

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.

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 zadataka 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.

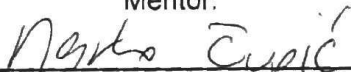
U okviru ovog završnog rada potrebno je proučiti načine za određivanje riječi i linija na temelju položaja individualnih znakova te njihovih omeđujućih pravokutnika. U okviru rada potrebno je pripremiti odgovarajući skup podataka za testiranje te napraviti prototipnu implementaciju sustava.

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.

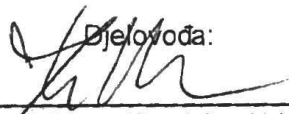
Rok za predaju rada: 15. lipnja 2018.

Mentor:



Doc. dr. sc. Marko Čupić

Djelovoda:



Doc. dr. sc. Tomislav Hrkać

Predsjednik odbora za
završni rad modula:



Prof. dr. sc. Siniša Srbljić

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 Smolkoviću i Ivanu Jurinu koji su me svojim bogatim znanjem i iskustvom usmjeravali u razvoju rada.

Tko hoće da među vama bude najveći, neka vam bude poslužitelj! I tko hoće da među vama bude prvi, neka bude svima sluga. - Mk 10,43-44

SADRŽAJ

1. Uvod	1
2. Optičko raspoznavanje znakova	2
2.1. Primjene	2
2.2. Proces izvođenja	3
3. Sustavi za određivanje strukture teksta	8
4. Zaključak	12
Literatura	13

1. Uvod

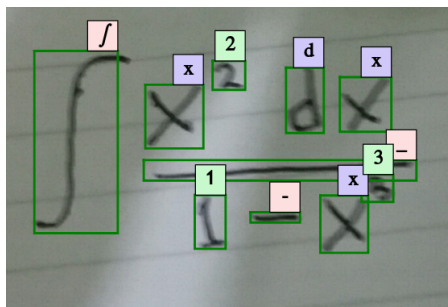
2. Optičko raspoznavanje znakova

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

2.1. Primjene

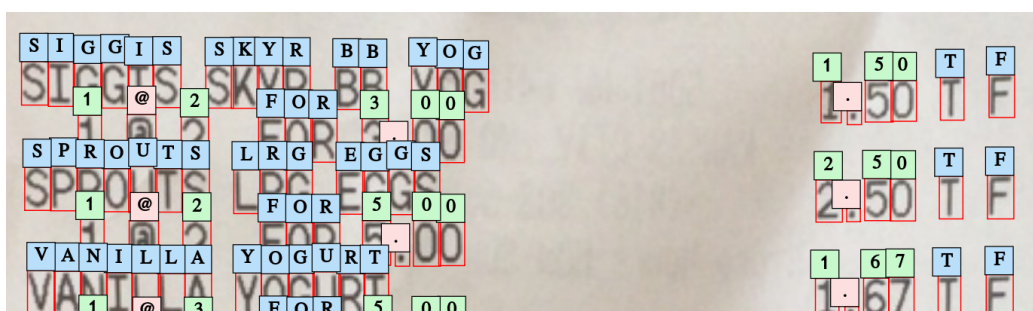
Osim tiskanog teksta, OCR sustavi koriste se i u prepoznavanju znakova rukom pisanog teksta. Prepoznavanje znakova rukom pisanog teksta je teži problem od prepoznavanja tiskanog teksta (Islam et al., 2017) zato jer se oblik znakova i njihov način pisanja razlikuje kod svake osobe (npr. rukopis odrasle 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, broj 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 kompliciranija (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 na tiskanim knjigama (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

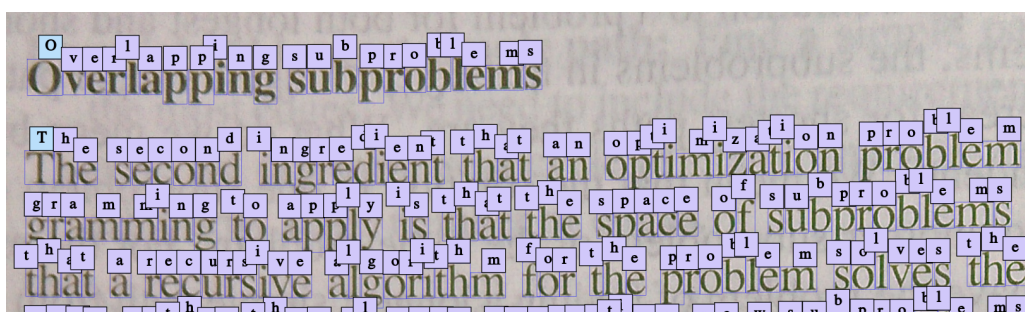


Slika 2.1: Rezultat *off-line* OCR sustava za detekciju znakova rukom pisanog teksta.

na tiskanim knjigama.



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. Proces izvođenja

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

1. pribavljanje slike,
2. pretprocesiranje,

- ## Pribavljanje slike

Walgreens

#04769 3599 N UNIVERSITY DR
SUNRISE, FL 33581
954-741-7751

239 8123 0021 02/10/2017 2:29 PM

EXTRA-DEPARTMENT GUM 5/V 155.00M
02300011248 A 2.99
RETURN VALUE 2.99

SUBTOTAL 2.99
SALES TAX A+6.0% 0.16

TOTAL 3.17
CASH 3.00
CHANGE 1.83

THANK YOU FOR SHOPPING AT WALGREENS

DID YOU KNOW THAT YOU CAN EARN POINTS
ON THOUSANDS OF ITEMS IN-STORE AND
ONLINE? SEE OUR WEEKLY AD FOR MORE
INFORMATION. ITEMS CHANGE WEEKLY.
RESTRICTIONS APPLY. FOR TERMS AND
CONDITIONS, VISIT WALGREENS.COM/BALANCE.

RFR# 0478-9218-1231-1732-1003



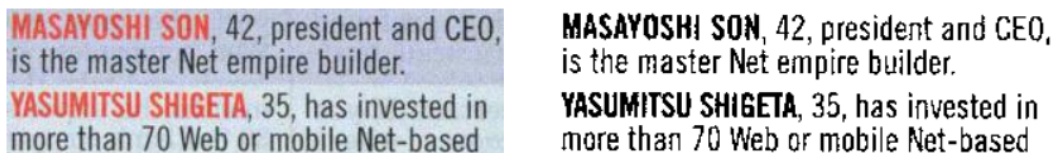

rewards

Get the flu shot that helps provide
a life-saving vaccine to a child in need.
Get a Shot. Give a Shot.® It's that easy.
Learn more at the pharmacy.

Preprocesiranje

4

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)

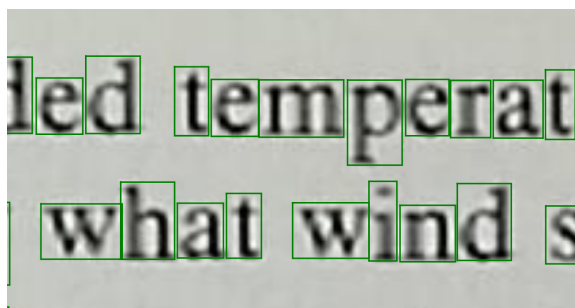


Slika 2.5: Prije binarizacije (lijevo) i nakon binarizacije (desno) (Vynckier, 2018).

Segmentacija znakova

Sljedeći korak, segmentacija znakova, je postupak segmentiranja slike u segmente unutar kojih se nalaze znakovi koje želimo klasificirati. Jedna od pristupa segmentacije izvodi se s vrha prema dnu gdje se najprije 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.



Slika 2.6: Segmentacija znakova.

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

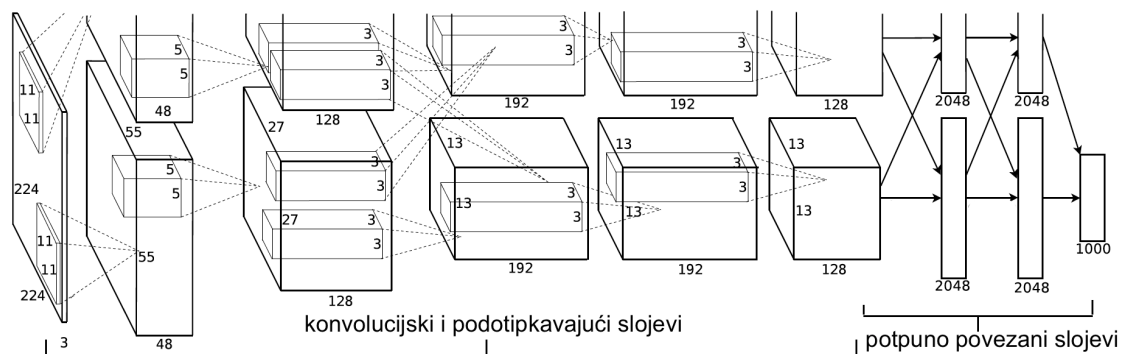
Izdvajanje značajki

Izdvajanje značajki pojedinog znaka podrazumjeva 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).

2012. godine Krizhevsky i suradnici (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)



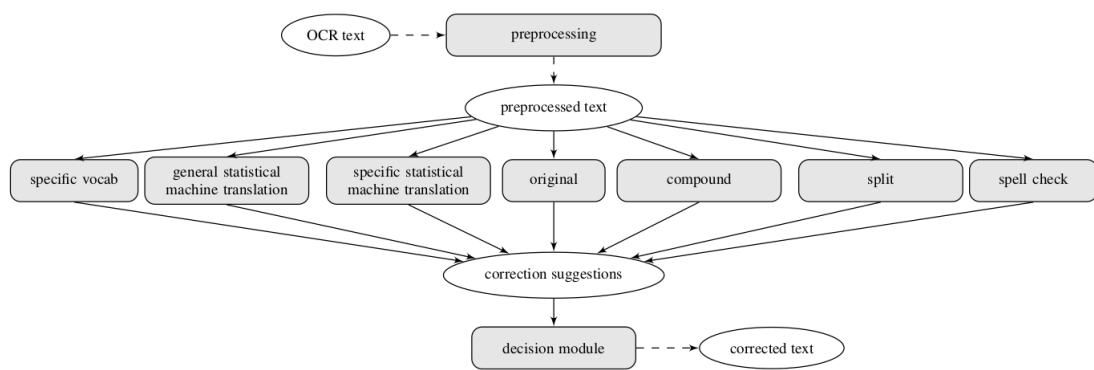
Slika 2.7: Arhitektura *AlexNet* (Jurin, 2017)

Postprocesiranje

Nakon klasifikacije znakova slijedi njihovo postprocesiranje koje se koristi kako bi se poboljšali rezultati OCR-a. 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 spomenuto u 2.2 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 suradnici (Schulz i Kuhn, 2017) 2017. godine predstavili su arhitekturu tzv. *post-correction* OCR sustava kojim su pokazati na koji način su adaptirali generički sustav za postprocesiranje OCR rezultata koristeći domensko znanje za konkretan problem koji su rješavali. Ovim pristupom ostvarili su bolje rezultate za konkretni problem nego što su ostvarili koristeći postojeći generički sustav za postprocesiranje OCR rezultata.



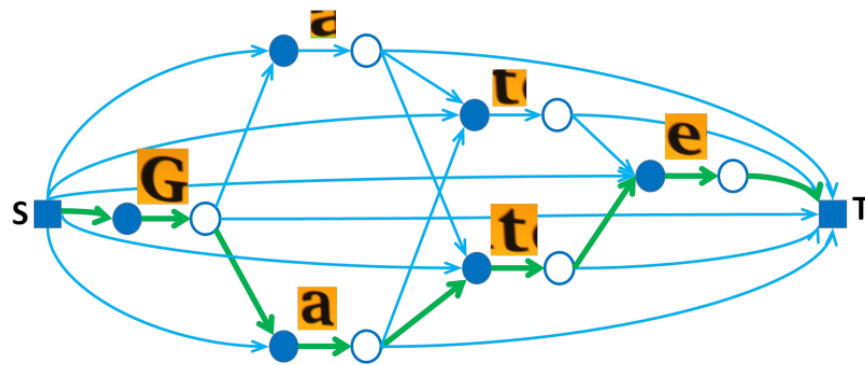
Slika 2.8: Arhitektura *post-correction* OCR sustava (Schulz i Kuhn, 2017)

3. Sustavi za određivanje strukture teksta

Sustavi za određivanje strukture teksta na temelju OCR rezultata sastavni su dio OCR sustava. Određivanje strukture teksta podrazumjeva segmentaciju linija i segmentaciju riječi unutar linije. Neke tehnike segmentacije OCR 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 na primjer matematički izrazi (Jurin, 2017). Nedostatak je što nakon klasifikacije moramo razviti sustav koji će OCR znakove dodatno procesirati da bismo dobili strukturu teksta (Jurin, 2017).

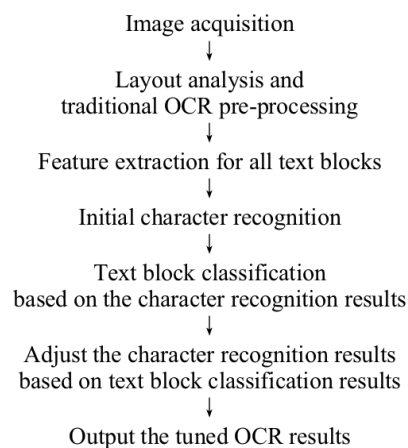
Tekst na slici može biti podijeljen na linije ili blokove, a u bloku tekst možemo podijeliti na linije. Unutar jedne linije znakove možemo grupirati riječi. Način na koji će se odrediti struktura teksta uvelike ovisi o problemu koji riješavamo i kakve rezultate želimo dobiti.

Tain i suradnici (Tian et al., 2016) predlažu 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 3.1) koji svoj problem modeliraju *min-cost flow network* modelom.



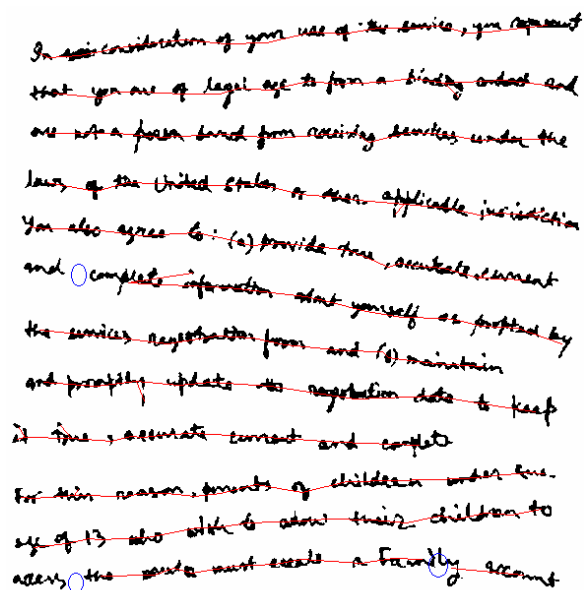
Slika 3.1: Težinski usmjereni graf temeljen na *min-cost flow network* modelu (Tian et al., 2016).

Zhu i suradnici (Zhu et al., 2016) predložili su novu arhitekturu (Slika 3.2) 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 novi 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 3.2: Arhitektura novog OCR sustava kojeg predlažu Zhu i suradnici (Zhu et al., 2016).

Yin i suradnici (Yin i Liu, 2007) pronalaze linije u tekstu povezujući znakove u težinski graf nad kojim provode Kruskalov algoritam za pronalazak minimalnog razapinjućeg 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 3.3 prikazuje rezultat pronalaska linija teksta u rukom pisanom dokumentu na Engleskom jeziku.

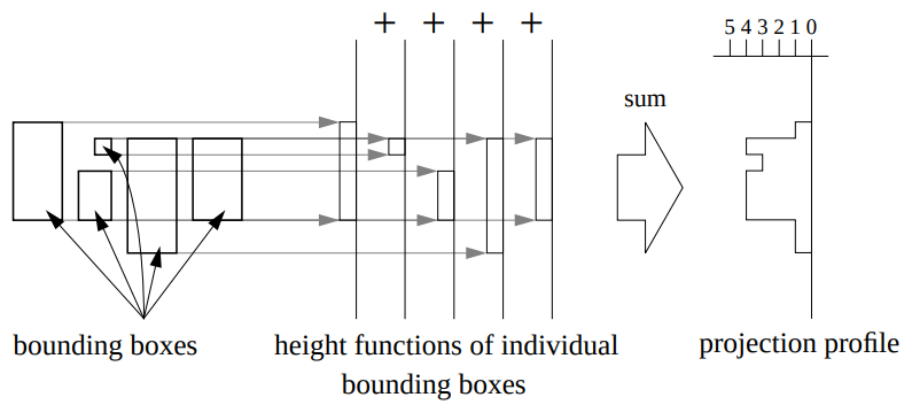


Slika 3.3: Rezultat pronalaska linija teksta u rukom pisanom dokumentu na Engleskom jeziku (Yin i Liu, 2007)

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

Još jedan pristup predložili su Yin i suradnici (Yin et al., 2013) koji koriste tehniku hijerarhiskog 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 3.4) omeđujućih pravokutnika na jednu ravninu i pronalazi vrhove i doline u histogramu koji prikazuje frekvencije pojavljivanja projektiranih pravokutnika. 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 i suradnici (Gupta et al., 2006) također koriste razne heuristike prema kojima povezuju susjedne omeđujuće pravokutnike.



Slika 3.4: Histogram dobiven horizontalnom projekcijom omeđujućih pravokutnika (Liang et al., 1996)

4. Zaključak

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 Niranjana, 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. U *Advances in neural information processing systems*, stranice 1097–1105, 2012.
- Rayson Laroca, Evair Severo, Luiz A. Zanolensi, 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 entitles. U *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. U *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. U *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>. Pristupano: 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

Ključne riječi:

Text Layout Analysis System Based on Individual Character Positions

Abstract

Keywords: