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

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

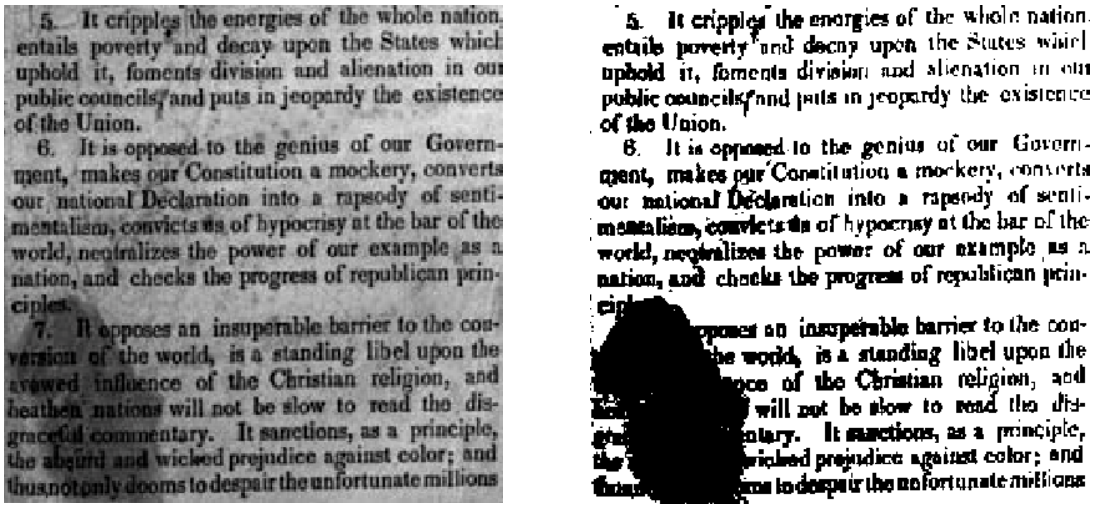
Početna slika dana je kao crno-bijela slika koja sadrži 256 nijansi sive boje, gdje je crna označena sa vrijednošću 0, a bijela sa 255. Prije nego li se započne bilo kakva analiza slike, potrebno je istu binarizirati, odnosno pretvoriti u oblik koji će sadržavati isključivo crne ili bijele elemente, bez ostalih nijansi sive. To je moguće učiniti na dva načina: korištenjem fiksno definiranog praga nakon kojega ćemo svako 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. Nedostatci 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 doljnjem 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 sakog 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]:

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

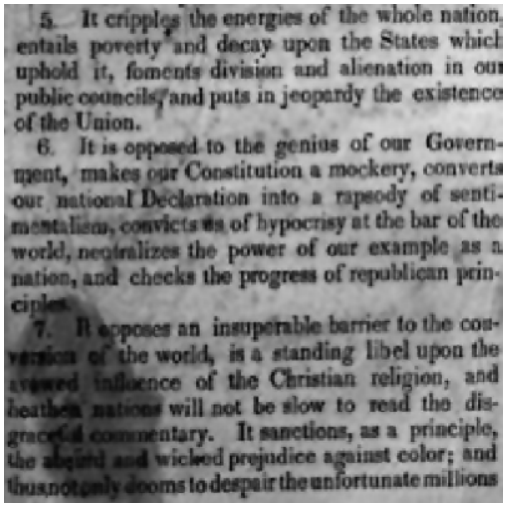
Sa $\mu(x, y)$ označena je aritmetička sredina vrijednosti slikovnih elemenata u okruženju veličine NxM, prema formuli:

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

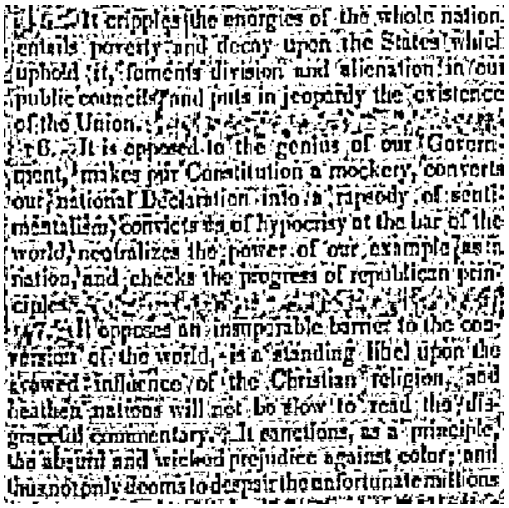
Sa σ^2 označena je varijanca vrijednosti slikovnih elemenata u okruženju veličine NxM, prema formuli:

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

Sa v^2 je označena srednja vrijednost svih lokalnih varijanci. Konačan rezultat filtriranja, korištenjem okruženja dimenzija 5x5 prikazan je na slici 2.2.



Slika 2.2: Slika dobivena filtriranjem šuma



Slika 2.3: Slika dobivena korištenjem Niblackovog algoritma adaptivne binarizacije

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 saržaju dokumenta. 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:

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

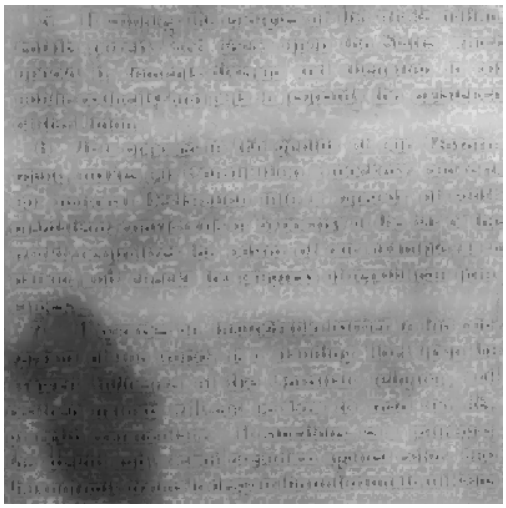
Rezultati ovog postupka, korištenjem kliznog prozora dimenzija 20x20, uz $k = -0.2$, prikazan je na slici 2.3. Na slici se primjećuje kako je sav tekst prepoznat i prikazan crnom bojom, ali je također prisutan i jako izražen šum.

2.2.3. Procjena pozadine

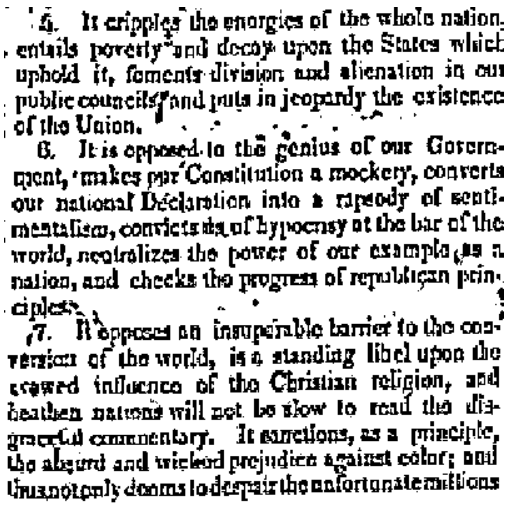
U ovom koraku pokušava se procijeniti izgled pozadine dokumenta, što je označeno sa B. Za potrebe toga koristi se obrađena početna slika I te prethodno dobivena slika N. Ako je neki slikovni element na slici N označen nulom, taj element predstavlja pozadinu slike i njegova vrijednost na slici B biti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

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

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



5. It cripples the energies of the whole nation, entails poverty and decay upon the States which uphold it, foments division and alienation in our public councils, and puts in jeopardy the existence of the Union.

6. It is opposed to the genius of our Government, makes our Constitution a mockery, converts our national Declaration into a rapsody of sentimentalism, convicts us of hypocrisy at the bar of the world, neutralizes the power of our example as a nation, and checks the progress of republican principles.

7. It opposes an insuperable barrier to the conversion of the world, is a standing libel upon the avowed influence of the Christian religion, and heathen nations will not be slow to read the disgraceful commentary. It sanctions, as a principle, the absurd and wicked prejudice against color; and thus not only dooms to despair the unfortunate millions

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

Kao što je opisano u radu [2], u ovom koraku se provodi konačna binarizacija, uspoređivanjem izračunate pozadine B i izvorne slike I. Postupak se temelji na tome da je razlika u boji

teksta i pozadine veća nego li razlika u boji šuma dobivenog procjenom sadržaja dokumenta i pozadine. Ovisno o intenzitetu sive boje okruženja, računa se prag binarizacije d, kako bi se sačuvao tekst u tamnim područjima. Kako bi se to postiglo, vrijednost praga d mora biti manja u područjima sa tamnom pozadinom. Konačna slika T određena je formulom:

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

Vrijednost praga, $d(B(x,y))$, moguće je procijeniti kao

$$d = q * \delta$$

, gdje je q konstanta koja je fiksno postavljena na 0.8, a δ se računa kao razlika intenziteta sive boje na početnoj slici za mjesta koja predstavljaju pozadinu i za mjesta koja predstavljaju tekst. Vrijednost δ možemo izračunati kao

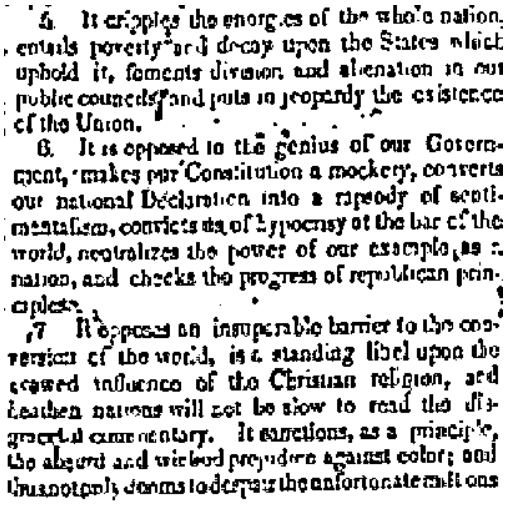
$$\delta = \frac{\sum_x \sum_y (B(x,y) - I(x,y))}{\sum_x \sum_y N(x,y)}$$

Konačan rezultat nakon primjene opisanog postupka binarizacije prikazan je na slici 2.5. Za razliku od rezultata korištenjem binarizacije sa fiksnim pragom, prikazanim na slici 2.1b, primjećuje se kako zatamnjeno područje u donjem lijevom kutu slike ne predstavlja problem prilikom raspoznavanja teksta od pozadine. Također je potrebno primjetiti i šumove koji su ostali prisutni nakon trenutno opisanog postupka, a koji su riješeni u nastavku.

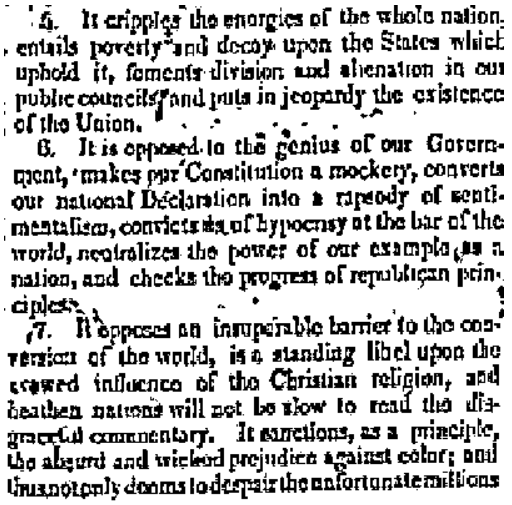
2.2.5. Dodatna obrada slike

Završni korak obrade slike koristi se kako bi se popravila kvaliteta konačne binarizirane slike. Problemi koji se mogu uočiti na danim primjerima uključuju prisutnost šuma, kao i potencijalne prekide slova, gdje je moguće da neko slovo nije u potpunosti zacrnjeno.

Prisutnost šuma se nastoji ukloniti tako što se pregledava cijela slika te se za svaki crni slikovni element provjerava je li on rezultat šuma. Za potrebe toga koristi se klizni prozor dimenzija nxn. U slučaju kada je središnji element kliznog prozora crne boje, prebrojavaju se svi pozadinski (bijeli) slikovni elementi unutar kliznog prozora, što je označeno sa P_{sh}. Ako je $P_{sh} > k_{sh}$, gdje je k_{sh} proizvoljno zadana konstanta, središnji slikovni element se postavlja u bijelu boju. Na slici 2.6 prikazan je rezultat obrade binarizirane slike navedenim postupkom uz $n = 5$ i $k_{sh} = 0.8$. Primjećuje se kako i dalje postoje smetnje na slici, no problem nastaje kod toga što bi niža vrijednost parametra k_{sh} osigurala bolje uklanjanje šumova,



Slika 2.6: Slika dobivena uklanjanjem crnih slikovnih elemenata



Slika 2.7: Slika dobivena uklanjanjem bijelih slikovnih elemenata

ali bi istovremeno povećala vjerojatnost da se uklone elementi teksta, što nije poželjno.

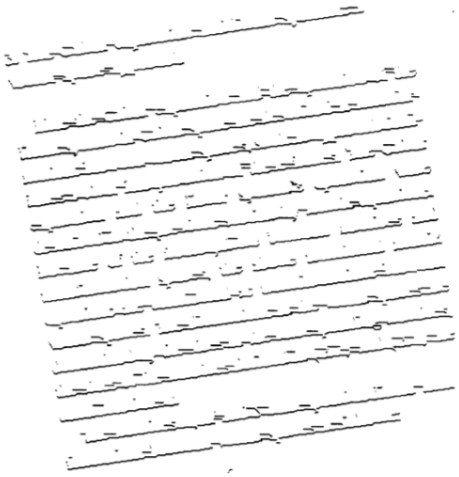
Drugi korak spomenut u ovom postupku provodi se slično kao i prethodni. Potrebno je prebrojati broj crnih slikovnih elemenata, P_{sw}, koji se nalaze unutar kliznog prozora dimenzija nxn. Ako je srednji element bijel i vrijedi $P_{sw} > k_{sw}$, središnji se element pretvara u crni. Postupak je demonstriran na slici 2.7, uz $n = 5$ i $k_{sw} = 0.6$.

3. Prepoznavanje zakrivljenosti slike

Jedan od problema koji nastaje prilikom digitalizacije dokumenata je mogućnost da dokument bude nakošen. Do toga može doći ili prilikom nepažnje tokom ubacivanja papira u optički čitač, ili zbog nesvjesnog zakrivljenja prilikom korištenja kamere mobilnog uređaja. Nakošenost digitalizirane slike može uzrokovati teže ili lošije prepoznavanje elemenata na slici, zbog čega je korisno i poželjno otkriti stupanj rotacije dokumenta i ispraviti ga.

3.1. Računanje kuta rotacije

Počevši od dokumenta kakav je prikazan na slici 3.1, može se uočiti kako se na njemu jasno prepoznaju vodoravne linije dokumenta. Unatoč činjenici da to nije opći slučaj koji uvijek vrijedi, u ovom radu nije razrađeno kako općeniti dokument svesti na ovakav oblik, već se kreće od pretpostavke da će dokument uvijek imati jasno izražene linije, bilo vodoravne, bilo okomite. Ova pretpostavka vrijedi isključivo iz razloga da se rad fokusira na prepoznavanje tablica na skeniranim dokumentima.



Slika 3.1: Početni nakošeni dokument[6]

Temeljna ideja koja se koristi prilikom izračuna kuta rotacije je ta da se dokument presječe okomitim linijama, što je predloženo u radu [6]. Za svaku okomitu liniju se računaju

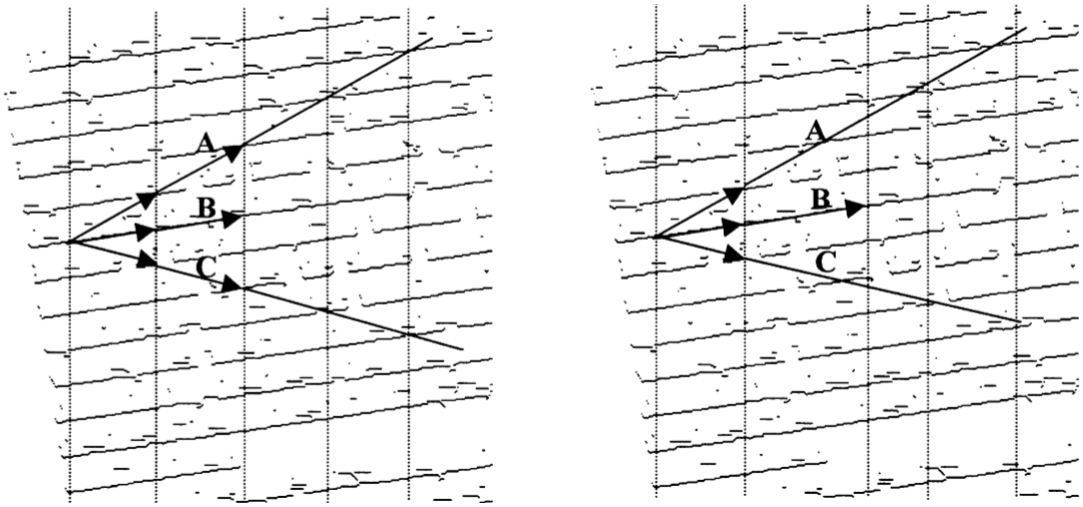
točke u kojima ona sječe linije osnovnog dokumenta. Jednom kada su određene točke sjecišta, moguće je odabrati jednu takvu točku najljevije okomite linije (T_1) te usporediti tu točku sa svim točkama koje se nalaze desno od nje (T_2). Dvije tako odabrane točke analiziraju se na način da se kroz njih povuče pravac, čija jednadžba glasi:

$$y - y_1 = k(x - x_1),$$

gdje je k koeficjent smjera:

$$k = \frac{y_2 - y_1}{x_2 - x_1}.$$

Pomoću koeficjenta smjera k moguće je izračunati kut pod kojim je nagnut dobiveni pravac, kao $\alpha = tan^{-1}(k)$. Uz pretpostavku da dokument nije rotiran te da se točke T_1 i T_2 nalaze na istim vodoravnim linijama dokumenta, kut α trebao bi iznositi $0°$. U slučaju da je dokument rotiran, ovako određen kut α biti će jednak stupnju rotacije dokumenta.



(a) Dokument podijeljen ekvidistantnim linijama[6]

(b) Dokument podijeljen linijama sa nasumičnim udaljenostima[6]

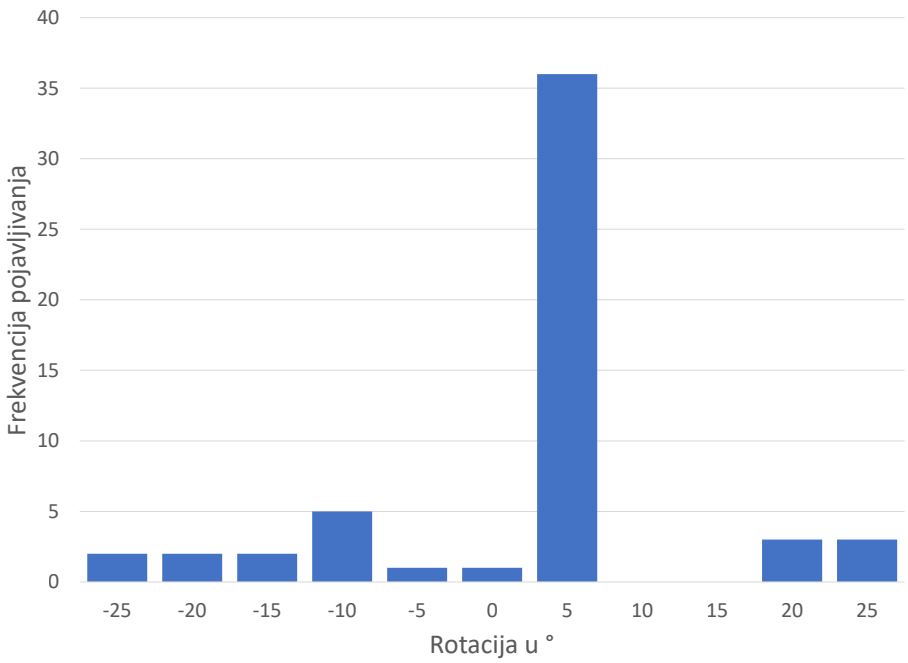
Slika 3.2: Primjer podjele dokumenta okomitim linijama

Jedan od problema koji preostaje je početna pretpostavka da se točke T_1 i T_2 nalaze na istoj vodoravnoj liniji dokumenta, što neće biti istinito u većini slučajeva. To se rješava na način da se izračunaju kutevi koje zatvaraju svi pravci koji počinju u točki T_1, a pritom se bilježi koliko puta se pojavio koji kut. Uz ovaj pristup rotaciju dokumenta više ne određuje kut α koji je dobiven za jedan pravac, već onaj kut koji je najviše puta zabilježen. Ovaj pristup prikazan je na slici 3.2, gdje je za određenu podjelu okomitim linijama i odabranu početnu točku prikazan izračun triju kuteva: **A**, **B**, **C**. Slika 3.2a također prikazuje i važnost načina odabira okomitih linija. Na navedenoj slici odabrane su ekvidistantne linije, zbog čega dolazi do toga da se sva tri kuta javljaju jednak broj puta, što ometa rad algoritma.

Za razliku od toga, na slici 3.2b koriste se linije sa nasumičnim međusobnim udaljenostima te se primjećuje kako se na njoj kut **B** javlja prilikom svakog presjeka okomite linije sa vodoravnom linijom na kojoj se nalazi početna točka T_1, dok se ostali kutevi javljaju samo jednom, čime se prepoznaje da oni dolaze radi krivo odaranih točaka.

3.2. Rezultati

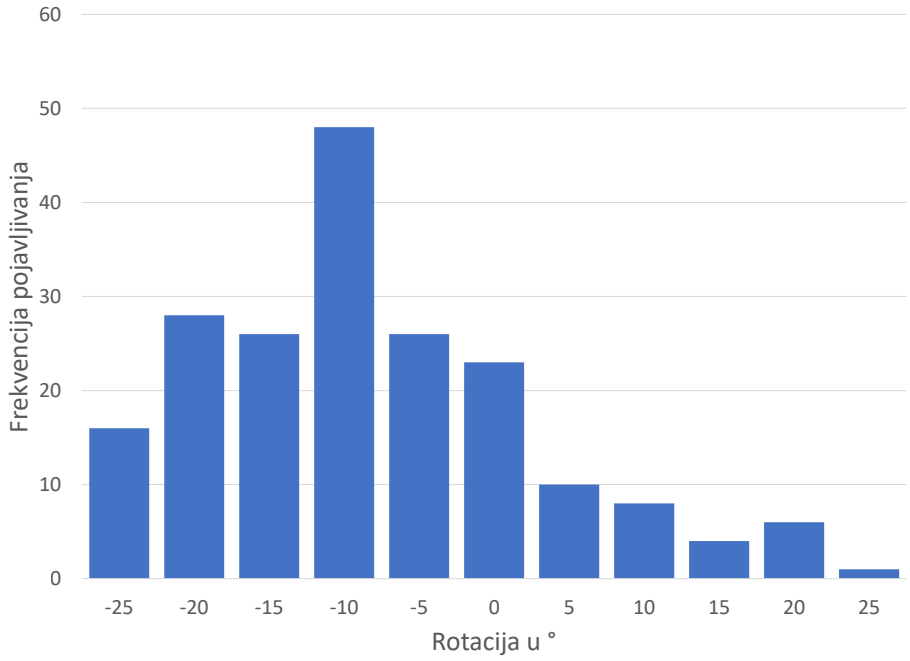
Kako bi se uštedjelo na vremenu izvođenja, navedeni algoritam je implementiran na način da nakon što odredi nasumične okomite linije odabire samo 4 početne točke, prethodno označene sa T_1. Za te točke se računaju kutevi sa svim ostalim točkama, a na kraju se provjerava postoji li kut koji prevladava. Primjer jednog takvog izvođenja prikazan je na slici 3.3.



Slika 3.3: Histogram s prikazom frekvencija pojavljivanja kuteva

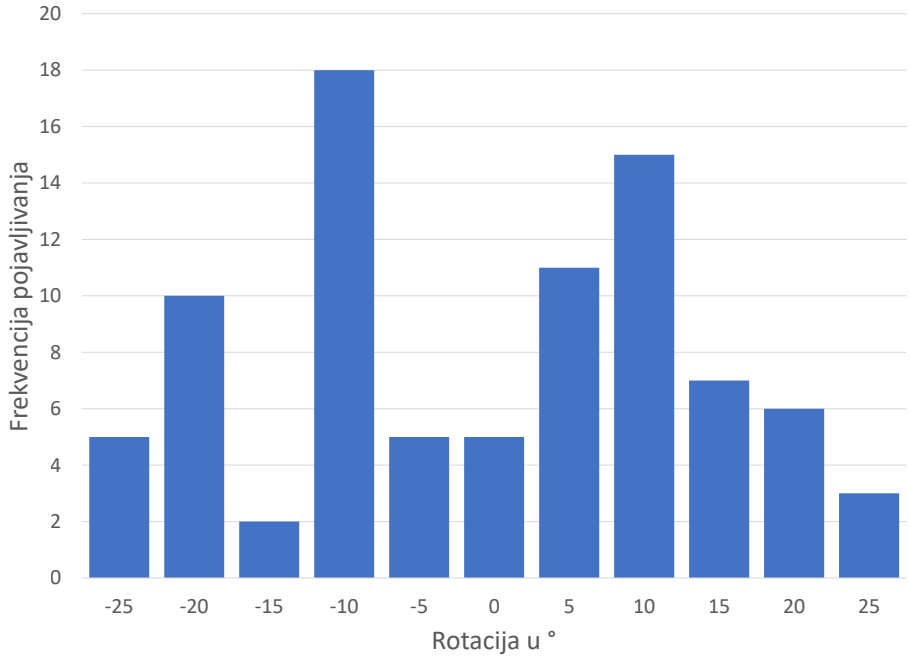
Iz slike se može lako primjetiti kako je najčešće izmjeren kut od $5°$ stupnjeva, iz čega se zaključuje da je slika rotirana za točno taj iznos. Isti postupak ponovljen je za sliku koja je bila rotirana za kut od $10°$ stupnjeva, a rezultati toga prikazani su histogramom na slici 3.4. Na ovoj slici se primjećuje kako se za veće iznose kuteva javlja veći šum u izračunatim kutevima, no i dalje je moguće razaznati dominantnu vrijednost kuta rotacije.

Problem se javlja kod izračuna koji je prikazan histogramom na slici 3.5, gdje nije moguće sa sigurnošću odrediti kako je slika rotirana. Do ovog problema je moglo doći iz više razloga, poput loše odabranih okomitih linija, ili problema zbog točaka dobivenih kao sjeci-



Slika 3.4: Histogram s prikazom frekvencija pojavljivanja kuteva

šta okomitih linija i teksta koji se nalazi u tablicama. Jednostavno i efikasno rješenje navedenog problema je ponovno određivanje okomitih linija te ponovan izračun kuteva sa nove proizvoljne 4 točke.



Slika 3.5: Histogram s prikazom frekvencija pojavljivanja kuteva

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.