E L E K T R O T E H N I Č K I F A K U L T E T
B A N J A L U K A
K A T E D R A Z A O P Š T U E L E K T R O T E H N I K U
D I G I T A L N A O B R A D A S L I K E

SEGMENTACIJA SKENIRANIH STRANICA NA BLOKOVE TEKSTA
(ANALIZA PROSTORNE ORGANIZACIJE DOKUMENATA)

profesor: dr Zdenka Babić
asistent: mr Vladimir Risojević

studenti:
Saša Blagojević , br. indeksa: 59/03
Janko Rosuljaš, br. indeksa: 30/04

UVOD

Cilj seminarskog rada je da se iz slike dokumenta izdvoje odnosno označe dijelovi slike dokumenta na kojima se nalaze slike odnosno tekst. Nakon razmatranja većeg broja metoda za rješavanje ovog problema uzet je metod za koji se smatralo da će dati najbolje rezultate u najvećem broju slučajeva.

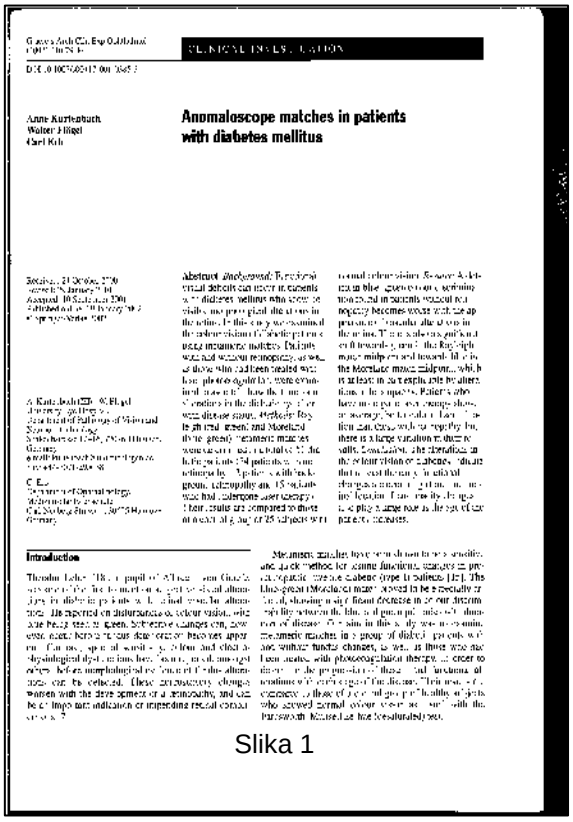
Nakon razmatranja problema i mogućnosti koje nam pruža programski paket MATLAB izbrana je realizacija koja podrazumijeva rad sa binarnim slikama, tako da se sve slike prvo moraju konvertovati u binarni oblik a zatim se na njima vrši obrada. Prvi problem koji se javio je pojava crnih piksela na rubovima slike koji se najčašće javljaju prilikom skeniranja dokumenata. Način realizacije koji je odabran nije dozvoljavao pojavu takvih piksela pa su isti morali biti uklonjeni. Nakon uklanjanja rubnih piksela slijedi lociranje i označavanje mjesta na kojima se nalaza slike odnosno tekst na orginalnoj slici. Mjesta na dokumentu na kojima se nalaze slike predstavljena su crvenom bojom, tekst plavom bojom a praznine bijelom bojom.

Za pozivanje algoritma potrebno je učitati sliku u promjenjivu **a** i pozvati funkciju ***segmentuj.***

Za problem pronalaženja i označavanja teksta obrađen je još jedan algoritam ali rezultati, iako su bili dobri, nisu bili vizuelno dopadljivi. Primjer će biti prikazan na kraju izvještaja.

REALIZACIJA

Problem pojave crnih rubnih piksela koji je pomenut ranije je riješen upotrebom funkcije ***kopija*** na sledeći način. Na osnovu veličine slike odabrane su sirine rubnih “traka” (gornja, donja i lijeva) koje će biti obojene u bijelo dok je širina desne “trake” računata na osnovu broja crnih rubnih piksela na desnoj strani. U odabranoj realizaciji pri rješavanju ovog problema može doći do gubitka određenog broja korisnih piksela ukoliko bi oni bili suviše blizu rubovima slike. Na sledeće dvije slike je prikazan rezultat ovog postupka.



Anne Kurtenbach
Walter Hügel
Karl Kih

Anomaloscope matches in patients with diabetes mellitus

Introduction

Slika 1

Anne Kurtenbach
Walter Hügel
Karl Kih

Anomaloscope matches in patients with diabetes mellitus

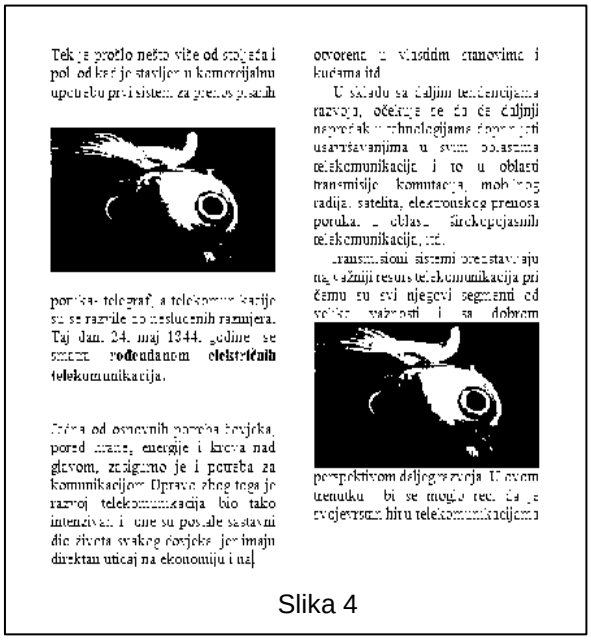
Introduction

Slika 2

Da ne bi bilo zabune prethodne slike ali i buduće su okvirene u Word-u i crne linije na ivicama potiču od toga.

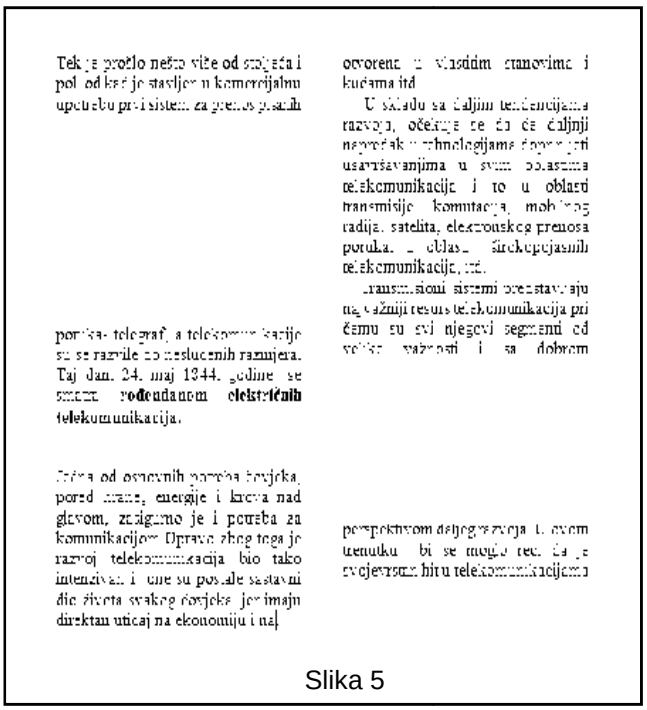
Nakon uklanjanja rubnih piksela pristupa se rješenju samog problema. Slika nad kojom se vrši obrada se konvertuje u binarnu čiji bijeli pikseli imaju vrijednost jedan a crni nula.

Slika 3

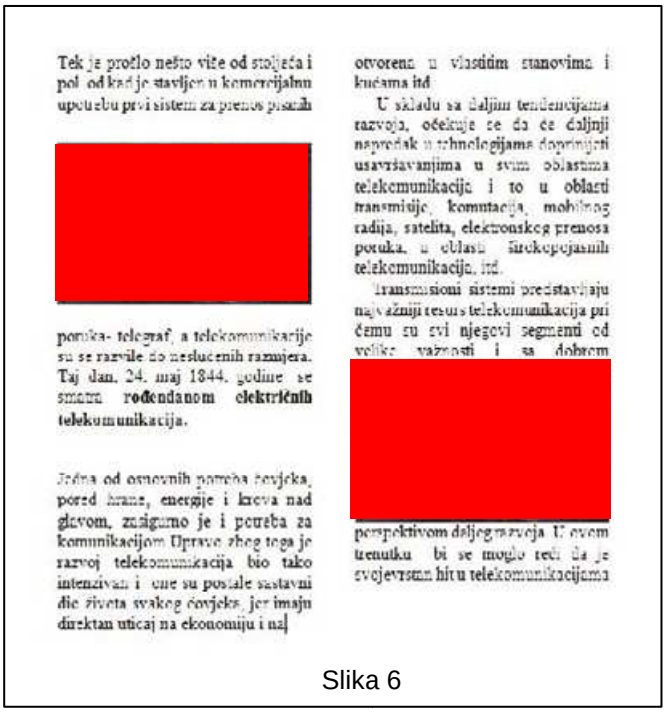


Slika 4

Sve operacije se vrše na binarnoj slici a na originalnoj slici se mjesta na kojim se nalaze objekti boje određenim bojama. Na osnovu veličine slike se formira vrednost promjenjive ***prag*** koja će biti korištena za lociranje slika na dokumentu. Komplement binarne slike se sumira po vrstama tako da se dobije vektor **S** čiji elementi predstavljaju sumu piksela svake vrste.Ukoliko su u k-toj vrsti svi pikseli bjeli, k-ti element vektora **S** će imati vrednost nula, a u zavisnosti od broja crnih piksela poprimiće neku drugu vrijednost. Ukoliko je broj uzastopnih elemenata vektora **S** različitih od nule veći od vrijednosti promjenjive ***prag*** , pretpostavlja se da se u tom djelu nalazi slika. Tada se isjeca taj dio slike, pa novodobijena slika se sumira po kolonama i formira se novi vector **S**. U djelu u kome se ponovo javi veći broj elemenata vektora **S** različitih od nule pretpostavlja se da je slika, na binarnoj slici tim pikselima se dodjeljuje vrijednost 1, a na originalnoj slici taj dio se oboji u crveno. Postupak se ponavlja dok se ne lociraju sve slike. Prethodni postupak je implementiran u funkciji ***izdvojsl***. Nedostaci ove metode su ti, da ukoliko bi slika bila manjih dimenzija od vrednosti promjenjive ***prag*** , ona ne bi bila tretirana kao slika. Takodje, ukoliko bi se na slici nalazila linija koja nije paralelna niti sa jednom od ivica dokumenta, ili ako bi tekst bio ispisan ukoso algoritam ne bi dobro radio. Rezultat prethodno opisanog postupka je dat na sledećim slikama.



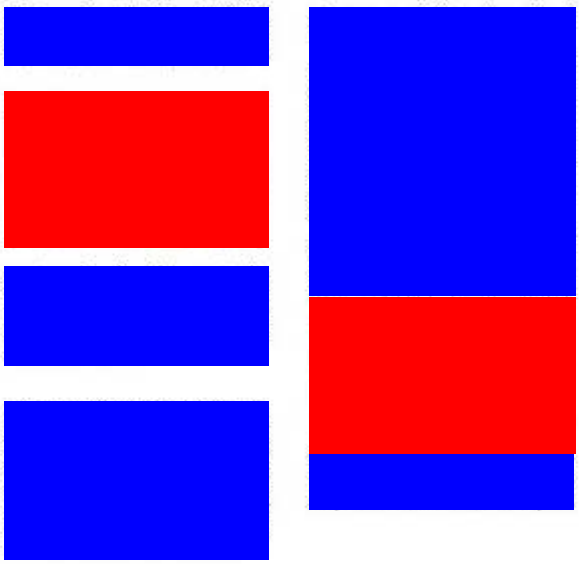
Slika 5



Slika 6

Nakon što se uklone sve slike sa dokumenta, potrebno je da se locira i izdvoji tekst, odnosno blokovi teksta (pasusi). Taj dio se obavlja pomoću funkcije ***tekst.*** Takodje na osnovu veličine slike promjenjivoj ***prag1*** se dodjeljuje odredjena vrijednost. Sumira se po vrstama komplement binarne slike sa koje su predhodno uklonjene slike. Algoritam za lociranje teksta je sličan predhodno objašnjenom za lociranje slika. Prvi put kada se pronađe element vektora **S** različit od nule, broji se broj uzastopnih elemenata vektora različitih od nule. Smatra se da je kraj pasusa tek kada broj uzastopnih elemenata jednakih nuli bude jednak ili veći od vrednosti promjenjive ***prag1***. Pamte se koordinate položaja tog pasusa i na binarnoj slici, koju smo predhodno zapamtili svim pikselima sem onim koji se nalaze u ravni sa selektovanim pasusom dodjeljujemo vrednost 1 (bijelo). Sada se binarna slika na kojoj se nalazi samo selektovani pasus sumira po kolonama. Na isti način se pronalazi početak i kraj pasusa (porebno je u slučaju da je tekst pisan u nekoliko kolona) . Tako se dobiju koordinate pasusa pa na originalnoj slici se taj dio oboji u plavo, a na binarnoj svim pikselima iz tog regiona se dodijeli vrijednost 1. Postupak se ponavlja sve dok se ne lociraju svi djelovi teksta.

Konačni rezultat kada se izvrše sve ove operacije je prikazan na sledećoj slici. Radi poređenja data je i originalna slika.



Slika 7

Tek je prošlo nešto više od stoljeća i pol od kad je stavljen u komercijalnu upotrebu prvi sistem za prenos pisanih

poruka- telegraf, a telekomunikacije su se razvile do neslućenih razmjera. Taj dan, 24. maj 1844. godine se smatra **rođendanom električnih telekomunikacija.**

Jedna od osnovnih potreba čovjeka, pored hrane, energije i krova nad glavom, zasigurno je i potreba za komunikacijom Upravo zbog toga je razvoj telekomunikacija bio tako intenzivan i one su postale sastavni dio života svakog čovjeka, jer imaju direktan uticaj na ekonomiju i na

otvorena u vlastitim stanovima i kućama itd

U skladu sa daljim tendencijama razvoja, očekuje se da će daljnji napredak u tehnologijama doprinijeti usavršavanjima u svim oblastima telekomunikacija i to u oblasti transmisije, komutacija, mobilnog radija, satelita, elektronskog prenosa poruka, u oblasti širokopojasnih telekomunikacija, itd.

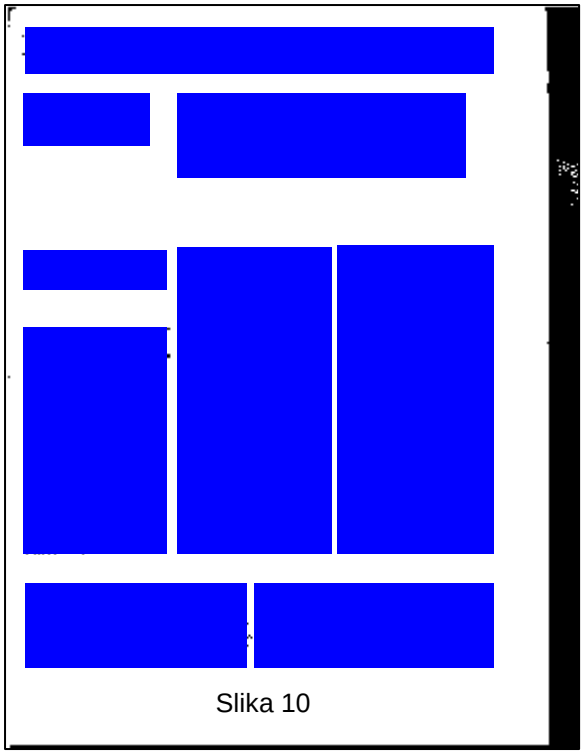
Transmisioni sistemi predstavljaju najvažniji resurs telekomunikacija pri čemu su svi njegovi segmenti od velike važnosti i sa dobrom

perspektivom daljeg razvoja U ovom trenutku bi se moglo reći da je svojevrstan hit u telekomunikacijama

Slika 8

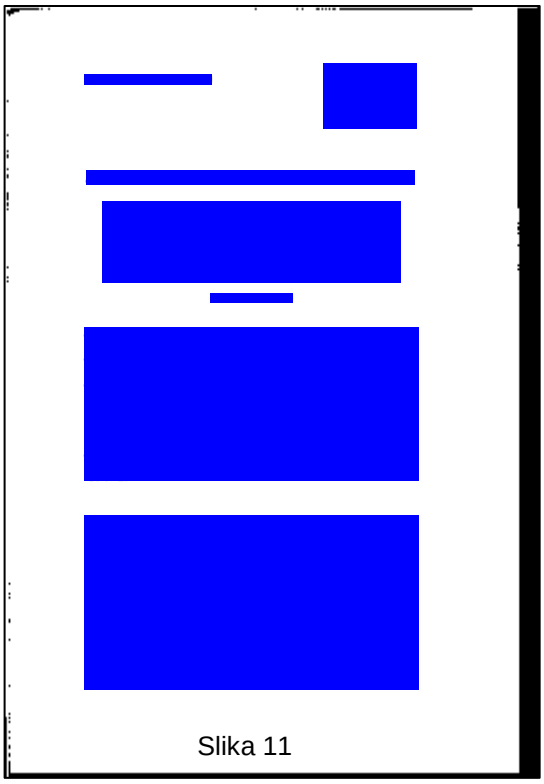
Sada je dat rezultat za sliku na kojoj je pokazano odstranjivanje rubnih piksela. Takođe je prikazana i orginalna slika radi poređenja. Na konačnoj slici se vide crni rubni pikseli koji su ostavljeni tako sa namjerom da se pokaže da oni nisu stvarni djelovi dokumenta i da su rezultat greške.

Slika9

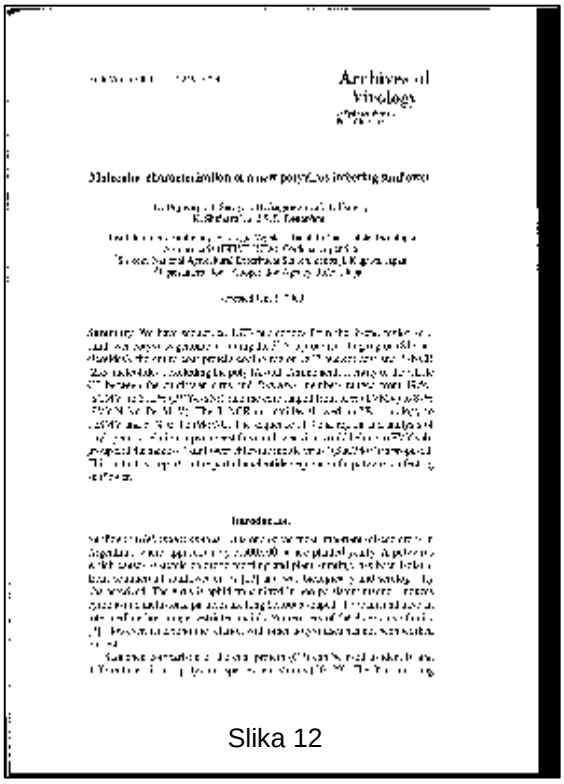


Slika 10

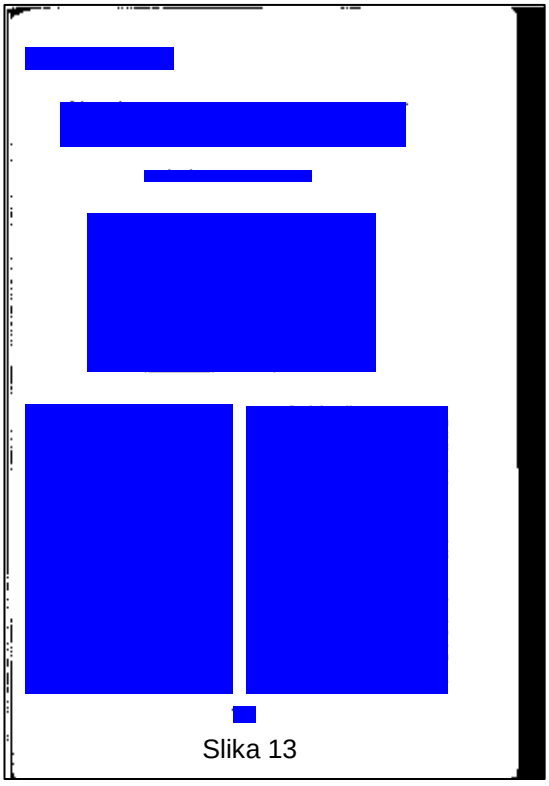
Na sledećim slikama su prikazani primjeri na kojima opisane procedure daju zadovoljavajuće rezultate.



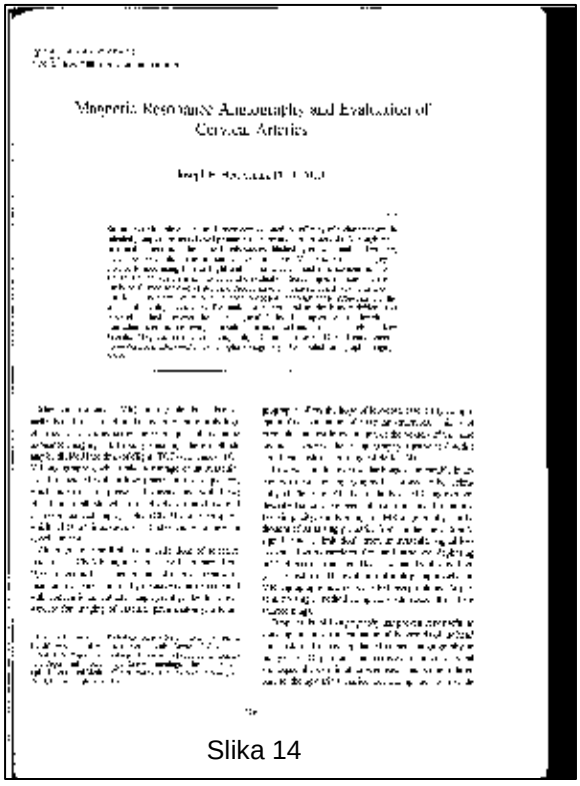
Slika 11



Slika 12

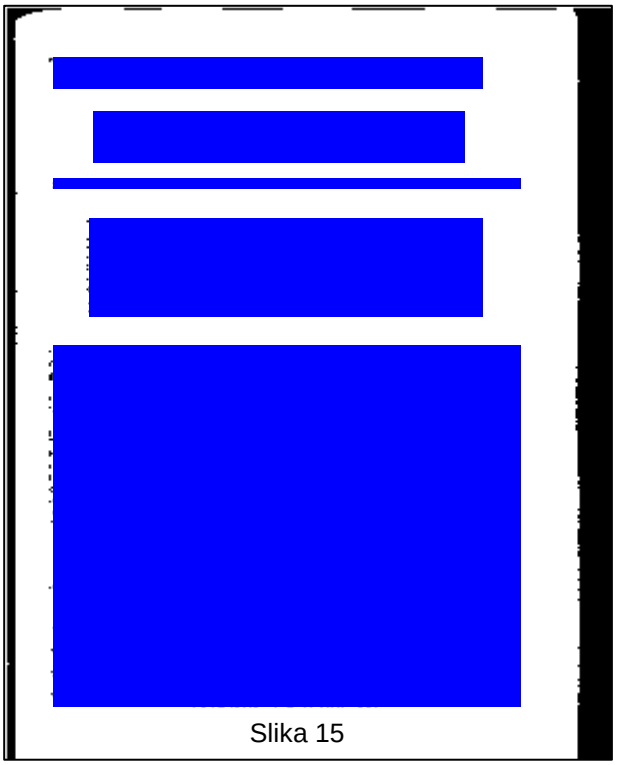


Slika 13

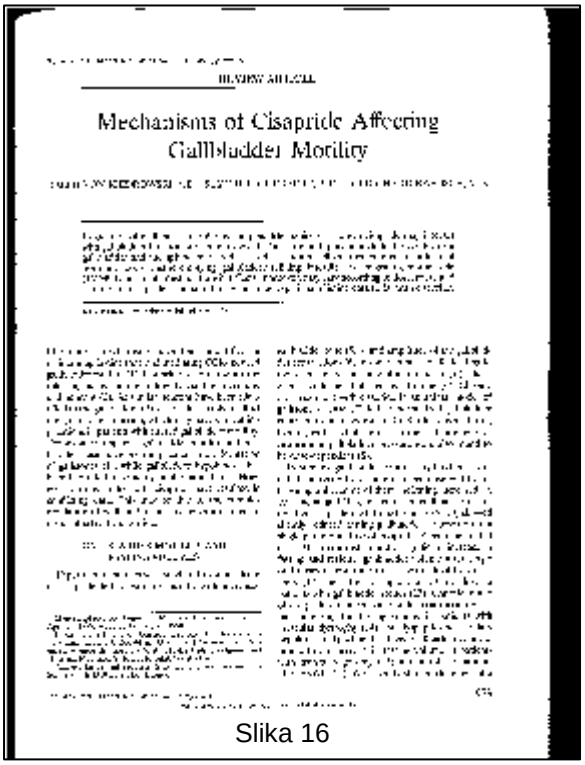


Slika 14

Na sledećim slikama su prikazani rezultati koji pokazuju da postoje nedostaci ove metode a oni su najčešće posljedica greške pri izboru pragova ali i razmjestaja pasusa teksta i njihovog oblika.

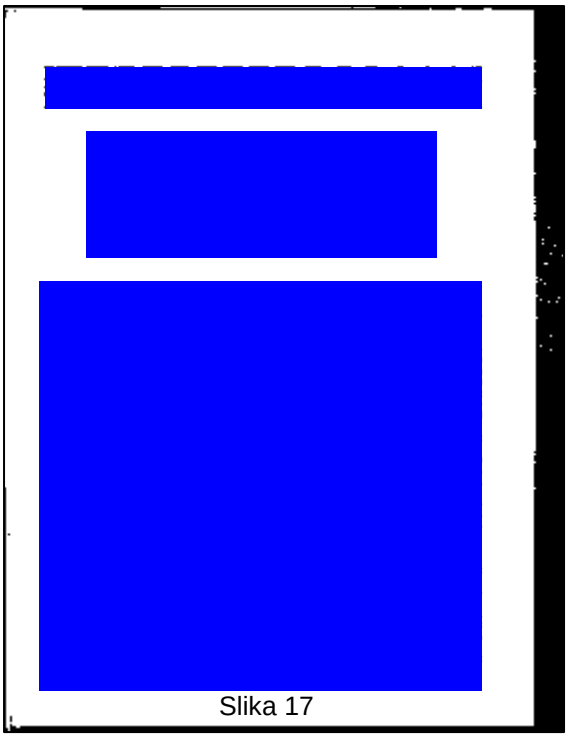


Slika 15

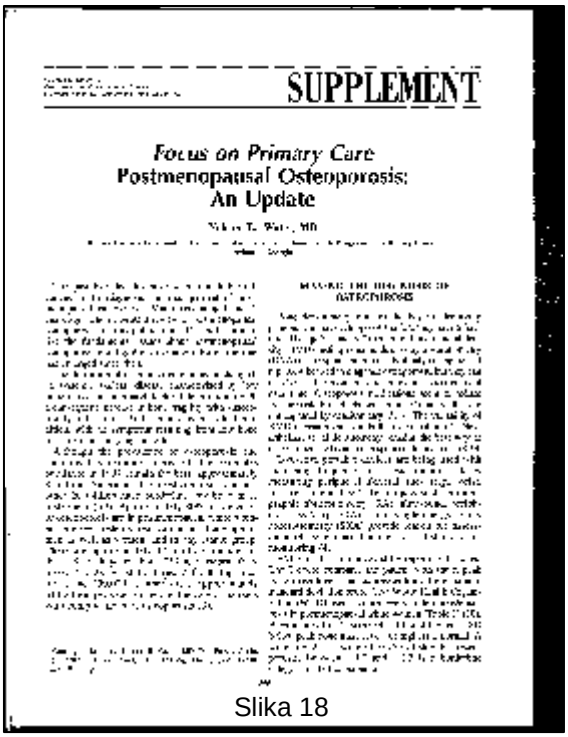


Mechanisms of Cisapride Affecting Gallbladder Motility

Slika 16



Slika 17



SUPPLEMENT

Focus on Primary Care

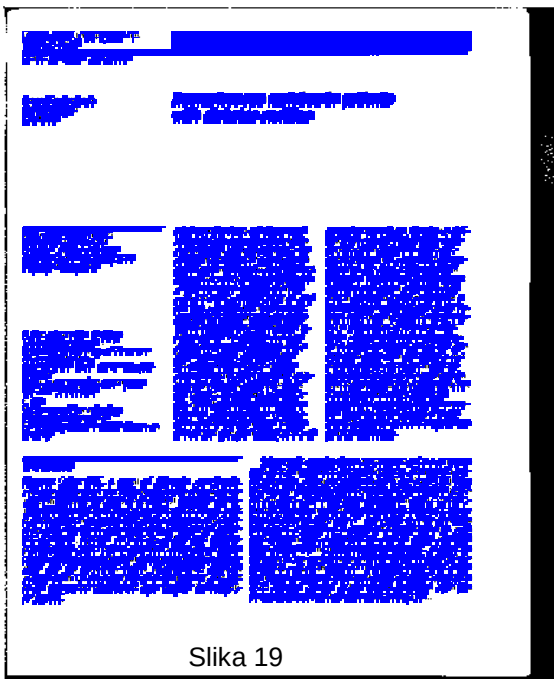
Postmenopausal Osteoporosis: An Update

Slika 18

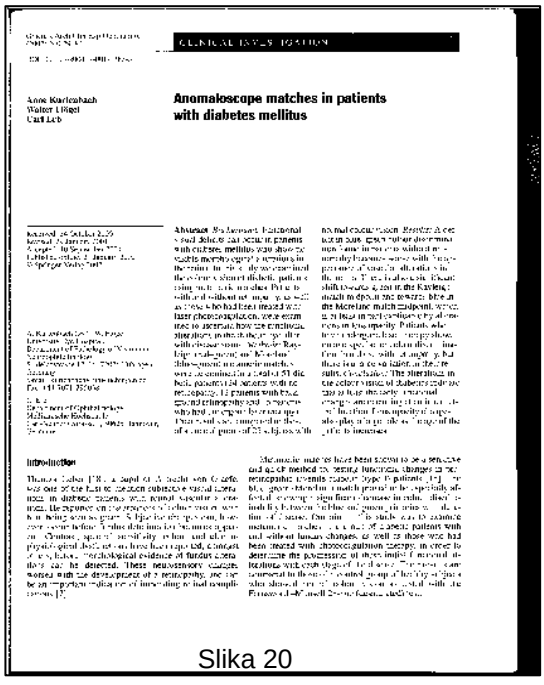
Algorotam koji je opisan, iako se pokazao kao dobar u velikom broju slučajeva mogao bi se popraviti. Ukoliko bi prilikom kopiranja dokument odnosno linije teksta bile nakošene, to bi stvaralo problem. Ovaj problem bi se mogao rješiti ispitivanjem da li su ivice teksta paralelne sa ivicama dokumenta, pa zatim ukoliko je potrebno njihovim poravnanjem.

Takođe bi se moglo poboljšati vreme izvršavanja algoritma. Naime kada algoritam na primjer pronađe sliku na dokumentu, neće se tu zaustaviti, već će nastaviti pretragu i ukoliko pronađe još jednu sliku, koordinate prve slike će biti izgubljene. To će se ponavljati sve dok se ne dođe do poslednje slike. Kada se ona obradi, kreće se ispočetka i time se usporava algoritam. Isti problem se javlja i pri obradi teksta. Ovo bi se moglo rješiti tako da kada se jednom pronađe objekt od interesa, te koordinate se pamte, nastavlja se sa skeniranjem dokumenta, pa tek kada se lociraju svi objekti, pristupa se potrebnoj obradi.

Dodatni algoritam koji je bi testiran za lociranje teksta će ukratko biti objašnjen u narednom djelu. Pomoću funkcije ***scenh*** i ***scenv*** se binarna slika dokument, bez slika horizontalno odnosno vertikalno skenira matricom dimenzija NxN (u našem primjeru N=10). Ukoliko se vrednost barem jednog od 100 piksela razlikuje od nule, čitav region se boji odredjenom bojom. Rezultati su dobri ali ne vizuelno dopadljivi.



Slika 19



Anomaloscope matches in patients with diabetes mellitus

Slika 20

LITERATURA:

[1] Help Matlab
[2] Materijali sa vježbi iz predmeta *Digitalna obrada slike*
[3] Diploma thesis : Document Layout Analysis