/?utm_autosource=google&utm_automedium=cpc&utm_autocampaign=EEU_FineReader_Search_OCR&gclid=Cj0KCQjwreT8BRDTARIsAJLI0KLDwAcs5rceRU8610GS-mYf33sOuaeFSx2kGKirAKUK4d67eD65Wf4aAsfsEALw_wcB (2019-08-08)

5. S kolikom uspješnošću se mogu optički prepoznati slijedeći dokumenti:

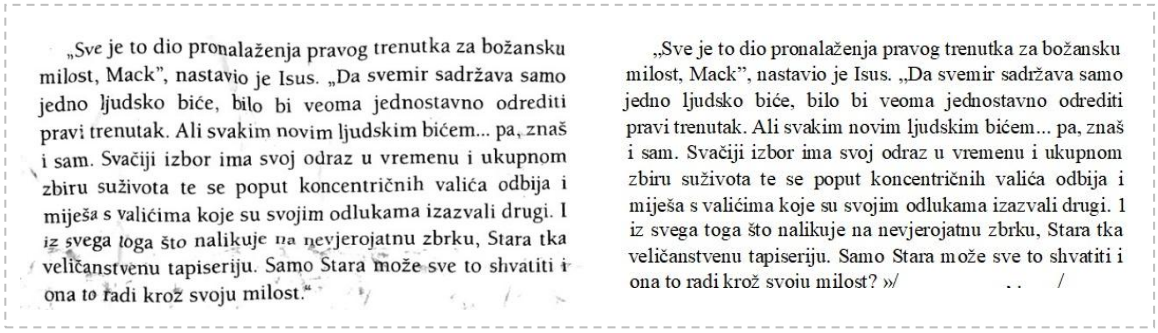
A) Tekst iz knjige



Slika 3. Tekst iz knjige prije i nakon procesa OCR – a (ABBYY FineReader 15)

Uz pomoć aplikacije ABBYY FineReader 15, tekst iz knjige[140] prepoznat je sa 100% - tnom točnošću, odnosno u potpunosti je prepoznato 437 slova kao i 16 interpunkcijskih znakova.

B) Zgužvani tekst iz knjige

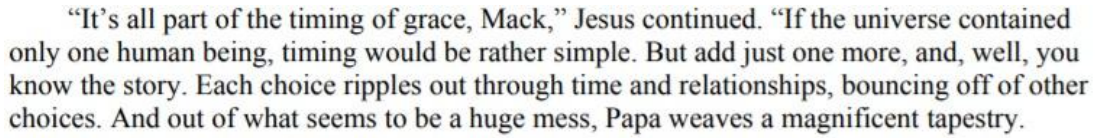


Slika 4. Zgužvani tekst iz knjige prije i nakon procesa OCR-a (ABBYY FineReader 15)

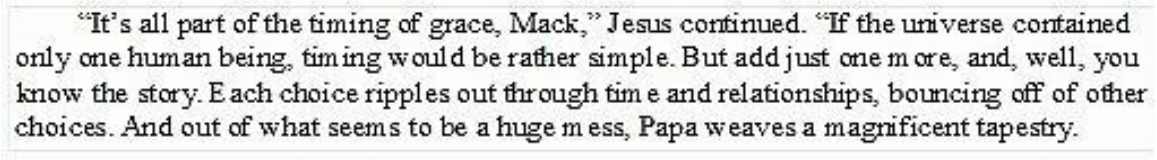
Što se tiče optičkog prepoznavanja slova, zgužvani tekst iz knjige prepoznat je sa točnošću od 99.54% odnosno samo dva slova od 437 slova su krivo prepoznata. Prilikom optičkog prepoznavanja interpunkcijskih znakova, samo jedan znak od 16 interpunkcijskih znakova (93.75%) je krivo prepoznat te su neke od nečistoća na papiru prepoznate kao 4 dodatna interpunkcijska znaka, što je navedeno samo kao zanimljiv podatak ali nije statistički analizirano. Ukupna točnost prilikom prepoznavanja cijelog teksta (slova i interpunkcijski znakovi) je 99.33 %.

[140] Young, William Paul. Koliba. Zagreb: Naklada Ljevak, 2008. , str 34.

C) Tekst iz knjige na engleskom jeziku


"It's all part of the timing of grace, Mack," Jesus continued. "If the universe contained only one human being, timing would be rather simple. But add just one more, and, well, you know the story. Each choice ripples out through time and relationships, bouncing off of other choices. And out of what seems to be a huge mess, Papa weaves a magnificent tapestry.

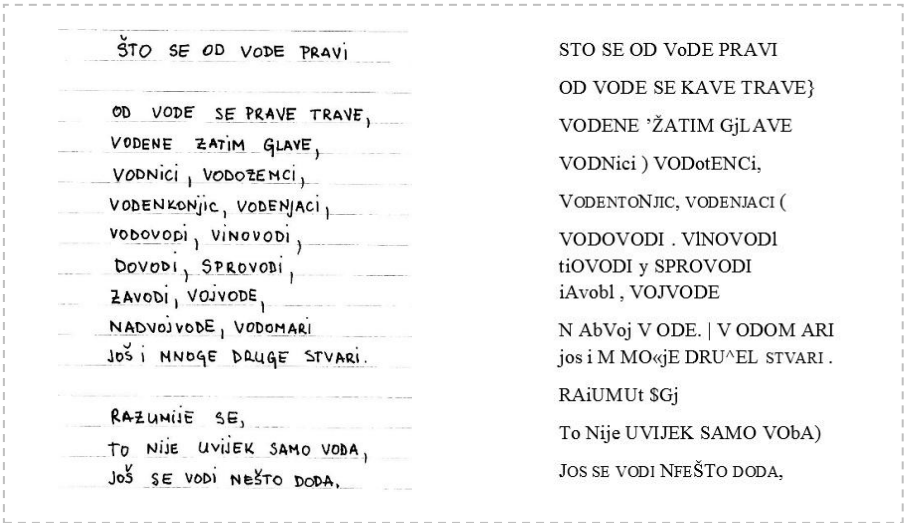
Slika 5. Tekst iz knjige na engleskom prije procesa OCR-a (ABBYY FineReader 15)


"It's all part of the timing of grace, Mack," Jesus continued. "If the universe contained only one human being, timing would be rather simple. But add just one more, and, well, you know the story. Each choice ripples out through time and relationships, bouncing off of other choices. And out of what seems to be a huge mess, Papa weaves a magnificent tapestry.

Slika 6. Tekst iz knjige na engleskom nakon procesa OCR-a (ABBYY FineReader 15)

Optičkim prepoznavanjem znakova teksta na engleskom jeziku, uz pomoć aplikacije ostvarena je 100% - tna točnost prilikom prepoznavanja i slova i interpunkcijskih znakova, odnosno u potpunosti su prepoznata sva 283 slova i 16 interpunkcijskih znakova.

D) Tekst pisan rukom



Slika 7. Tekst pisan rukom prije i nakon procesa OCR-a (ABBYY FineReader 15)

Pjesma Z. Baloga, pisana rukom prethodno je nekoliko puta obrađivan u Gimp aplikaciji za stvaranje i obradu rasterske grafike, dok promjenom kontrasta, svjetline i zasićenosti nije dobivena verzija teksta čije je optičko prepoznavanje izvršeno s najmanje grešaka. Ista ta verzija teksta korištena je i prilikom optičkog prepoznavanja znakova u sklopu svih drugih analiziranih aplikacija. Od 213 slova i 17 interpunkcijskih znakova, uz pomoć aplikacije

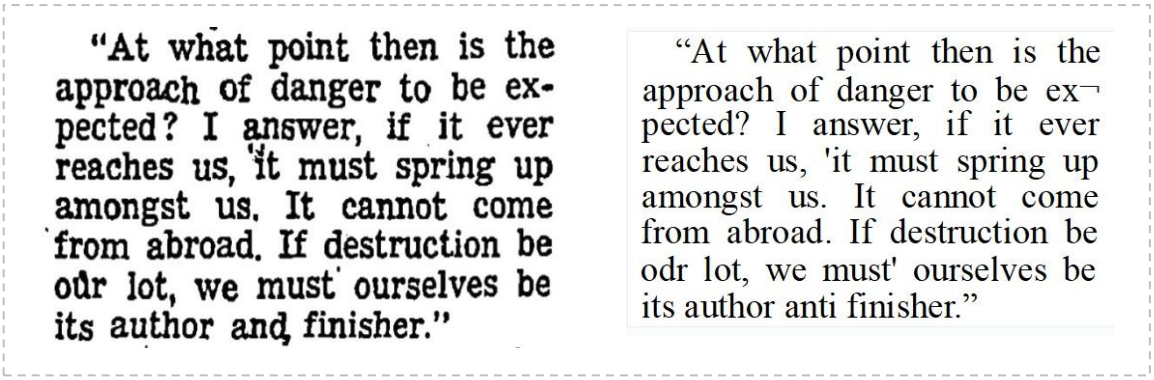
ABBYY Fine Reader 15, uspješno je prepoznato 186 slova (87.32 %) i 4 interpunkcijska znaka (23.52 %). Sveukupna točnost prilikom optičkog prepoznavanja znakova je 82.60 %.

E) Članak iz novina

Slika 8. Članak iz novina na hrvatskom jeziku prije i nakon OCR-a (ABBYY FineReader 15)

Upotrebom aplikacije ABBYY FineReader 15 članak iz novina prepoznat je sa 100% - tnom točnošću, odnosno u potpunosti je prepoznato 198 slova i 8 interpunkcijskih znakova (u to su uključene i spojnice nastale prijelomom riječi iz jednog retka u drugi).

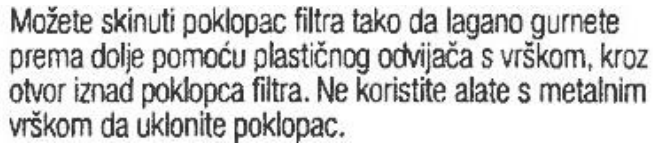
F) Članak iz novina pisan engleskim jezikom



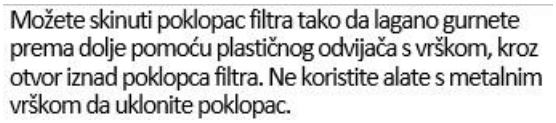
Slika 9. Članak iz novina na engleskom jeziku prije i nakon procesa OCR-a (ABBYY FineReader 15)

Prilikom optičkog prepoznavanja znakova članka iz novina pisanog na engleskom jeziku samo 2 od 175 slova su krivo prepoznata odnosno postotak točnog optičkog prepoznavanja je 98.85%. Postotak optičkog prepoznavanja interpunkcijskih znakova u ovom primjeru je 90% (9 od 10). Sveukupni postotak točnosti optičkog prepoznavanja (slova i interpunkcijski znakovi) je 98.37 %. Neke od nečistoća, točnije dvije, na originalnom uzorku nakon optičkog prepoznavanja znakova prepoznate su kao interpunkcijski znakovi koji se zapravo ne pojavljuju u originalnoj verziji teksta.

G) Tekst pisan vrlo sitnim fontom



Slika 10. Tekst pisan sitnim fontom prije procesa OCR-a (ABBYY FineReader 15)

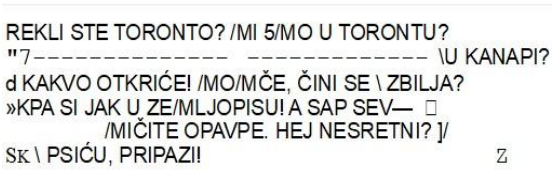


Slika 11. Tekst pisan sitnim fontom nakon procesa OCR-a (ABBYY FineReader 15)

*Napomena: Tekst je povećan nakon optičkog prepoznavanja znakova zbog bolje preglednosti.

Primjer teksta pisanog sa vrlo sitnim fontom (upute za rad s perilicom rublja) prepoznat je sa 100% - tnom točnošću odnosno u potpunosti su prepoznata sva 223 slova i svih 6 interpunkcijskih znakova.

H) Tekst sa neuobičajenim fontom – strip

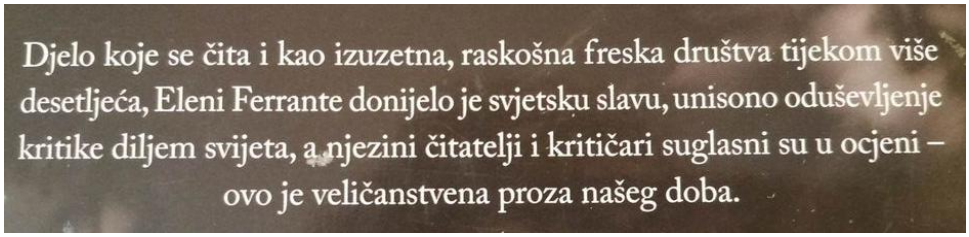


Slika 12. Isječak iz stripa prije i nakon procesa OCR-a (ABBYY FineReader 15)

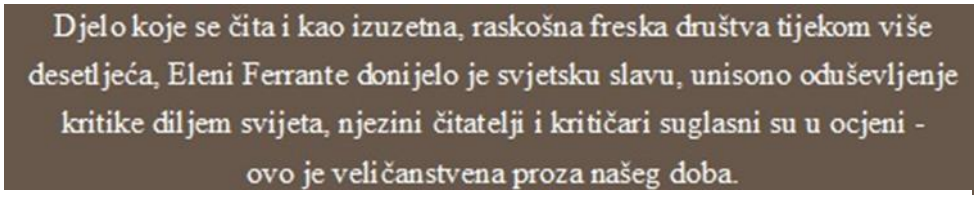
Primjer stripa uzet je kako bi se vidjelo koliko je uspješno optičko prepoznavanje znakova za neuobičajene fontove. Iako se naočigled radi o vrlo preglednom fontu, prilikom optičkog prepoznavanja teksta iz oblačića, prepoznato je tek 115 od 121 slova (95.04 %). Interpunkcijski znakovi su prepoznati sa 100% - tnom točnošću (10 od 10). Sveukupna točnost prilikom optičkog prepoznavanja svih znakova u tekstu je 87.78 %. Ono što je problematično kod prepoznavanja ovog primjera je velik broj „prepoznatih" interpunkcijskih

znakova koji uopće nisu prisutni u originalnoj inačici (45 sveukupno). Također, aplikacija nije uspjela prepoznati da se radi o dva zasebna oblačića teksta te je spojila tekst iz prvog reda tekstualnog oblačića sa posljednjim redom prvog tekstualnog oblačića.

I) Svijetli tekst na tamnoj pozadini



Slika 13. Svijetli tekst na tamnoj pozadini prije procesa OCR-a (ABBYY FineReader 15)



Slika 14. Svijetli tekst na tamnoj pozadini nakon OCR-a (ABBYY FineReader 15)

Uz pomoć programa za optičko prepoznavanje znakova ABBYY FineReader 15 uspješno je prepoznato svih 218 slova i svih 6 interpunkcijskih znakova odnosno ostvarena je 100% - tna točnost pri optičkom prepoznavanju slova i znakova teksta sa stražnjih korica knjige Elene Ferrante[141]. Ovaj primjer je posebno važan jer ne samo da se radi o primjeru svijetlog teksta na tamnoj pozadini već se radi o tekstu sa dosta zamućenom i nejasnom pozadinom unatoč kojoj je ipak ostvarena 100% - tna točnost prilikom optičkog prepoznavanja znakova.

[141] Ferrante, Elena. Genijalna prijateljica. Zagreb: Profil, 2020.

5.4.2. Free OCR[142]

1. Da li je analizirana aplikacija otvorenog koda?

Free OCR aplikacija je otvorenog koda što znači da ga se može besplatno preuzeti sa Interneta te ga se može i uređivati od strane korisnika.

2. Kakve mogućnosti pohrane nudi analizirana aplikacija?

U sklopu ove aplikacije dokumenti se mogu pohraniti u .txt i .rtf formatu.

3. Koje su dodane vrijednosti svake od aplikacija (faktor iznenađenja)?

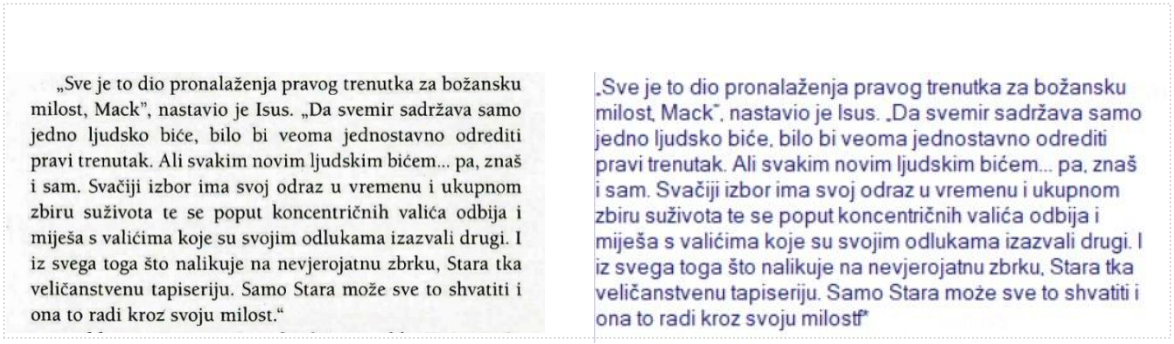
Ovaj program ima samo funkcije prepoznavanja slikovnih i .pdf formata.

4. Koliko jezika podržava analizirana aplikacija i da li je hrvatski jedan od njih?

U sklopu Free OCR-a pohranjeno je 11 jezika te hrvatski jezik nije među njima. No, ukoliko se prije optičkog prepoznavanja znakova postavi poljski jezik kao jezik za prepoznavanje, program uspješno prepoznaje i tekstove na hrvatskom jeziku.

5. S kolikom uspješnošću se mogu optički prepoznati slijedeći dokumenti :

A) Tekst skeniran iz knjige na hrvatskom jeziku

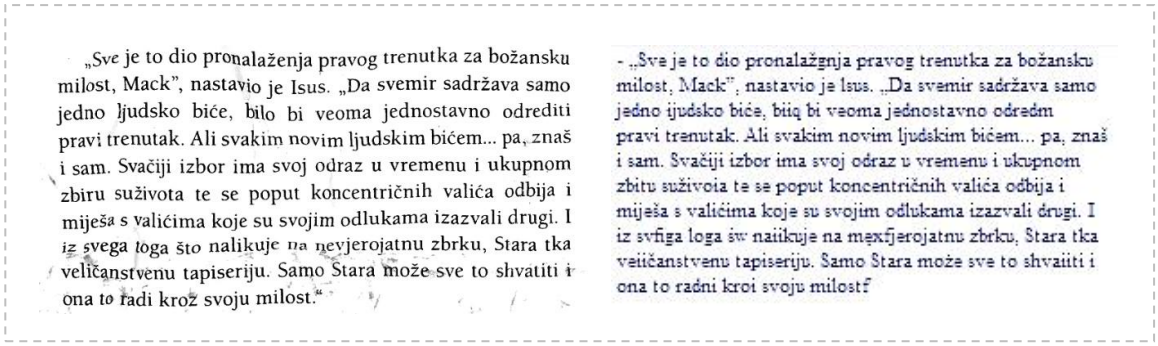


Slika 15. Tekst iz knjige prije i nakon procesa OCR – a (Free OCR)

Uz pomoć aplikacije Free OCR 15, tekst iz knjige prepoznat je sa 100% - tnom točnošću, odnosno u potpunosti je prepoznato 437 slova kao i 16 interpunkcijskih znakova. Napravljena je samo jedna greška i to na samom kraju teksta gdje je nadodano jedno slovo koje se ne pojavljuje u prvotnoj verziji teksta.

[142] Free OCR Software. URL: http://www.paperfile.net/index.html (2019-08-08)

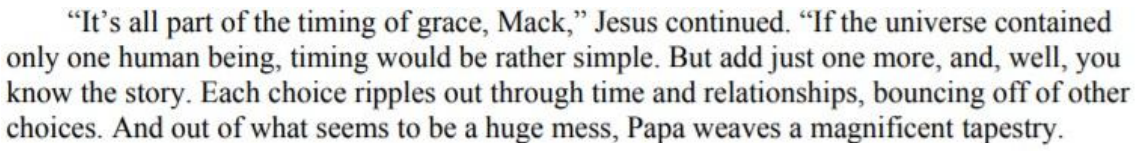
B) Zgužvani tekst iz knjige



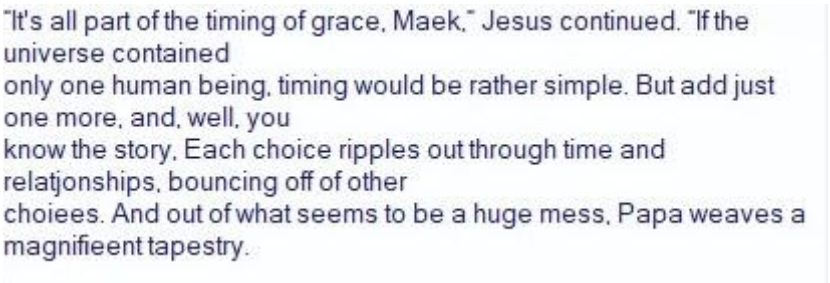
Slika 16. Zgužvani tekst iz knjige prije i nakon procesa OCR-a (Free OCR)

Optičkim prepoznavanjem zgužvanog teksta iz knjige, uz pomoć programa Free OCR 15, krivo je prepoznato 16 slova od 437, odnosno postignuta je točnost optičkog prepoznavanja od 96.33%. Samo jedan od 16 interpunkcijskih znakova nije prepoznat, što čini 93.75% izraženo u postotcima. Sveukupni postotak točnosti optičkog prepoznavanja znakova 93.15%.

C) Tekst iz knjige na engleskom jeziku



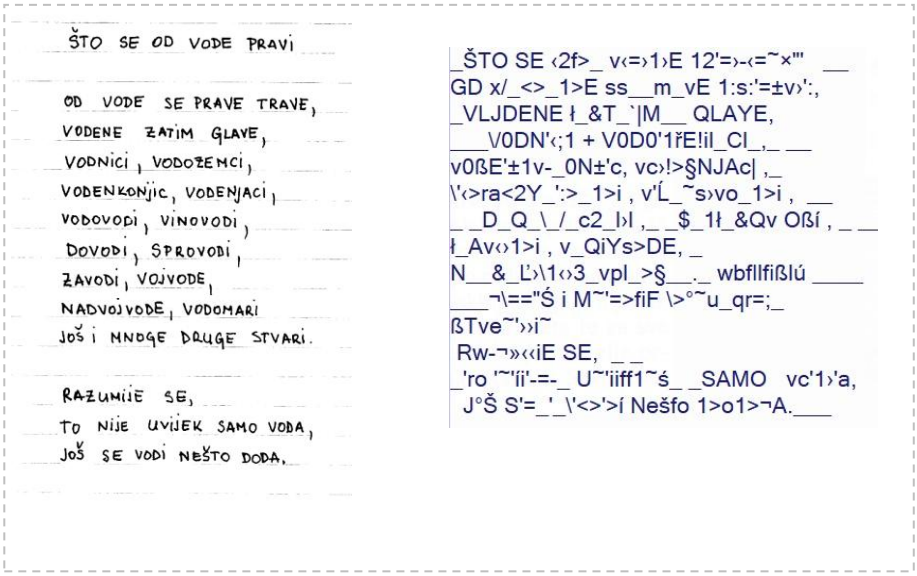
Slika 17. Tekst iz knjige na engleskom prije procesa OCR-a (Free OCR)



Slika 18. Tekst iz knjige na engleskom nakon procesa OCR-a (Free OCR)

Kako bi se zadržala određena dosljednost, analiziran je isti ulomak iz iste knjige ali na engleskom jeziku (Koliba, William P. Young). Prepoznata su 272 od 283 slova, odnosno u primjeru je ostvarena točnost optičkog prepoznavanja slova od 96.11%. Od 16 interpunkcijskih znakova samo ih je 11 točno prepoznato (68.75%). Sveukupni postotak točnosti optičkog prepoznavanja znakova u ovom primjeru iznosi 94.64%.

D) Tekst pisan rukom

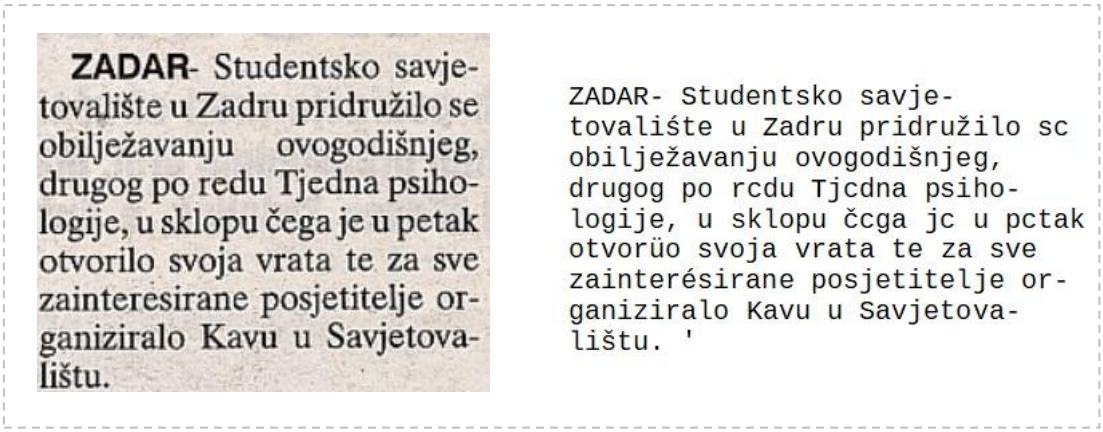


Slika 19. Tekst pisan rukom prije i nakon procesa OCR-a (Free OCR)

Kako bi se zadržala određena dosljednost, analiziran je isti ulomak iz iste knjige ali na engleskom jeziku (Koliba, William P. Young). Prepoznata su 272 od 283 slova, odnosno u primjeru je ostvarena točnost optičkog prepoznavanja slova od 96.11%. Od 16 interpunkcijskih znakova samo ih je 11 točno prepoznato (68.75%). Sveukupni postotak točnosti optičkog prepoznavanja znakova u ovom primjeru iznosi 94.64%.

S obzirom da je jedan od ciljeva optičkog prepoznavanja znakova što manja potreba za ljudskom intervencijom nakon provođenja procesa optičkog prepoznavanja znakova, rezultat optičkog prepoznavanja ovog primjera u potpunosti je beskoristan. Od 213 slova i 17 interpunkcijskih znakova točno su prepoznata samo 74 slova (34.74 %) te 14 interpunkcijskih znakova (41.31%) iz čega proizlazi da bi ljudska intervencija nakon prethodno provedenog procesa trebala biti veća od 60%, što ovaj primjer čini potpuno neiskoristivim. Sveukupni postotak točnosti optičkog prepoznavanja slova i interpunkcijskih znakova u ovom primjeru je 38.26 %. Također, treba napomenuti kako se zbog brojnih krivo prepoznatih te „nadodanih" slova zapravo ne može sa 100% - tnom sigurnošću reći jesu li 74 i 14 točni brojevi slova i interpunkcijskih znakova no za potrebe ovog rada brojevi su nakon višebrojnog prebrojavanja odabrani kao konačni za detaljniju analizu.

E) Članak iz novina na hrvatskom jeziku



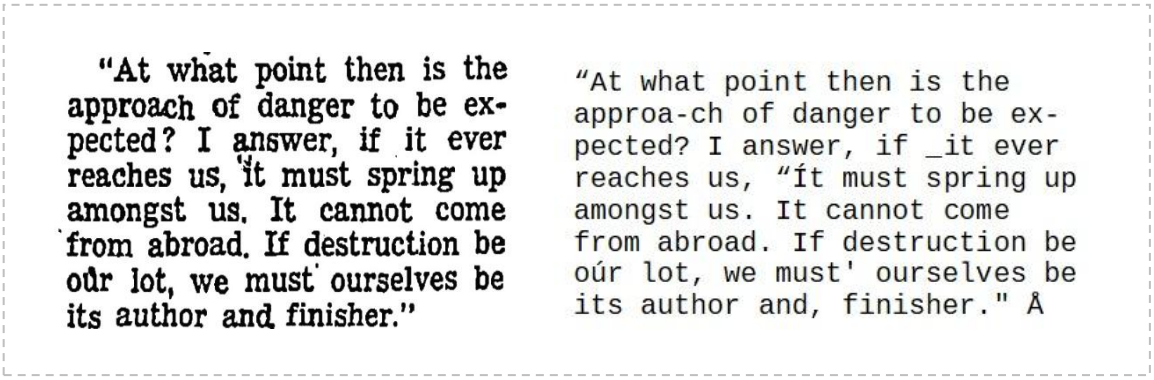
ZADAR- Studentsko savjetovalište u Zadru pridružilo se obilježavanju ovogodišnjeg, drugog po redu Tjedna psihologije, u sklopu čega je u petak otvorilo svoja vrata te za sve zainteresirane posjetitelje organiziralo Kavu u Savjetovalištu.

```
ZADAR- Studentsko savje-
tovalište u Zadru pridružilo sc
obilježavanju ovogodišnjeg,
drugog po rcdu Tjcdna psiho-
logije, u sklopu čcga jc u pctak
otvorüo svoja vrata te za sve
zainterésirane posjetitelje or-
ganiziralo Kavu u Savjetova-
lištu. '
```

Slika 20. Članak iz novina na hrvatskom jeziku prije i nakon OCR-a (Free OCR)

Uslijed optičkog prepoznavanja znakova članka iz novina na hrvatskom jeziku točno je prepoznato 188 od 198 (94.94 %) dok su interpunkcijski znakovi prepoznati sa 100% - tnom točnošću (8 od 8) te je na kraju paragrafa nadodan jedan znak (apostrof) koji se ne nalazi u originalnoj inačici. Sveukupni postotak točnosti optičkog prepoznavanja znakova je 95.14 %.

F) Članak iz novina na engleskom jeziku



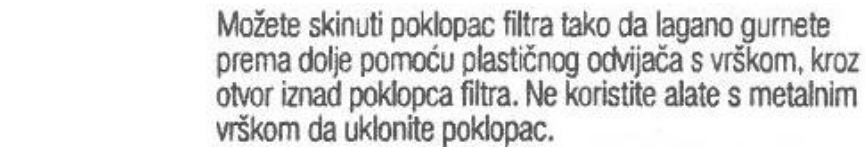
"At what point then is the approach of danger to be expected? I answer, if it ever reaches us, it must spring up amongst us. It cannot come from abroad. If destruction be our lot, we must ourselves be its author and finisher."

```
“At what point then is the
approa-ch of danger to be ex-
pected? I answer, if _it ever
reaches us, “Ít must spring up
amongst us. It cannot come
from abroad. If destruction be
oúr lot, we must' ourselves be
its author and, finisher." Å
```

Slika 21. Članak iz novina na engleskom jeziku prije i nakon procesa OCR-a (Free OCR)

Optičkim prepoznavanjem znakova članka iz novina na engleskom jeziku točno su prepoznata 173 znaka od 175 (98.66 %) i to se radi o točno prepoznatim slovima ali su im zbog nečistoća na tekstu nadodani naglasci. Svih 10 interpunkcijskih znakova su prepoznati sa 100% - tnom točnošću. U prepoznatom tekstu nadodana su tri interpunkcijska znaka te jedno slovo. Sveukupni postotak točnosti optičkog prepoznavanja znakova je 93.51 %.

G) Tekst pisan vrlo sitnim fontom



Slika 22. Tekst pisan sitnim fontom prije procesa OCR-a (Free OCR)

```
Možeše skinuti poklopac filtra tako da Iagann gurnete
prema dolāe pormću nlastičnug odvijmü-1 8 wškom, km:
olvor iznad poidopca filtra. Ne koristite alate s metalnim
wškom da uklonite pnklopac.
```

Slika 23. Tekst pisan sitnim fontom nakon procesa OCR-a (Free OCR)

Prilikom prepoznavanja znakova iz primjera sa vrlo sitnim fontom slova (upute za rad sa perilicom rublja) uspješno je prepoznato 141 od 162 slova (87.03 %). Sva tri interpunkcijska znaka prepoznata su sa 100 % - tnom točnošću. Ukupni postotak točnosti (slova i interpunkcijski znakovi) je 85.45 %. U OCR - iziranoj inačici teksta nadodan je jedan znak kojeg nema u originalnoj inačici teksta.

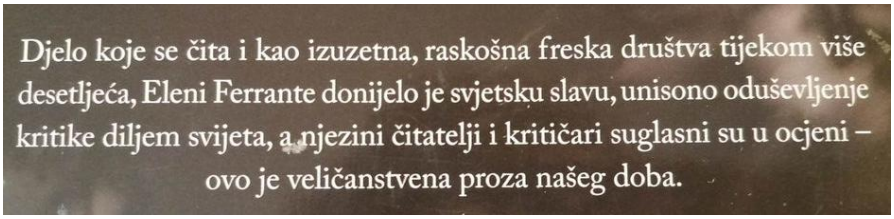
H) Tekst sa neuobičajenim fontom- strip


```
\
“l
7'
\
MJ
```

Slika 24. Isječak iz stripa prije i nakon procesa OCR-a (Free OCR)

Primjerak teksta sa neuobičajenim fontom odnosno stripa u potpunosti je neprepoznat (0%).

I) Svijetli tekst sa tamnom pozadinom



Slika 25. Svijetli tekst na tamnoj pozadini prije procesa OCR-a (Free OCR)

```
Djclo kojc sc čita i kao izuzetna, raskošna freska društva tíje  `više'
desctljcća, Eleni Fcrrantc donijclo je svjctsku slavu, urúsono od
krifikc díljem svijcta, ąupjczini čitatelji i -krítičari suglasní su u ocjeni -
ovo je veličanstvena proza našeg doba. -
```

Slika 26. Svijetli tekst na tamnoj pozadini nakon OCR-a (Free OCR)

Optičkim prepoznavanjem znakova primjera svijetlog teksta na tamnoj pozadini, točno su prepoznata 182 od 218 slova (83.48 %). Svih 6 interpunkcijskih znakova je točno prepoznato. Sveukupni postotak točnosti optičkog prepoznavanja znakova je 81.25 %.

5.4.3. Google Drive OCR[143]

1. Da li je analizirana aplikacija otvorenog koda?

Aplikacija Google Drive OCR u sklopu svoje arhitekture sadrži komponente neke od aplikacija koje su otvorenog koda, kao što je Tesseract ili OCRopus.

2. Kakve mogućnosti pohrane nudi analizirana aplikacija?

Procesirani dokumenti se mogu pohraniti u .doc, .odt, .rtf, .pdf, .txt, .html i .epub formatu.

3. Koje su dodane vrijednosti ove aplikacije (faktor iznenađenja)?

Dodane vrijednosti ove aplikacije su što se optički prepoznat tekst može uređivati kao u word dokumentu jer imaju sličnu traku izbornika. (Slika 20.) Tekst se može odmah preoblikovati u smislu da se može mijenjati font, veličina fonta, proredi, umetati slike, baš kao i u aplikaciji Word te dijeliti finalni tekst mrežno ili spremiti na Google disk.

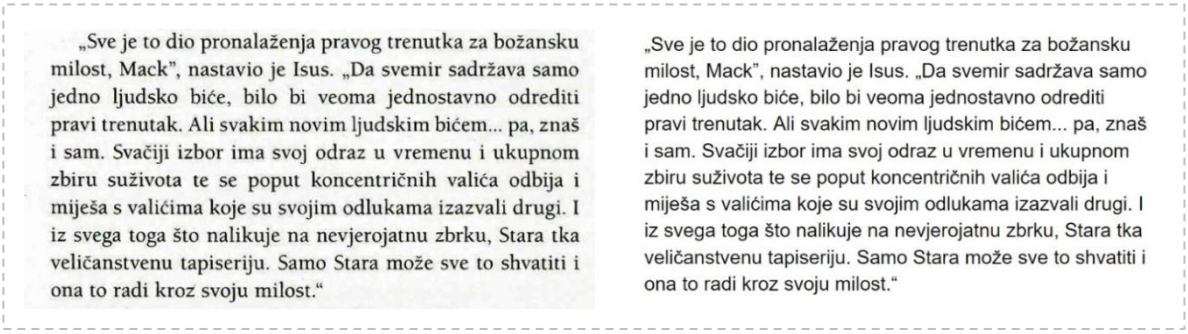
Slika 27. Prikaz izborne trake Google Drive OCR aplikacije

4. Koliko jezika podržava analizirana aplikacija i da li je hrvatski jedan od njih?

Google Drive OCR podržava aktivno 193 jezika i 36 jezika na čijem se optičkom prepoznavanju tek radi, a među podržanim jezicima je i hrvatski jezik.

5. S kolikom uspješnošću se mogu optički prepoznati slijedeći dokumenti :

A) Tekst iz knjige


„Sve je to dio pronalaženja pravog trenutka za božansku milost, Mack", nastavio je Isus. „Da svemir sadržava samo jedno ljudsko biće, bilo bi veoma jednostavno odrediti pravi trenutak. Ali svakim novim ljudskim bićem... pa, znaš i sam. Svačiji izbor ima svoj odraz u vremenu i ukupnom zbiru suživota te se poput koncentričnih valića odbija i miješa s valićima koje su svojim odlukama izazvali drugi. I iz svega toga što nalikuje na nevjerojatnu zbrku, Stara tka veličanstvenu tapiseriju. Samo Stara može sve to shvatiti i ona to radi kroz svoju milost."

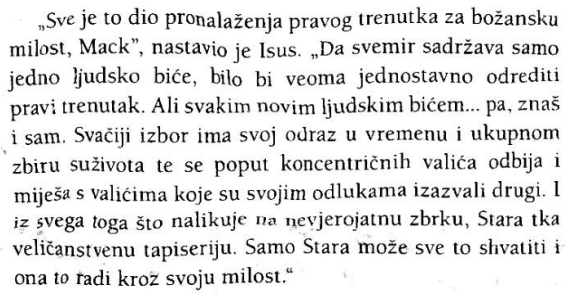
„Sve je to dio pronalaženja pravog trenutka za božansku milost, Mack", nastavio je Isus. „Da svemir sadržava samo jedno ljudsko biće, bilo bi veoma jednostavno odrediti pravi trenutak. Ali svakim novim ljudskim bićem... pa, znaš i sam. Svačiji izbor ima svoj odraz u vremenu i ukupnom zbiru suživota te se poput koncentričnih valića odbija i miješa s valićima koje su svojim odlukama izazvali drugi. I iz svega toga što nalikuje na nevjerojatnu zbrku, Stara tka veličanstvenu tapiseriju. Samo Stara može sve to shvatiti i ona to radi kroz svoju milost."

Slika 28. Tekst iz knjige prije i nakon procesa OCR – a (Google Drive OCR)

Uz pomoć aplikacije Google Drive Docs, ulomak iz knjige Koliba, na hrvatskom jeziku, prepoznat je sa 100 % - tnom točnošću, odnosno u potpunosti je prepoznato 437 slova i 16 interpunkcijskih znakova.

[143] Google Disk. URL: https://www.google.com/intl/hr_HR/drive/ (2019-10-10)

B) Zgužvani tekst iz knjige


„Sve je to dio pronalaženja pravog trenutka za božansku milost, Mack", nastavio je Isus. „Da svemir sadržava samo jedno ljudsko biće, bilo bi veoma jednostavno odrediti pravi trenutak. Ali svakim novim ljudskim bićem... pa, znaš i sam. Svačiji izbor ima svoj odraz u vremenu i ukupnom zbiru suživota te se poput koncentričnih valića odbija i miješa s valićima koje su svojim odlukama izazvali drugi. I iz svega toga što nalikuje na nevjerojatnu zbrku, Stara tka veličanstvenu tapiseriju. Samo Stara može sve to shvatiti i ona to radi kroz svoju milost."

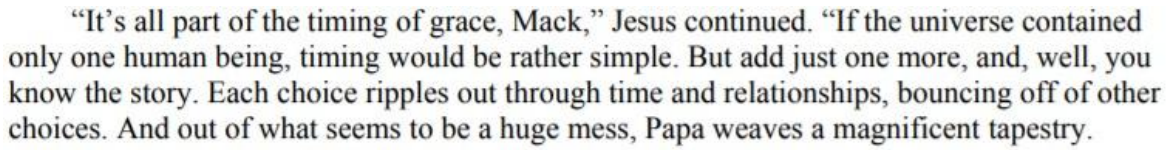
Slika 29. Zgužvani tekst iz knjige prije procesa OCR-a (Google Drive OCR)

„Sve je to dio pronalaženja pravog trenutka za božansku milost, Mack", nastavio je Isus. „Da svemir sadržava samo jedno ljudsko biće, bilo bi veoma jednostavno odrediti pravi trenutak. Ali svakim novim ljudskim bićem... pa, znaš i sam. Svačiji izbor ima svoj odraz u vremenu i ukupnom zbiru suživota te se poput koncentričnih valića odbija i miješa s valićima koje su svojim odlukama izazvali drugi. I iz svega toga što nalikuje na nevjerojatnu zbrku, Stara tka veličanstvenu tapiseriju. Samo Stara može sve to shvatiti i ona to radi kroz svoju milost."

Slika 30. Zgužvani tekst iz knjige nakon procesa OCR-a (Google Drive OCR)

Prethodno zgužvani ulomak iz knjige Koliba, uz pomoć aplikacije Google Drive u potpunosti je prepoznat odnosno 437 slova i 16 interpunkcijskih znakova u potpunosti su prepoznati.

C) Tekst iz knjige na engleskom jeziku


"It's all part of the timing of grace, Mack," Jesus continued. "If the universe contained only one human being, timing would be rather simple. But add just one more, and, well, you know the story. Each choice ripples out through time and relationships, bouncing off of other choices. And out of what seems to be a huge mess, Papa weaves a magnificent tapestry.

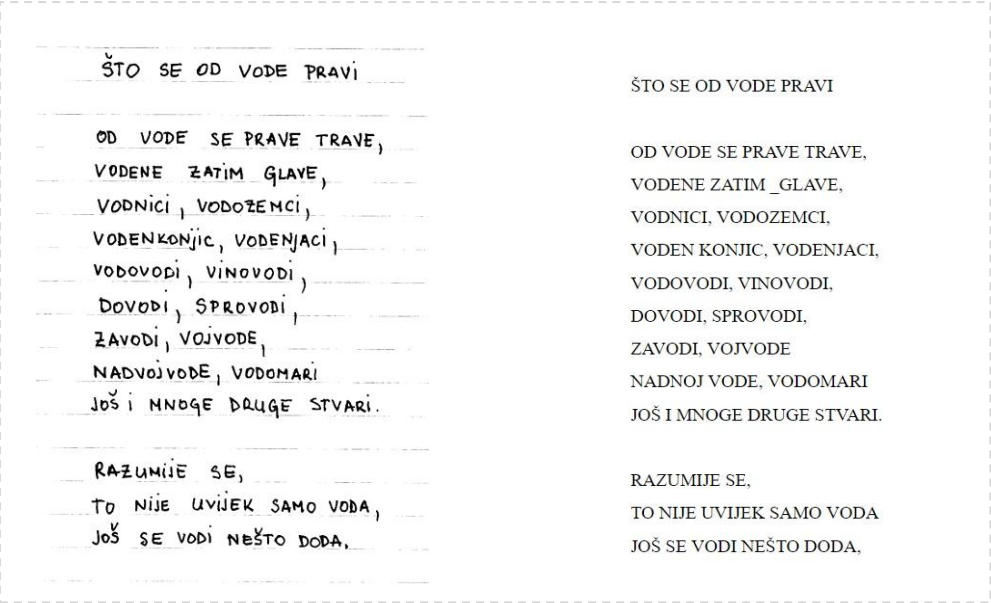
Slika 31. Tekst iz knjige na engleskom prije procesa OCR-a (Google Drive OCR)

"It's all part of the timing of grace, Mack," Jesus continued. "If the universe contained only one human being, timing would be rather simple. But add just one more, and, well, you know the story. Each choice ripples out through time and relationships, bouncing off of other choices. And out of what seems to be a huge mess, Papa weaves a magnificent tapestry.

Slika 32. Tekst iz knjige na engleskom nakon procesa OCR-a (Google Drive OCR)

Ulomak iz knjige na engleskom, uz pomoć aplikacije Google Drive, optički je prepoznat sa postotkom točnosti od 100 %, odnosno sva 283 slova i 16 interpunkcijskih znakova u potpunosti su prepoznati.

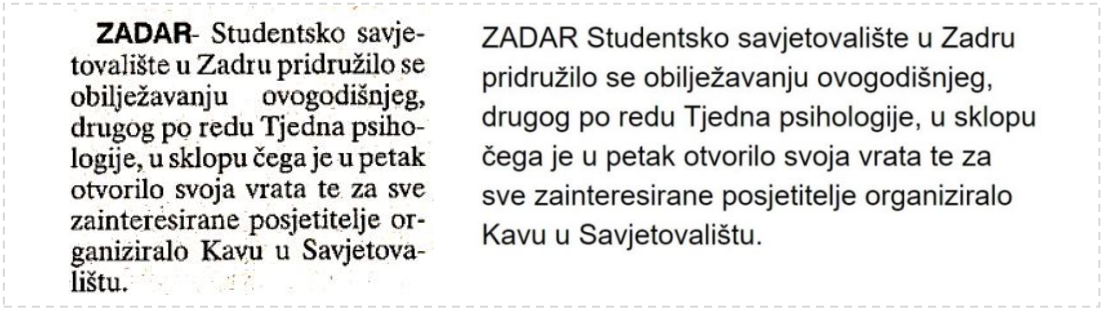
D) Tekst pisan rukom



Slika 33. Tekst pisan rukom prije i nakon procesa OCR-a (Google Drive OCR)

Pjesma Z. Baloga, Što se od vode pravi, pisana rukom, optički je prepoznata s jednim pogrešnim slovom i jednim izostavljenim interpunkcijskim znakom, što znači da je od 213 slova točno prepoznato 212 a od 17 interpunkcijskih znakova njih 16. Dakle, u ovom primjeru, prilikom prepoznavanja slova ostvarena je točnost od 99.53 %, a prilikom optičkog prepoznavanja interpunkcijskih znakova 94.11 %. Sveukupno je ostvarena točnost od 99.13 %.

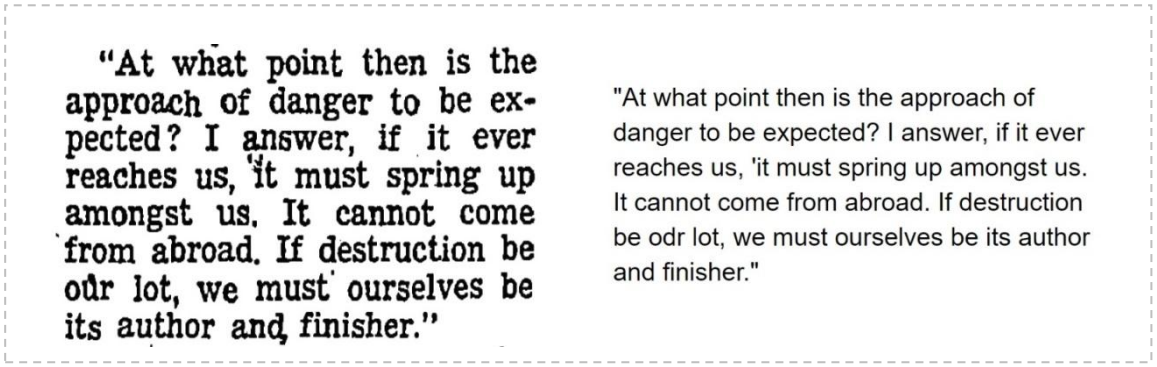
E) Članak iz novina na hrvatskom jeziku



Slika 34. Članak iz novina na hrvatskom jeziku prije i nakon OCR-a (Google Drive OCR)

Članak iz novina na hrvatskom jeziku prepoznat je sa jednom pogreškom odnosno zanemarena je spojnica nakon prve riječi u tekstu. Svih 198 slova prepoznato je sa 100 % -tnom točnošću dok su interpunkcijski znakovi prepoznati sa točnošću od 87.5 % (7 od 8 znakova. Sveukupna točnost optičkog prepoznavanja znakova je 99.51%.

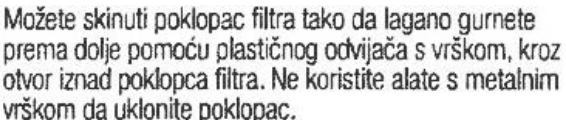
F) Članak na engleskom jeziku



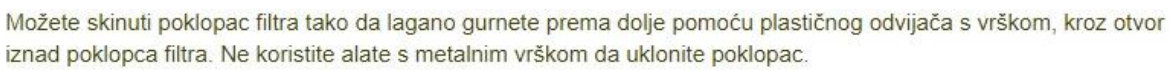
Slika 35. Članak iz novina na engleskom jeziku prije i nakon procesa OCR-a (Google Drive OCR)

Ulomak članka na engleskom jeziku sastoji se od 175 slova i 10 interpunkcijskih znakova od kojih su, uz pomoć aplikacije Google Drive, uspješno optički prepoznata 174 slova (99.42 %) te svi interpunkcijski znakovi (100 %) što čini postotak uspješnosti optičkog prepoznavanja svih znakova (slova i interpunkcijski znakovi) od 94.05%. Kao i kod prethodne aplikacije, greška prilikom optičkog prepoznavanja se dogodila na istom mjestu zbog postojanja nejasnoće u tekstu nastale prilikom skeniranja. (our – odr)

G) Tekst pisan vrlo sitnim fontom



Slika 36. Tekst pisan sitnim fontom prije procesa OCR-a (Google Drive OCR)



Slika 37. Tekst pisan sitnim fontom nakon procesa OCR-a (Google Drive OCR)

Tekst pisan vrlo sitnim fontom prepoznat ja sa 100% - tnom točnošću odnosno u potpunosti su prepoznata sva slova i svi interpunkcijski znakovi.

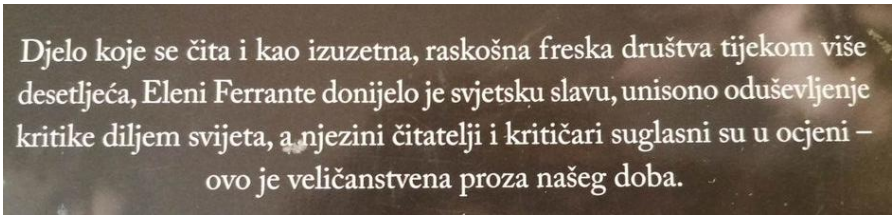
H) Tekst sa neuobičajenim fontom- strip

REKLI STE TORONTO? MI SMO U TORONTU?
U KANADI? KAKVO OTKRIĆE! MOMČE,
ČINI SE ZBILJA? DA SI JAK U ZEMLJOPISU! A SAD SE
MIČITE ODAVDE, HEJ NESRETNI PSIĆU, PRIPAZI!

Slika 38. Isječak iz stripa prije i nakon procesa OCR-a (Google Drive OCR)

Odmah nakon prvog pregleda teksta dobivenog nakon optičkog prepoznavanja znakova lako se uočava kako aplikacija ne prepoznaje da se radi o različitim tekstualnim jedinicama ili blokovima, već tekst povezuje doslovno kako je prikazan što je vidljivo u trećem retku rezultirajućeg teksta gdje je spojeno „čini se" iz jednog oblačića i „zbilja" iz drugog. Iako su i tekst i interpunkcijski znakovi prepoznati sa 100% - tnom točnošću, ovakav rezultat u praksi bi se mogao smatrati neuspješnim.

I) Svijetli tekst sa tamnom pozadinom



Slika 39. Svijetli tekst na tamnoj pozadini prije procesa OCR-a (Google Drive OCR)

Djelo koje se čita i kao izuzetna, raskošna freska društva tijekom više desetljeća, Eleni Ferrante donijelo je svjetsku slavu, unisono oduševljenje kritike diljem svijeta, a njezini čitatelji i kritičari suglasni su u ocjeni - ovo je veličanstvena proza našeg doba.

Slika 40. Svijetli tekst na tamnoj pozadini nakon OCR-a (Google Drive OCR)

U prethodno navedenom primjeru, sva slova i interpunkcijski znakovi, 218 slova i 6 interpunkcijskih znakova, prepoznati su sa 100% - tnom točnošću.

5.4.4. I2OCR[144]

1. Da li je analizirana aplikacija otvorenog koda?

Aplikaciji I2OCR moguće je besplatno mrežno pristupiti i nema ograničen broj dokumenata s kojima se može raditi, kao neke od aplikacija koje su analizirane u radu.

2. Kakve mogućnosti pohrane nudi analizirana aplikacija?

Dokumenti procesirani u sklopu aplikacije I2OCR se mogu pohraniti u .txt, .doc, docx, .pdf i .html formatu.

3. Koje su dodane vrijednosti svake od aplikacija (faktor iznenađenja)?

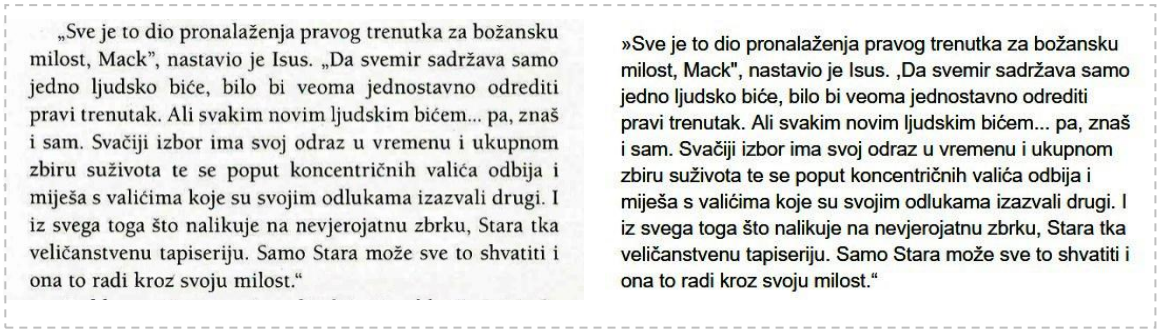
Aplikacija I2OCR nudi samo nekolicinu dodatnih opcija za optičko prepoznavanje dokumenta te mogućnost pohrane u četiri prethodno navedena formata. Te dodatne opcije su mogućnost direktnog prijevoda dobivenog teksta putem Google ili Bing prevoditelja; obradu dokumenta u Google Docs aplikaciji te mogućnost kopiranja teksta u drugi dokument. Posebno treba naglasiti da u sklopu mrežne stranice I2OCR postoji usluga putem koje možete mrežno poslati dokumente s većim brojem stranica, kao što su knjige u pdf formatu, koje želite optički prepoznati te će to za vas, za određeni novčani iznos, učiniti administratori stranice.

4. Koliko jezika podržava analizirana aplikacija i da li je hrvatski jedan od njih?

Podržano je 117 jezika te je među ponuđenim jezicima i hrvatski jezik.

5. S kolikom uspješnošću se mogu optički prepoznati slijedeći dokumenti :

A) tekst iz knjige

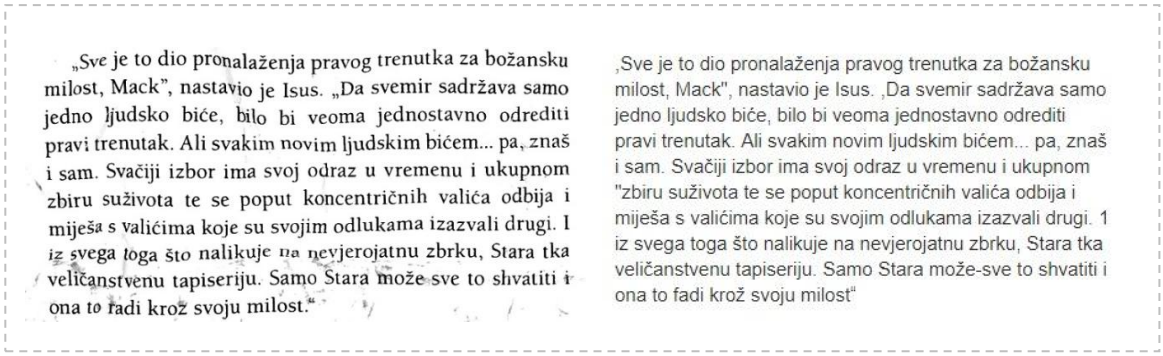


Slika 41. Tekst iz knjige na engleskom prije i nakon procesa OCR-a (I2OCR)

Skenirani tekst iz knjige sadrži 437 slova i 16 interpunkcijskih znakova, a uz pomoć aplikacije I2OCR u potpunosti su prepoznata sva slova i svi interpunkcijski znakovi.

[144] I2OCR. URL: http://www.i2ocr.com/ (2019-09-19)

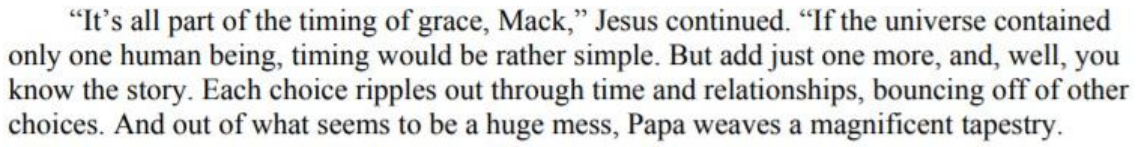
B) Zgužvani tekst iz knjige na hrvatskom jeziku



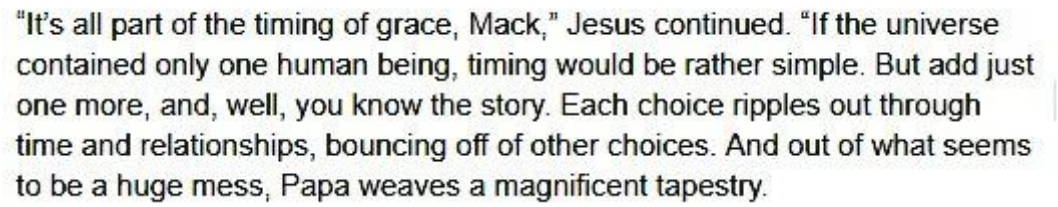
Slika 42. Zgužvani tekst iz knjige prije i nakon procesa OCR-a (I2OCR)

Optičkim prepoznavanjem zgužvanog ulomka iz knjige u potpunosti je prepoznato svih 437 slova i 16 interpunkcijskih znakova. Od specifičnosti treba izdvojiti da je program „prepoznao" 3 interpunkcijska znaka i jedno slovo koji se uopće ne pojavljuju u originalnoj inačici teksta.

C) Tekst iz knjige na engleskom jeziku



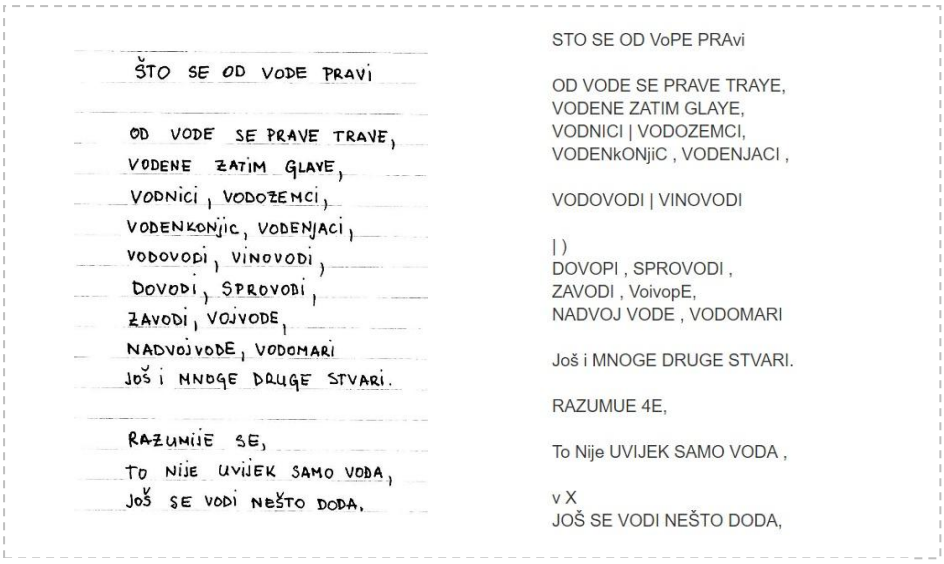
Slika 43. Tekst iz knjige na engleskom prije procesa OCR-a (I2OCR)



Slika 44. Tekst iz knjige na engleskom nakon procesa OCR-a (I2OCR)

U sklopu procesa optičkog prepoznavanja znakova uz pomoć aplikacije I2OCR sa 100% - tnom točnošću su prepoznata sva 283 slova i svih 16 interpunkcijskih znakova.

D) Tekst pisan rukom



Slika 45. Tekst pisan rukom prije i nakon procesa OCR-a (I2OCR)

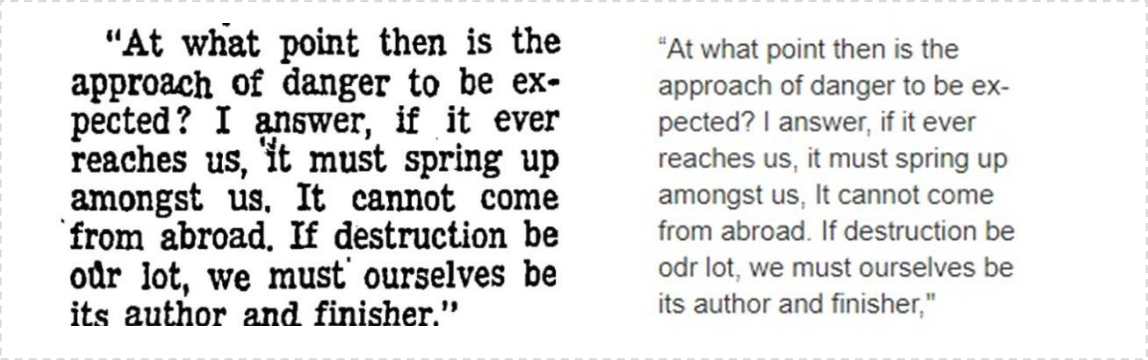
Tekst sadrži 213 slova te 17 interpunkcijskih znakova, a uz pomoć aplikacije I2OCR uspješno je prepoznato 200 slova i 13 interpunkcijskih znakova što znači da je ostvaren postotak uspješnosti optičkog prepoznavanja slova od 93.89 %. te postotak prepoznavanja interpunkcijskih znakova od 76.47 %. Sveukupni postotak uspješnosti optičkog prepoznavanja slova i interpunkcijskih znakova je 92.60 %. Zanemareno je to što je analizirana aplikacija u nekim situacijama velika tiskana slova optički prepoznala kao mala, s obzirom da se radi o točnom rezultatu. No, u praksi bi ovakva greška predstavljala problem s obzirom da je za ono što je trebala „odraditi" aplikacija, potrebna ljudska intervencija i dodatni utrošak vremena.

E) Članak iz novina na hrvatskom jeziku

Slika 46. Članak iz novina na hrvatskom jeziku prije i nakon OCR-a (I2OCR)

Prilikom optičkog prepoznavanja znakova sva slova (198) i svi interpunkcijski znakovi (8) prepoznati su sa 100 % - tnom točnošću.

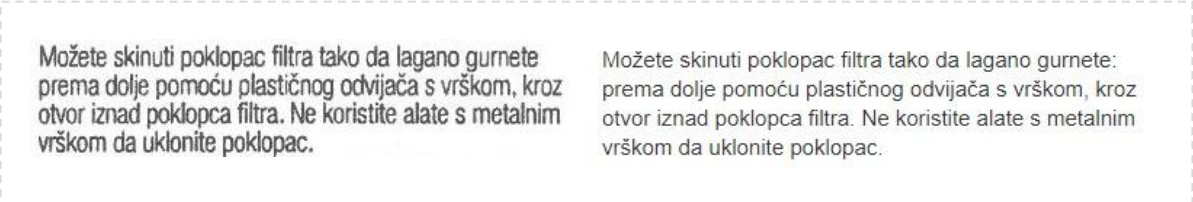
F) Članak iz novina na engleskom jeziku



Slika 47. Članak iz novina na engleskom jeziku prije i nakon procesa OCR-a (I2OCR)

Od 175 slova u sklopu ulomka članka na engleskom jeziku prepoznata su 174 slova (99.42%) dok je od 10 interpunkcijskih znakova prepoznato njih 9 (90%). Sveukupni postotak točnosti rezultata optičkog prepoznavanja ovog teksta je 98.91%. Kao i kod rezultata prethodno analiziranih aplikacija, pogrešno prepoznavanje slova nastalo je uslijed nečistoće na papiru zbog čega se slovo *u* u gotovo svim aplikacijama redovito prepoznaje kao d (prva riječ u sedmom retku – our).

G) Tekst pisan vrlo sitnim fontom



Slika 48. Tekst pisan sitnim fontom prije i nakon procesa OCR-a (I2OCR)

Prilikom prepoznavanja znakova iz primjera sa vrlo sitnim fontom slova (upute za rad sa perilicom rublja) sva slova i svi interpunkcijski znakovi prepoznati su sa 100% - tnom točnošću (162 slova i 3 interpunkcijska znaka). Iznimka je što je nakon provedbe optičkog prepoznavanja znakova „nadodan" jedan interpunkcijski znak kojeg nema u prvotnoj inačici teksta.

H) Tekst sa neuobičajenim fontom - strip

| da
M2

"

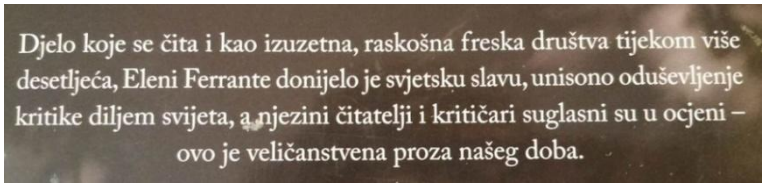
xa

A

Slika 49. Isječak iz stripa prije i nakon procesa OCR-a (I2OCR)

Tekst sa neuobičajenim fontom, odnosno strip u potpunosti je neprepoznat uz pomoć aplikacije I2OCR. Za pretpostaviti je da aplikacija nije namijenjena optičkom prepoznavanju stripa odnosno slijeda crteža obično popraćenih tekstom u oblačićima.

I) Svijetli tekst sa tamnom pozadinom



Slika 50. Originalni tekst sa stražnjih korica knjige (I2OCR)

IRAN Re RU REZ EN Ke a EI ZV a
RESE Bars e RR Ja SIDE Bua Kai

ei Ra OVER (zine z ORE sute tso a bE nisu PSS =
[ON VO EEE SEE ZEO ge)

Slika 51. Tekst sa stražnjih korica knjige nakon OCR procesa (I2OCR)

Za potrebe ovog rada preuzet je tekst sa stražnjih korica knjige Elene Ferrante[145]. Kao i u prethodnom primjeru, svijetli tekst na tamnoj pozadini je u potpunosti neprepoznat odnosno u inačici teksta koja bi trebala prikazati rezultat nakon optičkog prepoznavanja znakova prikazan je samo vrlo besmislen tekst koji se u niti jednom dijelu niti približno ne poklapa sa prvotnom inačicom teksta.

[145] Ferrante, Elena. Genijalna prijateljica. Zagreb: Profil knjiga d.o.o., 2016.

5.4.5. Convertio[146]

1. Da li je analizirana aplikacija otvorenog koda?

Aplikaciji Convertio je moguće besplatno pristupiti mrežno te optički prepoznati najviše deset dokumenata dok je za daljnju upotrebu potrebno platiti.

2. Kakve mogućnosti pohrane nudi analizirana aplikacija?

U sklopu aplikacije I2OCR procesirani dokumenti se mogu pohraniti u .doc, .xlsx, .xls, .pptx, .pdf, .txt, .rtf, .csv, .epub, .fb2 i .djvu formatima.

3. Koje su dodane vrijednosti ove aplikacije (faktor iznenađenja)?

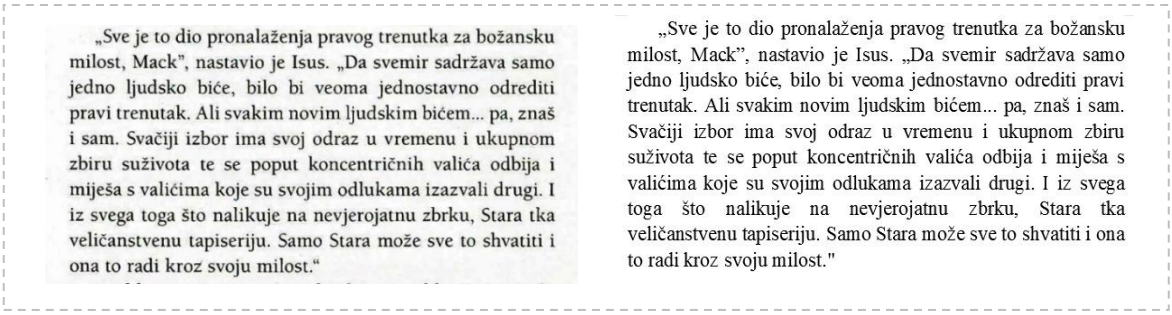
Aplikacija Convertio nudi osnovne funkcije optičkog prepoznavanja dokumenta, pohrane u 11 formata ali i prepoznavanje višestupčanog teksta.

4. Koliko jezika podržava analizirana aplikacija i da li je hrvatski jedan od njih?

Podržano je 76 jezika te je među njima i hrvatski jezik.

5. S kolikom uspješnošću se mogu optički prepoznati slijedeći dokumenti :

A) Tekst skeniran iz knjige

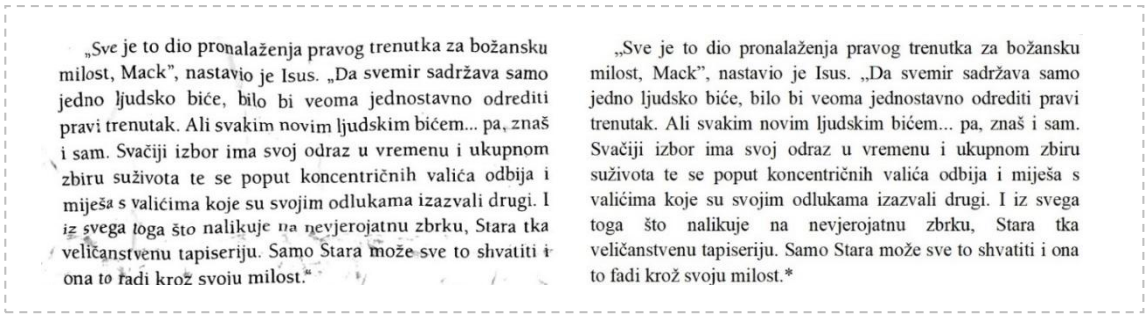

„Sve je to dio pronalaženja pravog trenutka za božansku milost, Mack", nastavio je Isus. „Da svemir sadržava samo jedno ljudsko biće, bilo bi veoma jednostavno odrediti pravi trenutak. Ali svakim novim ljudskim bićem... pa, znaš i sam. Svačiji izbor ima svoj odraz u vremenu i ukupnom zbiru suživota te se poput koncentričnih valića odbija i miješa s valićima koje su svojim odlukama izazvali drugi. I iz svega toga što nalikuje na nevjerojatnu zbrku, Stara tka veličanstvenu tapiseriju. Samo Stara može sve to shvatiti i ona to radi kroz svoju milost."

Slika 52. Tekst iz knjige prije i nakon procesa OCR – a (Convertio)

Uz pomoć aplikacije Convertio, tekst skeniran iz knjige prepoznat je sa 100% - tnom točnošću, odnosno u potpunosti je prepoznato 437 slova i 16 interpunkcijskih znakova.

[146] Convertio. URL: https://convertio.co/ (2019-09-10)

B) Zgužvani tekst iz knjige



„Sve je to dio pronalaženja pravog trenutka za božansku milost, Mack", nastavio je Isus. „Da svemir sadržava samo jedno ljudsko biće, bilo bi veoma jednostavno odrediti pravi trenutak. Ali svakim novim ljudskim bićem... pa, znaš i sam. Svačiji izbor ima svoj odraz u vremenu i ukupnom zbiru suživota te se poput koncentričnih valića odbija i miješa s valićima koje su svojim odlukama izazvali drugi. I iz svega toga što nalikuje na nevjerojatnu zbrku, Stara tka veličanstvenu tapiseriju. Samo Stara može sve to shvatiti i ona to radi krož svoju milost.*

„Sve je to dio pronalaženja pravog trenutka za božansku milost, Mack", nastavio je Isus. „Da svemir sadržava samo jedno ljudsko biće, bilo bi veoma jednostavno odrediti pravi trenutak. Ali svakim novim ljudskim bićem... pa, znaš i sam. Svačiji izbor ima svoj odraz u vremenu i ukupnom zbiru suživota te se poput koncentričnih valića odbija i miješa s valićima koje su svojim odlukama izazvali drugi. I iz svega toga što nalikuje na nevjerojatnu zbrku, Stara tka veličanstvenu tapiseriju. Samo Stara može sve to shvatiti i ona to radi krož svoju milost.*

Slika 53. Zgužvani tekst iz knjige prije i nakon procesa OCR-a (Convertio)

Od 437 slova i 16 interpunkcijskih znakova, uz pomoć Convertio programa, u sklopu zgužvanog teksta na hrvatskom jeziku točno je prepoznato 435 slova i 15 interpunkcijskih znakova što znači da je prilikom optičkog prepoznavanja slova postignuta točnost od 99.54 %, a interpunkcijskih znakova 93.75 %. Sveukupni postotak točnosti rezultata optičkog prepoznavanja ovog teksta je 99.33 %.

C) Tekst iz knjige na engleskom jeziku

"It's all part of the timing of grace, Mack," Jesus continued. "If the universe contained only one human being, timing would be rather simple. But add just one more, and, well, you know the story. Each choice ripples out through time and relationships, bouncing off of other choices. And out of what seems to be a huge mess, Papa weaves a magnificent tapestry.

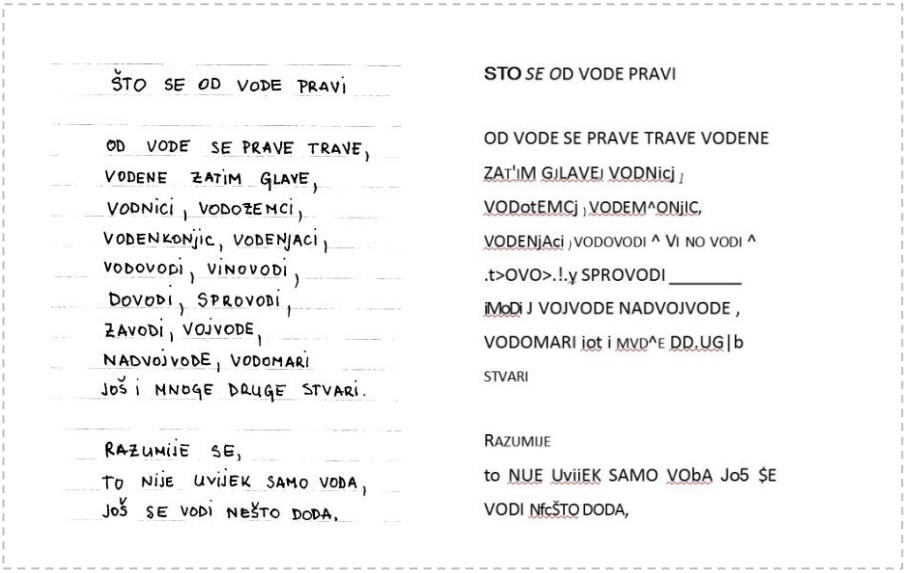
Slika 54. Tekst iz knjige na engleskom prije procesa OCR-a (Convertio)

"It's all part of the timing of grace, Mack," Jesus continued. "If the universe contained only one human being, timing would be rather simple. But add just one more, and, well, you know the story. Each choice ripples out through time and relationships, bouncing off of other choices. And out of what seems to be a huge mess, Papa weaves a magnificent tapestry.

Slika 55. Tekst iz knjige na engleskom nakon procesa OCR-a (Convertio)

Optičkim prepoznavanjem ulomka iz knjige Koliba, ali na engleskom jeziku (The Shack), postignuta je 100%-tna točnost, odnosno sva 283 slova i 16 interpunkcijskih znakova u potpunosti je prepoznato.

D) Tekst pisan rukom



Slika 56. Tekst pisan rukom prije i nakon procesa OCR-a (Convertio)

Pjesma Zvonimira Baloga sastoji se od 213 slova i 17 interpunkcijskih znakova od čega je točno prepoznato samo 187 slova i 2 interpunkcijska znaka iz čega proizlazi da prilikom optičkog prepoznavanja slova postotak uspješnosti iznosi 88.20 %, a interpunkcijskih znakova 11.76 %. Sveukupni postotak točnosti iznosi 82. 53 %. Iz rezultata je vidljivo kako je najveći broj pogrešaka napravljen prilikom optičkog prepoznavanja interpunkcijskih znakova, naročito zareza koji su najvećim dijelom prepoznati kao zagrade ili su jednostavno zanemareni. Nakon optičkog prepoznavanja pjesme pisane rukom, nadodano je još podosta slova i interpunkcijskih znakova koji ne postoje u prvotnoj inačici.

E) Članak iz novina na hrvatskom jeziku

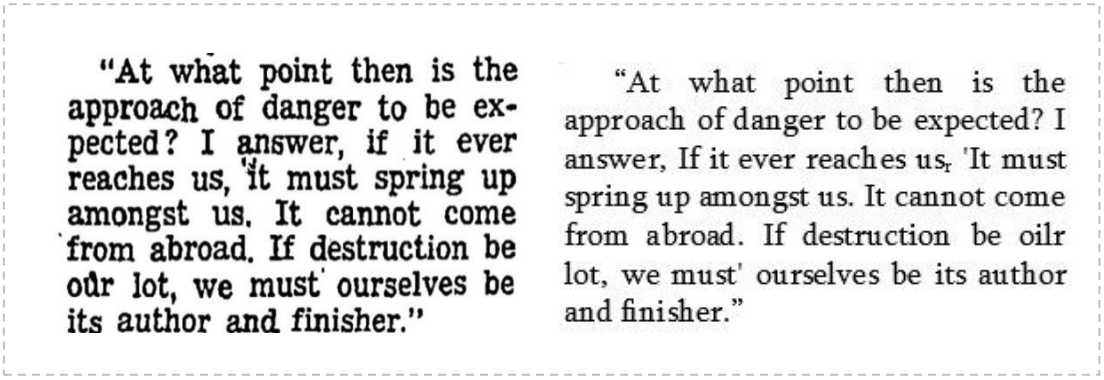
ZADAR- Studentsko savjetovalište u Zadru pridružilo se obilježavanju ovogodišnjeg, drugog po redu Tjedna psihologije, u sklopu čega je u petak otvorilo svoja vrata te za sve zainteresirane posjetitelje organiziralo Kavu u Savjetovalištu.

ZADAR- Studentsko savjetovalište u Zadru pridružilo se obilježavanju ovogodišnjeg, drugog po redu Tjedna psihologije, u sklopu čega je u petak otvorilo svoja vrata te za sve zainteresirane posjetitelje organiziralo Kavu u Savjetovalištu.

Slika 57. Članak iz novina na hrvatskom jeziku prije i nakon OCR-a (Convertio)

Optičko prepoznavanje članka na hrvatskom jeziku, uz pomoć aplikacije Convertio, provedeno je sa 100 % - tnom točnošću odnosno svih 198 slova i 8 interpunkcijskih znakova u potpunosti je prepoznato.

F) Članak na engleskom jeziku



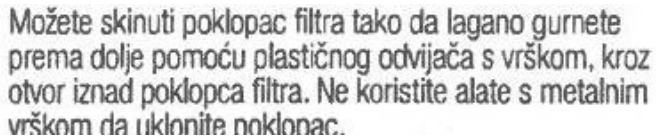
"At what point then is the approach of danger to be expected? I answer, if it ever reaches us, it must spring up amongst us. It cannot come from abroad. If destruction be our lot, we must ourselves be its author and finisher."

"At what point then is the approach of danger to be expected? I answer, If it ever reaches us, 'It must spring up amongst us. It cannot come from abroad. If destruction be oilr lot, we must' ourselves be its author and finisher."

Slika 58. Članak iz novina na engleskom jeziku prije i nakon procesa OCR-a (Convertio)

Analizirani ulomak članka na engleskom u originalu se sastoji od 175 slova i 10 interpunkcijskih znakova. Prilikom optičkog prepoznavanja znakova ovog ulomka na engleskom, koristeći aplikaciju Convertio, napravljena je samo jedna pogreška što znači da postotak uspješnosti optičkog prepoznavanja znakova iznosi 99.42 % za slova te 100 % za interpunkcijske znakove. Sveukupan postotak uspješnosti je 99.45 %. Treba napomenuti kako su i u ovom primjeru „nadodana" dva interpunkcijska znaka viška.

G) Tekst pisan vrlo sitnim fontom



Slika 59. Tekst pisan sitnim fontom prije procesa OCR-a (Convertio)

Možete skinuti poklopac filtra tako da lagano gurnete
prema dolje pomoću plastičnog odvijača s vrškom, kroz
otvor iznad poklopca filtra. Ne koristite alate s metalnim
vrškom da uklonite poklopac.

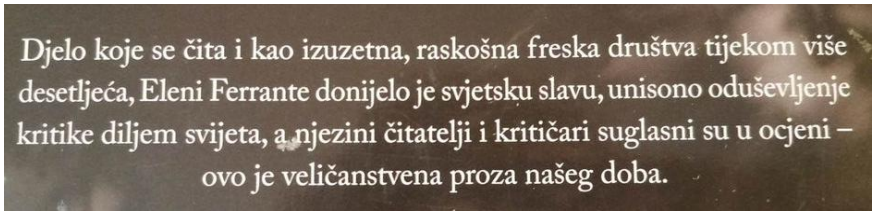
Slika 60. Tekst pisan sitnim fontom nakon procesa OCR-a (Convertio)

Prilikom optičkog prepoznavanja znakova iz primjera sa vrlo sitnim fontom slova (upute za rad sa perilicom rublja) sva slova i svi interpunkcijski znakovi prepoznati su sa 100% - tnom točnošću (162 slova i 3 interpunkcijska znaka).

H) Tekst sa neuobičajenim fontom – strip

Optičko prepoznavanje znakova teksta sa neuobičajenim fontom odnosno stripa pokazalo se u potpunosti neuspješnim zato što aplikacija nije dala nikakav rezultat.

I) Svijetli tekst sa tamnom pozadinom



Slika 61. Svijetli tekst na tamnoj pozadini prije procesa OCR-a (Convertio)

Djelo koje se čita i kao izuzetna, raskošna freska društva tijekom više
desetljeća, Eleni Ferrante donijelo je svjetsku slavu, unisono oduševljenje
kritike diljem svijeta, njezini čitatelji i kritičari suglasni su u ocjeni -
ovo je veličanstvena proza našeg doba.

Slika 62. Svijetli tekst na tamnoj pozadini nakon OCR-a (Convertio)

Optičkim prepoznavanjem znakova primjera svijetlog teksta na tamnoj pozadini, točno je prepoznato svih 218 slova te svih 6 interpunkcijskih znakova.

5.5. Rezultati istraživanja prikazani tablično[147]

Slova	Tekst iz knjige na hrv. jeziku	Zgužvani tekst na hrv./polj. jeziku	Tekst iz knjige na eng. jeziku	Tekst pisan rukom	Članak na hrv. jeziku	Članak na eng. jeziku	Vrlo sitni tisak	Neuobi - čajeni font strip	Svijetli tekst na tamnoj pozadini
ABBYY	100	99.54	100	87.32	100	98.85	100	95.04	100
Free OCR (poljski)	100	96.33	96.11	34.74	94.94	98.66	87.03	/	83.48
Google Drive	100	100	100	99.53	100	99.42	100	100*	100
I2OCR	100	100	100	93.89	100	99.42	100	/	/
Convertio	100	99.54	93.75	99.33	100	99.42	100	/	/

Tablica 1. Rezultati optičkog prepoznavanja **slova** u svim aplikacijama

Znakovi	Tekst iz knjige na hrv. jeziku	Zgužvani tekst na hrv. jeziku	Tekst iz knjige na eng. jeziku	Tekst pisan rukom	Članak na hrv. Jeziku	Članak na eng. jeziku	Vrlo sitni tisak	Neuobi - čajeni font strip	Svijetli tekst na tamnoj pozadini
ABBYY	100	93.75	100	23.52	100	90	100	100	100
Free OCR	100	93.75	68.75	82.35	100	100	100	/	100
Google Drive	100	100	100	94.11	87.5	100	100	100*	100
I2OCR	100	100	100	76.47	100	90	100	/	100
Convertio	100	93.75	100	11.76	100	100	100	/	100

Tablica 2. Rezultati optičkog prepoznavanja interpunkcijskih **znakova** u svim aplikacijama

[147] Rezultati su prikazani u postotcima.

Slova i znakovi	Tekst iz knjige na hrv. jeziku	Zgužvani tekst na hrv. jeziku	Tekst iz knjige na eng. jeziku	Tekst pisan rukom	Članak na hrv. Jeziku	Članak na eng. jeziku	Vrlo sitni tisak	Neuobi - čajeni font strip	Svijetli tekst na tamnoj pozadini
ABBYY	100	99.33	100	82.6	100	98.37	100	87.78	100
Free OCR	100	93.15	94.64	38.26	95.14	93.51	85.45	/	81.25
Google Drive	100	100	100	99.13	99.51	94.05	100	100*	100
I2OCR	100	100	100	92.60	100	98.91	100	/	/
Convertio	100	99.33	100	82.53	100	99.45	100	/	100

Tablica 3. Sveukupni rezultati optičkog prepoznavanja **slova i interpunkcijskih znakova** svih aplikacija

5.6. Rasprava

U istraživačkom dijelu rada odabrano je pet aplikacija za optičko prepoznavanje znakova te se istraživalo njihove osobine i uspješnost postupka optičkog prepoznavanja znakova. U sklopu ovog dijela rada uspoređuju se odgovori na pet postavljenih istraživačkih pitanja te donose određeni zaključci.

Istraživačka pitanja su bila:

1. Da li je analizirana aplikacija otvorenog koda i kako ju se može koristiti?
2. Kakve mogućnosti pohrane nudi analizirana aplikacija?
3. Koje su dodane vrijednosti svake od aplikacija (faktor iznenađenja)?
4. Koliko jezika podržava analizirana aplikacija i da li je hrvatski jedan od njih?
5. S kolikom uspješnošću se mogu optički prepoznati slijedeći dokumenti :

(A) tekst iz knjige,
(B) zgužvani tekst,
(C) tekst pisan rukom,
(D) članak iz novina,
(E) tekst pisan vrlo sitnim fontom,
(F) tekst sa neuobičajenim fontom = strip,
(G) tekst koji je mnogo puta kopiran,
(H) svijetli tekst sa tamnom pozadinom
Odgovori na istraživačka pitanja:

1) Pokazalo se da od svih aplikacija samo ABBYY FineReader 15 i Convertio nisu otvorenog koda odnosno nisu besplatno dostupne te ih nije moguće slobodno distribuirati i mijenjati. Convertio aplikacija mrežno je dostupna samo za rad sa deset dokumenata dok se za daljnje optičko prepoznavanje znakova treba platiti, a ABBYY FineReader 15 aplikacija mogla se koristiti samo u probnom roku od trideset dana, što je bilo dovoljno za potrebe istraživačkog dijela rada.

2) S obzirom na mogućnost odabira formata za pohranu prednjači aplikacija Convertio koja podržava mogućnost pohrane izlaznog rezultata u 11 formata (.doc, .xlsx, .xls, .pptx, .pdf, .txt, .rtf, .csv, .epub, .fb2 i .djvu), a slijedi ju aplikacija ABBYY FineReader 15 koja ima mogućnost pohrane u 9 formata (.pdf, .doc, .xlsx, .epub, .epub, .txt, .djvu, .fb2 i .csv.), dok je odmah iza njega Google Drive OCR s mogućnošću pohrane u 7 formata (.doc, .odt, .rtf, .pdf,

.txt, .html, .epub). Aplikacija I2OCR nudi mogućnost pohrane samo u četiri formata (.txt, .doc, .docx, .pdf i .html), dok se u sklopu Free OCR aplikacije prepoznati tekst može spremiti samo u .txt i .rtf formatu.

3) Analizom dodanih vrijednosti pojedinačnih aplikacija dolazi se do zaključka kako ABBYY FineReader i Google Drive OCR prednjače već pri samom pregledu korisničkog sučelja. Osim što ima mogućnost odvajanja tekstualnog dijela od slikovnog, s ABBY FineReaderom moguće je raditi s više dokumenata istovremeno, dobiti pregled finalnog rezultata u navedenim formatima u bilo kojem trenutku procesa optičkog prepoznavanja znakova te su sve preostale funkcije vrlo pregledne u korisničkom sučelju. Google Drive OCR ima čak i više dodatnih funkcija koje se ponajviše odnose na uređivanje teksta za finalni prikaz. S obzirom da ima formu aplikacije Word, u sklopu aplikacije Google Drive OCR ponuđene su i slične funkcije kao u aplikaciji Word. Također, rad u ovoj aplikaciji omogućuje spremanje dokumenta u bilo kojem stadiju procesa na Google Drive mjesto za pohranu, do slijedećeg otvaranja dokumenta. Preostale tri aplikacije, Free OCR, I2OCR i Convertio, pružaju mogućnost korištenja samo osnovnih funkcija, odnosno može se odabrati jezik teksta koji se želi optički prepoznati, uređivati prepoznati tekst te odabrati izlazni format.

4) Odgovorom na četvrto istraživačko pitanje saznalo se da četiri od pet analiziranih aplikacija podržavaju rad s hrvatskim jezikom i to su: ABBYY FineReader 15, I2OCR, Google Drive OCR i Convertio. Preostala aplikacija, Free OCR, podržava znatno manji fond jezika s kojima se može raditi, a kako bi se prepoznao tekst na hrvatskom jeziku aplikacija je postavljena na poljski jezik uz pomoć kojega su prepoznati hrvatski dijakritički znakovi. Prema broju jezika koje podržavaju prednjači Google Drive OCR sa 193 jezika i 36 jezika na čijem se optičkom prepoznavanju radi, a slijede ga ABBYY FineReader 15 sa 192 jezika i I2OCR sa 117 jezika dok aplikacija Free OCR podržava tek 11 jezika.

5a) U sklopu ABBYY FineReader 15 aplikacije, bez greške su optički prepoznati tekst iz knjige, članak na hrvatskom jeziku, tekst sa vrlo sitnim fontom, svijetli tekst na tamnoj pozadini te tekst iz knjige na engleskom jeziku. Zgužvani tekst i članak na engleskom prepoznati su sveukupno s po 3 greške što je relativno uspješan rezultat. Prepoznavanje rukom pisanog teksta, s obzirom na 40 grešaka (slova i interpunkcijski znakovi), može se smatrati neuspjelim optičkim prepoznavanjem znakova. Razlog tomu je što analizirana aplikacija nije strojno naučena na prepoznavanje različitih stilova rukopisa odnosno nije predviđena za prepoznavanje rukom pisanog teksta, te se na temelju implementiranih algoritama i pohranjenih baza znakova dolazi do približnih, ali ne i točnih rezultata. Kako se

radi o aplikaciji koja je dosta popularna, prema navodima više različitih autora stručnih članaka kao što su Tafti[148]i Helinski[149], ne bi bilo naodmet da se porazmisli o implementaciji ove opcije. Također, prilikom prepoznavanja teksta sa neuobičajenim fontom odnosno stripa napravljeno je sveukupno 16 grešaka što i ovaj rezultat čini neuspjelim.

5b) Prilikom optičkog prepoznavanja znakova koristeći Free OCR aplikaciju, u potpunosti je prepoznat samo tekst skeniran iz knjige. Razlog tomu je vjerojatno velika osjetljivost aplikacije na šum u tekstu i izostanak ili zakazanost funkcije različitih filtera kao što je medijan ili nekih od algoritama u sklopu aplikacije. Zgužvani tekst prepoznat je s 23 greške, tekst na engleskom jeziku s 6, članak na hrvatskom jeziku s 10, članak na engleskom s 12 i tekst pisan sitnim fontom s 24 greške. Iako ovi rezultati nisu optimalni, uzeti su kao prihvatljivi. Optičko prepoznavanje rukom pisanog teksta i svijetlog teksta na tamnoj pozadini prepoznato je s po 42 greške te se može smatrati neuspjelim. Dok je za potrebe optičkog prepoznavanja teksta iz knjige pokazala iznenađujuću točnost, ova mrežno dostupna aplikacija u potpunosti je zakazala što se tiče optičkog prepoznavanja rukom pisanog teksta, što se vjerojatno može obrazložiti bazom prethodno pohranjenih znakova koji ne obuhvaćaju različite stilove pisanja rukom.

5c) Google Drive OCR aplikacija, globalno je pokazala najbolje rezultate optičkog prepoznavanja znakova za gotovo sve analizirane vrste dokumenata. U potpunosti su optički prepoznati tekst iz knjige, zgužvani tekst iz knjige, tekst iz knjige na engleskom jeziku, tekst pisan vrlo sitnim fontom, strip te svijetli tekst na tamnoj pozadini. Ova aplikacija predstavlja potpuno iznenađenje što se tiče optičkog prepoznavanja rukom pisanog teksta prilikom čijeg prepoznavanja su napravljene samo 2 greške što znači da je aplikacija dobro strojno izučena za prepoznavanje i teksta na hrvatskom jeziku i rukom pisanog teksta odnosno u sklopu nje dobro su definirane metode za normalizaciju teksta te potom i metode za poučavanje i prepoznavanje, o kojima se u teoretskom dijelu rada ponajviše govorilo pozivajući se na radove Chaudhuri[150], a potom i drugih autora. Tijekom optičkog prepoznavanja članka iz novina na hrvatskom jeziku napravljena je samo jedna greška što znači da kada bi se radilo o stvarnom projektu, ljudska intervencija i dodatni utrošak vremena bili bi minorni. Najviše grešaka (11) napravljeno je prilikom prepoznavanja članka na engleskom jeziku.

[148] Tafti, Ahmad P.; Baghaie, Ahmadreza; Assefi, Mehdi; Arabnia, Hamid R.; Yu, Zeyun; Peissig, Peggy. Op. Cit.
[149] Helinski, Marcin; Kmieciak, Milosz; Parkola, Tomasz. Op, cit.
[150] Usp. Chaudhuri, Arindam…[et al.]. Op. Cit.

5d) Uz pomoć aplikacije I2OCR u potpunosti točno su prepoznati tekst iz knjige, zgužvani tekst, tekst iz knjige na engleskom jeziku, članak na hrvatskom jeziku te tekst pisan vrlo sitnim fontom. Prilikom optičkog prepoznavanja rukom pisanog teksta napravljeno je 17 grešaka što u usporedbi s dosta lošijim rezultatima nekih drugih aplikacija predstavlja uspjeh. Članak na engleskom jeziku prepoznat je sa 2 greške. U sklopu ove aplikacije u potpunosti su neprepoznati tekst s neuobičajenim fontom (strip) te svijetli tekst na tamnoj pozadini. Razlog ovomu može biti što aplikacija nema predviđene algoritme za prepoznavanje slikovnih prikaza s tekstom kao ni za prepoznavanje teksta na tamnoj pozadini, što je zapravo šteta s obzirom na dobre preostale rezultate.

5e) Aplikacija Convertio pokazala je točne rezultate prilikom optičkog prepoznavanja teksta iz knjige, teksta na engleskom jeziku, članka na hrvatskom jeziku, teksta pisanog vrlo sitnim fontom te svijetlog teksta na tamnoj pozadini. Zgužvani tekst iz knjige prepoznat je s 3 greške, a članak na engleskom jeziku s jednom greškom. Tekst s neuobičajenim fontom, odnosno strip, u potpunosti je neprepoznat. Aplikacija Convertio je prilikom prepoznavanja rukom pisanog teksta dala rezultat s 31 pogreškom što prepoznavanje čini neuspješnim.

6. Zaključak

Proces optičkog prepoznavanja znakova još uvijek nije toliko poznat no teško da bi se mnogi od poznatih procesa danas odvijali bez njega te je upravo iz tog razloga tema ovog rada. Ovim istraživanjem nastojalo se odgovoriti što se odvija u pozadini procesa optičkog prepoznavanja znakova što je analizirano u teoretskom dijelu rada u kojem se najvećim dijelom opisuje funkcioniranje osam komponenti optičkog prepoznavanja znakova. Došlo se do saznanja kako se u pozadini svakog od procesa optičkog prepoznavanja znakova nalaze unaprijed definirane kompleksne naredbe kao što su npr. algoritmi i metode kao thresholding. Sažeto rečeno, optičko prepoznavanje znakova je proces pretvorbe analognih dokumenata u digitalizirane pretražive dokumente, što je detaljno analizirano u ovom dijelu rada.

U istraživačkom dijelu rada provedena je analiza rada pet aplikacija za optičko prepoznavanje znakova i to ABBYY FineReadera 15, Free OCR, I2OCR, Google Drive OCR i Convertio. Prema odgovorima na pet istraživačkih pitanja, aplikacija Google Drive OCR se pokazala najboljom opcijom za optičko prepoznavanje svih vrsta dokumenata analiziranih u radu. Radi se o aplikaciji otvorenog koda, s dovoljno različitih formata za pohranu, mogućnošću rada sa više od 248 jezika, bogatim korisničkim sučeljem i kvalitetnim dodanim vrijednostima. Uz pomoć ove aplikacije uspješno je prepoznato 6 od 9 analiziranih vrsta dokumenata. Posebno treba naglasiti kako su prilikom optičkog prepoznavanja teksta pisanog rukom napravljene samo dvije greške što znači da su u sklopu ove aplikacije implementirane prethodno bogate formirane baze slova i znakova te dobro osmišljeni algoritmi. Slijedi aplikacija ABBYY FineReader 15 u sklopu koje je u potpunosti uspješno prepoznato 5 od 9 analiziranih vrsta dokumenata. Iako ova aplikacija nije otvorenog koda ima mogućnost rada sa 193 jezika, pohrane u 9 formata te bogato korisničko sučelje. Statistički gledano, vrlo slične rezultate imala je aplikacija Convertio, no s obzirom da aplikacija ABBY FineReader 15 nudi brojne dodatne opcije zaključno je odabrana kao bolji izbor. Pomoću I2OCR, male mrežno dostupne aplikacije otvorenog koda sa mogućnošću pohrane u četiri formata, korisničkim sučeljem samo s osnovnim funkcijama i mogućnošću prepoznavanja tekstova na 137 jezika, uspješno je prepoznato 5 od 9 vrsta dokumenata no nažalost tekst sa neuobičajenim fontom te svijetli tekst na tamnoj pozadini u potpunosti su neprepoznati. Pomoću aplikacije Free OCR u potpunosti je prepoznat samo tekst iz knjige na hrvatskom jeziku. Tekst pisan rukom vrlo loše je prepoznat a aplikacija, kao i I2OCR, ne nudi mogućnost prepoznavanja svijetlog teksta na tamnoj pozadini.

Iz rezultata se može zaključiti kako i neke od vrlo popularnih aplikacija, kao što je ABBYY FineReader 15, imaju nedostatke dok, s druge strane, ne tako popularna aplikacija, kao što je Convertio, može iznenaditi i u potpunosti točno prepoznati svijetli tekst na tamnoj pozadini. Generalno, sve analizirane aplikacije dale su besprijekorne rezultate prilikom optičkog prepoznavanja znakova teksta skeniranog iz knjige, iz čega možemo zaključiti kako je arhitektura svih analiziranih aplikacija primarno prilagođena prepoznavanju takve vrste dokumenta. Za gotovo sve aplikacije, osim Google Drive OCR i I2OCR, tekst pisan rukom predstavljao je najveći izazov prilikom optičkog prepoznavanja znakova. Tekst iz knjige na engleskom jeziku, članci na hrvatskom i engleskom jeziku, zgužvani tekst na hrvatskom jeziku i tekst pisan vrlo sitnim fontom relativno su dobro prepoznati u sklopu svih analiziranih aplikacija. Prilikom optičkog prepoznavanja teksta sa neuobičajenim fontom 3 od 5 analiziranih aplikacija nisu dale nikakve rezultate što ne začuđuje jer takva vrsta slikovno – tekstualnog prikaza zahtijeva višu razinu kompleksnosti koju je pokazala aplikacija ABBYY FineReader 15 te donekle i Google Drive OCR koja unatoč besprijekornom prepoznavanju znakova i slova nije uspjela prepoznati da se radi o tekstu u dvije zasebne cjeline (oblačića).

Završno, iako su rezultati nekih od analiziranih aplikacija za optičko prepoznavanje znakova vrlo optimistični, za potrebe ovog rada uzeti su primjeri s malim brojem znakova i slova kako bi se jednostavnije došlo do rezultata. Treba uzeti u obzir da veći uzorak znači i veći broj grešaka te veću ljudsku intervenciju i utrošak vremena. Veliko je pitanje koliko bi zapravo ove aplikacije dale uspješne rezultate kada bi se trebalo optički prepoznati čak i samo jednu tiskanu knjigu te naročito ako bi ih se koristilo za opsežne projekte digitalizacije. Odgovor na ovo pitanje postavlja se kao jedno od mogućih slijedećih istraživanja.

7. Popis literature

ABBYY FineReader 15. URL: https://pdf.abbyy.com/?utm_autosource=google&utm_automedium=cpc&utm_autocampaign=EEU_FineReader_Search_OCR&gclid=Cj0KCQjwreT8BRDTARIsAJLI0KLDwAcs5rceRU8610GS-mYf33sOuaeFSx2kGKirAKUK4d67eD65Wf4aAsfsEALw_wcB

Aparna, A. et al. Optical Character Recognition for Handwritten Cursive English characters. // International Journal of Computer Science and Information Technologies 5,1(2014), str 847-848.

Balog Vojak, Jelena; Šinkić, Zdenka. Projekt digitalizacije hemeroteke Hrvatskog povijesnog muzeja.//Informatica museologica 44, 1-4(2013) URL: https://hrcak.srce.hr/index.php?show=clanak&id_clanak_jezik=257178

Blažević, Antonela. Primjena skrivenih Markovljevih modela za modeliranje i predviđanje meteoroloških pojava. (dip. rad, Fakultet elektrotehnike i računarstvu u Zagrebu, 2017.)

Bojan, Šekoranja. Segmentacija slike. (Predavanje, Fakultet strojarstva i brodogradnje u Zagrebu) URL: http://www.sjever.fsb.hr/vizija/predavanja/Segmentacija%20slike-sekoranja.ppt

Carnet. Biometrija. URL: https://www.cert.hr/wp-content/uploads/2006/09/CCERT-PUBDOC-2006-09-167.pdf

Convertio. URL: https://convertio.co/

Cvisiontech. OCR software primer. Thresholding with OCR. URL: http://www.cvisiontech.com/resources/ocr-primer/thresholding-within-ocr.html

Čurković, Petar; Stipančić, Tomislav. Segmentacija slike. (Predavanje, Fakultet strojarstva i brodogradnje) URL: http://www.sjever.fsb.hr/vizija/predavanja/Segmentacija%20slike-sekoranja.ppt

Dalbelo Bašić, Bojana. Umjetna inteligencija. Uvod u strojno učenje. (Predavanje, Fakultet elektrotehnike i računarstva u Zagrebu) URL: http://degiorgi.math.hr/~singer/ui/ui_1415/UI_10_UvodUStrojnoUcenje.pdf

Digitalizacija. Leksikografski zavod Miroslava Krleže. Encikopedija. URL: http://enciklopedija.hr/Natuknica.aspxID=68025

Dimić, Marina. Biometrijski sustavi identifikacije. (Dip. rad, Studij Informatologije u Osijeku) URL: https://repozitorij.ffos.hr/islandora/object/ffos:2344/preview

Ferrante, Elena. Genijalna prijateljica. Zagreb: Profil, 2020.

Forsyth, David A.; Ponce, Jean. Computer Vision. A modern approach. New Jersey: Pearson, 2003. URL: https://eclass.teicrete.gr/modules/document/file.php/TM152/Books/Computer%20Vision%20-%20A%20Modern%20Approach%20-%20D.%20Forsyth,%20J.%20Ponce.pdf

Fu, K.C. Sequential Methods in Pattern Recognition and Machine Learning. New York: Academic Press, 1968. URL: https://books.google.hr/books?hl=hr&lr=&id=TrFoHng-H8MC&oi=fnd&pg=PP1&dq=pattern+recognition+and+machine+learning&ots=ZhzU5SuS-L&sig=GgFmnZAgWPopc6pNiol3pFD0xn8&redir_esc=y#v=onepage&q=pattern%20recognition%20and%20machine%20learning&f=false

Google Drive. URL: https://www.google.com/intl/hr_HR/drive/

Google Privacy and Terms. URL: https://policies.google.com/technologies/pattern-recognition?hl=hr

Gross, Ari. Understanding OCR Technology. Thresholding within OCR. URL: http://www.cvisiontech.com/resources/ocr-primer/thresholding-within-ocr.html

History of computer. The Reading Machine (first OCR device) of Gustav Tauschek. URL: https://history-computer.com/ModernComputer/Basis/OCR.html

Hmoumen, Marouane. A review of optical character recognition system. // Design of Machines and Structures 7, 2(2017)

Hrga, Milan. Računalni vid.// Zbornik radova Veleučilišta u Šibeniku 1-2(2018), str. 208. URL: https://hrcak.srce.hr/index.php?show=clanak&id_clanak_jezik=292457

I2OCR. URL: http://www.i2ocr.com/

IFLA Guidelines for digitization projects. URL: https://www.ifla.org/files/assets/preservation-and-conservation/publications/digitization-projects-guidelines.pdf

Image Processing. Image analysis. Brittanica. URL: https://www.britannica.com/technology/information-processing/Image-analysis#ref212121

Intan, Fariza Bt Haruman; Foong, Oi-mean. OCR signage recognition with skew & slant correction for visually impaired people. // 11th International Conference on Hybrid Intelligent Systems, HIS 2011. URL: https://www.researchgate.net/publication/220981126_OCR_signage_recognition_with_skew_slant_correction_for_visually_impaired_people

Islam, Norman; Islam, Zeeshan.; Noor, Nazia. A Survey on Optical Character Recognition System. // ITB Journal of Information and Communication Technology. 2015. URL: https://www.researchgate.net/publication/320442536_A_Survey_on_Optical_Character_Recognition_System

Karić, Miran; Krpić, Zdravko. Optičko prepoznavanje znakova na grid i višejezgrenim platformama. // Tehnički vjesnik 20, 4(2013.)

Kaur, Amendeep; Baghla, Seema; Kumar, Sunil. Study of various character segmentation techniques for handwritten off-line cursive words: a review. // International Journal of Advances in Science Engineering and Technology 3, 3(2015) , Str. 154. URL: http://www.iraj.in/journal/journal_file/journal_pdf/6-162-1440573382154-158.pdf

Križaić, Damjan. Postupci segmentacije objekata zadanih poligonalnom mrežom. (dip.rad, Fakultet elektrotehnike i računarstva u Zagrebu).Hrvatska znanstvena bibliografija. URL: https://bib.irb.hr/datoteka/714522.diplomski-0036442298.pdf

Lesić, Dragan; Oblučar, Bojan. Optičko prepoznavanje glazbenog crtovlja. Fakultet elektrotenike i računarstva. Repozitorij. URL: https://www.fer.unizg.hr/_download/repository/Opticko_prepoznavanje_glazbenog_crtovlja.pdf

Lu, Yue; Lim Tan, Chew. A nearest-neighbor chain based approach to skew estimation in document images. // Pattern Recognition Letters 24 (2003)

Lukovac, Bojan. Sazrijevanje računalnog vida: Automatsko pronalaženje korespondencija. (Sem. rad, Fakultet elektrotehnike i računarstva u Zagrebu), URL: https://pdfs.semanticscholar.org/a876/5aeef9ab286b40280948f5ffc89456673ecc.pdf

Marković, Darija. Osnove umjetne inteligencije. Strojno učenje. (Predavanje, Sveučilište Josipa Jurja Strossmayera u Osijeku) URL: http://www.mathos.unios.hr/oui/p11.pdf

Martinović, Anđelo. Raspoznavanje znakova na registarskim tablicama. (Zav. rad, Sveučilište u Zagrebu. Fakultet elektrotehnike i računarstva, 2008.) URL: http://www.zemris.fer.hr/~kalfa/ZR/Martinovic_ZR_2008.pdf

Nehta, Nikita; Doshi, Jyotika. A Review of Handwritten Character Recognition. //International Journal of Computer Applications 165, 4(2017)

Ong, Veronica; Suhatorno, Derwin. Using k-nearest neighbor in optical character recognition. // ComTech 7, 1(2016), str. 55. URL: https://media.neliti.com/media/publications/165987-EN-using-k-nearest-neighbor-in-optical-char.pdf

Panwar, Subhash; Nain, Neeta. A Novel Approach of Skew Normalization for Handwritten Text Lines and Words. 2012 Eighth International Conference on Signal Image Technology and Internet Based Systems. URL: https://www.researchgate.net/publication/235218361_A_Novel_Approach_of_Skew_Normalization_for_Handwritten_Text_Lines_and_Words/download

Patel, Craig; Patel, Atul; Patel; Dharmendra. Optical Character Recognition by Open Source OCR Tool Tesseract: A Case Study. // International Journal of Computer Applications 55, 10(2012) URL: https://www.researchgate.net/profile/Chirag_Patel27/publication/235956427_Optical_Character_Recognition_by_Open_source_OCR_Tool_Tesseract_A_Case_Study/links/00463516fa43a64739000000.pdf
Pattern recognition. Britannica. URL: https://www.britannica.com/technology/pattern-recognition-computer-science

Radošević, Danijel. Postupci i problemi optičkog prepoznavanja teksta.// Zbornik radova 21(1996). URL: https://hrcak.srce.hr/index.php?show=clanak&id_clanak_jezik=117350

Salopek, Damir. Sustav za prepoznavanje uzoraka pri varijantnom konstruiranju. (mag. rad, Fakultet strojarstva i brodogradnje u Zagrebu, 2002.) URL: https://www.bib.irb.hr/186661

Seiter-Šverko, Dunja; Križaj Lana. Digitalizacija kulturne baštine u Republici Hrvatskoj: Od trenutne situacije prema nacionalnoj strategiji.// Vjesnik bibliotekara Hrvatske 55, 2(2012), str. 29-40.

Smith, Ray (2007). "An Overview of the Tesseract OCR Engine", Ray Smith, Proc. Ninth Int. Conference on Document Analysis and Recognition (ICDAR), 2007, pp. 629-633. URL: https://static.googleusercontent.com/media/research.google.com/hr//pubs/archive/33418.pdf

Stančić, Hrvoje; Zanier, Katharina. Heritage Live. Upravljanje baštinom uz pomoć informacijskih alata. URL: https://www.had-info.hr/dokumenti/publikacije/Heritage%20live%20-%20Upravljanje%20bastinom%20uz%20pomoc%20informacijskih%20alata.pdf

Stipanović, Zoran. Primjena OCR-a u vektorizaciji katastarskih planova. (dip. rad, Geodetski fakultet Zagreb, 2004) URL: https://www.bib.irb.hr/147904
Strategija e-Hrvatska 2020. URL: https://uprava.gov.hr/UserDocsImages/Istaknute%20teme/e-Hrvatska/Strategija_e-Hrvatska_2020.pdf

Šapro-Ficović, Marica. Masovna digitalizacija knjiga: utjecaj na knjižnice. // Vjesnik bibliotekara Hrvatske 54, 1/2(2011), str. 217. URL: https://hrcak.srce.hr/80483

Škrabo, Katarina; Vrana, Radovan. Digitalne zbirke u narodnim knjižnicama u Hrvatskoj.// Vjesnik bibliotekara Hrvatke 60, 1(2017)

Tafti, Ahmad P.; Baghaie, Ahmadreza; Assefi, Mehdi; Arabnia, Hamid R.; Yu, Zeyun; Peissig, Peggy. OCR as a Service: An Experimental Evaluation of Google Docs OCR, Tesseract; ABBYY FineReader and Transym., URL: https://www.researchgate.net/publication/310645810_OCR_as_a_Service_An_Experimental_Evaluation_of_Google_Docs_OCR_Tesseract_ABBYY_FineReader_and_Transym

Vynckier, Ivo. How OCR works. URL: http://www.how-ocr-works.com/

Walls, John. OCR and Content Management with SAP and Imaging (2008) URL: https://www.slideshare.net/verbella/ocr-and-content-management-with-sap-and-imaging

Witten. Ian H., Frank, Eibe. Data Mining. Practical Machine Learning Tools and Techniques. San Francisco: Elsevier, 2005. URL: http://thuvien.thanglong.edu.vn:8081/dspace/bitstream/DHTL_123456789/4050/1/Data%20mining-1.pdf

Comparison of optical character recognition applications

Summary

The aim of this paper is to resolve what exactly lies behind the optical character recognition process, what its components are, and how it looks in practice. An optical character recognition, in simple words is translating an analog document into a searchable digital version and the concept of digitization is explained as the root part of the theoretical part of the paper, and what are the projects and initiatives of digitization that would not exist without optical character recognition. This process is interwoven with several related disciplines, of which the pattern recognition, computer vision, and machine learning are briefly explained in the paper. The theoretical part of the paper is formulated around a description of eight major components of the optical character recognition process: optical scanning, segmentation of image regions, preprocessing, normalization, segmentation, representation, feature extraction, and teaching and recognition. Although most of the literature is mathematically based, the components of optical character recognition are mathematically illustrated to the limit understandable to all. Optical Character Recognition applications were then introduced and a comparison study was performed on ABBYY FineReader 15, Free OCR, Google Drive OCR, I2OCR and Convertio applications. The methodology used in the research is a qualitative comparison of five different optical character recognition applications and a quantitative comparison of how many errors were made and how many languages were included in the architecture of each application. After the research, a discussion of the results was conducted and conclusions were drawn based on the research and the theory analyzed.

Keywords: optical character recognition, digitization, ABBYY FineReader 15, Tesseract Google Drive OCR, Free OCR, Convertio, I2OCR