

SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

ZAVRŠNI RAD br. 5187

Ekstrakcija tablica na skeniranim dokumentima

Kristijan Vulinović

Zagreb, svibanj 2017.

Umjesto ove stranice umetnite izvornik Vašeg rada.
Kako biste uklonili ovu stranicu, obrišite naredbu \izvornik.

Zahvaljujem svima onima koji su odvojili dio svojeg vremena na ispunjavanje primjeraka tablica, kao i svima koji su pomogli prilikom prikupljanja istih.

SADRŽAJ

1. Uvod	1
2. Binarizacija slike	2
2.1. Binarizacija fiksnim pragom	2
2.2. Adaptivna binarizacija	2
2.2.1. Filtriranje šuma	3
2.2.2. Procjena sadržaja	3
2.2.3. Procjena pozadine	4
2.2.4. Binarizacija	5
2.2.5. Dodatna obrada slike	5
3. Prepoznavanje zakrivljenosti slike	6
4. Detekcija tablice	7
4.1. Detekcija vrhova ćelija	7
4.2. Rekonstrukcija tablice	7
5. Primjena na automatskom prepoznavanju rukom pisanih simbola	8
6. Zaključak	9
Literatura	10

1. Uvod

U današnje vrijeme postoje izuzetno velike količine papirnatih dokumenata. Samo u Sjedinjenim Američkim Državama nastaje više od milijarde novih papirnatih dokumenata svakog radnog dana. Mogućnost digitalizacije takvih dokumenata može biti od velike koristi prilikom pohrane, slanja ili pretraživanja istih. [6] Digitalizaciju dokumenata možemo podijeliti u dva dijela: prepoznavanje teksta te prepoznavanje grafičkih objekata. [1] Za prepoznavanje teksta dostupan je velik broj alata koji omogućuju optičko prepoznavanje znakova (engl. *optical character recognition*). Prepoznavanje grafičkih objekata dokumenta mnogo je manje zastupljeno u odnosu na prepoznavanje teksta te je postalo popularnije tek u novije vrijeme. U to spada prepoznavanje linija, oblika, slika, simbola, tablica i raznih drugih objekata koji se mogu nalaziti na skeniranim dokumentima. Najveći razvoj ovoga područja nastupio je zahvaljujući razvoju dubokih neuronskih mreža i sklopovlja koje omogućuje velike brzine izračuna koje prije nisu bile moguće.

Ovaj rad se fokusira isključivo na prepoznavanje tablica, što je prethodno već opisano u radovima poput [1] i [3]. Taj postupak se dijeli na prepoznavanje položaja tablice u odnosu na ostatak dokumenta, prilikom čega je potrebno u dokumentu izdvojiti tablicu od ostatka teksta i ostalih grafičkih objekata, a što je opisano u radu [4]. Nakon što je tablica pronađena određuje se njezin izgled, odnosno broj redaka i stupaca, odnosno koordinate pojedine ćelije, a što je detaljnije opisano u nastavku rada.

Predstavljeno rješenje počinje od slike u sivim tonovima (engl. *gray-scale*), koja se binarizira kako bi se dobila slika koja se sastoji od isključivo crne i bijele boje. Dobivena crno-bijela slika koristi se u daljnjoj obradi te se provjerava je li slika rotirana, odnosno kut rotacije iste, nakon čega se slika po potrebi rotira kako bi tablica stajala okomito. Ovako obrađena slika koristi se dalje za detekciju tablica, postupkom koji se temelji na prepoznavanju kuteva ćelija, te kasnijoj rekonstrukciji istih a koji je detaljnije opisan u nastavku rada.

2. Binarizacija slike

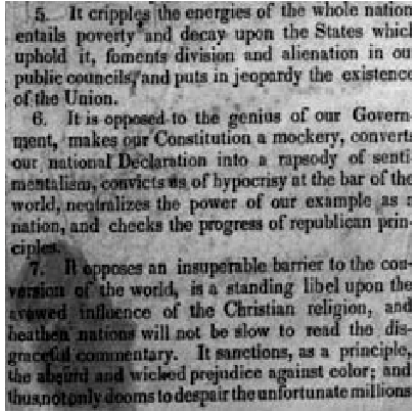
Početna slika dana je kao crno-bijela slika koja sadrži 256 nijansi sive boje, gdje je crna označena sa vrijednošću 0, a bijela sa 255. Prije nego li se započne bilo kakva analiza slike, potrebno je istu binarizirati, odnosno pretvoriti u oblik koji će sadržavati isključivo crne ili bijele elemente, bez ostalih nijansi sive. To je moguće učiniti na dva načina: korištenjem fiksno definiranog praga nakon kojega ćemo svaku vrijednost proglasiti crnom, ili korištenjem adaptivne binarizacije koja se temelji na usporedbi trenutnog intenziteta sive sa intenzitetom sive u okruženju.

2.1. Binarizacija fiksnim pragom

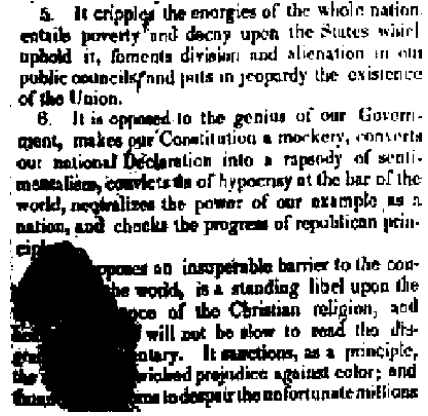
Najjednostavniji oblik binarizacije je korištenje fiksno definiranog praga. U tom se slučaju gleda svaki pojedini slikovnog elementa te ukoliko je njegova vrijednost manja od zadanog praga, element se postavlja na crnu boju, dok se u protivnom postavlja na bijelu. Prednosti ove metode su izrazito jednostavna implementacija, ali i velika brzina izvođenja. Nedostaci se primjećuju u slučajevima lošeg ili nejednoličnog osvjetljenja gdje se događa da se neki dijelovi dokumenta u potpunosti prepoznaju kao crni, unatoč činjenici da je prethodno bilo moguće razlikovati i prepoznati pozadinu od sadržaja dokumenta. Slika 2.1 prikazuje opisani problem te se na istoj može primjetiti kako je u donjem lijevom kutu slike tamno područje, koje nakon binarizacije postaje u potpunosti crno. Također se primjećuje i kako je gornji desni kut slike slabije osvjetljen, zbog čega u binariziranoj slici slova postaju tanja i slabije vidljiva.

2.2. Adaptivna binarizacija

Problemi prikazani u prethodnom postupku rješavaju se primjenom adaptivne binarizacije koja vrijednost svakog pojedinog slikovnog elementa ne određuje samo na osnovu njegove boje, već u obzir uzima i boju okoline. U nastavku je opisan postupak koji je predložen u [2]. Prikazani primjeri koriste sliku 2.1a kao početnu.



(a) Početna crno-bijela slika



(b) Slika dobivena binarizacijom fiksnim pragom

Slika 2.1: Primjer binarizacije fiksnim pragom

2.2.1. Filtriranje šuma

Ovisno o stanju dokumenta i načinu digitalizacije istoga moguće je da se na dobivenoj slici pojavljuje šum, kojega je potrebno otkloniti. Za potrebe opisanoga koristi se niskopropusni Wiener filter [5], koji se temelji na statističkoj procjeni temeljenoj na okruženju svakog pojedinog slikovnog elementa. [2] Označimo sa $I_s(x, y)$ vrijednost slikovnog elementa početne slike, a sa $I(x, y)$ vrijednost slikovnog elementa filtrirane slike. Tada se filtrirana slika I može izračunati pomoću formule opisane u knjizi [5]:

$$I(x, y) = \mu(x, y) + \frac{\sigma(x, y)^2}{(\sigma(x, y)^2 + v^2)}(I_s(x, y) - \mu(x, y))$$

Sa $\mu(x, y)$ označena je aritmetička sredina vrijednosti slikovnih elemenata u okruženju veličine $N \times M$, prema formuli:

$$\mu(x, y) = \frac{1}{NM} \sum_{i=x-\frac{N}{2}}^{x+\frac{N}{2}} \sum_{j=y-\frac{M}{2}}^{y+\frac{M}{2}} I_s(i, j)$$

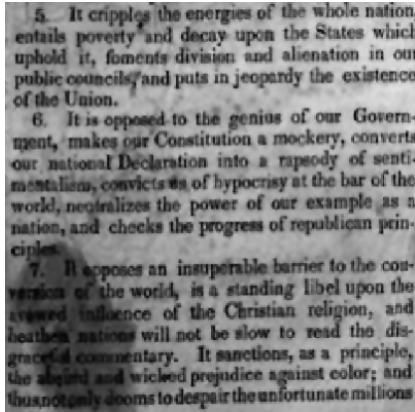
Sa σ^2 označena je varijanca vrijednosti slikovnih elemenata u okruženju veličine $N \times M$, prema formuli:

$$\sigma(x, y)^2 = \frac{1}{NM} \sum_{i=x-\frac{N}{2}}^{x+\frac{N}{2}} \sum_{j=y-\frac{M}{2}}^{y+\frac{M}{2}} (I_s(i, j)^2 - \mu^2)$$

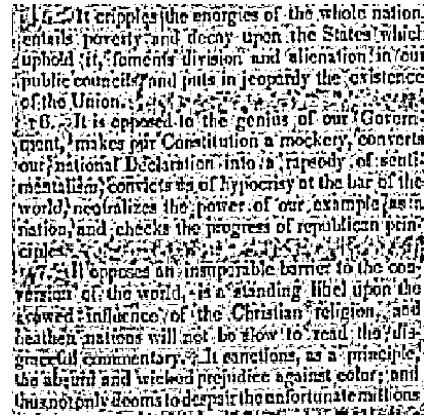
Sa v^2 je označena srednja vrijednost svih lokalnih varijanci. Konačan rezultat filtriranja, korištenjem okruženja dimenzija 5×5 prikazan je na slici 2.2.

2.2.2. Procjena sadržaja

Sljedeći korak binarizacije temelji se na procjenjivanju sadržaja dokumenta. Cilj ovog koraka je procijeniti koji elementi slike pripadaju pozadini, a koji pripadaju sadržaju dokumenta.



Slika 2.2: Slika dobivena filtriranjem šuma



Slika 2.3: Slika dobivena korištenjem Niblackovog algoritma adaptivne binarizacije

Pritom je procijenjeni sadržaj zapravo nadskup stvarnog sadržaja, odnosno na dobivenoj slici biti će prisutan šum. Za potrebe ovoga koristi se Niblackov algoritam adaptivne binarizacije. [2]

Algoritam se temelji na ideji kliznog prozora određenih dimenzija, pomoću kojega se računa lokalni median vrijednosti m te varijanca s . Kako bi se ubrzao izračun, umjesto mediana se računa aritmetička sredina μ . Konačan prag binarizacije, T , određuje se kao:

$$T = m + ks$$

gdje je k proizvoljna konstanta koja određuje koliko će okolina trenutnog slikovnog elementa utjecati na prag binarizacije. Korištena vrijednost je $k = -0.2$. Konačna slika N , dobivena je od početne slike I na sljedeći način:

$$N(x, y) = \begin{cases} 1, & \text{ako je } I(x, y) > T \\ 0, & \text{inače} \end{cases}$$

Rezultati ovog postupka, korištenjem kliznog prozora dimenzija 20x20, uz $k = -0.2$, prikazan je na slici 2.3. Na slici se primjećuje kako je sav tekst prepoznat i prikazan crnom bojom, ali je također prisutan i jako izražen šum.

2.2.3. Procjena pozadine

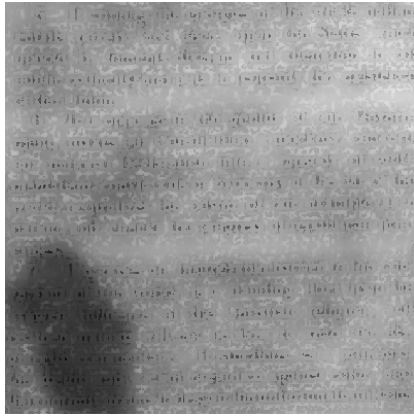
U ovom koraku pokušava se procijeniti izgled pozadine dokumenta, što je označeno sa B . Za potrebe toga koristi se obrađena početna slika I te prethodno dobivena slika N . Ako je neki slikovni element na slici N označen nulom, taj element predstavlja pozadinu slike i njegova vrijednost na slici B biti će jednaka onoj sa slike I . U protivnome će vrijednost tog

slikovnog elementa biti određena interpolacijom vrijednosti susjednih slikovnih elemenata.

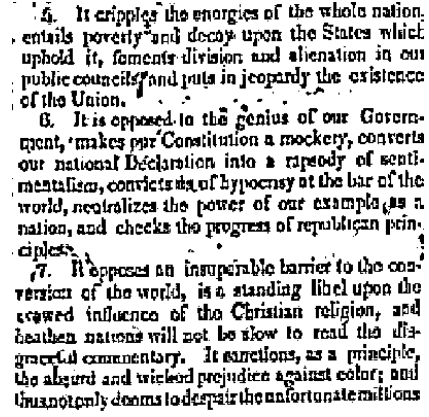
Konačna formula za elemente slike B glasi:

$$B(x, y) = \begin{cases} I(x, y), & \text{ako je } N(x, y) = 0 \\ \frac{\sum_{i=x-dx}^{x+dx} \sum_{j=y-dy}^{y+dy} (I(i, j)(1 - N(i, j)))}{\sum_{i=x-dx}^{x+dx} \sum_{j=y-dy}^{y+dy} (1 - N(i, j))}, & \text{ako je } N(x, y) = 1 \end{cases}$$

Pri čemu dx i dy određuju dimenzije okruženja koje se gleda prilikom interpolacije. Postupak je u cijelosti objašnjen u radu [2].



Slika 2.4: Slika dobivena procjenom pozadine



Slika 2.5: Slika dobivena binarizacijom

Rezultati ovog postupka, primjenjenog na slici 2.1a prikazani su na slici 2.4, uz $dx = 3$ i $dy = 3$. Moguće je uočiti kako se na slici i dalje primjećuju obrisi slova, premda sama slova nisu prisutna.

2.2.4. Binarizacija

2.2.5. Dodatna obrada slike

3. Prepoznavanje zakrivljenosti slike

4. Detekcija tablice

4.1. Detekcija vrhova ćelija

4.2. Rekonstrukcija tablice

5. Primjena na automatskom prepoznavanju rukom pisanih simbola

6. Zaključak

Zaključak.

LITERATURA

- [1] S. Deivalakshmi, K. Chaitanya, i P. Palanisamy. Detection of table structure and content extraction from scanned documents. U *Communications and Signal Processing (ICCSP)*, stranice 270–274. IEEE, apr 2014.
- [2] Basilios Gatos, Ioannis Pratikakis, i Stavros J. Perantonis. An adaptive binarization technique for low quality historical documents. *Lecture Notes in Computer Science*, (3163):102–113, sep 2004.
- [3] Basilios Gatos, Dimitrios Danatsas, Ioannis Pratikakis, i Stavros J. Perantonis. Automatic table detection in document images. U *International Journal of Document Analysis*, svezak 8, stranica 172–182, aug 2005.
- [4] Jianying Hu, Ram Kashi, Daniel Lopresti, i Gordon Wilfong. Medium-independent table detection. *Proceedings of SPIE - The International Society for Optical Engineering*, dec 1999.
- [5] Lim J.S. *Two-Dimensional Signal and Image Processing*. PH, 1989. ISBN 0139353224,9780139353222.
- [6] Peng-Yeng Yin. Skew detection and block classification of printed documents. *Image and Vision Computing*, 19(8):567–579, may 2001.

Ekstrakcija tablica na skeniranim dokumentima

Sažetak

Sažetak na hrvatskom jeziku.

Ključne riječi: Ključne riječi, odvojene zarezima.

Table Extraction on Scanned Documents

Abstract

Abstract.

Keywords: Keywords.