

SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

ZAVRŠNI RAD br. 5709

**Sustav za određivanje strukture
teksta na temelju položaja
pojedinih znakova**

Herman Zvonimir Došilović

Zagreb, lipanj 2018.

SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA
ODBOR ZA ZAVRŠNI RAD MODULA

Zagreb, 14. ožujka 2018.

ZAVRŠNI ZADATAK br. 5709

Pristupnik: Herman Zvonimir Došilović (0036480275)
Studij: Računarstvo
Modul: Računarska znanost

Zadatak: Sustav za određivanje strukture teksta na temelju položaja pojedinih znakova

Opis zadatka:

Sustavi za automatsko očitavanje teksta sa skeniranih dokumenata imaju nekoliko zadaća koje uključuju lokalizaciju, segmentaciju i prepoznavanje pojedinih znakova te slaganje prepoznatih znakova u složenije strukture poput riječi i linija. To je u praksi vrlo težak problem.

U okviru ovog završnog rada potrebno je proučiti načine za određivanje riječi i linija na temelju položaja individualnih znakova te njihovih omeđujućih pravokutnika. U okviru rada potrebno je pripremiti odgovarajući skup podataka za testiranje te napraviti prototipnu implementaciju sustava.

Radu priložiti izvorni kod razvijenih postupaka uz potrebna objašnjenja i dokumentaciju. Predložiti pravce budućeg razvoja. Citirati korištenu literaturu i navesti dobivenu pomoć.

Zadatak uručen pristupniku: 16. ožujka 2018.

Rok za predaju rada: 15. lipnja 2018.

Mentor:

Doc. dr. sc. Marko Čupić

Predsjednik odbora za
završni rad modula:

Prof. dr. sc. Siniša Srblijić

Djelatnica:

Doc. dr. sc. Tomislav Hrkać

Zahvaljujem svom mentoru doc. dr. sc. Marku Čupiću na dozvoli za odabir vlastite teme i na strpljenju, poticaju i savjetima u razvoju rada.

Zahvaljujem tvrtki Microblink na danim sredstvima bez kojih ovaj rad ne bi bio moguć. Posebno zahvaljujem kolegama Jurici Cerovecu, Nenadu Mikši, Borisu Trubiću, Igoru Smolkoviču i Ivanu Jurinu koji su me svojim bogatim znanjem i iskustvom usmjeravali u razvoju rada.

SADRŽAJ

1. Uvod	1
2. Optičko raspoznavanje znakova	2
2.1. Primjene	2
2.2. Komponente OCR-sustava	4
3. Određivanje strukture teksta	8
4. Određivanje strukture teksta na temelju položaja pojedinih znakova	12
4.1. Željena funkcionalnost	15
4.2. Skup podataka za testiranje	17
4.2.1. Slike	17
4.2.2. Ulazne datoteke	19
4.2.3. Očekivane izlazne datoteke	21
4.3. Korištenje skupa podataka za testiranje	21
5. Algoritmi za određivanje strukture teksta	23
5.1. Algoritmi za određivanje linija	25
5.1.1. Algoritam temeljen na maksimalnom preklapanju znakova . .	25
5.2. Algoritmi za rastavljanje riječi	31
5.2.1. Algoritam temeljen na prosječnoj širini znaka	31
5.2.2. Algoritam temeljen na prosječnoj relativnoj udaljenosti . . .	34
5.2.3. Algoritam temeljen na prosječnoj udaljenosti centara	37
6. Rezultati i analiza	40
6.1. Mjere točnosti algoritama	40
6.1.1. Mjera točnosti algoritama za određivanje linija	42
6.1.2. Mjera točnosti algoritama za rastavljanje riječi	43
6.1.3. Daljnja poboljšanja u određivanju točnosti	44

6.2. Rezultati i analiza algoritama za određivanje linija	45
6.3. Rezultati i analiza algoritama za rastavljanje riječi	47
7. Zaključak	50
Literatura	51

1. Uvod

Optičko raspoznavanje znakova sve je više zastupljeno u našem svakodnevnom životu. Koristi se u provjeri osobnih dokumenata, prepoznavanju registarskih tablica, digitalizaciji starih tiskanih knjiga, digitalizaciji raznih tiskanih dokumenata, automatskom unosu podataka s uplatinica itd. U nekim primjenama optičkog raspoznavanja znakova potrebna nam je i struktura očitanog sadržaja nad kojom bi se mogla provesti daljnja analiza. Na primjer, u digitalizaciji knjiga potrebna nam je struktura sadržaja u kojoj su znakovi grupirani u linije i riječi kako bismo nad takvim strukturiranim sadržajem mogli provesti neku drugu obradu. U provjeri osobnih dokumenata potrebna nam je struktura sadržaja kako bismo mogli povratiti informacije o osobnim podacima korisnika kao što su npr. njegovo ime i prezime.

Sustavi za određivanje strukture teksta sastavni su dio optičkog raspoznavanja znakova. U ovom završnom radu predložit ćemo nekoliko načina za određivanje strukture teksta koji će u nestrukturiranom sadržaju odrediti linije i riječi samo na temelju položaju pojedinih znakova. Sustav za određivanje strukture teksta na temelju položaja pojedinih znakova, implementiran u skopu ovog rada, rješavat će problem određivanje strukture teksta u sadržaju s računa iz trgovine i u sadržaju iz knjiga. Sustav će biti podjeljen na dva podsustava od kojih će prvi odrediti linije, a drugi će rastaviti riječi u svakoj liniji.

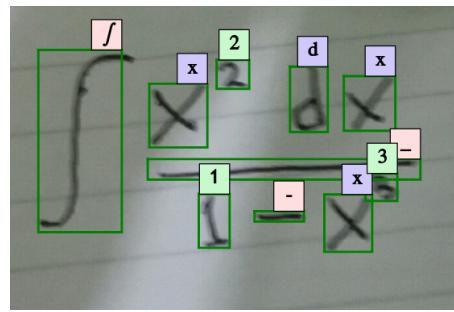
U drugom poglavlju opisane su primjene i način rada sustava za optičko raspoznavanje znakova. Treće poglavlje daje uvid u dosadašni rad u području određivanja strukture teksta, a četvrto poglavlje detaljno opisuje problem koji rješava ovaj rad. Također, u okviru četvrtog poglavlja bit će opisan skup podataka za testiranje razvijenog sustava. Poglavlje 5 formalno opisuje način rada algoritama za određivanje linija i algoritama za rastavljanje riječi koji zajedno čine sustav za određivanje strukture teksta. Rezultati i analiza algoritama, kao i korištene mjere točnosti opisane su u šestom poglavlju. U sklopu analize dane su smjerince za budući rad.

2. Optičko raspoznavanje znakova

Sustav za optičko raspoznavanje znakova (engl. *optical character recognition*) (u daljnjem tekstu: *OCR-sustav*) pretvara sliku tiskanog teksta u digitalizirani format kojim možemo jednostavno manipulirati na računalu. Iako je to ljudima jednostavan zadatak, računalima nije lako prepoznati tekst i pojedine znakove teksta sa slike zbog velike raznolikosti jezika, fonta i stila kojim tekst može biti napisan. Optičko raspoznavanje znakova je stoga vrlo zahtjevan problem i mnogo je istraživačkog truda uloženo u pokušaju da se slike teksta pretvore u format koji računalo razumije. (Islam et al., 2017)

2.1. Primjene

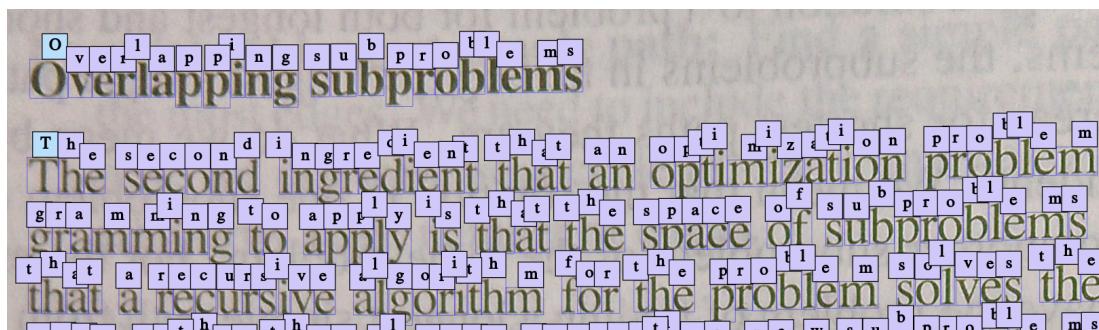
Osim tiskanog teksta, OCR-sustavi koriste se i u prepoznavanju znakova rukom pisanih teksta. Prepoznavanje znakova rukom pisanih teksta je teži problem od prepoznavanja tiskanog teksta (Islam et al., 2017) zato jer se oblik znakova i njihov način pisanja razlikuje kod svake osobe (npr. rukopis odrasle osobe potpuno je drugačiji od rukopisa djeteta). OCR-sustave za detekciju rukom pisanih znakova možemo podijeliti na dvije potkategorije: *on-line* i *off-line*. *On-line* OCR-sustavi detektiraju znakove dok ih korisnici unose i to im omogućuje praćenje parametara poput: brzine pisanja, broja napravljenih poteza, smjer pisanja, itd. *Off-line* OCR-sustavi izvode se nad jednom slikom na kojoj se nalazi sav sadržaj nad kojim je potrebno napraviti detekciju. Takvi sustavi nemaju dodatne informacije koje imaju *on-line* sustavi i zato je detekcija znakova komplikiranija (Islam et al., 2017). Slika 2.1 prikazuje primjer rezultata *off-line* OCR-sustava za detekciju rukom pisanih znakova. OCR-sustavi imaju široku primjenu i možemo ih pronaći primjerice u detekciji znakova na registarskim pločicama (Saghaei, 2016), (Laroca et al., 2018), u detekciji znakova sadržaja knjiga (Wick et al., 2018), (Christy et al., 2017) i detekciji znakova na raznim dokumenatima (Harraj i Raissouni, 2015) (Verma et al., 2016). Na slici 2.2 prikazan je primjer rezultata korištenja OCR-sustava za detekciju znakova na računima iz trgovine. Slika 2.3 prikazuje rezultat OCR-sustava za detekciju znakova na tiskanim knjigama.



Slika 2.1: Rezultat *off-line* OCR-sustava za detekciju znakova rukom pisanog teksta.



Slika 2.2: Rezultat OCR-sustava za detekciju znakova na računima iz trgovine.



Slika 2.3: Rezultat OCR-sustava za detekciju znakova na tiskanim knjigama.

2.2. Komponente OCR-sustava

Optičko raspoznavanje znakova provodi se u nekoliko koraka (Islam et al., 2017) (Kaur i Rani, 2016):

1. pribavljanje slike,
2. predobrada,
3. segmentacija znakova,
4. izdvajanje značajki znakova,
5. klasifikacija znakova i
6. naknadna obrada.

Pribavljanje slike

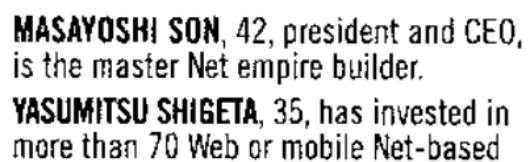
U prvom koraku OCR-a, pribavljanju slike, potrebno je pribaviti sliku nad kojom ćemo provesti ostale korake. Sliku možemo pribaviti s raznih uređaja poput kamere fotoaparata, mobilnog uređaja ili nekog drugog uređaja za digitalizaciju dokumenata (engl. *scanner*). Nakon prvog koraka, slika dokumenta nad kojim provodimo raspoznavanje znakova sastoji se samo od slikovnih elemenata (engl. *pixels*) (Vynckier, 2018). Slika 2.4 prikazuje primjer slike nad kojom možemo provesti postupak raspoznavanja znakova. Slika može sadržati pozadinu koju bi OCR-sustav trebao zanemariti.



Slika 2.4: Ulazna slika u OCR-sustav pribavljena kamerom mobilnog uređaja.

Predobrada

U predobradi slike OCR-sustavi često provode niz morfoloških transformacija i filtra nad pribavljenom slikom. Cilj ovog koraka je povećati kvalitetu slike i smanjiti informacije na slici. Binarizacija je jedan od potkoraka predobrade koji slike u boji ili u nijansama sive pretvara u crno-bijele. Osim binarizacije koriste se neke morfološke transformacije poput dilatacije, rezanja i skaliranja. Slika 2.5 prikazuje primjer slike prije i nakon binarizacije. (Gulan, 2016), (Islam et al., 2017), (Jurin, 2017)



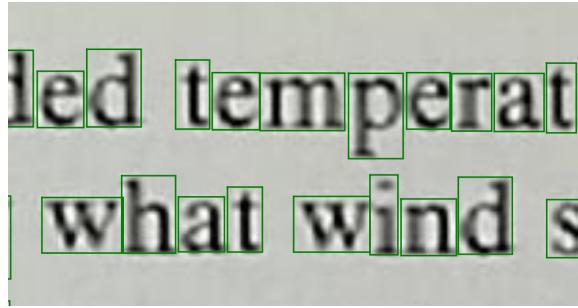
Slika 2.5: Prije binarizacije (lijevo) i nakon binarizacije (desno) (Vynckier, 2018).

Segmentacija znakova

Sljedeći korak, segmentacija znakova, je postupak segmentiranja slike u segmente unutar kojih se nalaze znakovi koje želimo klasificirati. Jedan od pristupa segmentacije izvodi se s vrha prema dnu gdje se prvo segmentiraju linije, zatim riječi i na kraju pojedini znakovi (Jurin, 2017), (Vynckier, 2018). Prednost ovakvog pristupa je da uz lokaciju svakog znaka dobivamo i strukturu cijelog teksta, odnosno, znamo kojoj liniji i kojoj riječi znak pripada. Nedostatak ovakvog pristupa je da ne postoje korekcijski mehanizmi kojima bismo znak pridružili nekoj drugoj liniji ili riječi ako su prva dva koraka segmentacije linije ili riječi neispravni. (Jurin, 2017)

Drugi pristupi poput *ZICER OCR*¹ sustava izravno izvode segmentaciju cijele slike na području koji predstavljaju znakove. Prednost takvog pristupa je da možemo detektirati znakove teksta u kojemu nema riječi i linija, kao što je na primjer matematički izraz. Nedostatak takvog pristupa je da gubimo informaciju o strukturi teksta i zato postoji potreba za razvojem dodatnog sustava koji bi znakove grupirao u riječi, a riječi u linije (Jurin, 2017). Slika 2.6 prikazuje rezultat segmentacije pojedinih znakova.

¹OCR-sustav tvrtke *Microblink*, <https://microblink.com>



Slika 2.6: Segmentacija znakova.

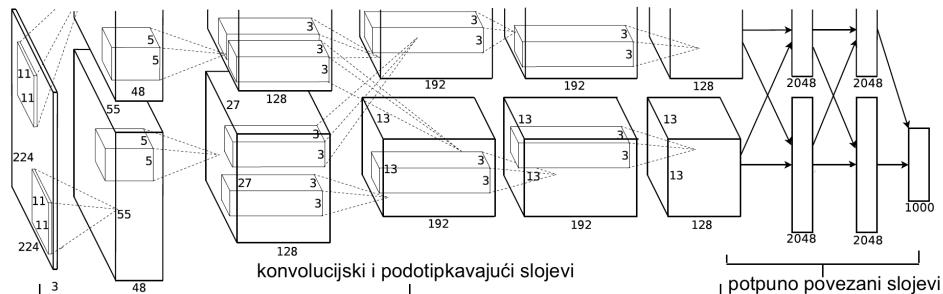
Izdvajanje značajki

Izdvajanje značajki pojedinog znaka podrazumijeva odabir značajki prema kojima će se jedinstveno klasificirati svaki znak. Značajke poput geometrijskog oblika ili statističkih svojstava mogu biti uzete u obzir prilikom klasifikacije. Važno područje istraživanja pripada razmatranju koje i koliko značajki je potrebno uzeti u obzir za kvalitetnu i ispravnu klasifikaciju. (Islam et al., 2017)

Klasifikacija

Klasifikacija je najvažniji korak optičkog raspoznavanja znakova (Verma i Ali, 2012) (Zhu et al., 2016) koji koristi izdvojene značajke za određivanje klase pojedinog znaka (Lehal i Singh, 1999) (Kaur i Rani, 2016). Statistički pristupi klasifikacije koriste diskriminativne funkcije za određivanje klase znaka (Islam et al., 2017), a u novije vrijeme koriste se duboke neuronske mreže (Jurin, 2017). Neki od statističkih pristupa su: Bayesov klasifikator, klasifikator stablom odluke, umjetne neuronske mreže i metoda k-najbližih susjeda (Islam et al., 2017).

(Krizhevsky et al., 2012) objavili su rad koji je označio prekretnicu u klasifikaciji i lokalizaciji objekata (Jurin, 2017). Slika 2.7 prikazuje arhitekturu *AlexNet* koja je pobijedila na natječaju *ImageNet 2012* u području klasifikacije objekata. (Jurin, 2017)

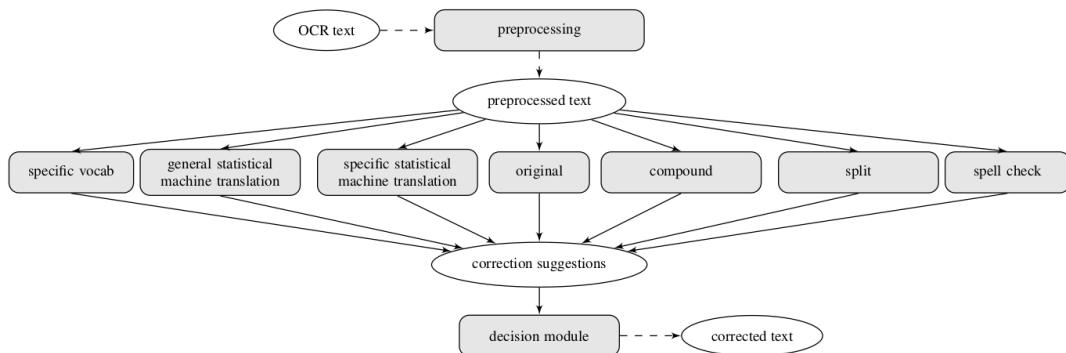


Slika 2.7: Arhitektura *AlexNet* (Jurin, 2017).

Naknadna obrada

Nakon klasifikacije znakova slijedi njihova naknadna obrada koja se koristi kako bi se poboljšali OCR-rezultati. Jedan od pristupa postprocesiranja koristi rezultate više različitih klasifikatora koji mogu biti korišteni slijedno, paralelno ili hijerarhijski. Nakon toga rezultati klasifikatora se kombiniraju različitim pristupima (Islam et al., 2017). Kao što je prije spomenuto, segmentacija koja se ne provodi s vrha prema dnu nema informaciju o strukturi teksta i zato je potrebno razviti dodatan **sustav za određivanje strukture teksta na temelju položaja pojedinih znakova**.

(Schulz i Kuhn, 2017) predstavili su arhitekturu tzv. *post-correction* OCR-sustava kojim su pokazati na koji su način adaptirali generički sustav za naknadnu obradu OCR-rezultata koristeći domensko znanje za konkretni problem koji su rješavali. Ovim pristupom ostvarili su bolje rezultate za konkretni problem nego što su ostvarili koristeći postojeći generički sustav za naknadnu obradu OCR-rezultata.

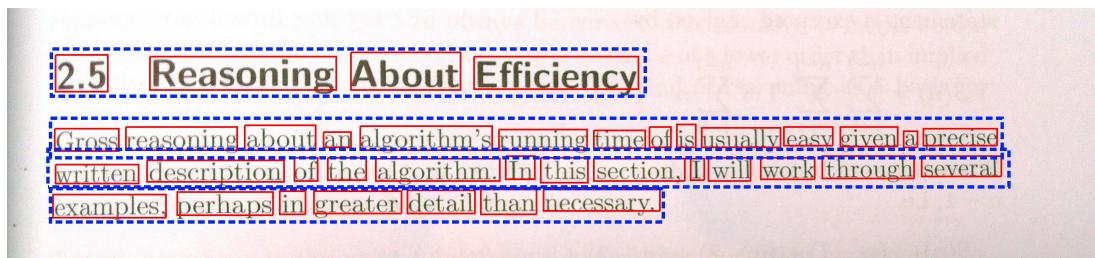


Slika 2.8: Arhitektura *post-correction* OCR sustava (Schulz i Kuhn, 2017).

3. Određivanje strukture teksta

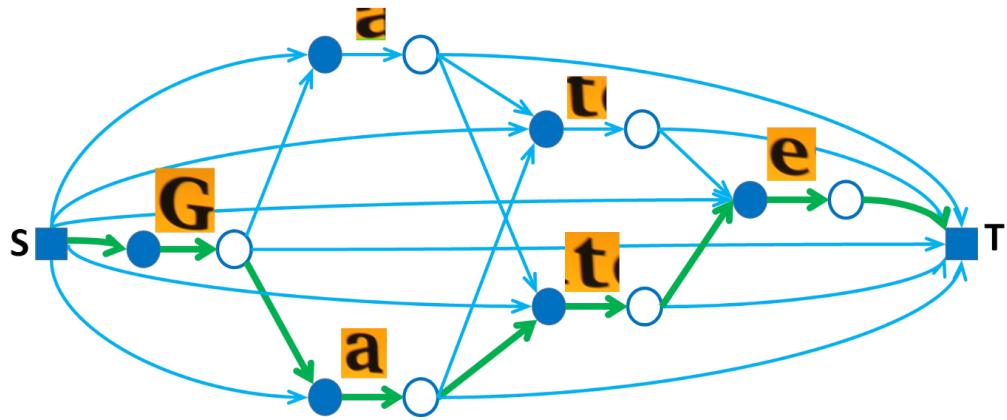
Sustavi za određivanje strukture teksta na temelju OCR-rezultata sastavni su dio OCR-sustava. Određivanje strukture teksta podrazumijeva segmentaciju linija i segmentaciju riječi unutar linije. Neke tehnike segmentacije znakova i njihove klasifikacije nemaju informaciju o tome kojoj liniji i riječi pojedini znak pripada. Prednost takvog pristupa je da takav OCR-sustav možemo koristiti nad slikama koje ne sadrže linije, kao što su na primjer slike matematičkih izraza (Jurin, 2017). Nedostatak je što nakon klasifikacije moramo razviti sustav koji će znakove naknadno obraditi da bismo odredili strukturu teksta (Jurin, 2017).

Tekst na slici može biti podijeljen na linije ili blokove, a u bloku tekst možemo podijeliti na linije. Unutar jedne linije znakove možemo grupirati u riječi. Način na koji će se odrediti struktura teksta uvelike ovisi o problemu koji rješavamo i kakve rezultate želimo dobiti. Tekst na slici 3.1 podijeljen je na linije (plavo), a unutar svake linije na riječi (crveno).



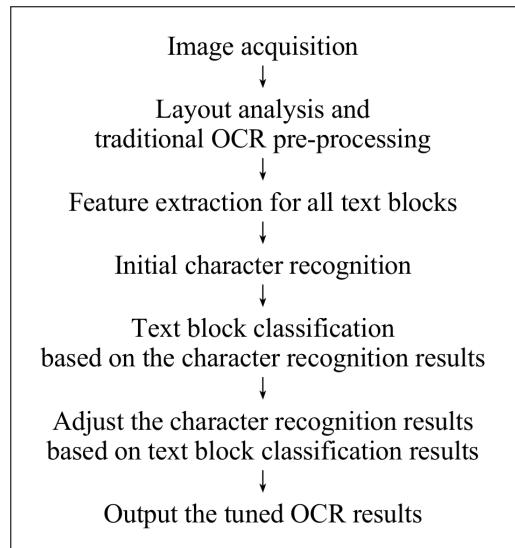
Slika 3.1: Segmentacija linija (plavo) i riječi (crveno) unutar linije.

(Tian et al., 2016) predložili su sustav za određivanje strukture teksta koji će osim određivanja kojoj liniji pojedini znak pripada znati izbaciti tzv. *false positive* znakove odnosno znakove koje je OCR-sustav prepoznao, a zapravo u tekstu ne postoje. Njihov sustav temelji se na *min-cost flow network* modelu koji objedinjuje izbacivanje *false positive* znakova i pronalazak strukture teksta. Na temelju međusobne pozicije između dva prepoznata znaka i dodatnog parametra kojeg dobivaju od klasifikatora, a koji označava vjerojatnost ispravne detekcije, grade težinski usmjereni graf (slika 3.2).



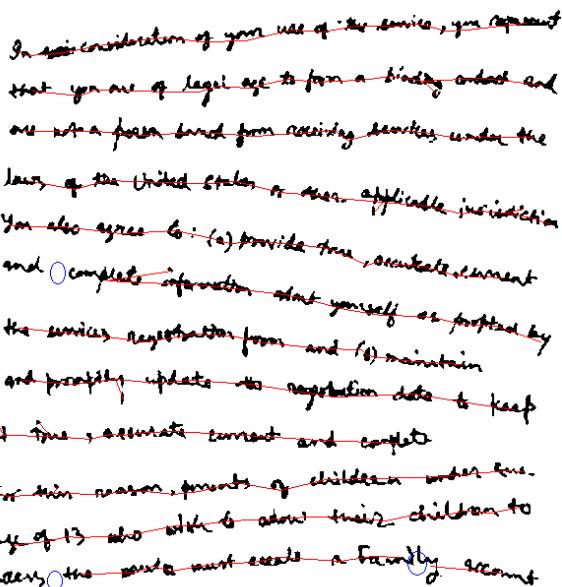
Slika 3.2: Težinski usmjereni graf (Tian et al., 2016).

(Zhu et al., 2016) predložili su novu arhitekturu (slika 3.3) OCR-sustava koji se temelji na empirijskim rezultatima koji su pokazali da sadržaj riječi ne ovisi samo o dijelu teksta u kojem se ta riječ nalazi nego i o susjednim dijelovima teksta. Njihov OCR-sustav radi dvostruku analizu strukture teksta – prije klasifikacije i nakon klasifikacije. Prva analiza strukture teksta omogućuje im da odrede strukturu teksta u blokovima, a druga analiza strukture teksta im omogućuje da poprave pogreške u klasifikaciji. Njihova nova arhitektura predstavlja hibridni OCR-sustav koji iskorištava rezultate analize strukture teksta.



Slika 3.3: Arhitektura OCR-sustava kojeg predlažu (Zhu et al., 2016).

(Yin i Liu, 2007) pronalaze linije u tekstu povezujući znakove u težinski graf nad kojim provode Kruskalov algoritam za pronalazak minimalnog razapinjujućeg stabla. Njihov pristup ne koristi rezultate klasifikacije, nego koriste povezane komponente koje im predstavljaju znakove i koje pronalaze koristeći algoritam temeljen na praćenju kontura (engl. *contour tracing*). Slika 3.4 prikazuje rezultat pronalaska linija teksta u rukom pisanim dokumentu na Engleskom jeziku.

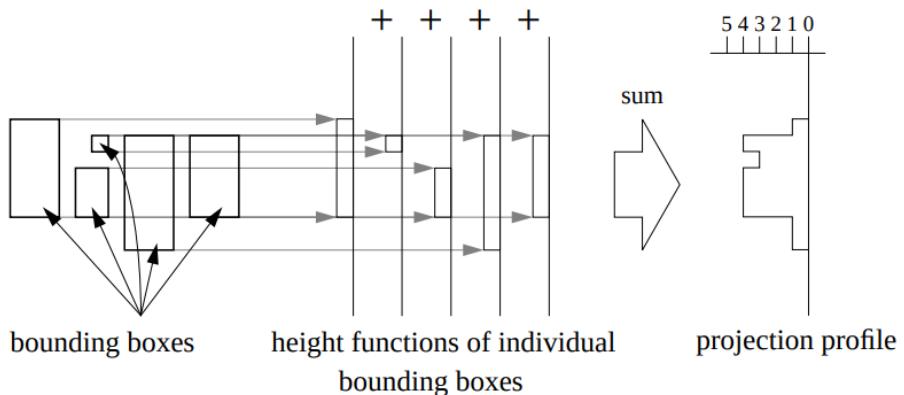


Slika 3.4: Rezultat pronalaska linija teksta u rukom pisanim dokumentu na Engleskom jeziku (Yin i Liu, 2007)

Motivirani njihovim radom (Pan et al., 2011) predstavili su sličan pristup koji u težinama grafa uzima u obzir dodatne težine koje su učene mjerom MCE (engl. *minimum classification error*).

Još jedan pristup predložili su (Yin et al., 2013) koji koriste tehniku hijerahijskog grupiranja koji postupno spaja linije koje dijele znakove dok god postoje linije koje se mogu spojiti. (Tian et al., 2016)

Liang i suradnici (Liang et al., 1996) predlažu heuristički algoritam za određivanje strukture teksta. Algoritam radi horizontalnu projekciju (slika 3.5) omeđujućih pravokutnika na jednu ravninu i pronalazi vrhove i doline u histogramu koji prikazuje frekvencije pojavljivanja projektiranih pravokntnika. Osim ovog pristupa predložili su još jedan koji spaja dva znaka u jednu cjelinu ako i samo ako su dva znaka dovoljno blizu da ih ima smisla spojiti. (Gupta et al., 2006) također koriste razne heuristike prema kojima povezuju susjedne omeđujuće pravokutnike.

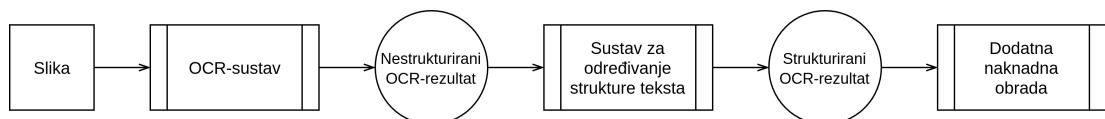


Slika 3.5: Histogram dobiven horizontalnom projekcijom omeđujućih pravokutnika (Liang et al., 1996)

Određivanje strukture teksta je teško i uvelike ovisi o problemu koji rješavamo. U ovom poglavlju pokazano je da strukturiranje teksta može biti izvođeno u raznim fazama izvođenja OCR-sustava. Trenutak u kojem ćemo pokrenuti analizu strukture teksta ovisi o načinu na koji smo označili segmente znakova, koje informacije o segmentu imamo i koji problem rješavamo. Pokazano je da su neki pristupi postigli bolje rezultate kada se iskoristilo domensko znanje i kada su se u nekim heurističkim pristupima koristili parametri koji su bili pomno izabrani za dani problem. Ovaj rad predložit će nekoliko pristupa za određivanje strukture teksta na temelju položaja pojedinih znakova nakon njihove klasifikacije.

4. Određivanje strukture teksta na temelju položaja pojedinih znakova

U kontekstu određivanja strukture teksta na temelju položaja pojedinih znakova, radi preglednije analize, razmatra se OCR-sustav koji nema korak naknadne obrade. Nakon završetka klasifikacije OCR-sustav posjeduje nestrukturirani OCR-rezultat u kojem se nalaze svi segmentirani i klasificirani znakovi. Takav nestrukturirani OCR-rezultat prosljeđuje se sustavu za određivanje strukture teksta na naknadnu obradu. U praksi je sustav za određivanje strukture teksta sastavni dio naknadne obrade OCR-sustava, međutim, u ovaj rad razdvojiti će ta dva sustava radi preglednije analize njihove suradnje koja je prikazana na slici 4.1.



Slika 4.1: Suradnja OCR-sustava i sustava za određivanje strukture teksta.

Sustav za određivanje strukture teksta na temelju položaja pojedinih znakova (u dalnjem tekstu: *Sustav*) na ulaz od OCR-sustava prima nestrukturirani OCR-rezultat koji u sebi sadrži sve znakove koje je OCR-sustav prepoznao. Svaki znak u OCR-rezultatu posjeduje sljedeće informacije:

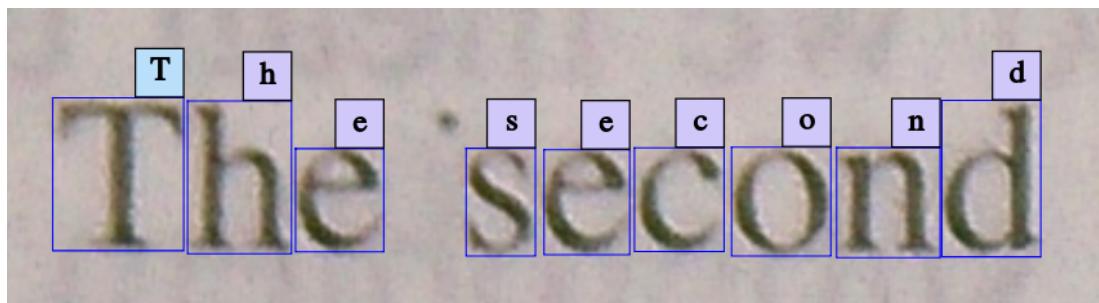
- x - horizontalnu poziciju gornjeg lijevog kuta,
- y - vertikalnu poziciju gornjeg lijevog kuta,
- $width$ - širinu znaka,
- $height$ - visinu znaka i
- $value$ - Unicode vrijednost znaka.

Izlaz iz sustava je strukturirani OCR-rezultat u kojemu su znakovi grupirani u linije, a unutar svake linije riječi su odvojene znakom bjeline. Strukturirani OCR-rezultat se zatim prosljeđuje dodatnim naknadnim obradama koje ovise o problemu koji se rješava.

U sklopu ovog rada potrebno je razviti sustav za određivanje strukture teksta koji će rješavati problem određivanja strukture teksta na računima iz trgovine i sadržaja iz knjiga. Stup podataka za testiranje sustava detaljno je objašnjen u odjeljku 4.2.

Slika 4.2 prikazuje vizualizaciju rezultata OCR-sustava koji je za svaki znak vratio omeđujući pravokutnik (označeno plavom bojom) i vrijednost znaka odnosno klasu kojoj znak pripada. Slike 2.2 i 2.3 također prikazuju vizualizaciju rezultata OCR-sustava i primjere s kakvim će se sustav susresti.

Primjer pojednostavljenog (detaljnije u odjeljku 4.2) nestrukturiranog OCR-rezultata u formatu JSON koji će sustav za određivanje strukture teksta dobiti kao svoj ulaz prikazan je isječkom 4.1. Ulazni nestrukturirani OCR-rezultat u formatu JSON uvijek se sastoji od jedne linije u koju su smješteni svi segmentirani i klasificirani znakovi u nasumičnim redoslijedom.



Slika 4.2: Vizualizacija OCR-rezultata.

Isječak 4.1: Pojednostavljeni nestrukturirani OCR-rezultat u formatu JSON.

```
1 {
2     "ocr_result": {
3         "lines": [
4             {
5                 "chars": [
6                     {
7                         "x": 25.95604, "y": 17.30562,
8                         "width": 10.64438, "height": 16.60289,
9                         "value": 48
10                    },
11                    {
12                        "x": 19.77133, "y": 1.28793,
13                        "width": 16.07777, "height": 10.76925,
14                        "value": 77
15                    },
16                    {
17                        "x": 5.50248, "y": 2.84320,
18                        "width": 12.13375, "height": 15.60966,
19                        "value": 73
20                    },
21                    {
22                        "x": 3.19550, "y": 19.67606,
23                        "width": 14.94088, "height": 20.78798,
24                        "value": 91
25                    },
26                    {
27                        ]
28                    }
29                ]
30            }
31        }
```

4.1. Željena funkcionalnost

Od sustava se očekuje da za dobiveni nestrukturirani OCR-rezultat u formatu JSON vrati novi strukturirani OCR-rezultat u istom formatu koji će znakove grupirati u linije i koji će unutar linija biti poredani s lijeva na desno. Također, linije moraju biti sortirane tako da se najviša linija u dokumentu nalazi na prvom mjestu.

Osim grupiranja linija od sustava se očekuje da između dva znaka, gdje smatra da završava prethodna i započinje nova riječ, ubaci novi znak bjeline čija je vrijednost (engl. *value*) 32, a ostale informacije mogu biti proizvoljne. Dodatan zahtjev je da sustav sve grupirane linije smjesti u jedan blok (detaljnije u pododjeljku 4.2.2).

Isječak 4.2 prikazuje primjer izlaza iz sustava za dani ulaz iz isječka 4.1. Sustav je znakove grupirao u dvije linije i između prvog i zadnjeg znaka u drugoj liniji je ubacio znak bjeline. Vrijednost znaka bjeline je zahtijevana vrijednost 32. Ostale informacije znaka bjeline mogle su biti proizvoljne, međutim, sustav im je dodijelio sljedeće smislenije vrijednosti:

- x - horizontalna pozicija gornjeg desnog kuta lijevog znaka,
- y - vertikalna pozicija gornjeg lijevog kuta desnog znaka,
- $width$ - horizontalna udaljenost između gornjeg desnog kuta lijevog znaka i gornjeg lijevog kuta desnog znaka,
- $height$ - visina lijevog znaka.

Pod *lijevi znak* podrazumijeva se na znak koji se nalazi prije znaka bjeline, a pod *desni znak* podrazumijeva se na znak koji se nalazi nakon znaka bjeline.

Isječak 4.2: Izlaz iz sustava za određivanje strukture teksta u formatu JSON.

```
1 {
2     "ocr_result": {
3         "lines": [
4             {
5                 "chars": [
6                     {
7                         "x": 5.50248, "y": 2.84320,
8                         "width": 12.13375, "height": 15.60966,
9                         "value": 73
10                    },
11                    {
12                        "x": 19.77133, "y": 1.28793,
13                        "width": 16.07777, "height": 10.76925,
14                        "value": 77
15                    }
16                ]
17            },
18            {
19                "chars": [
20                    {
21                        "x": 3.19550, "y": 19.67606,
22                        "width": 14.94088, "height": 20.78798,
23                        "value": 91
24                    },
25                    {
26                        "x": 18.13638, "y": 19.67606,
27                        "width": 7.81966, "height": 20.78798,
28                        "value": 32
29                    }
30                ]
31            },
32            {
33                "x": 25.95604, "y": 17.30562,
34                "width": 10.64438, "height": 16.60289,
35                "value": 48
36            }
37        ]
38    }
39}
```

4.2. Skup podataka za testiranje

Skup podataka za testiranje (u dalnjem tekstu: *podatci*) sustava sastoji se od slika, ulaznih datoteka u formatu JSON i očekivanih izlaznih datoteka. U sklopu ovog rada razvijeni sustav za određivanje strukture teksta rješavat će problem određivanja strukture teksta na računima iz trgovine i sadržaja iz knjiga.

4.2.1. Slike

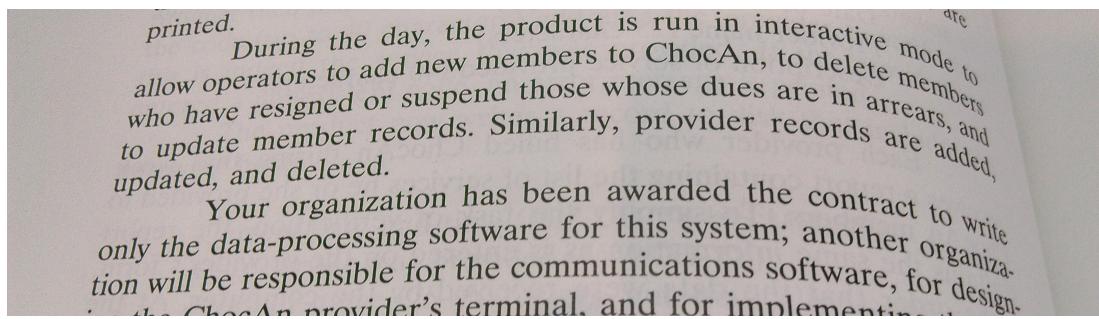
Podatci za testiranje sustava sadrže 100 slika računa (primjer na slici 4.3) i 34 slike sadržaja iz knjiga (primjer na slici 4.4).

Svaki znak na svakoj slici je **ručno** označen i klasificiran. Na slikama računa iz trgovine označeno je ukupno 85068 znakova, a na slikama sadržaja iz knjiga označeno je ukupno 25092 znaka. Ručno označeni podatci oponašaju nestrukturirane OCR-rezultate koji su ulaz u sustav za određivanje strukture teksta.

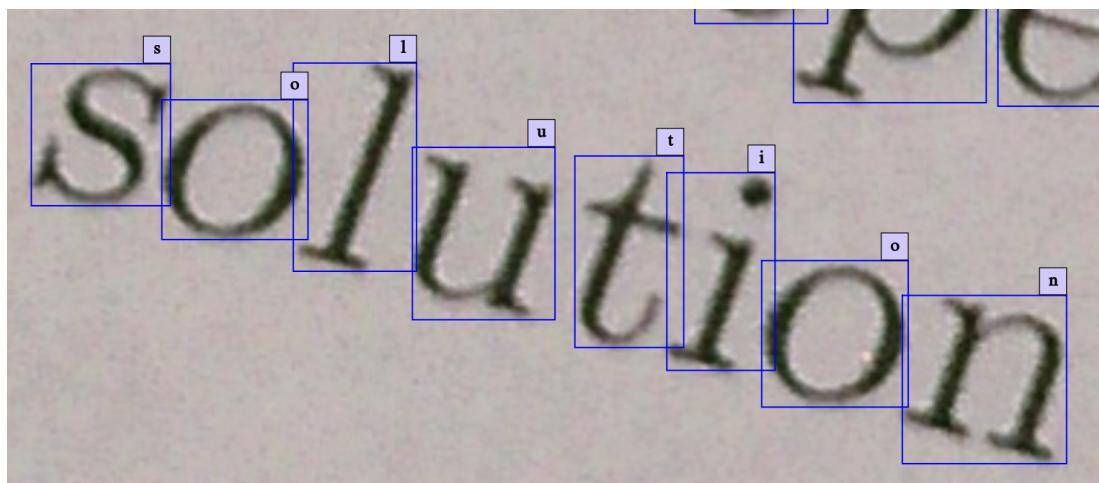
Omeđujući pravokutnici označenih znakova predstavljaju područje koje označeni znak zauzima na slici. Stranice omeđujućih pravokutnika su uvijek paralelne sa rubovima slike. Slika 4.5 prikazuje isječak slike, sadržaja iz knjige, na kojoj su znakovi ukošeni, a stranice njihovih omeđujućih znakova paralelne su sa rubovima slike. Možemo uočiti kako je moguće da se dva susjedna omeđujuća pravokutnika preklapaju.



Slika 4.3: Primjer slike računa iz trgovine.



Slika 4.4: Primjer slike sadržaja iz knjige.



Slika 4.5: Primjer slike s kosim tekstom.

4.2.2. Ulazne datoteke

Slike opisane u pododjelu 4.2.1 **ne predstavljaju** ulaz u sustav za određivanje strukture teksta. Nakon označavanja slika podaci o označavanju svake slike se izvoze i pohranjuju u datoteke u formatu JSON. Datoteke u formatu JSON predstavljaju nesstrukturnirani OCR-rezultat i ulaz u sustav. Ulazna datoteka u formatu JSON u kojoj su zapisani podaci o označavanju slike 4.3 prikazana je u isječku 4.3. Zbog specifičnosti sustava za označavanje slika i načina na koji pohranjuje informacije o označavanju, svi označeni znakovi bit će smješteni u jednu liniju u nasumičnom poretku i ta linija će biti smještena u jedan blok. Za svaki znak dostupna je informacija o Unicode vrijednosti znaka koja je smještena pod ključem `value`. Dodatno, za svaki znak dostupna je informacija o poziciji i veličini njegovog omeđujućeg pravokutnika (engl. *bounding box*). Za svaki omeđujući pravokutnik poznate su sljedeće informacije:

- x - horizontalna pozicija gornjeg lijevog kuta,
- y - vertikalna pozicija gornjeg lijevog kuta,
- $width$ - širina i
- $height$ - visina.

Kako vrijednost x omeđujućeg pravokutnika raste tako je znak bliže desnom rubu slike i kako vrijednost y omeđujućeg pravokutnika raste tako je znak bliže donjem rubu slike. Sve informacije o omeđujućem pravokutniku su vrijednosti iz skupa nenegativnih realnih brojeva.

Sustav za označavanje slika izvozi još neke dodatne informacije o znakovima i o označenoj slici. Sve informacije koje nisu ovdje navede mogu se zanemariti.

Isječak 4.3: JSON datoteka s podatcima o označavanju slike 4.3.

```
1  {
2      "ocr_result": {
3          "blocks": [
4              {
5                  "lines": [
6                      {
7                          "chars": [
8                              {
9                                  "value": 83,
10                                 "bounding_box": {
11                                     "x": 4.548218,
12                                     "y": 271.68826,
13                                     "width": 12.136101,
14                                     "height": 22.48648
15                                 }
16                             },
17                             {
18                                 "value": 65,
19                                 "bounding_box": {
20                                     "x": 4.244385,
21                                     "y": 247.67685,
22                                     "width": 12.59581,
23                                     "height": 21.750519
24                                 }
25                             },
26                             // ostalih 388 znakova
27                         ]
28                     }
29                 ]
30             }
31         ]
32     }
33 }
```

Izlaz iz sustava za određivanje strukture teksta

U odjeljku 4.1 navedene su željene funkcionalnosti sustava i opisan je pojednostavljeni format izlaza iz sustava. Od sustava se očekuje da rezultat nakon određivanja strukture teksta prikaže u istom formatu u kojemu je prikazan ulaz u sustav opisan u ovom odjeljku. Sve grupirane linije potrebno je smjestiti u jedan blok. Sustav ne treba isporučiti informacije koje su se mogu zanemariti na ulazu.

4.2.3. Očekivane izlazne datoteke

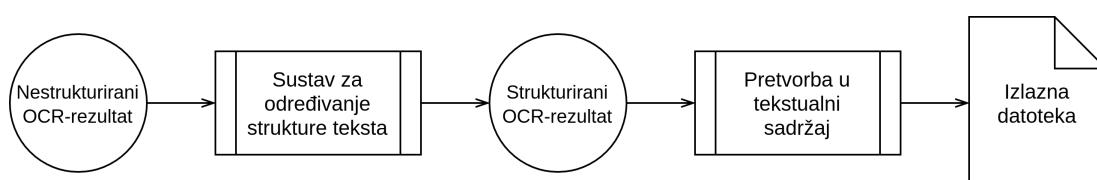
Očekivane izlazne datoteke predstavljaju ispravni strukturirani tekstualni sadržaj sa slike. Isječak 4.4 predstavlja ispravni strukturirani tekstualni sadržaj računa sa slike 4.3. U tekstu u slike 4.3. ne postoje bjeline prije početka linije koje postoje u sadržaju na slici. Osim toga, višestruke bjeline u sadržaju na slici predstavljene su točno jednom bjelinom u tekstu u slike 4.3.

Isječak 4.4: Tekstualni sadržaj slike 4.3.

```
1 POINTS TO $35 REWARD 8770
2 BALANCE REWARDS ACCT # ****2463
3 OPENING BALANCE 20820
4 EVERYDAY POINTS - RETAIL 410
5 CLOSING BALANCE 21230
6 ****
7 Walgreens 01875
8 ACCT 7681
9 SEQUENCE 1875220350
10 PAYMENT FROM PRIMARY
11 Get the flu shot that helps provide
   a lifesaving vaccine to a child in need
12 Get a Shot. Give a Shot.® It's that easy
13 Learn more at the pharmacy.
```

4.3. Korištenje skupa podataka za testiranje

Slika 4.6 prikazuje postupak dobivanja izlazne datoteke iz strukturiranog OCR-rezultata kojeg je vratio sustav za određivanje strukture teksta. Ulaz u sustav je ulazna datoteka u formatu JSON koja predstavlja nestrukturirani OCR-rezultat kao što je opisano u pododjeljku 4.2.2.



Slika 4.6: Postupak dobivanja izlazne datoteke iz strukturiranog OCR-rezultata.

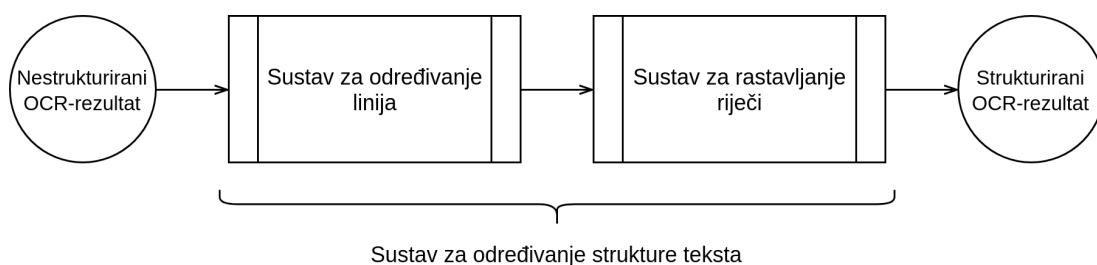
Izlaz iz sustava je strukturirani OCR-rezultat u formatu JSON u kojemu se nalaze znakovi grupirani u linije i gdje su između riječi ubačeni znakovi bjeline. Strukturirani OCR-rezultat potrebno je zapisati u izlaznu datoteku u formatu opisanom u odjeljku 4.2.3. Isječak 4.5 prikazuje pseudokôd algoritma za ispis OCR-rezultata u format opisan u pododjeljku 4.2.3. Izlazne datoteke stvorene na temelju strukturiranih OCR-rezultata i očekivane ulazne datoteke dostupne u skupu podataka za testiranje uspoređuju se metodom objašnjrenom u odjeljku 6.1.

Isječak 4.5: Pseudokôd algoritma za ispis OCR-rezultata.

```
1 def ispisi(ocrRezultat)
2     for linija u ocrRezultat.linije
3         for znak u linija
4             print(znak.value)
5         end
6         print("\n")
7     end
8 end
```

5. Algoritmi za određivanje strukture teksta

Sustav za određivanje strukture teksta podijeljen je u dva podsustava: **sustav za određivanje linija** i **sustav za rastavljanje riječi**. Slika 5.1 prikazuje navedene komponente sustava i njihovu suradnju. Ulaz u sustav za određivanje linija je nestrukturirani OCR-rezultat dobiven od OCR-sustava koji je ujedno i ulaz u sustav za određivanje strukture teksta. Izlaz iz sustava za određivanje linija je OCR-rezultat u kojem su znakovi grupirani u linije kojima pripadaju. Izlaz iz sustava za određivanje linija je ulaz u sustav za rastavljanje riječi koji u svakoj liniji ubacuje na odgovarajuća mesta znakove bjeline koji odvajaju riječi. Izlaz iz sustava za rastavljanje riječi je ujedno i izlaz iz sustava za određivanje strukture teksta.



Slika 5.1: Komponente sustava za određivanje strukture teksta.

U formalnoj definiciji algoritama koriste se sljedeće oznake:

- C - označava skup svih znakova u OCR-rezultatu,
- A i B - označavaju pojedini znak u OCR-rezultatu koji ima svoju: širinu A_w , visinu A_h , horizontalnu poziciju gornjeg lijevog kuta A_x , vertikalnu poziciju gornjeg lijevog kuta A_y i Unicode vrijednost A_v ,
- v - označava Unicode vrijednost,
- L - označava skup svih linija u OCR-rezultatu,

- l i k - označavaju pojedinu liniju u OCR-rezultatu,
- l_{-i} - označava i -ti znak od kraja linije (npr. l_{-1} označava zadnji znak u liniji),
- l_i - označava i -ti znak od početka linije (npr. l_1 označava prvi znak u liniji).

Linije su skupovi, pa će $|l|$ označavati duljinu linije, a za znak A reći ćemo da pripada liniji ako vrijedi $A \in l$. Za dva znaka A i B kažemo da su jednaki ako i samo ako vrijedi:

$$A = B \iff A_w = B_w \wedge A_h = B_h \wedge A_x = B_x \wedge A_y = B_y \wedge A_v = B_v$$

U definiciji algoritama koristit će se sljedeće horizontalne udaljenosti između dva znaka A i B :

$$d(A, B) = |\max(A_x, B_x) - \min(A_x + A_w, B_x + B_w)| \quad (5.1)$$

$$d_l(A, B) = |A_x - B_x| \quad (5.2)$$

$$d_c(A, B) = |A_x + \frac{A_w}{2} - (B_x + \frac{B_w}{2})| \quad (5.3)$$

$$\hat{d}(A, B) = \frac{d(A, B)}{\min(A_w, B_w)} \quad (5.4)$$

$$\hat{d}_l(A, B) = \frac{d_l(A, B)}{\min(A_w, B_w)} \quad (5.5)$$

$$\hat{d}_c(A, B) = \frac{d_c(A, B)}{\min(A_w, B_w)} \quad (5.6)$$

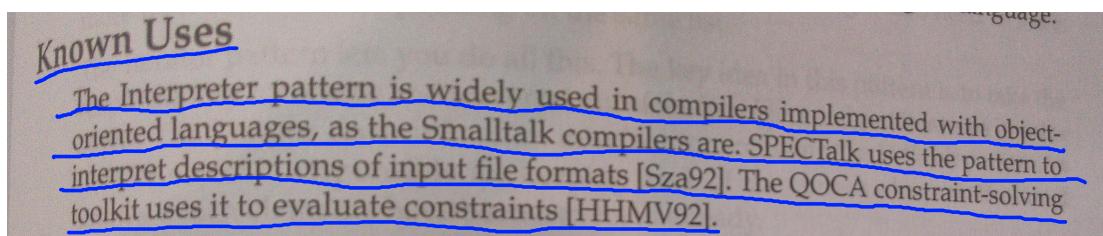
Udaljenost d predstavlja udaljenost između desnog ruba lijevog znaka i lijevog ruba desnog znaka. Udaljenost d_l predstavlja udaljenost lijevih rubova znakova, a udaljenost d_c predstavlja udaljenost centara znakova. Udaljenosti \hat{d} , \hat{d}_l i \hat{d}_c predstavljaju normalizirane udaljenosti d , d_l i d_c .

S ovako definiranim oznakama možemo npr. definirati skup svih znakova koji imaju Unicode vrijednost jednaku v :

$$S(v) = \{A | A \in C, A_v = v\}$$

5.1. Algoritmi za određivanje linija

Algoritmi za određivanje linija trebaju na temelju omeđujućih pravokutnika svakog znaka znakove grupirati u linije. Slika 5.2 predstavlja vizualizaciju očekivanog rezultata algoritma za određivanje linija.



Slika 5.2: Vizualizacija detektiranih linija u sadržaju iz knjige.

Algoritam na ulazu prima nestrukturirani OCR-rezultat koji je ujedno i ulaz u sustav za određivanje strukture teksta. Izlaz iz algoritma je OCR-rezultat u kojem su znakovi grupirani u linije, sortirani s lijeva na desno i u kojem su linije sortirane tako da se najviša linija na slici nalazi na prvom mjestu.

U ovom odjeljku bit će opisan jedan algoritam za određivanje linija temeljen na maksimalnom preklapanju znakova.

5.1.1. Algoritam temeljen na maksimalnom preklapanju znakova

Algoritam za određivanje linija temeljen na maksimalnom preklapanja znakova (u dalnjem tekstu: *algoritam*) temelji se na prepostavci da dva susjedna znaka koja se nalaze u istoj liniji ostvaruju maksimalno vertikalno preklapanje. Vertikalno preklapanje *overlap* dva znaka A i B definira se izrazom:

$$\text{overlap}(A, B) = \frac{\max(0, \min(A_y + A_h, B_y + B_h) - \max(A_y, B_y))}{\min(A_h, B_h)} \quad (5.7)$$

Na početku svog rada algoritam uzlazno sortira sve znakove po horizontalnoj x vrijednost, zatim iterira po svakom znaku i konstruira linije gledajući s kojom linijom promatrani znak ostvaruje maksimalno preklapanje. Preklapanje s linijom definira se kao preklapanje sa zadnjim znakom u toj liniji. Promatrani znak pripadne onoj liniji s kojom ostvari maksimalno preklapanje:

$$l_{max} = \arg \max_{l \in L} \{overlap(A, l_{-1})\} \quad (5.8)$$

Ako je promatrani znak s nekom linijom ostvario preklapanje vrijednosti 0 u skup L dodaje se nova linija i promatrani znak postaje početak te linije. Na ovaj način algoritam konstruira linije i skup svih linija L koji je na početku prazan.

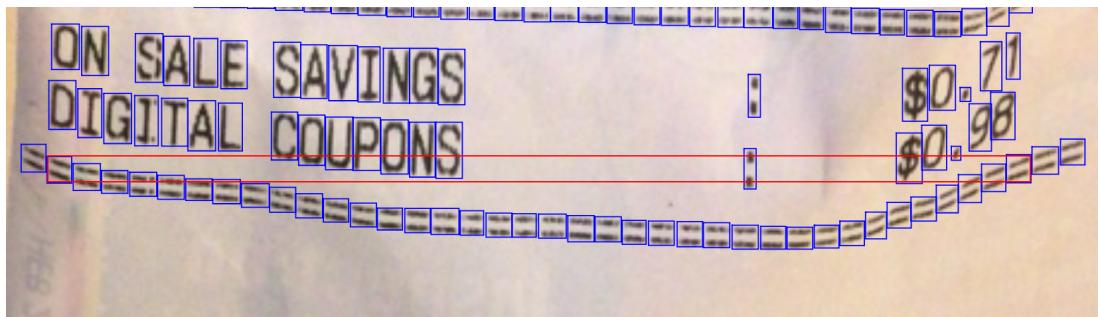
Rješavanje problema valovitih linija

Budući da linije u sadržaju promatranog skupa podataka mogu biti valovite (slika 5.3), ponekad je poželjno izmjeriti preklapanje ne samo sa zadnjim znakom u liniji nego i sa zadnjih nekoliko:

$$l_{max} = \arg \max_{l \in L, i \in [1, \min(|l|, c_1)]} \{overlap(A, l_{-i})\} \quad (5.9)$$

Za razliku od izraza 5.8 koji uzima u obzir samo zadnji znak u svakoj liniji, izraz 5.9 uzet će u obzir zadnjih $\min(|l|, c_1)$ znakova u svakoj liniji, gdje c_1 predstavlja **parametar algoritma**. Eksperimentalno je utvrđeno da se za vrijednost 1 parametra c_1 ostvaruju najbolji rezultati za sadržaj s računa iz trgovine i da se za vrijednost 2 parametra c_1 ostvaruju najbolji rezultati za sadržaj iz knjiga.

Slika 5.3 prikazuje kako znak = na desnoj strani crvenog pravokutnika ostvaruje maksimalno preklapanje sa znakom = na lijevoj strani crvenog pravokutnika koji nije zadnji znak u liniji u tom trenutku. Ovakvi i još mnogi drugi primjeri opravdavaju uvođenje parametra c_1 .

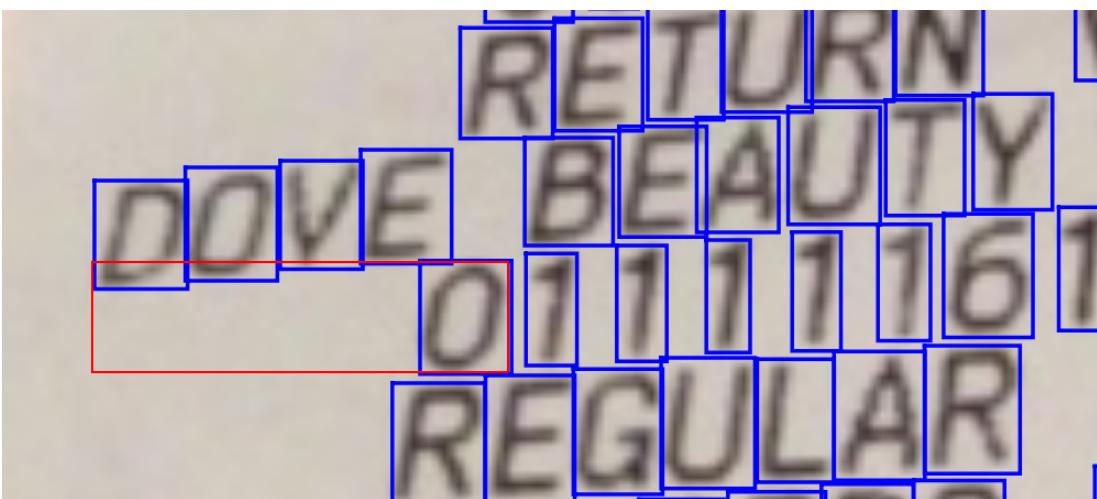


Slika 5.3: Valovite linije otežavaju određivanje linija.

Rješavanje problema preklapanja početaka dviju linija

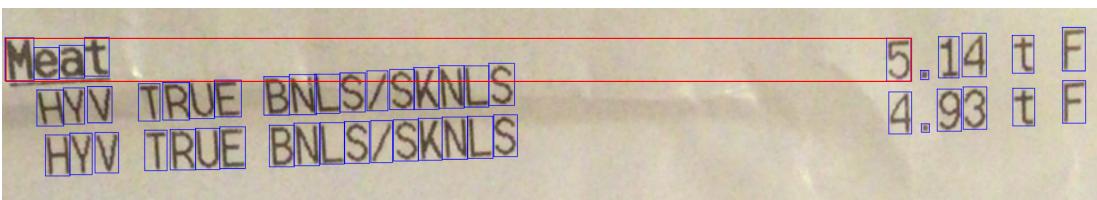
Slika 5.4 prikazuje kako je moguće preklapanje početaka dviju linija. Tako znakovi 0 i D ostvaruju nezanemarivo preklapanje čija će posljedica biti krivo određene linije. Za uspješno izbjegavanje ovakvih slučajeva uvodi se dodatan uvijet koji će odlučiti hoće li promatrani znak pripasti liniji s kojom ostvaruje maksimalno preklapanje:

$$\max_{i \in [1, \min(\lfloor l_{\max} \rfloor, c_1)]} \{overlap(A, l_{\max-i})\} > c_2 \quad (5.10)$$



Slika 5.4: Preklapanje početaka dviju linija.

Nakon što je izrazom 5.9 određeno s kojom linijom promatrani znak ostvaruje maksimalno preklapanje potrebno je provjeriti zadovoljava li iznos tog preklapanja uvjet naveden izrazom 5.10. Uvodi se novi parametar algoritma c_2 koji predstavlja donju granicu preklapanja koju znak mora ostvariti s linijom da bi joj se pripojio. Eksperimentalno je utvrđeno da se za vrijednost 0,13 parametra c_2 ostvaruju najbolji rezultati za sadržaj s računa iz trgovine i da se za vrijednost 0,05 parametra c_2 ostvaruju najbolji rezultati za sadržaj iz knjiga.



Slika 5.5: Lažno pozitivno preklapanje.

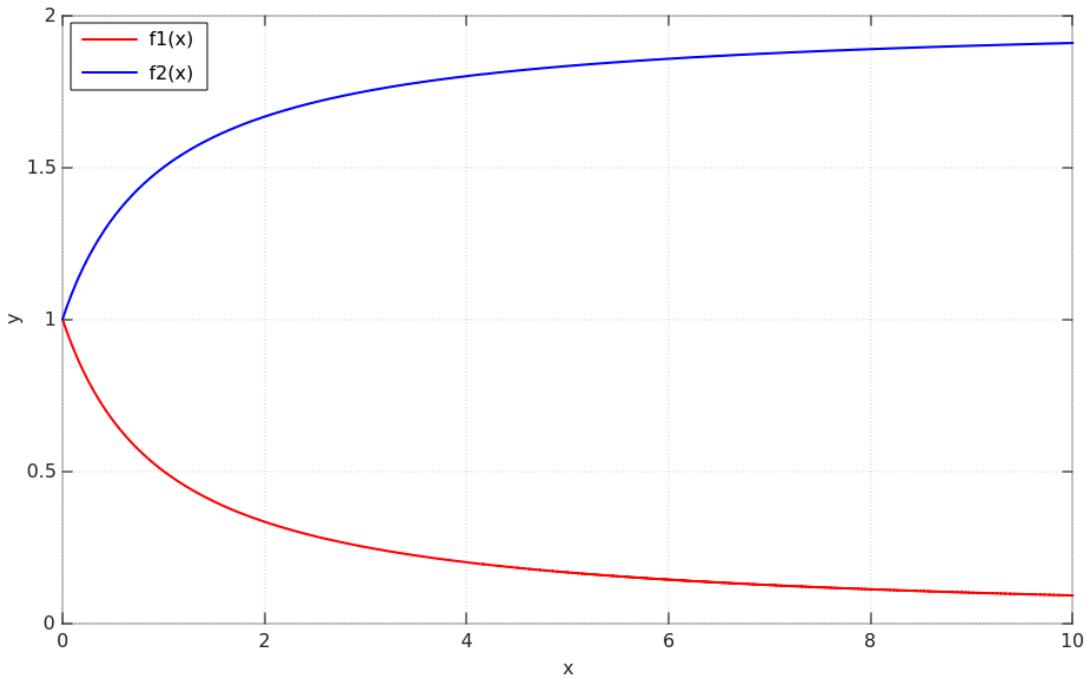
Rješavanje problema lažno pozitivnog preklapanja

Slika 5.5 prikazuje slučaj kada znak pripadne krivoj liniji zato jer s njom ostvaruje maksimalno preklapanje zbog zakriviljenosti slike i sadržaja. U ovom primjeru znak 5 pripast će prvoj liniji `Meat`, a trebao bi pripasti drugoj liniji. Promatraljući skup podataka za koji se rješava problem zaključeno je kako uvijek treba dati prednost onoj liniji čiji je zadnji znak bliži znaku kojeg promatramo. Tako će u primjeru sa slike 5.5 znak 5 pripasti drugoj liniji zato jer je njezin zadnji znak bliži znaku 5 nego znak `t` iz prve linije i zato jer s zadnjim znakom u drugoj liniji ipak postiže nezanemarivo preklapanje.

Za rješavanje ovog problema definiraju se dvije funkcije (slika 5.6) koje koriste dva nova parametra algoritma c_3 i c_4 :

$$f_1(x) = \frac{1}{1 + c_3 \cdot x} \quad (5.11)$$

$$f_2(x) = 1 + \frac{c_4 \cdot x}{1 + c_4 \cdot x} \quad (5.12)$$



Slika 5.6: Grafovi funkcije f_1 (crveno) i funkcije f_2 (plavo) za parametre $c_3 = c_4 = 1$.

Idea je koristiti navedene funkcije f_1 i f_2 kako bi se linijama olakšalo odnosno otežalo postizanje preklapanja u ovisnosti o udaljenosti između njihovih zadnjih znakova. Recimo da u smo u postupku traženja maksimalnog preklapanja u jednom trenutku ostvarili maksimalno preklapanje s linijom l iznosa p_l . S idućom linijom k izmjerimo preklapanje p_k . Da bi linija k postigla novo maksimalno preklapanje sa promatranim znakom mora vrijediti $p_k > p_l$. Međutim, recimo da smo uočili da je zadnji znak linije k bliži promatranom znaku nego zadnji znak linije l . Koristeći navedenu opservaciju možemo liniji k olakšati postizanje novog maksimalnog preklapanja tako da ostvareno preklapanje linije l umanjimo za neki iznos. Budući da je zadnji znak linije k bliži promatranom znaku novi uvijet koji linija k mora zadovoljiti da bi ju smatrali novim maksimalnim preklapanjem glasi:

$$p_k > p_l \cdot f_1(\hat{d}_l(l_{-1}, k_{-1})) \quad (5.13)$$

Funkcija f_1 će liniji k pomoći tim više što su zadnji znakovi linija l i k udaljeniji. Da je zadnji znak linije k bio dalje od promatranog znaka nego zadnji znak linije l onda bismo liniji k trebali otežati postizanje novog maksimalnog preklapanja čak i ako vrijedi $p_k > p_l$. U tom slučaju novi uvijet za liniju k glasi:

$$p_k > p_l \cdot f_2(\hat{d}_l(l_{-1}, k_{-1})) \quad (5.14)$$

Funkcija f_2 će liniji k otežati tim više što su zadnji znakovi linija l i k udaljeniji. Eksperimentalno je utvrđeno da se za vrijednosti parametra $c_3 = c_4 = 0,13$ ostvaruju najbolji rezultati za sadržaj s računa iz trgovine i za sadržaj iz knjiga.

Nakon što je svaki znak smješten u odgovarajuću liniju, linije je potrebno sortirati uzlazno po y vrijednosti prvih znakova linije. Naime, u početnom koraku sortiranja znakova po x vrijednosti moguće je da prvi znak u sortiranim znakovima započinje neku liniju u sadržaju koja nije prva u sadržaju. Taj znak će u algoritmu biti detektiran kao početak nove linije koja će se zatim ubaciti u skup svih linija L .

Konačno, pseudokôd algoritma za određivanje linija na temelju maksimalnog preklapanja znakova prikazan je u isječku. U navedenom isječku funkcije f_1 i f_2 implicitno koriste vrijednost 0,13 za parametare c_3 i c_4 . U pseudokôdu funkcija dln označava funkciju udaljenosti \hat{d}_l .

Isječak 5.1: Pseudokôd algoritma za određivanje linija.

```
1  def odrediLinije(nestrukturiraniOcrRezultat)
2      znakovi = sortirajPoX(nestrukturiraniOcrRezultat.sviZnakovi)
3      linije = []
4
5      for znak u znakovi
6          maxPreklapanje = 0
7          linijaSaMaxPreklapanjem = null
8          for linija u linije
9              p = 0
10             tezina = 1
11             for i u [0, min(c1, linija.duljina)]
12                 p = max(p, overlap(znak, linija[-i]))
13             end
14             if linijaSaMaxPreklapanjem != null
15                 if linijaSaMaxPreklapanjem[-1].x < linija[-1].x
16                     tezina = f1(dln(linijaSaMaxPreklapanjem[-1], linija[-1]))
17                 else
18                     tezina = f2(dln(linijaSaMaxPreklapanjem[-1], linija[-1]))
19                 end
20             end
21             if p > maxPreklapanje * tezina
22                 maxPreklapanje = p
23                 linijaSaMaxPreklapanjem = linija
24             end
25         end
26
27         if maxPreklapanje > c2
28             linijaSaMaxPreklapanjem.nadodaj( znak )
29         else
30             linije.nadodaj([znak])
31         end
32     end
33
34     sortiraneLinije = sortirajPoY(linije)
35
36     return new OcrRezultat(sortiraneLinije)
37 end
```

Vremenska složenost algoritma

U petlji koja počinje linijom 6 prolazi se kroz sve znakove koji provjeravaju s kojom linijom ostvaruju maksimalno preklapanje. U najgorem slučaju potrebno je svaki znak smjestiti u zasebnu liniju što znači da će zadnji (n -ti) znak provjeriti preklapanje sa svim

kom linijom (odnosno sa svakim znakom jer je svaki znak u zasebnoj liniji) što daje složenost $\mathcal{O}(n^2)$, gdje je n ukupan broj znakova. Parametar c_1 na prvi pogled utječe na složenost, ali jednostavnom analizom možemo se uvjeriti da to nije slučaj. Pretpostavimo da parametar c_1 postavimo na vrijednost puno veću od ukupnog broja znakova n . Zadnji znak u petlji s početkom linije 6 treba provjeriti preklapanje sa svakom linijom. Dodatno, zbog petlje koja počinje linijom 11 promatrani znak mora izmjeriti preklapanje sa svakim znakom u liniji što nas opet dovodi do složenosti $\mathcal{O}(n^2)$. Budući da sortiranja u linijama 2 i 36 imaju složenost $\mathcal{O}(n \cdot \log(n))$ one ne utječu na ukupnu složenost algoritma.

5.2. Algoritmi za rastavljanje riječi

Algoritmi za rastavljanje riječi na ulaz primaju OCR-rezultat u kojemu su znakovi grupirani u linije u prethodnom koraku opisanom u odjeljku 5.1. Zadaća algoritama za rastavljanje riječi je da između znakova, za koje smatra da su završetak prethodne odnosno početak iduće riječi, ubaci znak bjeline čija će vrijednost (engl. *value*) biti 32, a ostale informacije mogu biti proizvoljne.

Ovaj odjeljak opisat će tri algoritma za rastavljanje riječi:

- algoritam temeljen na prosječnoj širini znaka
- algoritam temeljen na prosječnoj relativnoj udaljenosti
- algoritam temeljen na prosječnoj udaljenosti centara

5.2.1. Algoritam temeljen na prosječnoj širini znaka

Algoritam temeljen na prosječnoj širini znaka (u dalnjem tekstu: *algoritam*) najjednostavniji je od svih algoritama u ovom odjeljku. Algoritam se temelji na prepostavci da je širina razmaka između riječi proporcionalna sa prosječnom širinom znakova u promatranoj liniji. Označimo prosječnu širinu znaka u liniji l sa \overline{w}_l :

$$\overline{w}_l = \frac{\sum_{A \in l} A_w}{|l|} \quad (5.15)$$

Sada možemo postaviti uvijet koji će odlučiti treba li ubaciti bjelinu između dva znaka A i B :

$$d(A, B) > \overline{w}_l \cdot c_1 \quad (5.16)$$

Parametar c_1 je jedini **parametar algoritma** za koji je eksperimentalno utvrđeno da najbolje rezultate za sadržaj na računima iz trgovine ostvaruje vrijednost 0,8, a najbolje rezultate za sadržaj iz knjiga ostvaruje vrijednost 0,44. Isječak 5.2 prikazuje pseudokôd algoritma za rastavljanje riječi temeljenog na prosječnoj širini znaka.

Isječak 5.2: Pseudokôd algoritma za rastavljanje riječi temeljenog na prosječnoj širini znaka.

```

1  def rastaviRijeci(ocrRezultat)
2      for linija u ocrRezultat.linije
3          prosjecnaSirina = 0
4          for znak u linija
5              prosjecnaSirina += znak.width
6          end
7          prosjecnaSirina /= linija.duljina
8
9          i = 0
10         j = 1
11         while j < linija.duljina
12             udaljenost = d(linija[i], linija[j])
13             if udaljenost > prosjecnaSirina * c1
14                 bjelina = new Znak()
15                 bjelina.value = 32
16                 bjelina.x = linija[i].x + linija[i].width
17                 bjelina.y = linija[i].y
18                 bjelina.width = udaljenost
19                 bjelina.height = linija[i].height
20                 linija.ubaciZnakPrijePozicije(j, znak)
21                 i += 2
22                 j += 2
23             else
24                 i++
25                 j++
26             end
27         end
28     end
29
30     return ocrRezultat
31 end
```

Algoritam prikazan pseudokôdom 5.2 ubacuje znak bjeline čija je vrijednost jednak 32, a ostale informacije su određene na sljedeći način:

- x - horizontalna pozicija gornjeg desnog kuta lijevog znaka,
- y - vertikalna pozicija gornjeg lijevog kuta desnog znaka,
- $width$ - horizontalna udaljenost između gornjeg desnog kuta lijevog znaka i gornjeg lijevog kuta desnog znaka,
- $height$ - visina lijevog znaka.

Pod *lijevi znak* misli se na znak u liniji na poziciji i , a pod *desni znak* misli se na znak u liniji na poziciji j . U liniji 20 ubacuje se novi znak bjeline prije znaka na poziciji j što ima za posljedicu pomicanje svih elemenata desno od j , uključujući i element na poziciji j , za jedno mjesto udesno.

Vremenska složenost algoritma

Ako su u najgorem slučaju svi znakovi smješteni u istu liniju vremenska složenost ovog algoritma bit će $\mathcal{O}(n)$, gdje je n ukupan broj znakova. U analizi vremenske složenosti nije uzeta u obzir vremenska složenost ubacivanja elementa u liniji 20 zato jer postoje strukture podataka (npr. povezana lista) nad kojima je moguće izvesti operaciju ubacivanja u konstantnoj složenosti $\mathcal{O}(1)$.

5.2.2. Algoritam temeljen na prosječnoj relativnoj udaljenosti

Algoritam temeljen na prosječnoj relativnoj udaljenosti (u dalnjem tekstu: *algoritam*) temelji se na pretpostavci da je udaljenost znaka A s vrijednosti (engl. *value*) A_v proporcionalna s prosječnom udaljenosti koju svi znakovi s vrijednosti A_v ostvaruju sa svojim susjedima.

Skup svih susjeda znaka A definiramo kao skup svih znakova B koji su različiti od A , koji se nalaze u istoj liniji kao i A , i za koje vrijedi:

$$\hat{d}_c(A, B) < c_1. \quad (5.17)$$

Skup svih susjeda znaka A označit ćemo s $S(A)$. Parametar c_1 prvi je parametar algoritma za koji je eksperimentalno utvrđeno da se najbolji rezultati za sadržaj s računa iz trgovina postižu za vrijednost 4,0 i da se najbolji rezultati za sadržaj iz knjiga postižu za vrijednost 1,5. Parametar c_1 kontrolira koliko najviše dva znaka smiju biti udaljena da bi se smatrali susjedima.

Skup svih susjeda vrijednosti v definiramo kao uniju svih skupova susjeda znakova A za koje vrijedi $A_v = v$. Formalno, skup svih susjeda vrijednosti v definiramo kao:

$$s(v) = \bigcup_{A \in C} \{S(A) | A_v = v\} \quad (5.18)$$

Konačno, prosječnu udaljenost znaka A od susjednih znakova definiramo kao prosječnu udaljenost koju imaju svi znakovi B , za koje vrijedi $A_v = B_v$, sa svojim susjedima:

$$\overline{d}_c(A) = \frac{\sum_{B \in C, B_v = A_v} \left[\sum_{D \in S(B)} d_c(B, D) \right]}{|s(A_v)|} \quad (5.19)$$

Budući da funkcija \overline{d}_c ne ovisi o A u izrazu 5.19 definiramo ju u ovisnosti o vrijednosti v :

$$\overline{d}_c(v) = \frac{\sum_{B \in C, B_v = v} \left[\sum_{D \in S(B)} d_c(B, D) \right]}{|s(v)|} \quad (5.20)$$

Funkcijom \overline{d}_c dobivamo informaciju kolika je prosječna udaljenost između znakova A , za koje vrijedi $A_v = v$, i njihovih susjeda. Sada možemo dodatnim uvjetom odrediti postoji li razmak između dva susjedna znaka A i B .

Algoritam detektira razmak između znakova A i B ako i samo ako vrijedi:

$$d_c(A, B) > \overline{d}_c(A_v) \cdot c_2 \quad \vee \quad d_c(A, B) > \overline{d}_c(B_v) \cdot c_2. \quad (5.21)$$

Parametar c_2 označava drugi parametar ovog algoritma za koji je eksperimentalno utvrđeno da se najbolji rezultati za sadržaj s računa iz trgovina postižu za vrijednost 1,2 i da se najbolji rezultati za sadržaj iz knjiga postižu za vrijednost 1,7. Parametar c_2 kontrolira koliko minimalno udaljenost između znaka A i B treba biti veća od prosječne udaljenosti \overline{d}_c .

Ovaj algoritam pogodan je za održavanje veze između dva znaka koje bi algoritam opisan u pododjeljku 5.2.1 razdvojio znakom bjeline zbog krive procjene. Ideja ovog algoritma temelji se na pretpostavci da su znakovi koji imaju istu vrijednost u prosjeku jednako udaljeni od znakova s kojima su neposredni susjedi. Ovaj algoritam se u ovoj inačici može koristiti kao dobar pokazatelj da između neka dva znaka zapravo ne treba ubaciti znak bjeline iako bi možda neki drugi algoritam opisan u ovom poglavlju ubacio znak bjeline i time vrlo vjerojatno napravio pogrešku.

Isječak 5.3 prikazuje pseudokôd algoritma temeljenog na prosječnoj relativnoj udaljenosti znakova. U linijama 2 i 3 inicijaliziraju se mape u kojima će se čuvati informacije u ukupnoj udaljenosti susjednih znakova i o tome koliko susjednih znakova je uzeto u obzir za svaku Unicode vrijednost. U liniji 7 koristi se funkcija d_c , a u liniji 8 funkcija \hat{d}_c . U linijama 23 i 24 koriste se navedene mape s kojima se računa prosječna udaljenost koju Unicode vrijednost ostvaruje sa susjednim znakovima, zatim se u uvjetu u liniji 25 detektira razmak. U liniji 33 ubacuje se novi znak bjeline prije znaka na poziciji j što ima za posljedicu pomicanje svih elemenata desno od j , uključujući i element na poziciji j , za jedno mjesto udesno. Algoritam ostale informacije o znaku bjeline određuje na isti način kao i algoritam opisan u pododjeljku 5.2.1.

Isječak 5.3: Pseudokôd algoritma za rastavljanje riječi temeljenog na prosječnoj relativnoj udaljenosti.

```

1  def rastaviRijeci(ocrRezultat)
2      sumMap = map<unicode value, float>()
3      cntMap = map<unicode value, int>()
4
5      for linija u ocrRezultat.liniije
6          for i = 1; i < linija.duljina; i++
7              udaljenost = dc(linija[i], linija[i-1])
8              normUdaljenost = ndc(linija[i], linija[i-1])
9              if normUdaljenost < c1
10                 sumMap[linija[i].value] += udaljenost
11                 cntMap[linija[i].value]++
12                 sumMap[linija[i-1].value] += udaljenost
13                 cntMap[linija[i-1].value]++
14             end
15         end
16     end
17
18     for linija u ocrRezultat.liniije
19         i = 0
20         j = 1
21         while j < linija.duljina
22             udaljenost = dc(linija[i], linija[j])
23             avgRelDistI = sumMap[linija[i].value]/cntMap[linija[i].value]
24             avgRelDistJ = sumMap[linija[j].value]/cntMap[linija[j].value]
25             if udaljenost > avgRelDistI * c2 ||
26                 udaljenost > avgRelDistJ * c2
27                 bjelina = new Znak()
28                 bjelina.value = 32
29                 bjelina.x = linija[i].x + linija[i].width
30                 bjelina.y = linija[i].y
31                 bjelina.width = udaljenost
32                 bjelina.height = linija[i].height
33                 linija.ubaciZnakPrijePozicije(j, znak)
34                 i += 2
35                 j += 2
36             else
37                 i++
38                 j++
39             end
40         end
41     end
42
43     return ocrRezultat
44 
```

Vremenska složenost algoritma

Ako su u najgorem slučaju svi znakovi smješteni u istu liniju vremenska složenost ovog algoritma bit će $\mathcal{O}(n)$, gdje je n ukupan broj znakova. U analizi vremenske složenosti nije uzeta u obzir vremenska složenost ubacivanja elementa u liniji 33 zato jer postoje strukture podataka (npr. povezana lista) nad kojima je moguće izvesti operaciju ubacivanja u konstantnoj složenosti. U analizi složenosti isto tako nije uzeta u obzir vremenska složenost ubacivanja i dohvaćanja elemenata iz mape zato jer bismo takvu mapu mogli ostvariti jednim nizom u kojem bi pozicija ključa bila jednaka vrijednosti ključa. Korištenje takvog niza omogućilo bi nam da operacije ubacivanja i dohvaćanja radimo u konstantnom vremenu. Dodatno, korištenje takvog niza bilo bi memorijski skupo jer bismo u najgorem slučaju trebali imati mjesta za svaku Unicode vrijednost. Ako bismo htjeli izbjegći takav niz možemo koristiti mapu s vremenskom složenosti $\mathcal{O}(\log(n))$ za operacije ubacivanja i dohvaćanja, gdje je n ukupan broj znakova. Korištenjem takve mape ukupna složenost algoritma bila bi $\mathcal{O}(n \cdot \log(n))$.

5.2.3. Algoritam temeljen na prosječnoj udaljenosti centara

Algoritam temeljen na prosječnoj udaljenosti centara (u dalnjem tekstu: *algoritam*) temelji se na pretpostavci da je širina razmaka između riječi proporcionalna s prosječnom udaljenosti centara između susjednih znakova. Skup svih susjeda znaka A definira se na isti način kao i u algoritmu opisanom u odjeljku 5.2.2. Skup svih susjeda znaka A definiramo kao skup svih znakova B koji su različiti od A , koji se nalaze u istoj liniji kao i A , i za koje vrijedi izraz:

$$\hat{d}_c(A, B) < c_1. \quad (5.22)$$

Skup svih susjeda znaka A označit ćemo s $S(A)$. Parametar c_1 prvi je parametar algoritma za koji je eksperimentalno utvrđeno da se najbolji rezultati za sadržaj s računa iz trgovina postižu za vrijednost 3,0 i da se najbolji rezultati za sadržaj iz knjiga postižu za vrijednost 1,5. Parametar c_1 kontrolira koliko najviše dva znaka smiju biti udaljena da bi se smatrali susjedima.

Pomoću skupa $S(A)$ definirat ćemo prosječnu udaljenost centara između susjednih znakova u liniji l :

$$\overline{d}_c(l) = \frac{\sum_{A \in l} \left[\sum_{B \in S(A)} d_c(A, B) \right]}{\left| \bigcup_{A \in l} S(A) \right|} \quad (5.23)$$

Napokon, algoritam ubacuje znak bjeline između znakova A i B ako je ispunjen uvijet:

$$d_c(A, B) > \overline{d}_c(l) \cdot c_2. \quad (5.24)$$

Parametar algoritma c_2 kontrolira koliko minimalno udaljenost između znaka A i B treba biti veća od prosječne udaljenosti centara između susjednih znakova u liniji l da bi znakovi bili razdvojeni. Eksperimentalno je utvrđeno da se najbolji rezultati za sadržaj s računa iz trgovina postižu za vrijednost 1,2 i da se najbolji rezultati za sadržaj iz knjiga postižu za vrijednost 1,4.

Isječak 5.4 prikazuje pseudokôd algoritma temeljenog na prosječnoj udaljenosti centara. U liniji 5 koristi se funkcija \hat{d}_c , a u liniji 6 funkcija d_c . U liniji 24 ubacuje se novi znak bjeline prije znaka na poziciji j što ima za posljedicu pomicanje svih elemenata desno od j , uključujući i element na poziciji j , za jedno mjesto udesno. Algoritam ostale informacije o znaku bjeline određuje na isti način kao i algoritam opisan u pododjeljku 5.2.1.

Isječak 5.4: Pseudokôd algoritma za rastavljanje riječi temeljenog na prosječnoj udaljenosti centara.

```

1  def rastaviRijeci(ocrRezultat)
2      for linija u ocrRezultat.linije
3          brojac = 0
4          for i = 1; i < linija.duljina; i++
5              if ndc(linija[i], linija[i-1] < c1
6                  avgUdaljenost += dc(linija[i], linija[i-1])
7                  brojac++
8              end
9          end
10
11      avgUdaljenost /= brojac
12
13      i = 0
14      j = 1
15      while j < linija.duljina
16          udaljenost = dc(linija[i], linija[j])
17          if udaljenost > avgUdaljenost * c2
18              bjelina = new Znak()
19              bjelina.value = 32
20              bjelina.x = linija[i].x + linija[i].width
21              bjelina.y = linija[i].y
22              bjelina.width = udaljenost
23              bjelina.height = linija[i].height
24              linija.ubaciZnakPrijePozicije(j, znak)
25              i += 2
26              j += 2
27          else
28              i++
29              j++
30          end
31      end
32  end
33
34  return ocrRezultat
35 end
```

Vremenska složenost algoritma

Vremenska složenost algoritma za najgori slučaj je $\mathcal{O}(n)$ budući da će se svaki znak posjetiti samo jednom. U analizi vremenske složenosti nije uzeta u obzir vremenska složenost ubacivanja elementa u liniji 24 zato jer postoje strukture podataka (npr. povezana lista) nad kojima je moguće izvesti operaciju ubacivanja u konstantnoj složenosti.

6. Rezultati i analiza

Ovo poglavlje prikazuje i analizira rezultate rada svakog algoritma opisanog u poglavljju 5. U odjeljku 6.1 opisana je metoda za određivanje točnosti svakog algoritma. U odjeljcima 6.2 i 6.3 prikazani su i analizirani rezultati rada algoritama za određivanje linija i algoritama za rastavljanje riječi.

6.1. Mjere točnosti algoritama

U odjeljku 4.3 opisan je način uporabe skupa podataka za testiranje. Sustav za određivanje strukture teksta (u dalnjem tekstu: *Sustav*) na ulaz dobiva nestrukturirani OCR-rezultat, a na izlazu daje strukturirani OCR-rezultat koji je potrebno zapisati u izlaznu datoteku u formatu opisanom u pododjeljku 4.2.3. Sadržaj izlazne datoteke uspoređuje se sa sadržajem očekivane izlazne datoteke dostupne u skupu podataka za testiranje. Rezultat usporedbe odredit će točnost sustava za određivanje strukture teksta.

Za usporedbu sadržaja koristi se Levenshteinova udaljenost između niza znakova sadržaja očekivane izlazne datoteke i niza znakova sadržaja izlazne datoteke stvorene na temelju OCR-rezultata dobivenog iz sustava. Algoritam Levenshteinove udaljenosti dobro je opisan u literaturi i ovaj rad neće ulaziti u detalje njegova rada.

Isječak 6.1: Primjer sadržaja očekivane izlazne datoteke.

- 1 Lorem ipsum dolor sit amet,
- 2 consectetur adipiscing elit.

Način određivanja točnosti pokazat će se na primjeru isječaka 6.1 i 6.2. Isječak 6.1 prikazuje sadržaj očekivane izlazne datoteke, a isječak 6.2 prikazuje sadržaj izlazne datoteke stvorene na temelju OCR-rezultata dobivenog iz sustava.

Isječak 6.2: Primjer sadržaja izlazne datoteke stvorene na temelju OCR-rezultata dobivenog iz sustava.

- 1 Lorem dolors it amet ,
- 2 consecete tur ipsum ad ipiscing elit.

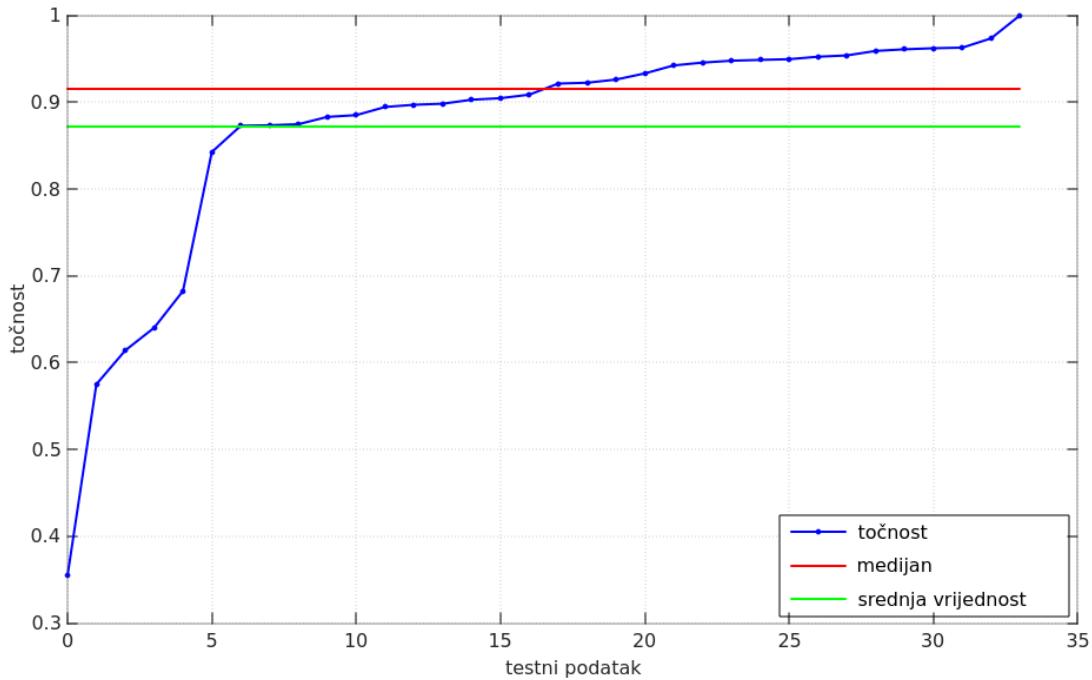
Sustav je progriješio u određivanju linija time što je riječ `ipsum` stavio u drugu liniju. Osim toga, progriješio je i u razdvajanju riječi u obje linije. Levenshteinova udaljenost sadržaja te dvije datoteke je 17, međutim, ta informacija ne govori ništa o točnosti sustava. Rezultat Levenshteinove udaljenosti određuje broj pogrešaka koje sustav radi u odnosu na referentne vrijednosti (očekivane izlazne datoteke). Maksimalan broj pogrešaka koje je sustav mogao napraviti je zapravo maksimalna moguća Levenshteinova udaljenost između dva niza znakova, koja je jednak duljini duljeg niza znakova. U primjeru, dulji niz znakova je niz znakova iz isječka 6.2 koji se sastoji od 59 znakova (uključujući i znak za novi redak `\n`). Sustav je napravio 17 pogrešaka od maksimalno 59, odnosno njegova greška iznosi $\frac{17}{59} \cdot 100\% = 29\%$. Iz pogreške se može izračunati točnost koja iznosi 71%.

Formalno, točnost ili dobrota (engl. *fitness*) sustava određuje se izrazom:

$$f(a, b) = 1 - \frac{d_{levenshtein}(a, b)}{\max(|a|, |b|)}. \quad (6.1)$$

U izrazu su s a i b označeni nizovi znakova, a s $d_{levenshtein}$ Levenshteinova funkcija udaljenosti. U nastavku ovog poglavlja točnost sustava iskazivat će se u vrijednostima između 0 i 1, a ne u postocima.

Slika 6.1 prikazuje graf točnosti (dobrote) sustava za određivanje strukture teksta za sadržaj računa iz trgovine. Za svaki testni primjer izmjerena je dobrota sustava, zatim su izmjerene dobrote uzlazno sortirane i tako su nanesene na graf dobrote te je zbog toga graf monotono rastući. Na ovaj će se način prikazivati grafove dobrote u rezultatima svih algoritama.



Slika 6.1: Graf točnosti sustava.

6.1.1. Mjera točnosti algoritama za određivanje linija

Mjera točnosti algoritama za određivanje linija svodi se na usporedbu sadržaja očekivane izlazne datoteke iz koje su izbrisani svi znakovi bjeline. Sadržaj takve (modificirane) datoteke smatra se referentnim sadržajem prilikom određivanja točnosti. Isječak 6.3 prikazuje referentni sadržaj, za određivanje točnosti algoritama za određivanje linija, dobiven brisanjem znakova bjeline iz sadržaja datoteke prikazane isječkom 6.1.

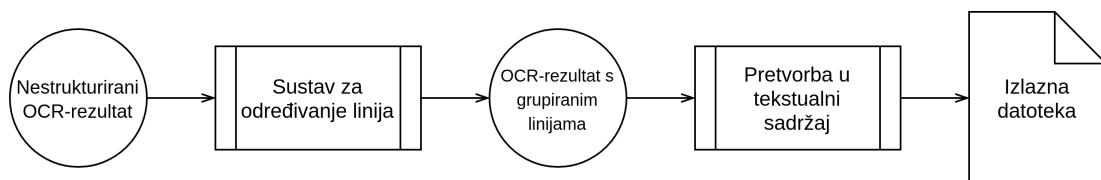
Isječak 6.3: Primjer referentnog sadržaja datoteke.

```

1 Lorem ipsum dolor sit amet,
2 consectetur adipisciing elit.

```

Slika 6.2 prikazuje postupak dobivanja izlazne datoteke iz OCR-rezultata s grupiranim linijama. Format izlazne datoteke opisan je u pododjeljku 4.2.3. Sadržaj dobivene izlazne datoteke uspoređuje se sa referentnim sadržajem koristeći izraz 6.1.



Slika 6.2: Postupak dobivanja izlazne datoteke iz OCR-rezultata dobivenog od sustava za određivanje linija.

Isječak 6.4 prikazuje primjer sadržaja izlazne datoteke stvorene na temelju OCR-rezultata dobivenog od sustava za određivanje linija. U sadržaju ne postoje razmaci zato jer OCR-rezultat nije još prošao kroz sustav za rastavljanje riječi. Točnost sadržaja izlazne datoteke u odnosu na referentni sadržaj iz isječka 6.3 iznosi 0,8.

Isječak 6.4: Primjer sadržaja izlazne datoteke.

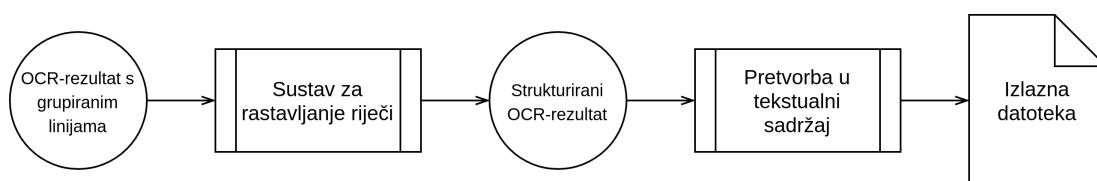
- | | |
|---|------------------------------------|
| 1 | Lorem dolor sit amet, |
| 2 | consectetur ipsum adipiscing elit. |

Na opisani način računat će se točnost za algoritme za određivanje linija, a funkcija točnosti prikazivat će se kao na slici 6.1.

6.1.2. Mjera točnosti algoritama za rastavljanje riječi

Za određivanje točnosti algoritama za rastavljanje riječi koriste se, kao referentne datoteke, očekivane izlazne datoteke koje će se uspoređivati sa sadržajem izlazne datoteke stvorene na temelju OCR-rezultata dobivenog od sustava za rastavljanje riječi. Slika 6.3 prikazuje postupak dobivanja izlazne datoteke iz OCR-rezultata dobivenog od sustava za rastavljanje riječi. Ulaz u sustav za rastavljanje riječi je OCR-rezultat, s grupiranim linijama, dobiven iz sustava za određivanje linija. Budući da sustav za određivanje linija ponekad neispravno određuje linije, ne može se od sustava za rastavljanje riječi očekivati bolja dobrota od dobre sustava za određivanje linija. Određivanje točnosti algoritama za rastavljanje riječi je time ovisno o točnosti algoritama za određivanje linija, što je najveći nedostatak ovog pristupa za mjerjenje točnosti.

Budući da je ulazni OCR-rezultat u sustava za rastavljanje riječi rezultat kojeg je vratio sustav za određivanje linija, mjera točnosti sustava za rastavljanje riječi je ujedno i mjera točnosti cijelog sustava za određivanje strukture teksta zato.



Slika 6.3: Postupak dobivanja izlazne datoteke iz OCR-rezultata
dobivenog od sustava za rastavljanje riječi.

6.1.3. Daljnja poboljšanja u određivanju točnosti

Prednost korištenog skupa za testiranje i opisanih metoda za određivanje točnosti je to što se lako može generirati veći skup podataka za testiranje koristeći metodu početnog podizanja (engl. *bootstrapping*). Kada su na raspolaganju samo označeni znakovi s omeđujućim pravokutnicima moguće je, na temelju nekoliko jednostavnih unaprijed ručno napisanih očekivanih izlaznih datoteka, napraviti prvu verziju sustava za određivanje strukture teksta. Analizom rezultata, sustav se poboljšava do zadovoljavajuće točnosti. Nakon toga, takav sustav se iskoristi za generiranje komplikiranijih očekivanih izlaznih datoteka u kojima je potrebno ispraviti pogreške koje je sustav napravio, ali zato ne treba ručno sastavlјati cijelu očekivanu izlaznu datoteku. Zatim se ponovo može raditi na razvoju sustava i interpretaciji rezultata da bi se sustav poboljšao dok se ne postigne zadovoljavajuća točnost. Ovom metodom, korištenom u ovom radu, može se brzo stvoriti skup podataka za testiranje.

Najveći nedostatak u navedenim metodama za određivanje točnosti je to što točnost sustava za rastavljanje riječi ovisi o točnosti sustava za određivanje linija. Osim toga, usporedbom sa sadržajem očekivane izlazne datoteke gubimo informaciju o tome koji znak s omeđujućim pravokutnikom predstavlja zapisani znak u datoteci. Oba problema mogu se rješiti boljim skupom podataka za testiranje koji bi se trebao sastojati od unaprijed strukturiranih OCR-rezultata u formatu JSON koji predstavljaju očekivani OCR-rezultat. S takvim skupom podataka može se nezavisno testirati i analizirati ponašanje sustava za određivanje linija i sustava za rastavljanje riječi. Dodatno, takav skup podataka omogućio bi korištenje bolje metode za mjerjenje točnosti koje bi se temeljile na usporedbi znakova iz OCR-rezultata, a ne znakova iz datoteka. Stvaranje takvog skupa za testiranje zahtjeva implementaciju posebnog sustava za označavanje linija i riječi na slici, odnosno sustava koji bi omogućio da se označeni znakovi (dostupni u trenutnom skupu podataka) spoje u odgovarajuće linije i riječi.

Za daljnji rad preporuča se sastavljanje novog skupa podataka i definiranje novih mjera točnosti koje bi dale bolji uvid u način rada pojedinih algoritama.

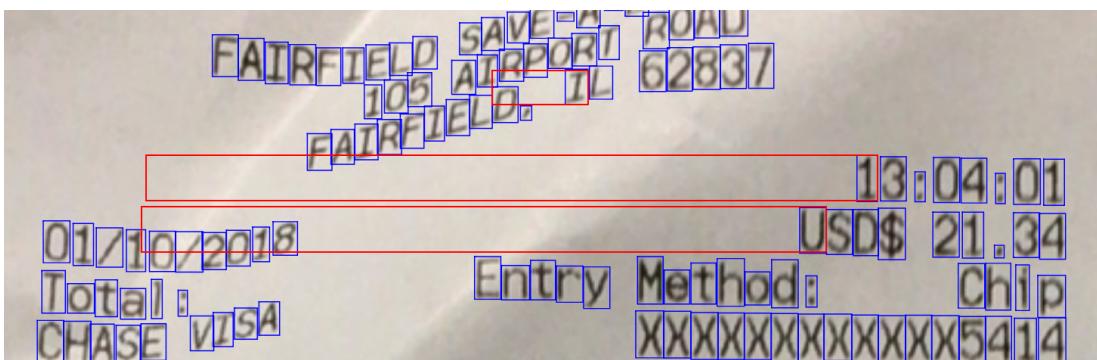
6.2. Rezultati i analiza algoritama za određivanje linija

Grafovi točnosti algoritma za određivanje linija temeljenog na maksimalnom preklapanju znakova prikazani su na slikama 6.5 i 6.6. Rezultati u tablici 6.1 pokazuju kako se za 57% primjera sadržaja računa iz trgovina i za 64% primjera sadržaja iz knjiga ostvaruje maksimalna točnost iznosa 1. Prikazani rezultati ostvareni su s algoritmom koji koristi optimalne parametre navedene u pododjeljku 5.1.1.

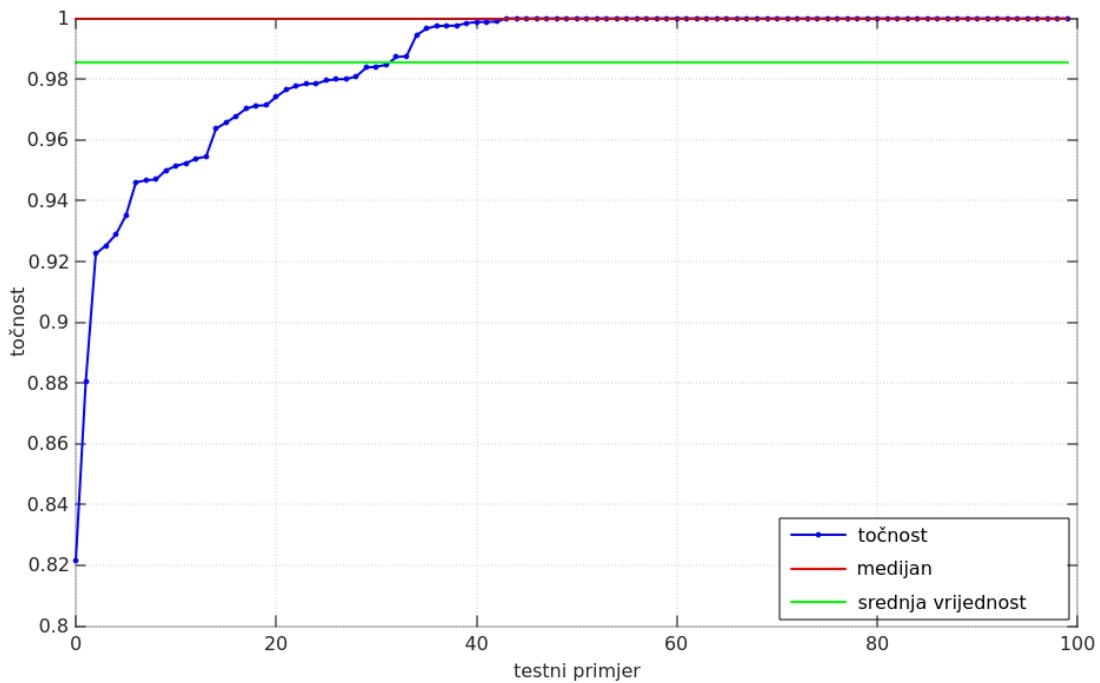
Tablica 6.1: Točnost algoritma za određivanje linija.

	Min.	Sred.	Med.	Maks.	Udio primjera s maks. točnosti
Računi	0,82	0,99	1	1	0,57
Knjige	0,67	0,98	1	1	0,64

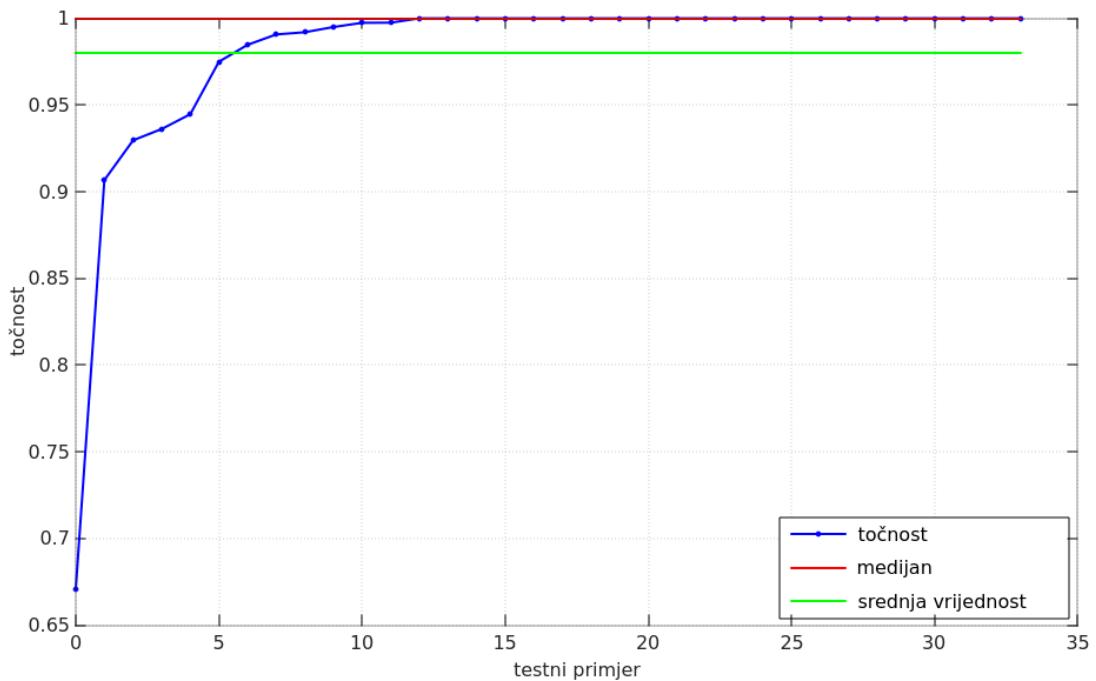
Na slici 6.4 prikazana su dva primjera za koje algoritam griješi u detekciji linija. U prvom slučaju algoritam nije mogao zadovoljiti minimalni iznos preklapanja da poveže znakove . i 1 u istu liniju. Ovaj problem se rješava povećanjem parametra c_1 na vrijednost 2, ali eksperimenti su pokazali kako to nije optimalan parametar za sadržaj s računa iz trgovina. Ovaj problem bi se mogao riješiti tako da algoritam mjeri preklapanje samo između znakova koji su relativno iste visine. U ovom primjeru to bi značilo da će algoritam zanemariti znak . i izmjerit će preklapanje sa znakom D. U drugom slučaju algoritam ne može spojiti lijevu i desnu stranu u istu liniju zbog prevelike zakriviljenosti sadržaja. Ovaj problem mogao bi se rješiti tako da algoritam ima informaciju o maksimalnoj x vrijednosti na kojoj se može nalaziti znak koji započinje liniju. S takvom informacijom može se forsirati spajanje znakova s krajnje desne strane s grupiranim linijama s lijeve strane.



Slika 6.4: Neispavna detekcija linija u računima zbog prevelike zakriviljenosti sadržaja.



Slika 6.5: Graf točnosti algoritma za određivanje linija u sadržaju s računa iz trgovine.



Slika 6.6: Graf točnosti algoritma za određivanje linija u sadržaju iz knjiga.

6.3. Rezultati i analiza algoritama za rastavljanje riječi

Slike 6.7 i 6.8 prikazuju rezultate algoritama za rastavljanje riječi koji koriste optimalne parametre navedene u pododjeljcima 5.2.1, 5.2.2 i 5.2.3. U tablicama 6.2 i 6.3 algoritmi iz odjeljka 5.2 imenovani su kraticama: *avgcharwidth*, *avgreldist* i *avgcenterdist*. Prikazani rezultati pokazuju točnost algoritama za rastavljanje riječi, a ujedno i točnost cijelog sustava za određivanje strukture teksta. Rezultati pokazuju da se s algoritmom temeljenog na prosječnoj udaljenosti centara može potpuno točno odrediti struktura teksta, u sadržaju s računa iz trgovina, za 54% testnih primjera. Isti algoritam postiže znatno lošije rezultate u određivanju strukture teksta u sadržaju iz knjiga.

Tablica 6.2: Točnost algoritama za rastavljanje riječi u sadržaju s računa iz trgovine.

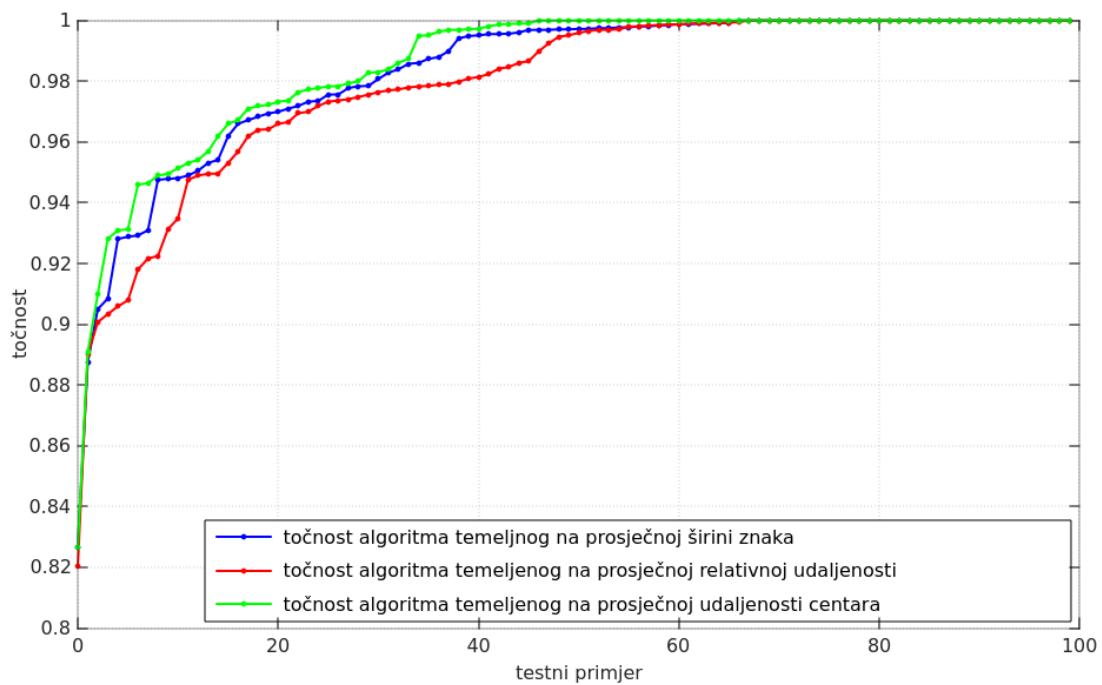
	Min.	Sred.	Med.	Maks.	Udio primjera s maks. točnosti
avgcharwidth	0,83	0,98	0,99	1	0,34
avgreldist	0,82	0,98	0,99	1	0,33
avgcenterdist	0,83	0,99	1	1	0,54

Ovakvi rezultati mogu se objasniti činjenicom da je sadržaj na računima iz trgovina tiskan fontom konstantne širine (engl. *monospaced font*), i to iskorištava algoritam *avgcenterdist*. Rezultati pokazuju kako je za rastavljanje riječi u sadržaju iz knjiga pogodnije koristiti algoritme koji se ne oslanjaju na udaljenost centara, nego na udaljenost rubova znakova. U budućem radu trebala bi se koristiti ta informacija kako bi se poboljšali rezultati.

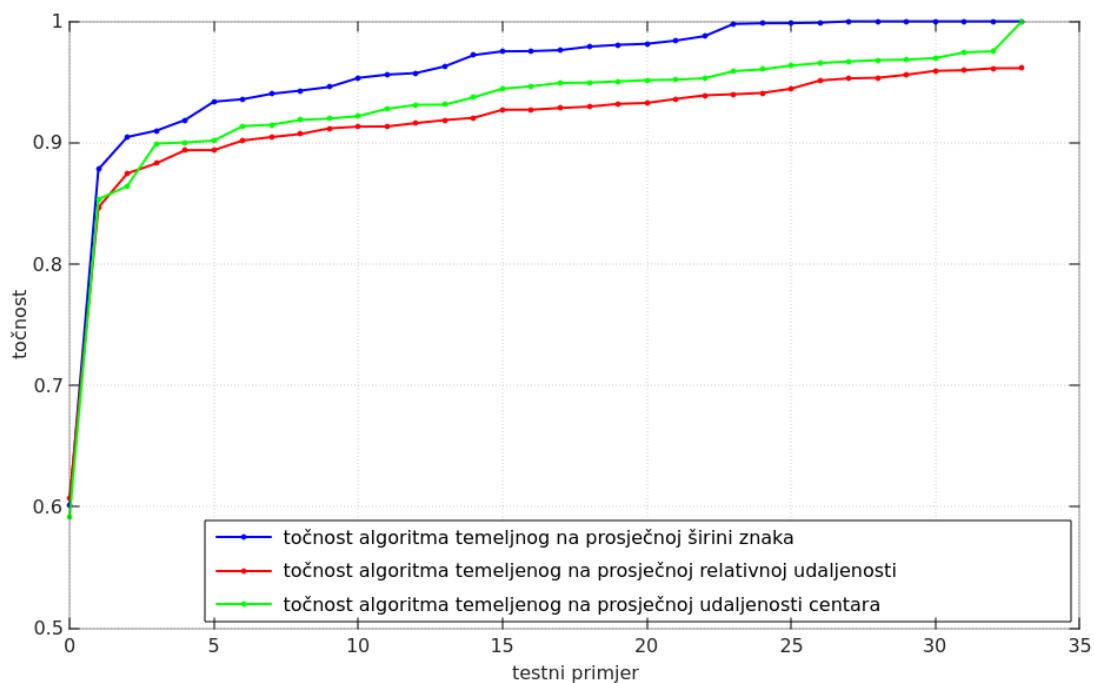
Tablica 6.3: Točnost algoritama za rastavljanje riječi u sadržaju iz knjiga.

	Min.	Sred.	Med.	Maks.	Udio primjera s maks. točnosti
avgcharwidth	0,6	0,96	0,97	1	0,2
avgreldist	0,61	0,92	0,93	0,961436	0,6
avgcenterdist	0,59	0,93	0,95	1	0,02

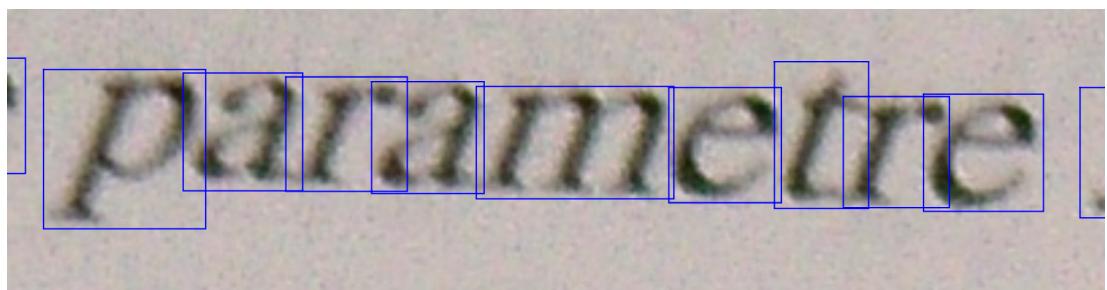
Najčešća pogreška koju sva tri algoritma rade je da ubacuju razmake između znakova koji se horizontalno preklapaju (slika 6.9). U budućem radu trebalo bi se uvesti mehanizam detekcije horizontalno preklapajućih znakova koji ne bi dozvolio njihovo razdvajanje. Algoritmi također grijše u razdvajanju uskih interpunkcijskih znakova koji su zbog svoje širine udaljeniji od svojih susjeda u odnosu na promatrani prosjek udaljenosti ostalih znakova. U budućem radu treba uzeti u obzir uske znakove koje ne treba razdvajati od susjeda ako su od njih udaljeniji od nekog promatranog prosjeka.



Slika 6.7: Graf točnosti algoritama za rastavljanje riječi u sadržaju s računa iz trgovine.



Slika 6.8: Graf točnosti algoritma za rastavljanje riječi u sadržaju iz knjiga.



Slika 6.9: Horizontalno preklapanje znakova.

7. Zaključak

Sustavi za određivanje strukture teksta sastavni su dio OCR-sustava koji se koriste za detekciju znakova na sadržaju strukturiranom u linije, riječi i blokove. Način izvedbe sustava ovisi o problemu koji se rješava i načinu integracije sa OCR-sustavom. U ovom radu predstavljeni su algoritmi za određivanje linija i razdvajanje riječi koji za određivanje strukture koriste položaj pojedinih znakova i zajedno čine sustav za određivanje strukture teksta.

Predstavljeni sustav za određivanje strukture teksta na temelju položaja pojedinih znakova namjenjen je i testiran za određivanje strukture na sadržaju s računa iz trgovina i na sadržaju iz knjiga. Rezultati i analiza pokazuju kako postoji još prostora za poboljšanje algoritama i mjera za određivanje točnosti. Algoritam za određivanje linija temelji se na pretpostavci da dva susjedna znaka, koja se nalaze u istoj liniji, ostvaruju maksimalno vertikalno preklapanje. Predstavljeni algoritmi za određivanje riječi na razne načine koriste statističke podatke o širini i udaljenosti znakova. Dodatno, svaki algoritam koristi parametre koji utječu na način određivanja linija i razdvajanja riječi. Optimalni parametri algoritama ovise o problemu koji se rješava. U ovom radu predstavljeni optimalni parametri pronađeni su ručnim eksperimentiranjem sa skupom podataka za testiranje.

U budućem radu predlaže se sastavljanje novog skupa podataka koji bi omogućio nezavisno testiranje dva podsustava u sustavu za određivanje strukture teksta i predlaže se korištenje evolucijskih algoritama za pronalaženje optimalnih parametara.

LITERATURA

Matthew Christy, Anshul Gupta, Elizabeth Grumbach, Laura Mandell, Richard Furuta, i Ricardo Gutierrez-Osuna. Mass digitization of early modern texts with optical character recognition. *J. Comput. Cult. Herit.*, 11(1):6:1–6:25, Prosinac 2017. ISSN 1556-4673. doi: 10.1145/3075645. URL <http://doi.acm.org/10.1145/3075645>.

Filip Gulan. Očitavanje rukom pisanih slova. Završni rad, Sveučilište u Zagrebu, Fakultet elektrotehnike i računarstva, Unska 3, 10000 Zagreb, Republika Hrvatska, Lipanj 2016.

Gaurav Gupta, Shobhit Niranjan, Ankit Shrivastava, i R Mahesh K Sinha. Document layout analysis and classification and its application in ocr. U *Enterprise Distributed Object Computing Conference Workshops, 2006. EDOCW'06. 10th IEEE International*, stranice 58–58. IEEE, 2006.

Abdeslam El Harraj i Naoufal Raissouni. OCR accuracy improvement on document images through a novel pre-processing approach. *CoRR*, abs/1509.03456, 2015. URL <http://arxiv.org/abs/1509.03456>.

Noman Islam, Zeeshan Islam, i Nazia Noor. A survey on optical character recognition system. *CoRR*, abs/1710.05703, 2017. URL <http://arxiv.org/abs/1710.05703>.

Ivan Jurin. Višeobjektni modeli detekcije za raspoznavanje teksta dubokim učenjem. Diplomski rad, Sveučilište u Zagrebu, Fakultet elektrotehnike i računarstva, Unska 3, 10000 Zagreb, Republika Hrvatska, Lipanj 2017.

Sukhpreet Kaur i Simpel Rani. A survey on feature extraction and classification techniques for character recognition of indian scripts. 2016.

Alex Krizhevsky, Ilya Sutskever, i Geoffrey E Hinton. Imagenet classification with

deep convolutional neural networks. In *Advances in neural information processing systems*, stranice 1097–1105, 2012.

Rayson Laroca, Evair Severo, Luiz A. Zanlorensi, Luiz S. Oliveira, Gabriel R. Gonçalves, William R. Schwartz, i David Menotti. A robust real-time automatic license plate recognition based on the YOLO detector. *CoRR*, abs/1802.09567, 2018. URL <http://arxiv.org/abs/1802.09567>.

Gurpreet S Lehal i Chandan Singh. Feature extraction and classification for ocr of gurmukhi script. *VIVEK-BOMBAY-*, 12(2):2–12, 1999.

Jisheng Liang, Jaekyu Ha, Robert M Haralick, i Ihsin T Phillips. Document layout structure extraction using bounding boxes of different entities. In *Applications of Computer Vision, 1996. WACV'96., Proceedings 3rd IEEE Workshop on*, stranice 278–283. IEEE, 1996.

Yi-Feng Pan, Xinwen Hou, i Cheng-Lin Liu. A hybrid approach to detect and localize texts in natural scene images. *IEEE Transactions on Image Processing*, 20(3):800–813, 2011.

Hamed Saghaei. Proposal for automatic license and number plate recognition system for vehicle identification. *CoRR*, abs/1610.03341, 2016. URL <http://arxiv.org/abs/1610.03341>.

Sarah Schulz i Jonas Kuhn. Multi-modular domain-tailored ocr post-correction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, stranice 2716–2726, 2017.

Shangxuan Tian, Yifeng Pan, Chang Huang, Shijian Lu, Kai Yu, i Chew Lim Tan. Text flow: A unified text detection system in natural scene images. *CoRR*, abs/1604.06877, 2016. URL <http://arxiv.org/abs/1604.06877>.

Abhishek Verma, Suket Arora, i Preeti Verma. Ocr-optical character recognition. In *7th International Conference on Recent Innovations in Science, Engineering and Management*, 2016.

Rohit Verma i Jahid Ali. A-survey of feature extraction and classification techniques in ocr systems. *International Journal of Computer Applications & Information Technology*, 1(3):1–3, 2012.

Ivo Vynckier. How ocr works, a close look at optical character recognition, 2018. URL <http://how-ocr-works.com/OCR/OCR.html>. Přistupano: 01.06.2018.

Christoph Wick, Christian Reul, i Frank Puppe. Improving OCR accuracy on early printed books using deep convolutional networks. *CoRR*, abs/1802.10033, 2018. URL <http://arxiv.org/abs/1802.10033>.

Fei Yin i Cheng-Lin Liu. Handwritten text line extraction based on minimum spanning tree clustering. U *Wavelet Analysis and Pattern Recognition, 2007. ICWAPR'07. International Conference on*, svezak 3, stranice 1123–1128. IEEE, 2007.

Xu-Cheng Yin, Xuwang Yin, i Kaizhu Huang. Robust text detection in natural scene images. *CoRR*, abs/1301.2628, 2013. URL <http://arxiv.org/abs/1301.2628>.

Weiheng Zhu, Yuanfeng Liu, i Liang Hao. A novel ocr approach based on document layout analysis and text block classification. U *Computational Intelligence and Security (CIS), 2016 12th International Conference on*, stranice 91–94. IEEE, 2016.

Sustav za određivanje strukture teksta na temelju položaja pojedinih znakova

Sažetak

U ovom radu predstavljen je sustav za određivanje strukture teksta na temelju položaja pojedinih znakova namjenjen za rješavanje problema određivanja strukture teksta u sadržaju s računa iz trgovina i sadržaju iz knjiga. Sustav je podjeljen na dva podsustava od kojih prvi određuje linije, a drugi rastavlja riječi u sadržaju. Algoritam za određivanje linija temelji se na pretpostavci da dva susjedna znaka, koja se nalaze u istoj liniji, ostvaruju maksimalno vertikalno preklapanje. Algoritmi za rastavljanje riječi na razne načine koriste statističke podatke o širini i udaljenosti znakova. Svaki algoritam koristi parametre koji utječu na način određivanja linija i razdvajanja riječi

Ključne riječi: sustav za određivanje strukture teksta, optičko raspoznavanje znakova, algoritmi za određivanje linija, algoritmi za razdvajanje riječi, segmentacija teksta, segmentacija linija

Text Layout Analysis System Based on Individual Character Positions

Abstract

This thesis describes text layout analysis system based on individual character positions intendend to solve the problem of determining text layout in receipts and book content. The system is divided into two subsystems, the first of which finds the lines in text and the other separates the words in lines. Algorithm for finding lines in text is based on the assumption that the two neighbouring characters in the same line have maximum vertical overlap. Algorithms for separating the words in lines use in various ways statistics of character width and distance of characters. Algorithms use parameters that can be tuned and affect on how the lines are found and the words are separated.

Keywords: text layout analysis system, optical character recognition, line finding algorithms, word separation algorithms, text segmentation, line segmentation