

hUNIVERZITET U BANJALUCI  
ELEKTROTEHNIČKI FAKULTET  
BANJALUKA

***DIGITALNA OBRADA SLIKE***  
*ANALIZA LAYOUT-a DOKUMENTA*

STUDENTI:  
MITROVIĆ SAŠA 5/99  
BLAGOJEVIĆ ŽELJKO 82/99

## UVOD

Već duži niz godina, postoje tvrdnje da će, nakon više hiljada godina dominacije papira, kao medijuma za prenos i čuvanje informacija, elektronski oblik zapisa preuzeti u potpunosti tu ulogu. Iako se niko ne usuđuje iznijeti tačnije prognoze, evidentno je da se svakim novim danom, sve više aktivnosti izvršava elektronskim putem, umjesto klasičnim zapisivanjem i dokumentovanjem. Shodno tome, javlja se snažna potreba za mogućnošću konvertovanja štampanih, klasičnih dokumenata u elektronski oblik. Jedan način, da se to izvrši je optičko prepoznavanje karaktera ili OCR (*optical character recognition*).

Važan korak u OCR sistemima je manipulacija *layout*-om dokumenata. Layout predstavlja zonu na kojoj je lociran tekst na dokumentu. Ako ne bismo poznavali okvire u kojima se tekst nalazi, bilo bi teško napraviti pravilnu rekonstrukciju skeniranog dokumenta, konvertovanog OCR softverom. Još jedan primjer je, kada bi trebalo prikazati na PDA displeju stranicu sa tri kolumne teksta i slike, dokument bi morao biti složen tako, da se slike nalaza na logički ispravnom mjestu, tj. pored paragrafa koji se na njih odnose.

Uzimajući u obzir ove i druge mnogobrojne primjere, postaje važno da se posveti pažnja rešavanju problema analize layout-a.

U samoj analizi layout-a, postoje mnoga područja istraživanja. Jedan od najvećih dijelova radova, na temu ove analize, posvećen je algoritmima, koji su korišćeni za izdvajanje homogenih oblasti na stranici dokumenta. Drugi dio se odnosi na logičku analizu layout-a, sa težnjom da se označe različiti segmenti teksta, vodeći se njihovom funkcijom u tekstu, tj. da li označeni tekst predstavlja naslov, autora, napomenu, i slično.

Što ce tiče testiranja performansi sistema za analizu, takođe se mnogo vodi računa. Testiranje ima za cilj da procijeni koliko algoritam za analizu kvalitetno funkcioniše, odnosno, koliko je automatski kreiran layout, ustvari, vjerodostojan originalu, napravljenom od strane čovjeka.

## „XY-cut” metoda

Postoji više metoda za analizu layout-a, i iako im je svima sličan cilj, ipak im se algoritmi međusobno razlikuju, pa i sami rezultati. To pruža korisniku mogućnost da primjeni u svom zadatku metodu koja najbolje odgovara njegovim zahtjevima.

Ovde će više riječi biti posvećeno tzv. „XY-cut” metodi.

Ideja ovog algoritma i njegova implementacija je krajnje jednostavna. Pikseli sadržaja na dokumentu su projektovani po vertikali i horizontali. Algoritam se bazira na traženju najvećeg bijelog proreda u sadržaju (između linija teksta) koji se uzima kao granica dva podsegmenta samog sadržaja. Zatim se to rekurzivno ponavlja na svaki podsegment, i vrši se „usitnjavanje” sve dok se ne zadovolje postavljeni zahtjevi, odnosno, dok se ne dođe do minimalno potrebnih proreda.

Ovim metodom ostvarujemo listu horizontalnih i vertikalnih dijelova, koji dijele kompletan sadržaj na različite regione. Ukoliko suzimo ove regione na najmanje moguće pravougaonike pokrivajući crne piksele, dobićemo različite blokove. U zavisnosti od kriterijuma zaustavljanja (zahtjev za širinom proreda), dobićemo različite blokove.

XY-cut metoda ima i svoje nedostatke. Postoje problemi sa razdvojnim linijama, kao i sa crnim linijama, koje nastaju pri skeniranju i kopiranju dokumenata, ili stranica knjige. U slučaju pojavljivanja takvih linija, metoda neće funkcionisati, jer praktično ne prepoznaje bijele prorede, koji su tim linijama narušeni, tj. prekriveni. Samim tim, softver „neće znati” gdje treba da presječe tekst, odnosno, na kojem mjestu treba razdvojiti segmente. Zbog toga je potrebno prvo ukloniti sjenke i linije, prije primjene XY-cut metode.

Još jedna mana je ta, što je metoda primjenljiva samo na tzv. „Menhetn” tip layout-a. To je tip layout-a koji je sačinjen isključivo od pravougaonih blokova.

Primjer koji pokazuje kako treba da izgleda strana dokumenta, na koju je primjenjen XY-cut metod je prikazana na sledećoj slici.



Sl.1. Originalni dokument



sl2. Dokument obrađen XY-cut Algoritmom

Praktična primjena XY-cut algoritma, ostvarena u matlabu, prema sledećem kodu, daje rezultate kao na slici 3.

```

b=imread('scan3.jpg');
a=im2bw(b);
pocetak=1;
pocetak1=1;
[vrsta kolona]=size(a);
imshow(a)

while(pocetak+1<vrsta),
pocetak=pocetak1;
while(sum(~a(pocetak,1:end)))==0,
if(pocetak==vrsta),break, end
pocetak=pocetak+1;
pocetak1=pocetak1+1;

```

```

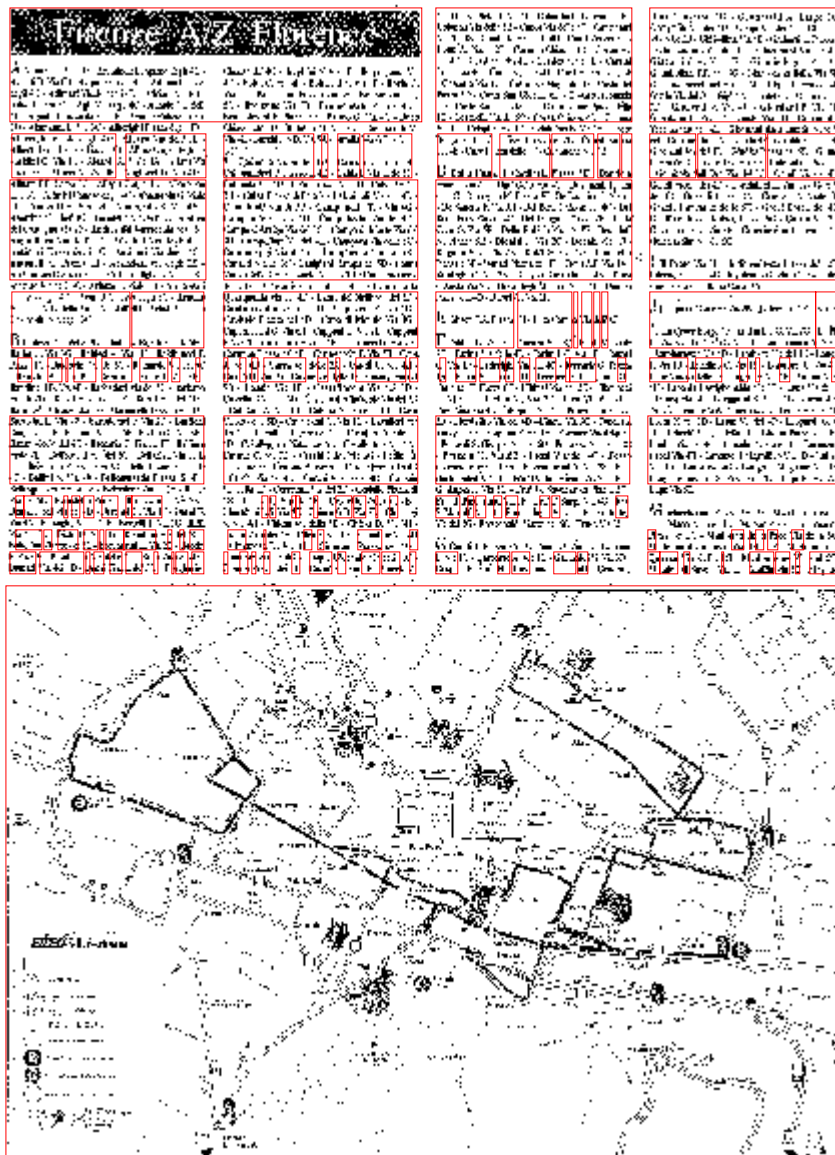
end
pocetak=pocetak-1;

while(sum(~a(pocetak1,1:end)))>0,
pocetak1=pocetak1+1;
end

pocetak;
pocetak1;
kraj=1;
kraj1=1;
while(kraj+1<kolona),
kraj=kraj1;

while(sum(~a(pocetak:20:pocetak1,kraj)))<=1,
if(kraj==kolona),break, end
kraj=kraj+1;
kraj1=kraj1+1;
end
kraj=kraj-1;
brojac=1;
while or((sum(~a(pocetak:pocetak1,kraj1)))>=1,brojac<0),
kraj1=kraj1+1;
end
kraj;
kraj1;
if(kraj1>kraj+1),
rectangle('Position',[kraj pocetak kraj1-kraj pocetak1-
pocetak],'edgecolor','r')
end
end
end

```



### Sl.3. Praktična realizacija