

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.

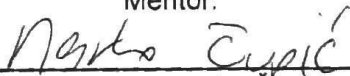
U okviru ovog završnog rada potrebno je proučiti načine za određivanje riječi i linija na temelju položaja individualnih znakova te njihovih omeđujućih pravokutnika. U okviru rada potrebno je pripremiti odgovarajući skup podataka za testiranje te napraviti prototipnu implementaciju sustava.

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 Srblić

*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 koji su me svojim bogatim znanjem i iskustvom usmjeravali u razvoju rada, a to su: Jurica Cerovec, Nenad Mikša, Boris Trubić, Igor Smolković i Ivan Jurin.*

*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

<b>1. Uvod</b>	<b>1</b>
<b>2. Optičko raspoznavanje znakova</b>	<b>2</b>
2.1. Primjene . . . . .	2
2.2. Proces izvođenja . . . . .	3
<b>3. Zaključak</b>	<b>8</b>
<b>Literatura</b>	<b>9</b>

# **1. Uvod**

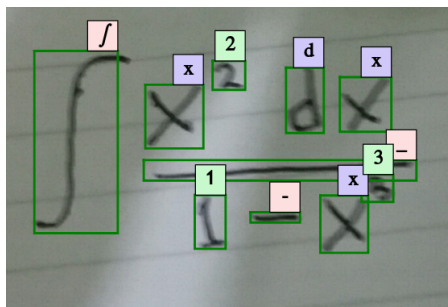
## 2. Optičko raspoznavanje znakova

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)

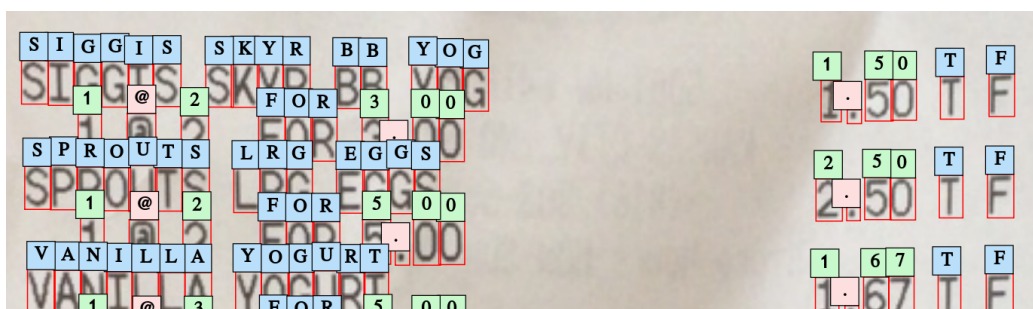
### 2.1. Primjene

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 *off-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). 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.

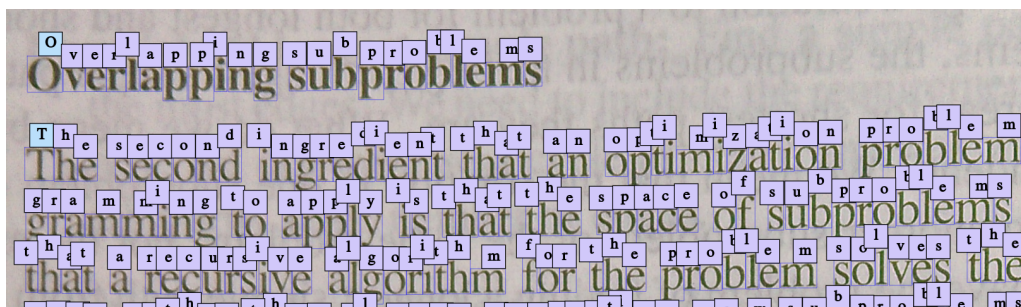


**Slika 2.2:** Rezultat OCR sustava za detekciju znakova na računima iz trgovine.

## 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,
3. segmentacija znakova,
4. izdvajanje značajki znakova,
5. klasifikacija znakova i



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



6. postprocesiranje.

## 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. Primjetimo da slika može sadržiti pozadinu koju bi OCR sustav trebao zanemariti.



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

## Preprocesiranje

U koraku pretprocesiranja slike provodimo 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 pretprocesiranja 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)

MASAYOSHI SON, 42, president and CEO,  
is the master Net empire builder.  
YASUMITSU SHIGETA, 35, has invested in  
more than 70 Web or mobile Net-based

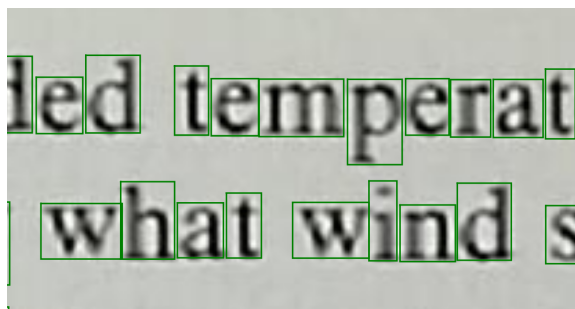
MASAYOSHI SON, 42, president and CEO,  
is the master Net empire builder.  
YASUMITSU SHIGETA, 35, has invested in  
more than 70 Web or mobile Net-based

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*<sup>1</sup> 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.

## Izdvajanje značajki

Izdvajanje značajki pojedinog znaka podrazumjeva odabir značajki prema kojima će se jedinstveno klasificirati svaki znak. Značajke poput geometrijskog oblika ili statis-

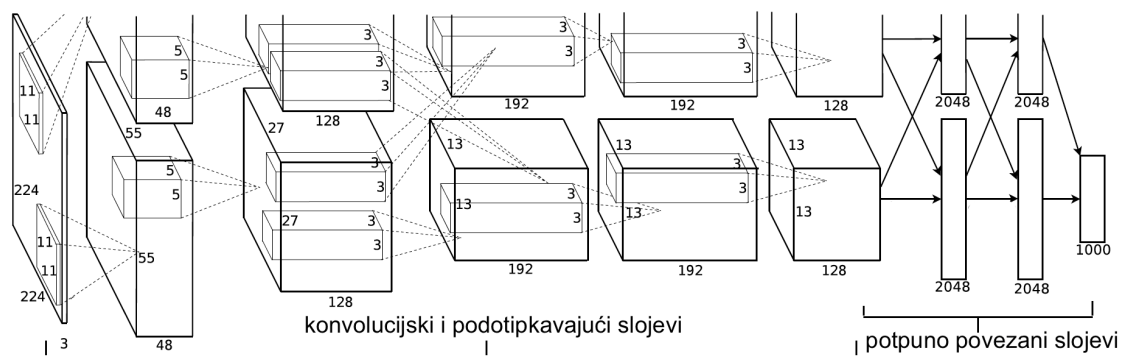
<sup>1</sup>OCR sustav tvrtke *Microblink*, <https://microblink.com>

tičkih 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) 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)

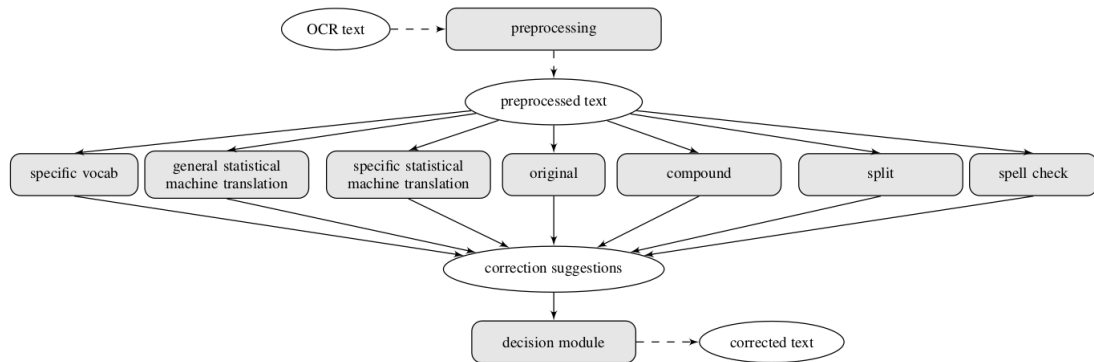


Slika 2.7: Arhitektura *AlexNet* (Jurin, 2017)

## 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. 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.
- 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.
- 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.
- 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>.

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

### **Sažetak**

#### **Ključne riječi:**

## **Text Layout Analysis System Based on Individual Character Positions**

### **Abstract**

#### **Keywords:**