Sveučilište u Zagrebu
Filozofski fakultet Zagreb
Odsjek za informacijske i komunikacijske znanosti
Akademska godina 2016./2017.

Ivan Taradi

Mogućnosti unapređenja programa za optičko prepoznavanje znakova (OCR programa)
Završni rad

Mentor: Dr. sc. Hrvoje Stančić, izv. prof.

Zagreb, rujan 2017.

Sadržaj

Popis slika

1. Uvod

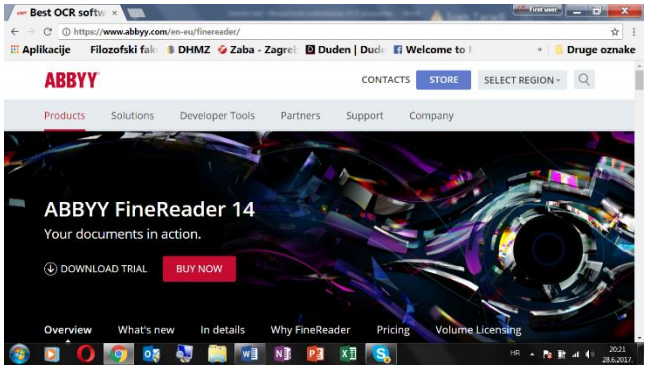
Programi za optičko prepoznavanje znakova (engl. *Optical Character Recognition, OCR)* su programi koji omogućuju pretvaranje odnosno konverziju različitih digitalnih slikovnih datoteka, zapravo teksta u njima u računalno čitljiv tekst. Primjerice, OCR programi omogućuju konvertiranje PDF dokumenta u Microsoft Word dokument. Dakle, uz pomoć OCR programa iz dokumenta u formatu PDF moguće je dobiti drugi tip digitalnog dokumenta koji se tada može uređivati, pretraživati i slično. Još jedna vrlo bitna funkcija OCR programa je i njihova široka primjena u digitalizaciji analognih (tiskanih) tekstova. Uz pomoć skeniranja i OCR programa analogno gradivo, dakle tekstove otisnute na papiru, možemo digitalizirati i odmah pretvoriti u tekstualnu datoteku (primjerice u Microsoft Word dokument) koja se može uređivati, pretraživati, indeksirati, konvertirati u neki drugi digitalni format i raditi razne druge radnje procesiranja.

U današnje vrijeme postalo je vrlo važno svo analogno (materijalno) gradivo informacijskih institucija imati i u digitalnom obliku. To je razumljivo s obzirom na činjenicu da suvremeni čovjek želi imati uvijek dostupne sve moguće informacije i sadržaje putem informacijske i komunikacijske tehnologije (engl. information and communication technology, ICT). Ljudi žele imati pristup gotovo svemu, i to u realnom vremenu i ne samo od kuće posredstvom osobnog računala, već svugdje putem pametnog telefona ili tableta koji nose u džepu. Dakle, digitalizacija analognog gradiva nameće se u suvremenom društvu sama po sebi kao standard. Digitaliziraju svi, informacijske institucije, poslovne organizacije, ali i sasvim obični ljudi i osobe s povećanim potrebama. Većina ljudi digitalizira iz sasvim praktičnih razloga kao što su izrada preslike neke potvrde, osobnih dokumenata i sl. Digitaliziranje iz takvih jednostavnih razloga uglavnom ne zahtijeva uporabu OCR programa. Kod osoba s povećanim potrebama, npr. osoba s određenim smetnjama i oštećenjima vida i motorike tijela situacija je ipak dosta drugačija. Takve osobe trebaju digitalizaciju tiskanih tekstova kako bi mogle funkcionirati u današnjem društvu, odnosno kako bi mogle ravnopravno sudjelovati u svim područjima života. No OCR programi nisu savršeni i njihova učinkovitost nije 100%-tna. Dakle, kod optičkog prepoznavanja znakova ima određenih problema. O tim nesavršenostima OCR programa i mogućnošću njihovog unapređenja bit će riječ u ovom završnom radu.

2. Povijest i razvoj programa za optičko prepoznavanje znakova

Kao što je već navedeno, uloga OCR programa pri digitalizaciji tiskanog teksta je vrlo važna. No prije nego što se pobliže analiziraju OCR programi i njihovi nedostaci, potrebno je reći nešto i o samoj digitalizaciji. Digitalizacija je, u najširem smislu, proces prevođenja analognog signala u digitalan oblik. U užem smislu, kojim se ovaj završni rad bavi, digitalizacija je pretvorba teksta, slike, zvuka, pokretnih slika (filmova i videa) ili trodimenzionalnog oblika nekog objekta u digitalni oblik. Taj digitalni oblik je u pravilu binarni kôd zapisan kao računalna datoteka sa sažimanjem ili bez sažimanja podataka. Te datoteke se mogu obrađivati, pohranjivati i prenositi putem računala i računalnih sustava. Digitalizacija teksta, pobliže, je zapravo postupak kojim skeniranjem od, na primjer, stupca otisnutog teksta dobijemo digitalnu sliku. Tu sliku može se prikazati na zaslonu računala i čitati, ali ne može se taj tekst uređivati i pretraživati. Da bi to bilo moguće potrebno je takve digitalne slike obraditi programima za optičko prepoznavanje znakova (OCR programima).

Ideja OCR sustava pojavila se ustvari i prije izuma elektroničkih računala i seže još u davne tridesete godine prošlog stoljeća. Prvi dokumentirani OCR sustav zamislio je Paul W. Handel. Radilo se o foto električnom mehanizmu i prepoznavanje znakova odvijalo bi se uz pomoć svjetlosne zrake koja je prolazila filterom s uzorcima znakova. Ako bi dovoljno svjetlosti prošlo kroz filter i podudaralo se sa znakom vraćalo bi se i signaliziralo podudarnost uzorka i znaka. Od tih prvih pokušaja sustavi za optičko prepoznavanje znakova se neprestano razvijaju. Jedna od prvih konkretnih primjena OCR programa vežu se uz Poštu Sjedinjenih Američkih Država (engl. U.S. Postal Service, USPS). Naime, ta je tvrtka 1965. godine počela eksperimentirati na području skeniranja i optičkog prepoznavanja tiskanog teksta (zapravo adresa i poštanskih brojeva na pošiljkama) kako bi mogla skenirati i verificirati otisnuti poštanski broj i adresu. Stvar je funkcionirala tako da bi nakon skeniranja računalo usporedilo skenirani poštanski broj i adresu s uzorcima koje je imalo pohranjeno u memoriji. U ranim osamdesetim godinama prošlog stoljeća u sklopu tih istraživanja u SAD-u razvijen je



Slika 1. OCR program Abbyy FineReader

Izvor: ABBY. Dostupno na: https://www.abbyy.com/en-eu/finereader/ (5.7.2017.)

uređaj koji je bio u stanju skenirati i prepoznati tri retka tiskanog teksta. Godine 1983. Pošta SAD-a je počela opremati svoje središnje poštanske urede OCR sustavima koji su znatno ubrzali sortiranje pošiljaka, pogotovo onih iz poslovnog sektora, i time je znatno ubrzala i pojednostavnila poslovanje te racionalizirala troškove (Britannica, 2017).

3. Suvremeni OCR sustavi i digitalizacija analognog gradiva

U svome radu pod nazivom *Postupci i problemi optičkog prepoznavanja teksta* iz 1996. godine Danijel Radošević objašnjava da se optičko prepoznavanje znakova pomoću OCR programa može podijeliti u tri osnovne faze.

Prva faza je dobivanje bitmape teksta koji se želi digitalizirati i učiniti obradivim. Ta faza uključuje skeniranje teksta uz pomoć ručnih ili stolnih skenera. Bolje je koristiti stolne skenere jer je njima moguće zahvatiti cijeli tekst na listu papira formata A4 i manja je mogućnost krivog nagiba teksta nakon skeniranja. Tekst se uglavnom skenira u jednobojnoj tehnici radi boljeg kontrasta. Skeniranjem dobivena slikovna datoteka je osnova i ključ za daljnji postupak optičkog prepoznavanja znakova.

Druga faza postupka prepoznavanja sastoji se od toga da se svakom znaku dodijele odgovarajući ASCII *(*akronim od engl. American Standard Code for Information Interchange*)*. To bi u prijevodu značilo *američki normirani kȍd za razmjenu informacija*, dakle, način kodiranja kojim se slovima, brojkama, interpunkcijskim znakovima te nekim grafičkim simbolima dodjeljuju brojčane vrijednosti (Hrvatska enciklopedija, 2017). U ovoj fazi je bitno da prepoznavanje bude dovoljno brzo kako bi OCR program bio uopće smatran učinkovitim. Ta brzina bi trebala odgovarati brzini kojom prosječna osoba čita tekst, a to je nekoliko stotina znakova u minuti. Današnji OCR programi bez problema ispunjavanju tu zadaću.

Treća faza postupka prepoznavanja teksta odnosi se na provjeru teksta dobivenog u drugoj fazi postupka prepoznavanja. Naime, potrebno je tekst provjeriti i ispraviti greške koje su eventualno nastale u prethodnoj fazi. Pri pronalaženju grešaka mogu pomoći programi za pronalaženje i ispravljanje gramatičkih grešaka. Zatim je potrebno posvetiti pažnju uređenju izgleda teksta, dakle, poravnava se tekst, uređuje se font i stil teksta te se podešavaju margine teksta. Ali i sve ostalo kako bi tekst dobio željeni izgled.

Budući da je danas prisutna sve veća potreba i potražnja za digitalnim sadržajima svih vrsta, razumljivo je da informacijske institucije poput arhiva, knjižnica i muzeja jednostavno moraju imati ponudu gradiva i predmeta u digitalnom obliku. Tako da je već postalo uobičajeno

da arhivi nude pristup digitaliziranoj arhivskoj građi, muzeji nude slobodan pristup digitalnim zbirkama muzejskih predmeta i virtualnim izložbama, a knjižnice sve više postaju hibridi između klasičnih i digitalnih knjižnica nudeći uz tiskane i elektroničke knjige (e-knjige). Tendencija je i da se digitalizira sve više književnih djela za koja su istekla autorska prava. Dakle, djela autora od čije je smrti prošlo 70 godina, i da se takvim elektroničkim knjigama omogući slobodan pristup (npr. internetska stranica besplatnih e-knjiga Project Gutenberg: https://www.gutenberg.org/). Digitalizacija arhivskog gradiva i slobodan pristup digitalnim arhivima smatraju se vrlo važnim u izgradnji demokratskog i transparentnog društva te je želja većine središnjih državnih arhiva digitalizirati što je moguće više gradiva i omogućiti mu slobodan pristup putem internetskih stranica i servisa. No digitalizacija je skupa i zahtjevna te ju nije moguće obaviti na brzinu.

S druge strane, ljudi digitaliziraju iz većinom praktičnih razloga. Mnogi jednostavno žele imati pohranjene digitalne verzija raznih analognih dokumenata, na primjer, pisama, rješenja, računa i sl. Ipak, neki ljudi koriste se digitalizacijom papirne građe odnosno tiskanih tekstova kako bi uopće mogli funkcionirati u društvu i sudjelovati u svim područjima života. Naime, već je navedeno da je slabovidnoj ili slijepoj osobi, ili osobi s oštećenjima motoričke prirode digitalizacija analognih sadržaja uz pomoć OCR programa izuzetno važna. Bez konvertiranja teksta na papiru u digitalni oblik, njihove obrade OCR programima i bez programa za čitanje ekrana koji tako obrađene sadržaje mogu pročitati svojim korisnicima (npr. program *JAWS*), gotovo bi im bilo nemoguće imati pristup mnogim sadržajima i informacijama, obrazovanju i sl. Primjerice, slijepa osoba ne može čitati knjigu, udžbenik ili bilo koji drugi tiskani sadržaj. Isto tako, osobe s invaliditetom čije je oštećenje takve prirode da ne mogu uzeti nikakav predmet u ruke, ne mogu čitati knjige, dokumente, pisma, izvještaje iz banke i sl. Iz navedenog je vidljivo koliko takvim osobama znače digitalizacija i OCR programi.

Kad je digitalizacija teksta u pitanju, stvari ipak nisu baš tako jednostavne kao što se to na prvi pogled čini. Moglo bi se pomisliti da je samo dovoljno skenirati neki papir, obraditi tu slikovnu datoteku OCR programom, ili ju odmah u startu skenirati u Microsoft Word dokument da bi se dobilio računalno obradive, pretražive i programima za čitanje ekrana čitljive datoteke. No tomu nije baš tako jer OCR programi nisu savršeni i optičko prepoznavanje znakova nije nikad 100%-tno učinkovito. Stančić u svojoj knjizi *Digitalizacija* iz 2009. godine navodi da kod digitalizacije današnjih tiskanih tekstova točnost prepoznavanja znakova može biti i do 99,95%. To se čini kao vrlo visok postotak ali je zapravo to donja granica isplativosti uporabe OCR programa. Svaki dokument koji nakon obrade OCR programom ima više od 4 do 5 grešaka na 1000 znakova nije isplativo tako obrađivati. Kako radi velikih troškova tako i radi

vremena trajanja obrade i ispravljanja grešaka. Kod nekih drugih vrsta tekstova grešaka može biti i dosta više.

OCR programi prepoznaju znakove na temelju kontrasta između podloge i znakova na toj podlozi. Do najviše problema dolazi kada su predlošci koje digitaliziramo nedovoljno kontrastni, kod tekstova s čestim tipografskim promjenama (primjerice rječnici i enciklopedije), tekstova sa znakovima svojstvenim za druge jezike (na primjer kod njemačkih prijeglasa: ä, ö i dr.) i tekstova sa zastarjelim oblicima pisma, a o rukopisima i loše i blijedo otisnutim tekstovima i preslikama da i ne govorimo. Greške su u takvim slučajevima česte i nekada je jeftinije i brže takav tekst ručno prepisati u digitalni oblik. Situaciju donekle popravlja podešavanje svjetline (engl. *brightness)* i rezolucije (engl. *resolution*) (Stančić, H. 2009).

Budući da OCR programi prepoznaju znakove zapravo kao neprekinuti niz povezanih točkica iste boje (najčešće se radi o crnoj boji na bijeloj podlozi odnosno papiru) velike probleme zadaju im loše i blijede kopije tekstova. Naime, događa se da kod blijedih i loših kopija tiskanih tekstova, kao i kod loših ispisa dokumenata i tekstova ima znakova koji imaju prekide u tom povezanom nizu crnih točkica. To je velik problem za OCR programe jer oni u svojoj memoriji jednostavno nemaju odgovarajući uzorak s kojim bi mogli takav „krnji“ znak usporediti. OCR programi rade uz pretpostavku da je jedan znak sačinjen iz jednog komada (mada ne mora biti uvijek tako, na primjer, znakovi *i, š ili %* sačinjeni su iz nekoliko komada) i u slučajevima loše skeniranih dokumenata i loših ispisa dolazi do pogrešaka u prepoznavanju znakova. Tako se često dogodi da OCR programi spoje dva susjedna znaka u jedan ili jedan znak podijele u više dijelova. Većina OCR programa vrši prepoznavanje slova, brojeva i sl. uz pomoć, kako pojašnjava Radošević, kombinacije više postupaka. Na primjer, jedan od tih postupaka je *prepoznavanjem na temelju svojstava oblika*. Kod *prepoznavanja na temelju svojstva oblika* bitno je da znak zadovolji određene kriterije.

Ti kriteriji su:

- *niska dimenzionalnost* (dakle, osnovne osobine znaka trebaju biti relativno malobrojne jer tada prepoznavanje traje kraće),
- *dovoljno informacije* (naime, osnovne osobine znakova trebaju sadržavati dovoljno informacija kako bi ih softver mogao uspješno raspoznati i razvrstati),
- *geometrijska postojanost* (geometrijska postojanost u prepoznavanju znakova polazi od pretpostavke da mala udaljenost među uzorcima u prostoru uzoraka znači i malu razliku u svojstvima objekata raspoznavanja mjerodavnim za raspoznavanje),

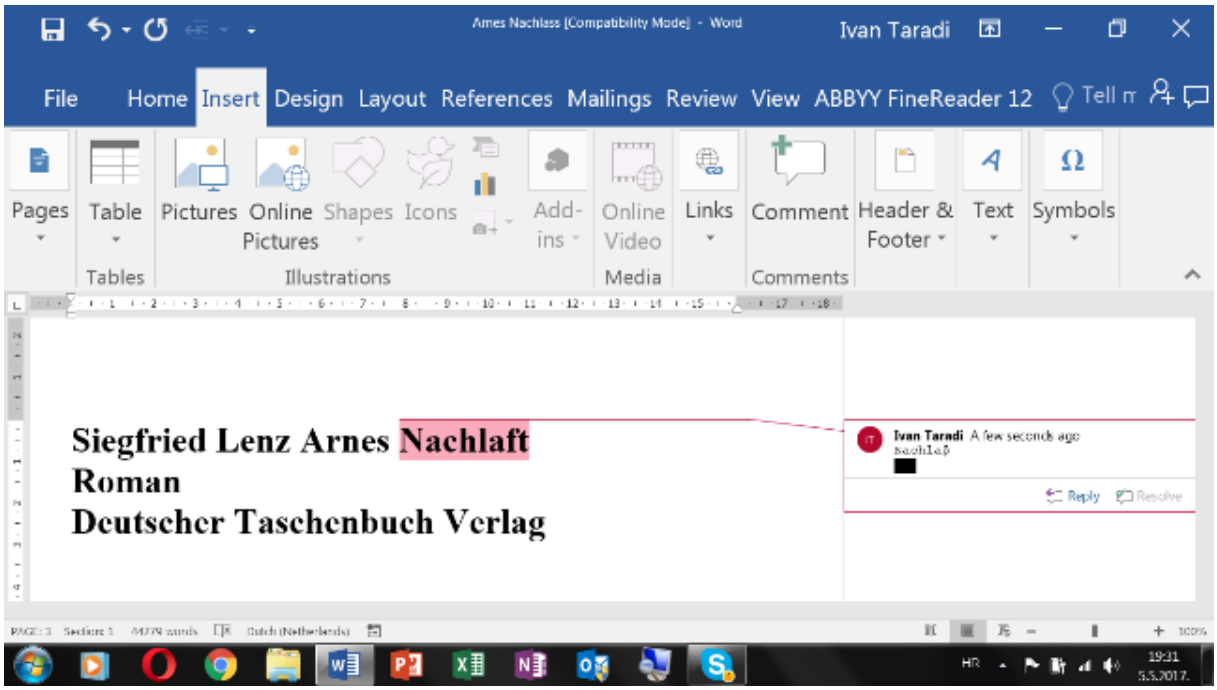
- *postojanost značajki* (postojanost značajki odnosi se uglavnom na uspoređivanje znakova odnosno uzoraka i njihovu vremensku postojanost ili nepostojanost; ali zapravo se rijetko koristi kod prepoznavanja teksta).

Iz svega navedenog vidljivo je zašto OCR programi moraju izdvojiti svaki pojedini oblik odnosno komad nekog znaka (ako ih je dva ili više kao kod, na primjer, znakova :, % i sl.) i u postupku sastavljanja znakova ih opet sastaviti. Kako bi proces optičkog prepoznavanja znakova bio uspješan sastavljene znakove treba uz pomoć odgovarajućih koordinata opet sastaviti i postaviti na pravu poziciju unutar reda teksta. Sastavljanje znakova nije uvijek jednostavno jer znakove treba pravilno rasporediti u redove teksta i s odgovarajućim razmakom među pojedinim riječima. Određene probleme ovdje stvaraju ukoso skenirani tekstovi, pogotovo tekstovi skenirani ručnim skenerima. Naime, u takvim slučajevima potrebno je ispraviti nagib teksta.

4. Greške OCR programa

Ono što me i ponukalo na istraživanje i pisanje o OCR programima su njihove greške kod prepoznavanja znakova svojstvenih stranim jezicima (konkretno kod znakova svojstvenih njemačkom jeziku), s kojima sam se susreo na početku studija. Naime, njemački jezik obiluje za hrvatski jezik netipičnim i neobičnim znakovima (slovima). Tu su, na primjer, znakovi za njemačke prijeglase (ä,ö i ü), a tu je i poznati i neobični znak za „šarfes S“ (β) koji se u pismu može zamijeniti i sa „ss“. Isto tako, znak *ö* može se zamijeniti s „*oe*“ i tako dalje. Na početku studija germanistike morao sam u sklopu kolegija *Suvremeni njemački jezik I* čitati i prevesti lektiru, dakle, prozno djelo od najmanje 100 stranica. U sklopu nekih drugih kolegija na germanistici trebalo je čitati isto tako dosta kraćih tekstova. Kod kraćih tekstova nije bilo toliko problema jer su se mogli čitati izravno na računalo ili ispisani uz pomoć specijalnih teleskopskih naočala. Ali, budući da sam slabovidna osoba, morao sam odabrani roman dati digitalizirati kako bih ga lakše i bez pomoći povećala i teleskopskih naočala mogao pročitati. Dakle, knjigu sam ostavio na skeniranje i rezultat je bio Microsoft Word dokument. Taj Microsoft Word dokument je bila moja lektira. Veličina fonta je bila je 24 i mogao sam ju dati ispisati kako bih ju mogao lakše i bez ikakvih pomagala čitati, bez naprezanja očiju dugotrajnim gledanjem u ekran računala. Iako je dokument (roman *Die verlorene Ehre der Katharina Blum* autora Heinricha Bölla) bio čitljiv i programom za čitanje ekrana kojeg sam imao (JAWS), to nije dolazilo u obzir. Razlog je bio taj što se svaku meni nepoznatu riječ u knjizi moralo potražiti u

rječniku i prevesti. Dakle, morao sam ju pogledati kako bih znao točno kako se ta riječ piše. Knjigu sam dao ispisati i dobio gotovo 200 listova papira. To je bilo radi velikog fonta i masno otisnutih slova. Tako je malo džepno izdanje romana od stotinjak stranica postala teška i velika hrpa listova. Počeo sam čitati roman, no vrlo brzo došlo je do problema. U tekstu su se nalazile neke vrlo neobične riječi, čudni znakovi i simboli, bilo na početku, u sredini ili na kraju pojedinih riječi. Nekada su ti čudni znakovi stajali i izolirano od ostatka teksta. Na početku sam čak i bezuspješno dugo u tiskanom rječniku ili na internetu pokušavao naći te čudne riječi, no bez uspjeha. Ubrzo sam shvatio da se radi o greškama i isprva sam mislio da je problem u ispisivanju. Ali kada sam ipak krenuo malo pogledati izvornu Microsoft Word datoteku koju sam primio elektroničkom poštom vidio sam da se iste greške i na istim mjestima nalaze i tamo. U slučajnom razgovoru sa starijim kolegom, studentom informacijskih i komunikacijskih znanosti saznao sam da su to greške OCR programa. Na slici donosim primjer greške OCR programa Abbyy FineReader. Na slici se vidi pogrešno otisnuta riječ *Nachlaft* iz moje druge



Slika 2. Primjer greške OCR programa

lektire, romana *Arnes Nachlaß,* autora Siegfrieda Lenza, a zapravo se radi o riječi *Nachlaß* (njem. *der Nachlaß*: hrv. o*stavština*). Pogrešno prepoznat njemački znak „šarfes S“ (β) OCR program prepoznao je kao *„ft“*. Kako sam bio tek na početku studija nisam raspolagao baš prevelikim vokabularom njemačkog jezika pa sam tu i još neke druge nepostojeće riječi bezuspješno u početku tražio u elektroničkim i tiskanim rječnicima. I to je bilo vrlo frustrirajuće i trajalo je sve dok nije došlo do već spomenutog slučajnog razgovora s kolegom studentom. Sreća u svemu tome je bila ta što se velika većina krivih znakova odnosno nepostojećih riječi

ponavljala. Otprilike do dvadesete stranice uspio sam dešifrirati gotovo sve krive znakove. Na primjer, slova *ft* su uvijek bila na mjestu znaka *β*. No bilo je i iznimaka i ponekad je bilo zaista teško dešifrirati pravu riječ te sam morao imati uz sebe i originalan roman u tiskanoj verziji kako bih provjerio poneku riječ i bio siguran o čemu se radi. Naime, to je bilo nužno jer je i smisao prijevoda lektire taj da se to napravi bez pogreške.

Inače, za vrijeme cijelog dosadašnjeg studija veliku pomoć oko digitalizacije literature dobio sam od udruge *Zamisli*. Udruge koja pruža pomoć i podršku mladim osobama s invaliditetom, naročito onima koji se obrazuju.

5. Unapređenja programa za optičko prepoznavanje znakova *Megaznak*

Program za optičko prepoznavanje znakova *Megaznak* napisan je u programskom jeziku *Turbo Pascal* koji radi *u* MS DOS-u. Autor programa je Danijel Radošević. OCR program *Megaznak* se bori s greškama u prepoznavanju znakova na nekoliko načina. Radošević pojašnjava da OCR program *Megaznak* ima proces optičkog prepoznavanja znakova podijeljen u šest glavnih faza. Treća faza ima i nekoliko podfaza. U sklopu tih faza u radu vidjet ćemo kako je poboljšana njegova učinkovitost u otkrivanju i ispravljanju grešaka pri prepoznavanju teksta.

1. Faza – Pretvaranje bitmape skeniranog teksta iz PCX formata (*PiCture eXchange – jedan od prvih široko prihvaćeni formata za slikovne datoteke u DOS-u)* u interni format. PCX format koristi se radi dobre mogućnosti kompresije datoteke i mogućnosti brze obrade.
2. Faza – Prikaz slike bitmape teksta na zaslonu računala. U ovoj fazi program pronalazi i izdvaja sve nakupine točaka. Nakupine točaka se zatim obrađuju i uklanjaju s ekrana ali i iz datoteke koja sadrži bitmapu teksta.
3. Faza – Obrada nakupina točaka (oblika). Ova faza sastoji se od nekoliko podfaza. Podfaze su opisane niže u tekstu.

3.1. Podfaza *izdvajanja nakupina točaka*. Izdvajanje znakova (oblika) iz okoline pomoću potprograma. Potprogrami prepoznaju znakove (oblike) kao međusobno povezanu nakupinu (neprekinuti niz) točaka iste boje te ih tako mogu razlikovati od podloge (najčešće bijelog papira).

3.2. Podfaza *uklanjanja smetnji*. Ova zanimljiva podfaza zadužena je za poboljšanje rada ovog OCR programa. OCR program *Megaznak* sve povezane nakupine točaka koje su manje od minimalnog broja točaka podešenog u programu isključuje kao smetnju iz daljnje obrade.

3.3. Podfaza *normiranje*. U ovoj podfazi *Megaznak* izvršava usporedbu izdvojenih znakova (oblika) s predlošcima koji su pohranjeni u memoriji. Izdvojeni oblici moraju biti usporedivi s predlošcima u sustavu i moraju stati u zadani standardni okvir (veličina je 16 * 16 točaka). Oblici koji ne odgovaraju toj veličini transformiraju se preračunavanjem koordinata točaka dok se ostali jednostavno kopiraju u standardni okvir.

3.4. Podfaza *izdvajanja osnovnih značajki oblika*. U ovoj podfazi izdvajaju se dvije osnovne značajke:

- *broj okvira znaka* i
- *veličina vanjskog znaka.*

Uz svaki znak koji služi kao predložak za prepoznavanje znakova, a koji je pohranjen u memoriji sustava, nalaze se podaci o ove dvije osnovne značajke. Izdvajanjem ovih osnovnih značajki postiže se to da se više ne mora utvrđivati stupanj sličnosti izdvojenih znakova i svih predložaka, već samo onih koji sadrže te iste dvije osnovne značajke. Ovdje je bitno reći da *broj okvira oblika* mora biti identičan, na primjer slovo A ima dva okvira. Jasno je da se radi o jednom vanjskom i jednom unutarnjem okviru. Značajka *veličina vanjskog okvira* koja je izražena brojem rubnih točaka koje čine okvir ne mora biti identična predlošku i može odstupati i do 20%. Ovime je značajno unaprijeđen rad ovog OCR programa, dakle, povećana je brzina i točnost prepoznavanja znakova odnosno teksta.

3.5. Podfaza *uspoređivanja izdvojenog oblika s predlošcima za prepoznavanje*. U ovoj podfazi prepoznavanja izdvojeni oblici uspoređuju se sa svim pohranjenim predlošcima s kojima dijele osnovne značajke oblika. Najprije se uspoređuje veličina, dakle, visina i širina izdvojenog oblika s veličinom predloška za prepoznavanje (u oba slučaja uspoređuje se veličina prije normiranja). Ako pri usporedbi odstupanje veličine izdvojenog znaka i predloška nije veće od 15% prelazi se u sljedeći korak. A taj korak je utvrđivanje stupnja podudarnosti izdvojenog znaka i predloška. Kôd predloška s najvećim stupnjem podudarnosti dodaje se izdvojenom obliku i on je sada prepoznat jer je najsličniji predlošku. Kôd se upisuje u tablicu prepoznavanja zajedno s koordinatama prepoznatog znaka u bitmapi teksta kako bi se znao točan položaj prepoznatog znaka u tekstu.

Ipak, u ovome dijelu procesa ima pogrešaka i neki slični znakovi, na primjer *„e“ i „o“* ili *„t“ i „f“* znaju biti pogrešno prepoznati i zamijenjeni jedni za druge. Program *Megaznak* za rješavanje ovog problema donosi jedno unapređenje optičkog prepoznavanja znakova. Naime, program *Megaznak* u slučaju problema kod razlikovanja znakova *„e“ i „o“* pokreće postupak za razlikovanje i usredotočuje se na dijelove tih znakova koji se najviše razlikuju. U ovome slučaju to je središnji dio znakova. Rezultat ovog postupka je, ako je došlo do greške u prethodnim koracima prepoznavanja, korekcija prepoznavanja.

4. Faza – *Sastavljanje znakova iz izdvojenih oblika*. U ovoj fazi potrebno je spojiti znakove koji se sastoje od više od jednog dijela. Kod znakova koji se sastoje od samo jednog dijela kao što je znak „A“ to nije potrebno provesti. Međutim, postoje znakovi koji se sastoje iz dva ili više dijelova (na primjer „i“, „č“ i „%“). Dakle, potrebno je da OCR program izdvoji i prepozna svaki dio znaka.

Prema tome od koliko su dijelova odnosno komada sastavljeni znakovi postoji nekoliko tipova znakova:

- tip 1 - znak se sastoji od 1 dijela,
- tip 2 - znak se sastoji od 2 dijela,
- tip 3 - znak se sastoji od 3 dijela, i
- tip 4 - specijalan znak, na primjer interpunkcije poput *točke* ili *zaraza*, odnosno znak koji može stajati samostalno u tekstu ili biti dio nekog drugog znaka. Interpunkcije *točka* i *zarez* mogu biti samostalni, a mogu biti i dio jednog istog znaka, primjerice znaka *točka sa zarezom* (;).

Kada se izvrši prepoznavanje i odredi tip znaka, s obzirom na broj dijelova, svi se dijelovi unose u tabelu prepoznatih znakova s pridruženim kodom predloška i koordinatama izdvojenih znakova u bitmapi teksta. To je potrebno kako bi se znakovi mogli ispravno sastaviti i smjestiti na odgovarajuće mjesto u tekstu. Radošević navodi da pri sastavljanju teksta može doći do poteškoća. Naime, ponekad se desi pogreška u sastavljanju redova teksta, a moguć je i krivi nagib teksta. Program *Megaznak* u ovakvim slučajevima nudi unapređenje odnosno korekciju nagiba. Dakle, potrebno je samo da korisnik unese faktor korekcije. Taj faktor može biti pozitivan ili negativan, a to ovisi o smjeru nagiba teksta.

Kada se uzme u obzir sve dosad navedeno o OCR programu *Megaznak*, može se reći da on donosi neka vrlo interesantna poboljšanja u optičkom prepoznavanju znakova. Njegove glavne prednosti su to što u procesu prepoznavanja znakova raspolaže s nekoliko različitih postupaka prepoznavanja i kombinira ih. U Radoševićevu istraživanju stoji da to dosta poboljšava samu točnost prepoznavanja. Isto tako *Megaznak* može učiti i mogu mu se zadavati pravila

sastavljanja znakova, brz je i omogućava naknadnu korekciju teksta. Uz ove objektivne prednosti *Megaznak* ima naravno i nedostataka.

Prema autoru programa Radoševiću, najveći nedostatak je neprepoznavanje slijepljenih susjednih znakova i neprepoznavanje znakova koji su sastavljeni od dva ili više dijelova (npr. *i, č* i sl.). Ovi slučajevi događaju se u fazi prepoznavanja u kojoj OCR program *Megaznak* treba izdvojiti pojedine oblike iz njihove okoline. To je prema autoru programa Megaznak i najosjetljivija faza postupka prepoznavanja. Daljnji nedostaci su i to što *Megaznak* ne može raditi direktno sa skenerom i što ne razlikuje fontove i stilove. Također ne može ni analizirati stranice u potpunosti. Moglo bi se još dodati da je nedostatak i to što je kod OCR programa *Megaznak*, koji je i napisan radi testiranja interoperabilnosti raznih postupaka prepoznavanja znakova, potrebno da korekcije grešaka izvrši sam korisnik programa.

Ovim nedostacima će se trebati i nadalje posvetiti velika pažnja i raditi na njihovu unapređenju. To je potrebno jer OCR programi igraju veliku ulogu u današnjoj sve većoj potrebi za unošenjem velikih količina podataka, posebice tekstova, u računala.

5. Unapređenje OCR programa uz pomoć lingvistike

Problemima optičkog prepoznavanja znakova može se baviti iz različitih kutova odnosno iz gledišta različitih znanosti. Kada je riječ o optičkom prepoznavanju znakova pri digitalizaciji tekstova jasno je da bi lingvistika mogla biti od pomoći. U mrežnom izdanju Enciklopedije Leksikografskog zavoda Miroslava Krleže Hrvatskoj enciklopediji[1] stoji kako je lingvistika odnosno jezikoslovlje znanstveno proučavanje ljudskog jezika. Dakle, proučavanje ljudske komunikacije unutar neke jezične zajednice. U jezikoslovlju se znanstveno promatraju, popisuju, opisuju, klasificiraju i objašnjavaju jezične činjenice. Kako to pomaže u unapređenju OCR programa možemo vidjeti na primjeru pokusa kojeg su proveli japanski znanstvenici Koichi Takeuchi i Yuji Matsumoto. Ovi su znanstvenici pokušali unaprijediti rad OCR programa koristeći se morfološkim analizama i modelom vjerojatnosti *n-grama*. Njemačka Wikipedija definira *n-gram* kao rezultat razlaganja nekog teksta na fragmente. Naime, fragmenti teksta mogu biti slova, slogovi, riječi itd. Ti se fragmenti sastavljaju ponovno u *n-grame*. Japanski znanstvenici su se u svome istraživanju mogućnosti poboljšanja OCR programa koristili vrstom *n-grama* poznatom pod nazivom *trigram*. *Trigrami* su dakle *n-grami*

[1] Hrvatska enciklopedija: http://www.enciklopedija.hr

koji se sastoje od po tri jezična elementa koji slijede jedan za drugim. Jednostavnije rečeno, to mogu biti tri slova, sloga ili riječi koje se nalaze neposredno jedna za drugom u tekstu. Metoda *n-grama*, kako navodi njemačka Wikipedija, koristi se, među ostalim i u *korpusnoj lingvistici*. *Korpusna lingvistika* se bavi ukupnim fondom (korpusom) riječi nekog jezika koje su u uporabi u govoru, ili velikim korpusima tekstova nekog jezika odnosno tekstova nekog znanstvenog područja određenog jezika. Japanski znanstvenici Matsumoto i Takeuchi su pokušali unaprijediti otkrivanje i ispravljanje grešaka u radu OCR programa uz pomoć jezičnih analiza tekstova, a sve na temelju velike količine suvremenih novinskih članaka.

6.1. Lingvističke metode i posebnosti japanskog pisma

Japanski jezik odnosno japansko pismo ima neke specifičnosti u usporedbi s većinom pisama, pa tako i u usporedbi s engleskim pismom. Postoje pokušaji unapređenja rada OCR programa u otkrivanju i ispravljanju grešaka engleskih tekstova uz pomoć računalnih algoritama. Računalni algoritmi, naime, otkrivaju pogrešnu riječ i pronalaze najpodobniju riječ uz pomoć analize riječi koje se nalaze uokolo *sumnjive* odnosno pogrešne riječi. Dakle, na temelju analize korpusa tekstova traži se riječ koja bi najbolje odgovarala danom kontekstu i okružju u kojem se nalazi. Ovakvu primjenu statističke analize korpusa tekstova omogućuje nam *korpusna lingvistika*. *Korpusna lingvistika* ustvari omogućuje primjenu statističkih metoda prikupljanjem velike količine tekstova i osiguravanjem dovoljne količine uzoraka. Međutim, ovakva analiza okružja i konteksta uokolo neke riječi nije moguća kada govorimo o japanskom pismu. Naime, u japanskom pismu nema, za većinu pisama, uobičajenog bijelog razmaka (engl. *space*) među riječima. Riječi u Japanskom se ne odvajaju razmakom i pišu se sve u jednome nizu bez razmaka. Tako da su japanski znanstvenici u svom istraživanju mogućnosti poboljšanja OCR programa morali napraviti odmak od otkrivanja i ispravljanja pogrešaka na temelju grafičkog oblika znakova. Stoga su primijenili metodu statističkih *n-grama* znakova. Fokus im je bio na greškama zamijene. Dakle, greškama kod kojih se pravi znak zamijeni krivim znakom na istome mjestu u nizu znakova (npr. u nizu riječi). Za otkrivanje i ispravljanje grešaka koristili su se statističkim *trigramom* znakova. To znači da su se pogreške i ispravci u tekstu iz određenog stručnog područja temeljili na pronalaženju najčešćih kombinacija tri znaka koji slijede jedan za drugim, a mogu se naći u velikom broju tekstova vezanih s nekom strukom. Po toj metodi pogrešan znak u nizu od tri znaka je onaj koji se ne može naći u takvoj kombinaciji u tekstovima određenog stručnog područja. Isto tako, znak *kandidat*, dakle, znak koji kao

ispravan predlaže program, je onaj koji se statistički često ili uvijek može naći u kombinaciji sa znakovima oko pogrešnog znaka.

Matsumoto i Takeuchi pojašnjavaju kako se sam postupak sastoji od tri dijela:

- otkrivanje pogrešnog znaka i mjesta u japanskom tekstu,
- generiranje znaka *kandidata i* zadržavanje riječi *kandidata* uz pomoć konzultiranja rječnika, a sve na temelju generiranog znaka *kandidata*, i
- odabiranje najpogodnijeg niza riječi na temelju generirane riječi *kandidata.*

Ovo zapravo znači da se ispravan znak odabire na temelju statističke vjerojatnosti *trigrama znakova* odnosno učestalosti pojavljivanja takvog *trigrama* u velikoj količini japanskih novinskih tekstova. Odabir ispravne riječi vrši se pak na temelju Japanskog rječnika morfološke analize (analize osnovnih oblika riječi) koji sadržava otprilike 170.000 riječi. Sukladno tome sustav označava neku riječ kao pogrešnu ako ona ne postoji u Japanskom rječniku morfološke analize. U skladu s tim, sustav odabire riječ odnosno niz riječi koje imaju najveći stupanj pojavnosti u japanskim novinskim tekstovima nekog stručnog područja, znači, u korpusu neke struke. U ovoj studiji japanski znanstvenici koristili su uz model *trigrama znakova* i modele *POS-trigram* i *trigram riječi*. Trigram riječi je, precizno, kombinacija od tri riječi u nizu koje dolaze uvijek skupa. *Pos-trigram* (engl. part of speech, POS) počiva na vrstama riječi i njihovom redanju u rečenici. Japanski jezik u tom smisli ima vrlo čvrstu strukturu rečenice. U japanskoj rečenici na prvom mjestu je subjekt pa dolazi objekt a zatim predikat (u većini indoeuropskih jezika redoslijed je subjekt-predikat-objekt). Po modelu *POS-trigrama* sustav prepoznaje krivi redoslijed vrsta riječi odnosno prepoznaje pogrešne nizove riječi, dakle prepoznaje konstrukcije koje sintaktički nisu ispravne i vrši korekcije.

Japanski znanstvenici zaključuju da je najbolje rezultate za vrijeme njihova istraživanja mogućnosti poboljšanja OCR programa davala kombinacija *POS-trigrama* i *trigrama riječi*. Uz kombinaciju tih dvaju modela uspjeli su tekstove s točnošću od 90% dići na 94,3%, a one s točnošću od 95% digli su na 96,9% točnosti. Ističu i da je na rezultate utjecao i ograničen korpus mrežno dostupnih tekstova iz područja biologije. Nadalje, Matsumoto i Takeuchi dodaju kako ovaj njihov sustav otkrivanja i ispravljanja pogrešaka može naći primjenu i u ispravljanju pravopisnih pogrešaka u tekstovima na japanskom jeziku.

7. Uloga pripreme u poboljšanju rada OCR programa

U gore navedenim primjerima pokusnog OCR programa *Megaznak* autora Radoševića i istraživanju koje su proveli japanski znanstvenici Matsumoto i Takeuchi vidjeli smo ustvari pokušaje poboljšanja OCR programa u smislu otkrivanja i ispravljanja grešaka. Idući primjer poboljšanja rada OCR programa odnosi se na samu pripremu postupka optičkog prepoznavanja znakova i sprječavanja grešaka. Dobrom pripremom pokušalo se dokazati da se optičko prepoznavanje znakova i učinkovitost softvera može značajno unaprijediti. Već je spomenuto u ovom radu da se na učinkovitost OCR softvera može utjecati podešavanjem svjetline (engl. *brightness*) i rezolucije (engl. *resolution*).

Slika 3. OCR sustav

Izvor: PCCHIP. Dostupno na: http://pcchip.hr/softver/novi-abbyy-finereader-12-uvelike-ubrzava-raspoznavanje-i-citiranje-teksta-iz-skenova-i-fotografija/ (7.7.2017.)

U istraživanju iz 2013. godine koje su proveli znanstvenici Kreković, Kukučka, Šprem i Zadro pokazalo se da priprema procesa obrade digitalne slike OCR programima ima veliku ulogu u njihovoj učinkovitosti. Naime, neprestan razvoj i napredak ICT tehnologije omogućuje da se skeniranjem dobiju sve kvalitetnije digitalne slike tekstova koje je potrebno obraditi softverima za optičko prepoznavanje znakova. U istraživanju se pokazalo da je rezultat obrade digitalne slike OCR programom puno bolji ako se obrađuje slika visoke rezolucije. Stoga je potrebno skeniranjem dobiti slikovnu datoteku rezolucije od minimalno 200 dpi-a (engl. dots per inch, dpi: točaka po inču). Takvu je sliku puno lakše i kvalitetnije obraditi OCR programima nego sliku kvalitete rezolucije od 100 dpi-a. Rezultat obrade slikovne datoteke OCR programom je još uvijek bolji čak i ako se slikovna datoteka slabije rezolucije naknadno poboljšavala.

Još jednu značajnu ulogu u poboljšanju rada OCR softvera kod sprječavanja pogrešaka igraju i *morfološke operacije*. Ovdje se to odnosi na *digitalnu morfologiju*. U radu Kreković, Kukučka, Šprem i Zadro navode da je *digitalna morfologija* ustvari opis i analiza digitalnih oblika. Iz ovoga slijedi da je, na primjer, digitalni objekt A skup točkica slike koje imaju neku zajedničku osobinu odnosno koje imaju vrijednost 1. Pozadina se po tome definira kao skup

točaka slike koje ne pripadaju objektu A. Te točke imaju vrijednost 0 (nula). Programskim jezikom *Matlab* u ovom istraživanju uspješno se otklanjaju greške u postupku prepoznavanja znakova pomoću *morfoloških operacija*. Ove operacije omogućuju da se svaka izolirana točka s vrijednošću 1 ukloni ako je okružena samim *nulama*. Program takvu izoliranu *jedinicu* koja ne pripada nekom skupu *jedinica* (npr. slovu A) prepoznaje kao grešku i uklanja ju *morfološkom operacijom clean*. Isto tako programski jezik *Matlab* uklanja izolirano slovo *morfološkom operacijom spur* ako ga je prepoznao kao izolirano među *nulama* odnosno ako ga ne smatra dijelom neke veće morfološke forme (riječi). Autori ove postupke u svome istraživanju nazivaju svojevrsnim *čišćenjem šumova*.

U sljedećem koraku postupka optičkog prepoznavanja znakova u programskom alatu *Matlab* autori istraživanja opisuju postupak prepoznavanja slova i izbjegavanje problema u tom procesu. Pomoću programskog alata *Matlaba* prepoznaje se slovo pomoću usporedbe sa znakovima odnosno predlošcima u bazi znakova koja je pohranjena u sustavu. Ta baza znakova predstavljena je skupom znakova S. Skup znakova S predstavljen je na sljedeći način: S = {A, B, C, Č, Ć, D, Đ, E, F, G, H, I, J, K, L, M, N, O, P, R, S, Š, T, U, V, Z, Ž, a, b, c, č, ć, d, đ, e, f, g, h, i, j, k, l, m, n, o, p, r, s, š, t, u, v, z, ž, 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, !, ?, “, ., ,, -, +, *, /, =, _, :, ;, , (,), {, }, [,], #, @, &, %, $, ~}. Softver slovo iz slikovne datoteke uspoređuje sa znakovima iz svih skupova S, a skupova S ima ustvari četiri i svaki sadrži isti 91 znak ali različitih fontova i veličina, te pronalazi najsličniji znak po fontu i veličini. Prepoznato slovo zatim je potrebno skalirati na veličinu znaka iz odgovarajućeg skupa S, dakle, na veličinu njemu najsličnijeg znaka iz skupa S. To je potrebno jer se rijetko dogodi da su veličine ulaznih znakova i predložaka identične.

Ovim postupkom izbjegavaju se problemi i moguće greške u optičkom prepoznavanju znakova. Ovime se ustvari izbjegavaju problemi koji mogu nastati kod velikih promjena u dimenzijama znakova. Znači, izbjegavaju se gubici piksela, ili pak pojavljivanje krivih piksela, ističu autori istraživanja te dodaju da ovaj dio procesa optičkog prepoznavanja znakova daje dobre rezultate ali ima i problema. U istraživanju se ističe da problemi nastaju kada ulazni znak i predložak nisu istog fonta i kada se ne nalaze na istome mjestu unutar okvira. Drugi problem je brzina postupka. Naime, postupak može biti vremenski zahtijevan.

Kako bi doskočili ovim problemima i poboljšali učinkovitost OCR programa autori istraživanja odredili su velik broj značajki znakova koje je potrebno izdvojiti da bi usporedba prošla što učinkovitije. Dakle, bilo je potrebno odrediti što veći broj značajki znakova koje zapravo mogu biti dosta različite kod različitih znakova. Potrebno je da te značajke budu takoreći otporne na različite fontove, dimenzije, ukošenost slova te njihovu zadebljanost ili

stanjenost. Kao najvažniju metodu izdvajanja značajki znakova u svome istraživanju autori Kreković, Kukučka, Šprem i Zadro ističu metodu *centralnih momenata*. Međutim, u radu se navodi da su istraživanje i praktična primjena ove metode dali zapravo najlošije rezultate od svih metoda uspoređivanja znakova s predlošcima. Naime, učinkovitost ove metode bila je niža od 50%. Iz navedenog proizlazi kolika je zapravo važnost ovakvih istraživanja i praktične primjene metoda u optičkom prepoznavanju znakova. Očito da teorija i praksa nisu uvijek jedno te isto.

8. Ubrzanje rada OCR programa

Već je ranije u radu navedeno da brzina rada OCR programa igra važnu ulogu u postupku digitalizacije i optičkog prepoznavanja znakova. U suvremenom digitalnom okružju i poslovnom svijetu brzina obrade podataka vrlo je važna. Da bi se OCR program uopće smatrao isplativim i korisnim trebao bi proces optičkog prepoznavanja znakova u slikovnoj datoteci koja predstavlja neki tekst obaviti brzinom kojom prosječna osoba tu istu količinu teksta može pročitati. Budući da se danas jako puno tekstova digitalizira i na taj način brzo i jednostavno unosi u računalne sustave postoje i razni pokušaji da se taj proces maksimalno ubrza. Potreba za digitalizacijom velike količine analognog gradiva pojavljuje se u novije vrijeme posebno u digitalnim knjižnicama ali i kod manjih organizacija i pojedinaca. Tim problemom bave se znanstvenici Miran Karić, Zdravko Krpić i Goran Martinović u radu *Optičko prepoznavanje znakova na grid i višejezgrenim platformama* iz 2013. godine.

Za potrebe ovog istraživanja znanstvenici Karić, Krpić i Martinović razvili su aplikaciju za optičko prepoznavanje znakova i testirali ju na CRO-NGI gridu (engl. Croatian national grid infrastructure, CRO-NGI) odnosno na Hrvatskoj nacionalnoj grid infrastrukturi. Cilj istraživanja bio je utvrditi performanse OCR programa u okruženju umreženih računala. Na mrežnim stranicama Sveučilišnog računskog centra (krat. SRCE) stoji da je CRO-NGI raspodijeljena računalna okolina, sastavljena od procesorskih i podatkovnih resursa, smještenih u čvorištima unutar Republike Hrvatske. Dakle, to je resurs koji stoji na raspolaganju znanstvenoj i akademskoj zajednici Republike Hrvatske upravo za ovakva istraživanja i za njeno povezivanje sa znanstvenim i akademskim krugovima širom svijeta.

Na ovome resursu je ispitan rad paralelnih OCR sustava kako bi se utvrdilo koliko računalna moć može ubrzati OCR programe i njihovu učinkovitost. U samome pokusu ispitana su i dva modela komunikacije među računalima umreženima radi rješavanja zajedničkog

zadatka odnosno radi zajedničkog rada na optičkom prepoznavanju znakova. Jedan model je model *koordinator*, a drugi je model *koordinator/radnik*. Nazivi modela odnose se na glavno računalo u gridu. U prvom je modelu glavno računalo samo koordinator i raspoređuje posao na ostale jedinice sustava (radnike). U drugom modelu glavno računalo je i koordinator i radnik. Dakle, ono raspoređuje posao ostalim radnim jedinicama grida ali i radi kada nema komunikacije s radnicima u sustavu. U modelu *koordinator* glavno računalo samo koordinira posao i ne sudjeluje u njemu. U oba tipa komunikacije posao optičkog prepoznavanja slikovnih datoteka može se ravnomjerno rasporediti na sve radnike, ali može se i podijeliti po različitim fazama posla na različite radnike. Također je moguće i odrediti koja količina posla će se slati kojoj radnoj jedinici u gridu.

Što se ovih dvaju tipova komunikacije među glavnim računalom i radnicima tiče, istraživanje je dalo sljedeće rezultate. Naime pokazalo se da model *koordinator/radnik* funkcionira bolje tj. da je brži. No, ta razlika u njegovu korist smanjuje se kako raste broj računala u gridu. Već kada je umreženo šest radnih jedinica ta prednost je vrlo mala. Kada je umreženo 10 ili 20 radnih jedinica u prednosti je model *koordinator*. To se objašnjava time da u mrežama s više računala glavno računalo u modelu koordinator/radnik i nema baš puno vremena za rad. To je i logično jer u takvom modelu glavno računalo mora mnogo komunicirati s drugim radnim jedinicama, a čak je i moguće da zaustavlja rad poneke radne jedinice u gridu jer ne stigne na vrijeme proslijediti idući zadatak.

Istraživanje je također pokazalo da model *koordinator/radnik* daje bolje rezultate kada radi sa srednje velikim paketima podataka za obradu. Znači, radi brže s paketima slika srednje veličine, dok s paketima slika manje i velike veličine radi sporije. Kod modela *koordinator* rezultati su bolji s manjim paketima slika. Ovaj model radi sporije što je paket slika za obradu veći.

Kada govorimo o točnosti prepoznavanja, pokazalo se da je ona bolja kada se koristi funkcija konkavnosti odnosno udubljenosti znakova. Isto tako pokazalo se da ta funkcija dosta usporava proces optičkog prepoznavanja znakova. Stoga se u radu ističe da se ona može primjenjivati kada brzina postupka optičkog prepoznavanja znakova ne igra presudnu ulogu u procesu.

9. OCR program otvorenog koda (Open source OCR)

OCR program otvorenog koda *OCRopus* zamišljen je kao platforma za korištenje i unapređenje OCR programa. *OCRopus* je, kako i sam pojam *otvoreni kȏd* u smislu softvera

govori, OCR program kojemu je moguće slobodno pristupiti, slobodno ga koristiti i unapređivati ga. Ovaj OCR softver napravio je Tom Breuel s Njemačkog istraživačkog centra za umjetnu inteligenciju (njem. Deutsches Forschungszentrum für Künstliche Intelligenz, DFKI) iz Kaiserslauterna. Projekt razvoja open source OCR programa počeo je 2007. godine podupirati i *google* s ciljem podrške slobodnom OCR softveru i digitalizaciji. Namjena ovog projekta je osigurati slobodno korištenje i razvoj OCR sustava za najjednostavnije postupke digitalizacije kao i za složene analize povijesnih tekstova i osiguranje pomoći pri čitanju tekstova osobama oštećenog vida. Iako je u početku zamišljen kao softver ograničenog vijeka korištenja i bio je namijenjen radu i istraživanju troje studenata, ipak je postao javno dostupan i sada je OCR program otvorenog koda. Glavna prednost ovog OCR programa je to što omogućuje unapređenje optičkog prepoznavanja znakova koji su svojstveni raznim jezicima. To je moguće zbog toga što ovaj softver ima ugrađene alate za modeliranje jezika koji se mogu dalje razvijati. Ove osobine *OCRopus-a* daju mogućnost njegovim korisnicima da sami kreiraju nove jezike i znakove u ovom OCR programu pomoću unošenja podataka u sustav. Danas je OCRopus softver gotovo jednako kvalitetan kao i komercijalni programi za optičko prepoznavanje znakova.

10. Zaključak

Imajući u vidu sve navedeno u ovome završnom radu može se zaključiti da programi za optičko prepoznavanje znakova čine važan dio sveopće digitalizacije u suvremenom digitalnom i informacijskom dobu u kojem živimo. Vidjeli smo da je digitalizacija analognog gradiva važna potreba suvremenog čovjeka i današnjih informacijskih institucija. Kada se govori o učinkovitosti OCR programa, pokazalo se da je u fokusu rad na prevenciji, otkrivanju i ispravljanju grešaka koje nastaju tijekom procesa optičkog prepoznavanja znakova. Dakle, istraživanja se provode s ciljem poboljšanja rada OCR sustava prema te tri navedene funkcije. Tu je naravno još i važna brzina procesa optičkog prepoznavanja teksta u slikovnim datotekama dobivenim skeniranjem. Iako većini prosječnih korisnika, u kakve i sam spadam, brzina rada OCR sustava nije presudan čimbenik, lako je zamisliti koju ona ulogu igra u današnjim informacijskim institucijama.

Suvremene informacijske institucije poput arhiva, knjižnica i muzeja možda su i najveći korisnici OCR sustava. Količine gradiva te muzejskih predmeta koje je potrebno digitalizirati i učiniti dostupnim svekolikoj javnosti su doista goleme. Dobrobit digitalizacije analognih

tekstova vrlo je velika. Ona znatno olakšava korištenje i upravljanje gradivom, olakšava i pospješuje poslovanje svim institucijama i organizacijama, a vrlo važnu ulogu ima i u očuvanju i zaštiti važnih i rijetkih izvornih dokumenata.

Daljnji razvoj i unapređenje OCR sustava, s obzirom na sve navedeno, nameće se sam po sebi kao logičan prioritet svih korisnika ove napredne tehnologije.

Literatura

1. Britannica. Dostupno na: https://www.britannica.com/topic/postal-system/Postal-services-in-the-developing-countries#ref367160 (18.9.2017.)
2. Eržišnik, D. INFORMATIZACIJA I UREDSKO POSLOVANJE - POVIJESNI PREGLED I PERSPEKTIVE. //Arhivski vjesnik (2000). URL: file:///C:/Users/korisnik/Downloads/AV_2000_43_07%20(1).pdf (17.9.2017.)
3. Haddej, Ben D.; O'Brien, S. Optical Character Recognition. 26.04.2012. *A Major Qualifying Project Report submitted to the faculty of the WORCESTER POLYTECHNIC INSTITUTE in partial fulfillment of the requirements for the Degree of Bachelor of Science.* 26.04.2012. URL: https://web.wpi.edu/Pubs/E-project/Available/E-project-042412-142927/unrestricted/sob_dbh_MQP_report.pdf (17.9.2017.)
4. Hrvatska enciklopedija, mrežno izdanje. Leksikografski zavod Miroslav Krleža. Dostupno na:http: http://www.enciklopedija.hr/ (18.9.2017.)
5. Karić, M.; Krpić, Z.; Martinović, G. Optičko prepoznavanje znakova na grid i višejezgrenim platformama – analiza performansi. //Tehnički vjesnik (2013). URL: file:///C:/Users/korisnik/Downloads/tv_20_2013_4_647_653%20(4).pdf (17.9.2017.)
6. Kreković, M.; Kukučka, J.; Šprem, J.; Zadro, I. *Ekstrakcija i prepoznavanje slova na digitalnim slikama - SLUČAJNI PROCESI U SUSTAVIMA*. sječanj 2013. URL: http://www.ieee.hr/_download/repository/tim_06_Ekstrakcija_i_prepoznavanje_slova_na_digitalnim_slikama.pdf (18.9.2017.)
7. Matsumoto, Y.; Takeuchi K. Japanese OCR Error Correction Using Stochastic Morphological Analyzer and Probabilistic Word N-gram Model. // International Journal of Computer Processing of Oriental Languages, (2000), str. 69–82. URL: http://web.b.ebscohost.com/ehost/pdfviewer/pdfviewer?vid=2&sid=447162bd-4a70-4e34-9c23-3fb30a947287%40sessionmgr103 (18.9.2017.)
8. Project Gutenberg. Dostupno na: https://www.gutenberg.org/ (18.9.2017.)
9. Radošević, D. Postupci i problemi optičkog prepoznavanja znakova. //Journal of Information and Organisational Sciences (1996). URL: http://hrcak.srce.hr/index.php?show=clanak&id_clanak_jezik=117350 (18.9.2017.)
10. SRCE. Dostupno na:http://www.srce.unizg.hr/cro-ngi (18.09.2017.)

11. Stančić, H. Digitalizacija ; obrada i kontrola kvalitete. Zagreb: Zavod za informacijske studije, 2009.
12. Wikipedia : Die freie Enzyklopädie. Dostupno na: https://de.wikipedia.org/wiki/Wikipedia:Hauptseite (18.9.2017.)
13. Wikipedia : The Free Encyclopedia. Dostupno na: https://en.wikipedia.org/wiki/Main_Page (18.9.2017.)

Sažetak

Digitalizacija tiskanih tekstova je postupak njihove pretvorbe u digitalni oblik. Dakle, njihovo pretvaranje u binaran kôd koji je zapisan kao računalna datoteka. Danas dolazi do izražaja rastuća potreba da se sve vrjednije tekstove koji nisu izvorno nastali u elektroničkom obliku digitalizira. To je potrebno radi njihovog očuvanja i lakšeg korištenja. Pri digitalizaciji tekstova značajnu ulogu igraju programi za optičko prepoznavanje znakova (OCR programi). Iako se učinkovitost OCR programa neprestano poboljšava, oni još uvijek nisu *„savršeni“*. Ovaj rad bavi se nastojanjima unaprjeđenja OCR programa.

Ključne riječi: digitalizacija, informacijske institucije, OCR programi, optičko prepoznavanje znakova, unapređenje OCR programa

Possibilities of improvement of software for optical character recognition (OCR software)

Summary

Digitalization of analogue texts is a process of their conversion into digital form. Texts are being converted into binary code which is written as a computer file. Today one can notice that there is a big need of conversion of all important analogue texts into digital form. This conversion should make preservation and more simple usage of texts possible. In process of digitalization very important role has optical character recognition software (OCR). Although the efficiency of OCR software is getting better all the time, they are still far from perfect. This thesis is dealing with OCR software errors and possibilities of their improvement.

Key words: digitalization, information institutions, OCR software, optical character recognition, improvement of OCR software.