SVEUČILIŠTE U ZAGREBU

FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

DIPLOMSKI RAD br. 1633

Zaštita teksta digitalnim vodenim žigom

Tihana Poljak

Voditelj: *Marin Golub*

Zagreb, studeni, 2007

1. Uvod

U današnje vrijeme postoji sve veća potreba za zaštitom osjetljivih informacija. Enkripcija, ograničavanje pristupa i zaštita dokumenata iza sigurnosne stijene (eng. *firewall*) neke su od uobičajenih tehnika zaštite osjetljivih informacija. Enkripcija je dobar način sprječavanja neovlaštene osobe od pregledavanja sadržaja osjetljivog dokumenta. Ali kada se dokument dekriptira tajnim ključem, ovlaštena osoba loših namjerna može spremiti, kopirati, ispisati ili proslijediti dekriptirani dokument.

Ograničavanje pristupa dokumentu nekolicini pojedinaca funkcionira kod pojedinaca vrijednih povjerenja. Nažalost, događa se da se povjerljive informacije nalaze izvan povjerljivih zona pa čak i u medijima. U tom slučaju želi se pronaći osobu koja je odala informacije, što nije uvijek jednostavan i ugodan proces.

Sigurnosna stijena je učinkovit način sprječavanja pristupa povjerljivoj mreži od strane vanjskih korisnika, koji nemaju prava pristupa. Ali to ne sprječava osobu unutar organizacije da spremi ili proslijedi osjetljiv dokument trećoj strani.

Rješenje koje osigurava zaštitu osjetljivih informacija ne može ovisiti o samo jednoj tehnologiji. Umjesto toga, efikasna sigurnost ostvaruje se svim prethodno spomenutima tehnologijama s time da mora ostaviti otisak na samom dokumentu. Pod ostavljanjem otisaka smatra se ugrađivanje jedinstvene informacije u dokument, koja identificira vlasnika ili primatelja dokumenta. Ugrađena informacija može se detektirati i dekodirati u bilo kojem trenutku, čak i nakon ispisa i skeniranja. Proces ostavljanja otisaka u dokumentu može se postići uporabom tehnika označavanja digitalnim vodenim žigom.

Označavanje digitalnim vodenim žigom je tehnika kojom se mogu zaštiti autorska prava različitih multimedijskih sadržaja. S obzirom da postoji više različitih formata: slike, audio podaci, video podaci, grafički objekti, potrebno je razviti posebne metode za svaki od njih. U usporedbi s istraživanjima o označavanju slika, video i audio podataka, istraživanja o označavanju teksta su malobrojna. Ipak pojavom novih primjena kao što su npr. digitalna knjižnica te knjige u elektroničkom formatu raste i interes za ovo područje.

U ovom radu se opisuju različiti načini označavanja teksta digitalnim vodenim žigom te njihova primjena. Poglavlje 2 je uvod u digitalne vodene žigove i njihove primjene. Poglavlje 3 opisuje kako se digitalna knjižnica može zaštititi uporabom otpornih digitalnih vodenih žigova. Poglavlje 4 opisuje neke od algoritama za označavanje teksta, a poglavlje 5 opisuje praktičnu implementaciju jednog od algoritama za označavanje teksta.

2. Uvod u digitalne vodene žigove i njihova primjena

2.1 Osnove označavanja digitalnim vodenom žigom

Osnovna ideja označavanja digitalnim vodenim žigom je stvaranje meta podataka koji sadrže informacije o digitalnom mediju koji se želi zaštititi. Meta podaci su vodeni žig koji se može neprimjetno ugraditi u željeni medij te treba biti otporan na namjerna i nenamjerna izobličenja signala.

Sustav za označavanje digitalnim vodenim žigom sastoji se od dva glavna dijela: ugrađivanje vodenog žiga i detekcija. Ugrađivanje kombinira medij C_o, audio vizualni signal u koji se ugrađuje informacija, i poruku (eng. *payload) P*, koja se dodaje mediju, čime se stvara označeni sadržaj C_w. Algoritam označavanja ima dva koraka. U prvom se koraku poruka *P* kodira u vodeni žig *W*. Vodeni žig *W* mora biti istog tipa i istih dimenzija kao i medij. Ako je npr. medij C_o slika, tada i vodeni žig mora biti uzorak slike istih dimenzija kao i originalna slika. Bolja sigurnost može se postići korištenjem ključa vodenog žiga *K* u procesu kodiranja. U drugoj fazi, vodeni žig *W* dodaje se mediju C_o kako bi se stvorio označeni medij C_w.

Postoje dvije vrste označavanja: slijepo i informirano. Vrsta označavanja ovisi o tome koristi li se medij C_o prilikom stvaranja vodenog žiga *W* ili ne.

Za slijepo označavanje nije potreban originalan medij, a može se opisati sljedećim izrazom:

$$C_W = E_1(C_O, W), \quad gdje \quad je \quad W = E_0(P, K) \tag{2.1}$$

gdje E_1 označava operaciju ugrađivanja vodenog žiga *W* u medij C_o. Vodeni žig *W* dobiva se kodiranjem (E_0) poruke *P* uz pomoć ključa vodenog žiga *K*.

Informirano označavanje koristi informacije iz originalnog medija prije kreiranja vodenog žiga *W* i može se opisati sljedećim izrazom:

$$C_W = E_1(C_O, W), \quad gdje \quad je \quad W = E_O(P, K, C_O) \tag{2.2}$$

gdje E_1 označava operaciju ugrađivanja vodenog žiga *W* u medij C_o. E_0 označava operaciju kodiranja, odnosno stvaranje vodenog žiga *W* korištenjem informacije iz originalnog medija C_o, poruke *P* i ključa vodenog žiga *K*.

Označeni medij može proći kroz različite operacije. Operacije mogu biti različita izobličenja uzrokovana uobičajenim transformacijama signala (kompresija, dekompresija, pretvorba iz analognog u digitalni i obrnuto) ili namjerni napadi. Primjenom tih operacija može se narušiti kvaliteta originalnog označenog medija, odnosno stvara se novi medij *Cw'*.

Detektori vodenih žigova također se dijele na dvije vrste, slijepe i informirane. Vrsta označavanja ovisi o tome koliko informacija o mediju je dostupno prilikom procesa detektiranja vodenog žiga.

Informirani detektor koristi originalni medij C_o u procesu detekcije te se može opisati sljedećim izrazom:

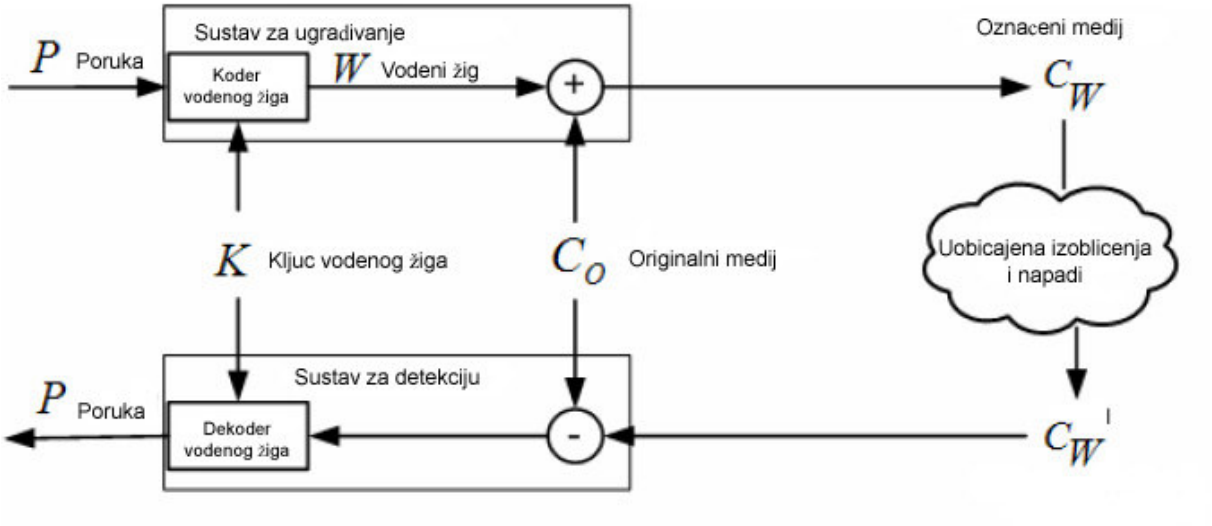
$$P = D(C_W', C_O, K) \tag{2.3}$$

gdje *D* označava proces detekcije poruke *P* korištenjem izmijenjenog medija *Cw'*, originalnog medija C_o i ključa vodenog žiga *K*.

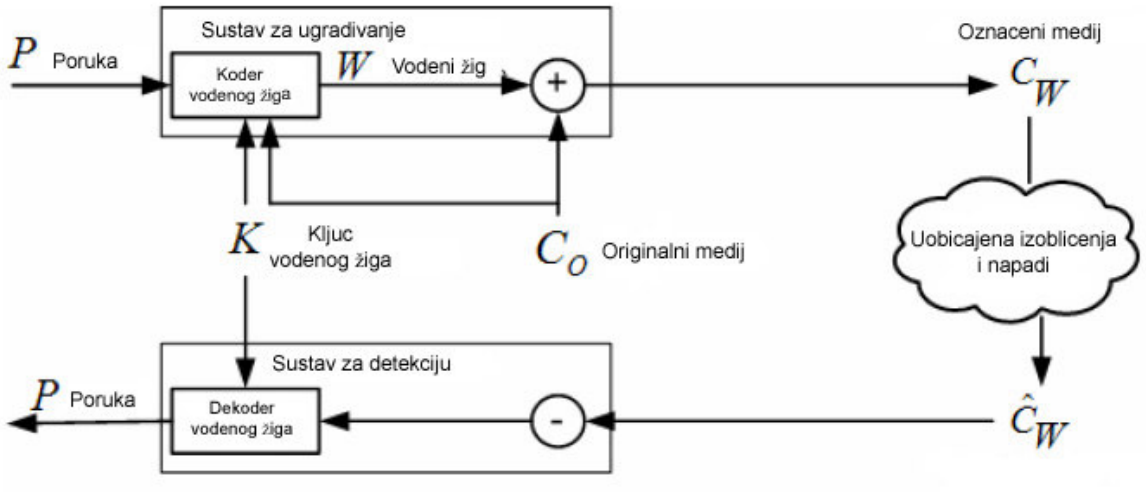
Slijepi detektor ne koristi originalni medij C_o te se može opisati sljedećim izrazom:

$$P = D(C_W', K) \tag{2.4}$$

gdje *D* predstavlja proces detekcije, *Cw'* izmijenjeni medij, *K* ključ vodenog žiga, a *P* poruku.



Slika 2.1 Primjer sustava za digitalni vodeni žig sa slijepim ugrađivanjem i informiranom detekcijom.



Slika 2.2 Primjer sustava za digitalni vodeni žig s informiranim ugrađivanjem i slijepom detekcijom.

Slika 2.1 prikazuje jedno od mogućih ostvarenja sustava za označavanje i detekciju digitalnog vodenog žiga koji u ovom slučaju koristi slijepo ugrađivanje i informiranu detekciju. Slika 2.2 prikazuje još jedno moguće ostvarenje koje ima informirano ugrađivanje i slijepu detekciju. Više o digitalnim vodenim žigovima dostupno je u [1].

2.2 Vrste digitalnih vodenih žigova

2.2.1 Lomljivi vodeni žigovi

Zovu se lomljivi jer je poželjno da se prilikom primjene većine tehnika obrade dokumenata izmjene ili unište.

Svojstva:

1. Vodeni žig je nevidljiv promatraču.
2. Vodeni žig se mijenja prilikom primjene većina tehnika za obradu dokumenata.
3. Neovlaštene osobe ne bi smjele moći ubaciti lažni vodeni žig.
4. Ovlaštene osobne mogu brzo izvaditi vodeni žig.
5. Očitani vodeni žig pokazuje gdje je došlo do promjena.

Svojstvo nevidljivosti vodenog žiga teško je izmjeriti jer ovisi o više faktora. Svojstvo 3 odnosi se na mogućnost da se vodeni žig može učitati iz označenog dokumenta i ubaciti u zamjenski. Kako bi ovo svojstvo bilo ostvareno potrebno je otežati potencijalnim napadačima saznanje je li dokument označen i kako je označen. Pogodni su vodeni žigovi koji se mogu učitati samo s posebnim ključem, a ključ se čuva u posebnoj, sigurnoj bazi podataka[2].

2.2.2 Otporni vodeni žigovi

Zovu se otporni žigovi jer se očekuje da budu postojani neovisno o napadima.

Svojstva:

1. Vodeni žig je nevidljiv promatraču.
2. Vodeni žig ostaje u dokumentu čak i nakon obrade dokumenta.
3. Neovlaštene osobe teško mogu detektirati vodeni žig.
4. Ovlaštene osobe mogu brzo izvaditi vodeni žig.
5. Nakon što je dokument ispisan i skeniran i dalje je moguće učitati vodeni žig.

Stvaranje algoritama koji posjeduju svojstva 3 i 5 težak je zadatak, ali otporan vodeni žig nije pretjerano koristan ako se može lagano ukloniti. Teško je razviti programski sustav koji će detektirati vodeni žig čak i nakon većine izmjena. Dakle, svojstva 2 i 4 su kontradiktorna. Za ostvarivanje svojstva 3 preporuča se korištenje vodenih žigova koji zahtijevaju poseban ključ za učitavanje[2].

2.3 Primjena digitalnih vodenih žigova

Primjene digitalnih vodenih žigova mogu se klasificirati na više različitih načina (ovisno o mediju, poruci itd.). Klasifikacija koja slijedi temelji se na otpornosti vodenog žiga na napade.

2.3.1 Dokazivanje autentičnosti sadržaja

Postoje različiti programski sustavi za uređivanje digitalnog sadržaja. S obzirom da je jednostavno mijenjati digitalni sadržaj bitno je naći način za dokazivanje integriteta i autentičnosti sadržaja. Rješenje ovog problema može se posuditi iz kriptografije, gdje se digitalni potpis koristi za dokazivanje autentičnosti. U slučaju označavanja digitalnim vodenim žigom digitalni potpis može biti vodeni žig koji će se ugraditi u sadržaj. Za dokazivanje autentičnosti preporuča se korištenje lomljivog vodenog žiga iz sljedećih razloga: lomljivi vodeni žig mora postati nevažeći u slučaju izmjena, korištenjem lomljivog vodenog žiga može se saznati kako je digitalni sadržaj izmijenjen ili koji je dio izmijenjen.

2.3.2 Praćenje emitiranja

Mnogo proizvoda svakodnevno se emitira preko televizijske mreže: vijesti, filmovi, sportska događanja, reklame, itd. Emitiranje je vrlo skupo i oglašivači moraju izdvajati značajna financijska sredstva za svako emitiranje kratkih reklama koje se pojavljuju za vrijeme pauza popularnih filmova, serija ili sportskih događaja. Mogućnost precizne naplate vrlo je bitna. Oglašivači žele biti sigurni da plaćaju samo za reklame koje su se emitirale.

Praćenje emitiranja (*Broadcast Monitoring*) obično se koristi za prikupljanje informacije o sadržaju koji se emitira. Prikupljene informacije koriste se za naplaćivanje i druge potrebe. Jednostavan način praćenja je korištenje ljudskih promatrača koji prate i bilježe sve što vide. Ova vrsta praćenja je skupa i sklona greškama. Automatizirano praćenje je očito bolji izbor. Postoje dvije vrste sustava za automatizirano praćenje: pasivni i aktivni. Pasivni sustav prati sadržaj koji se emitira i pokušava ga povezati s poznatim sadržajem pohranjenim u bazi. Implementacija pasivnih sustava nije jednostavna iz nekoliko razloga. Usporedba odaslanih signala sa sadržajem baze nije jednostavna. Održavanje velike baze sadržaja za usporedbu je skupo. Aktivni sustavi za praćenje oslanjaju se na dodatnu informaciju koja identificira sadržaj. Dodatna informacija emitira se zajedno sa sadržajem. Jedno od rješenja za aktivno praćenje je i označavanje digitalnim vodenim žigom. Vodeni žig koji sadrži informaciju za identifikaciju emitiranja ugrađuje se u sam sadržaj. Za ovu primjenu vodeni žigovi moraju biti otporniji na napade od lomljivih žigova te ih se mora moći lagano očitati.

2.3.3 Ostavljanje otisaka

Postoje određene primjene u kojima dodatna informacija o digitalnom sadržaju treba sadržavati informacije o krajnjem korisniku a ne o vlasniku sadržaja. Npr. okruženje u kojem se stvaraju filmovi. Za vrijeme produkcije filma, manji dijelovi rada na filmu obično se svaki dan distribuiraju određenom broju ljudi uključenom u stvaranje filma.

Ti dnevni dijelovi filmova su povjerljivi, te ako određena verzija procuri, studio želi imati mogućnost identificirati uzročnika curenja informacija. Problem identificiranja izvora curenja informacija može se riješiti distribuiranjem neznatno različitih kopija svakom primatelju. Svaka kopija jedinstveno je vezana uz osobu koja ju treba primiti.

Drugi primjer primjene je distribucija filmova kinima u digitalnom formatu umjesto korištenja poštanskih usluga i celuloidnih formata. Iako je ovakva distribucija fleksibilnija, efikasnija i jeftinija, producenti i distributeri ne prihvaćaju je jer se boje potencijalnog novčanog gubitka uzrokovanog ilegalnim kopiranjem i redistribucijom filmova. Rješenje ovog problema je da svako kino primi kopiju koja se jedinstveno veže uz kino. U slučaju pojave ilegalnih kopija, može se saznati koje je kino odgovorno te poduzeti potrebne pravne akcije protiv istog.

Povezivanje jedinstvene informacije o svakoj distribuiranoj kopiji digitalnog sadržaja zove se ostavljanje otisaka (eng. *Fingerprinting*). Označavanje vodenim žigovima je adekvatno rješenje za ovu primjenu jer je nevidljivo i nedjeljivo od sadržaja. Ovaj je tip primjene poznat i pod imenom praćenje izdajica (eng. *traitor tracing*) jer je korisno kod praćenja ilegalno proizvedenih kopija digitalnog sadržaja. Ova primjena zahtijeva visoku razinu otpornosti vodenog žiga od različitih vrsta obrade podataka i zlonamjernih napada.

2.3.4 Zaštita autorskih prava

Zaštita autorskih prava jedna je od prvih područja za koja je označavanje digitalnim vodenim žigom namijenjeno. Vodeni žig, u ovom slučaju, sadrži informaciju o vlasniku autorskog prava i neprimjetno se ugrađuje u za to namijenjeni sadržaj. Ako korisnici digitalnog sadržaja imaju lagani pristup detektorima vodenog žiga mogu prepoznati i interpretirati ugrađeni vodeni žig i identificirati vlasnika autorskog prava.

Bilo bi korisno kada bi se ugrađeni vodeni žig mogao koristiti i kao dokaz vlasništva. Može se zamisliti sljedeći scenarij: Vlasnik autorskog prava distribuira svoj digitalni sadržaj s ugrađenim vlastitim nevidljivim vodenim žigom. U slučaju spora oko vlasništva autorskog prava, legalni vlasnik trebao bi moći dokazati svoje vlasništvo. To se ostvaruje tako da stvarni vlasnik predoči originalni dokument i detektor vodenog žiga. Sporni sadržaj je originalni dokument u koji je ugrađen vodeni žig. Detekcijom vodenog žiga vlasnika u spornom dokumentu dokazuje se vlasništvo nad dokumentom. Nažalost gornji scenarij uz određene pretpostavke može biti pobijen a i označavanje vodenim žigom još nije dovoljno pouzdano za dokazivanje vlasništva. Jedan je potencijalni problem povezan s dostupnosti detektora vodenog žiga. Ako je detektor dostupan većem broju ljudi ne može se očuvati sigurnost vodenog žiga. U tom slučaju uvijek je moguće detektirati i ukloniti vodeni žig. To se može napraviti većim brojem neprimjetnih izmjena na označenom sadržaju sve dok detektor više ne može detektirati vodeni žig. Jednom kada je vodeni žig uklonjen originalni vlasnik ne može više dokazati svoje vlasništvo. Čak iako se vodeni žig ne ukloni u nekim uvjetima moguće je dodati novi vodeni žig preko postojećeg i to za sve kopije dokumenta, uključujući originalni dokument. Zbog toga je potrebno moći identificirati prvi, vodeni žig koji je stvarni vlasnik ugradio. Zbog svega toga za ovu primjenu potrebna je najviša razina otpornosti vodenog žiga.

Više o razinama otpornosti i primjenama u [1] i [3].

3. Zaštita digitalne knjižnice otpornim vodenim žigovima

Digitalni vodeni žigovi su neprimjetne, ili vrlo malo vidljive transformacije digitalnih podataka. Iako se digitalne slike najviše povezuju s digitalnim vodenim žigovima, mogu se označavati i drugi oblici digitalnih podataka kao što su video i audio zapisi te tekst.

Termin nevidljivi vodeni žigovi koristi se za opis digitalnih vodenih žigova koji su ljudskom oku nevidljivi, ali koji se mogu izvaditi pomoću računala. Često je za operacije otklanjanja vodenog žiga iz medija potrebno znati odgovarajući lozinku. Samo ovlašteni korisnici mogu otkloniti vodeni žig.

Jedna od najvećih primjena označavanja digitalnim vodenim žigom je zaštita informacije o vlasniku. Ova informacija ima dva oblika: vodeni žig koji identificira osobu koja je stvorila materijal ili korisnike kojima je materijal posuđen.

Ideja označavanja krajnjeg korisnika, odnosno osobe kojoj je određeni materijal posuđen u slučaju knjižnice, jedna je od najvećih primjena označavanja. Mnoge osobe smatraju označavanje primatelja kršenjem privatnosti. Ako primatelj poštuje pravila i dalje ne distribuira ili kopira materijal ne mora se bojati otkrivanja osobnih podataka. Označeni materijal treba biti privatan, odnosno treba ostati kod osobe koja ga je dobila (posudila), bez da ga vide druge osobe. Razotkrivanje identiteta primatelja događa se samo ako osoba ne poštuje pravila. Npr. objavljivanje ili distribucija materijala bez dozvole autora.

Za zaštitu autorskog prava, kao što je prije spomenuto, potrebni su veoma otporni digitalni vodeni žigovi, odnosno potrebno je što više otežati uklanjanje vodenog žiga od strane napadača.

Jedna od primjena je i sprečavanje kopiranja, pogotovo za video zapise. Tako se može svaki film označiti s vodenim žigom koji ima neku od sljedećih vrijednosti: zabranjeno kopiranje, dozvoljeno kopiranje jednom ili zabranjeno daljnje kopiranje. Svaki alat za snimanje morat će moći pročitati ovaj vodeni žig, te odbiti snimati bilo koji film koji ima oznaku zabranjeno snimanje. Velika prednost ove tehnologije je njena neovisnost o tehnologiji, protokolu i formatu distribucije. Vodeni žig je prisutan u bilo kojem trenutku gledanja filma.

Označavanje digitalnim vodenim žigom je područje zanimljivo muzejima, knjižnicama i za industriju zabave jer pruža mogućnost bolje zaštite multimedijskog sadržaja.

Važno je spomenuti da označavanje digitalnim vodenim žigom nije jedina tehnologija za zaštitu autorskog prava. Ona je jedna od 3 tehnologije (druge dvije su enkripcija i digitalni potpis) koje zajedno pružaju razumnu zaštitu autorskih prava za malu cijenu.

Više o sigurnosti digitalne knjižnice moguće je naći u [4].

4. Opis algoritama za označavanje teksta

4.1 Algoritmi za označavanje teksta

Većina organizacija ima potrebu za zaštitom osjetljivih dokumenata. Označavanje digitalnim vodenim žigom jedno je od rješenja ovog problema. Korištenjem digitalnog vodenog žiga moguće je ugraditi otisak u željeni dokument. Otisak može biti jedinstveni identifikacijski broj vlasnika ili primatelja dokumenta. Ugrađeni identifikacijski broj treba se moći detektirati i dekodirati u bilo kojem trenutku, čak i nakon ispisa i skeniranja.

Tehnike za označavanje slika mogu se lagano primijeniti na tekstualni dokument, ali one u tekstualni dokument unose bijeli šum koji se jako primjećuje. Taj šum nastaje zbog binarne (crno-bijele) prirode tekstualnog dokumenta i velike bijele pozadine. Kako bi se izbjegao prethodno spomenuti problem razvijeno je nekoliko tehnika označavanja vodenog žiga posebno za tekstualne dokumente.

Postoje četiri vrste tehnika za označavanje teksta: pomicanje linija teksta (eng. *line-shift coding*), pomicanje riječi unutar iste linije (eng. *word-shift coding*), označavanje značajki teksta (eng. *feature coding*) te jezično označavanje (eng. *natural language NL*) označavanje. Prvu i drugu metodu je opisao Brassil et. al. u [10] i [11].

Kod pomicanja linija teksta svaka parna linija neznatno se pomiče gore ili dolje, ovisno o vrijednosti informacije koja se ugrađuje. Ako je bit jedan odgovarajuća linija pomiče se gore, inače se linija pomiče dolje. Neparne linije su kontrolne linije i one se ne mijenjaju. Koriste se kao reference za mjerenja i uspoređivanje razmaka između linija za vrijeme dekodiranja. Dekodiranje se ostvaruje uspoređivanjem razmaka između baza linija ili razmaka između centroida linija. Baze linija u originalnom dokumentu su obično uniformno raspoređene dakle originalan dokument nije potreban ako se bazne linije koriste. Ali centroidi nisu nužno uniformno raspoređeni pa je potreban originalni dokument kod metoda koje koriste centroide.

Kod druge metode, pomicanja riječi, prvo se svaka linija dijeli u grupe riječi. Svaka grupa ima dovoljan broj znakova. Zatim se svaka parna grupa pomiče u lijevo ili desno, ovisno o vrijednosti specifičnog bita informacije koji se ugrađuje. Neparne grupe koriste se kao reference za mjerenje i uspoređivanje razmaka između riječi za vrijeme dekodiranja. Metoda korelacije i metoda centroida koriste se za detekciju vodenog žiga i obje metode zahtijevaju originalni tekst.

Treća metoda odnosi se na mijenjanje određenih značajki teksta (boje, fonta, veličine, itd.).

Kod četvrte metode, jezično označavanje, ugrađivanje se izvodi mijenjanjem sintakse ili semantike odabranih rečenica.

Poglavlje 4.2 opisuje algoritam koji modificira razmak između riječi i širinu riječi, tako da prosječni razmak svake linije predstavlja uzorak vala sinusa specifične faze i frekvencije. Poglavlje 4.3 opisuje algoritam koji mijenja razmak između riječi ili između linija. Poglavlje 4.4 opisuje primjer označavanje značajki teksta,a poglavlje 4.5 opisuje jezično označavanje.

4.2 Označavanje slika teksta pomoću valova sinusa koji reprezentiraju razmake između riječi

4.2.1 Uvod

Ova metoda [5] koristi jednu od značajki tekstualnog dokumenta, a to su razmaci između riječi za označavanje tekstualnog dokumenta. Tehnika kodiranja podešava razmake između riječi tako da srednji razmaci u različitim linijama pokazuju karakteristike funkcije sinus, a informaciju se može ugraditi u val ili valove sinusa. S obzirom da se označava u horizontalnom i vertikalnom smjeru ovakvo označavanje je otporno na vanjske utjecaje. Nadalje, do pohranjene informacije može se doći s ili bez originalnog dokumenta, a kontrolne liniji ili kontrolni blokovi nisu potrebni za proces detekcije.

4.2.2 Značajke razmaka i statistika

Stranica teksta u digitalnom obliku može biti prikazana sljedećom funkcijom:

$$f(x, y) \in [0,1], x = 0,1,...,W, y = 0,1,...,L \tag{4.1}$$

koja reprezentira bijele i crne piksele. U ovoj funkciji *W* predstavlja širinu stranice, dok *L* predstavlja duljinu stranice u pikselima.

U digitalnoj obradi slika razmak između riječi se detektira pomoću sljedeće vertikalne projekcije:

$$v(x) = \sum_{y=t}^{b} f(x, y) \tag{4.2}$$

koja je suma crnih piksela u vertikalnom stupcu od *t* (vrha) do *b* (dna) linije teksta. Ako ne postoji crni piksel u *x* uzastopnih piksela, odnosno:

$$v(x) = 0, x = k, k+1,..,k+c \tag{4.3}$$

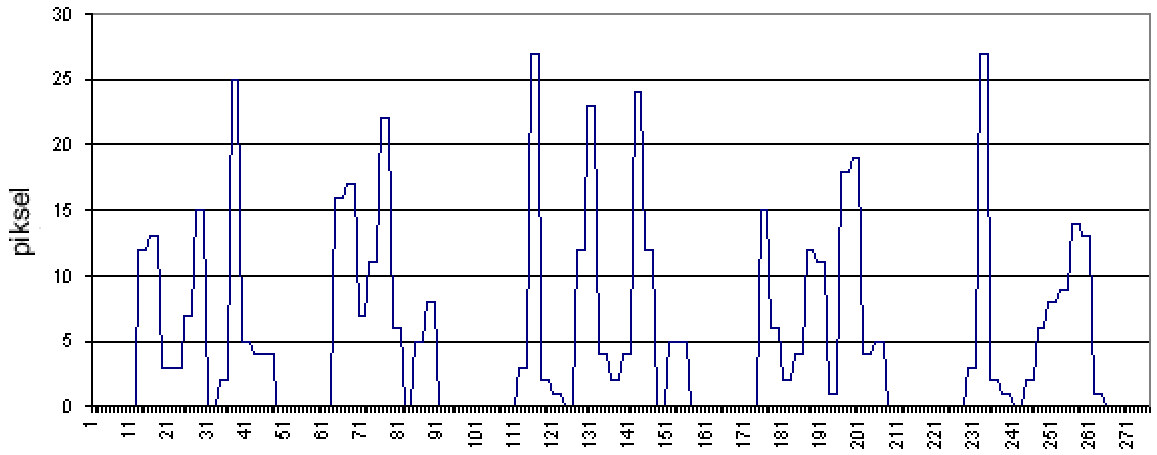
detektiran je razmak između riječi. Slika 4.1 prikazuje tipičan vertikalan profil pet riječi.

Prosječan razmak u liniji teksta može biti parametar za proučavanje značajki razmaka tekstualnog dokumenta. Za liniju s *d* riječi srednji razmak računa se kao:

$$S_a = S_t /(d-1), \quad d \neq 1 \tag{4.4}$$

gdje je S_t ukupna suma razmaka u liniji teksta, u pikselima

Učestale su dvije vrste teksta. Jedan je poravnat na lijevu marginu, dok je drugi jednoliko poravnat i s lijeve i s desne strane (eng. *justified*). Ovaj algoritam označava tekst poravnat s obje strane.



Slika 4.1 Vertikalan profil 5 riječi

4.2.3 Označavanje razmaka

S obzirom na slučajan raspored prosječnih razmaka linije teksta u tekstualnom dokumentu, definira se diskretna slučajna varijabla *X(n):*

$$X(n) = S_{an}, \quad n = 0,1,...,N-1 \tag{4.5}$$

gdje *n* predstavlja broj linije teksta u tekstualnom dokumentu s *N* linija. S_{an} predstavlja S_a (jednadžba (4.4)) n-te linije. Označavanje razmaka može se gledati kao označavanje slučajne varijable *X(n).*

Funkcija sinus koja se mijenja preko linija teksta ima neke zanimljive karakteristike:

1. Sinus se mijenja polako tako da se lokalne varijacije ne zamjećuju
2. Amplituda, frekvencija i početni kut sinusa mogu se koristiti za pohranu informacija
3. Periodična simetričnost sinusa čini proces detekcije lakšim i pouzdanim.

Mogu se koristiti različite linije teksta iz određenog dokumenta za ugrađivanje informacija. Vrijednosti S_a različitih linija teksta mogu se koristiti kao uzorci za vrijednosti sinusa.

Kako bi označavanje razmaka bilo neprimjetno razlike u razmacima između riječi moraju biti minimalne, odnosno promjena razmaka mora biti dovoljno velika da se može pravilno detektirati. Zbog ovih ograničenja postoji uski raspon amplituda sinusnog vala za označavanje.

Za ispravnu rekonstrukciju vala sinusa, frekvencija uzorkovanja mora biti bar dva puta veća od frekvencije sinusa. Postoje određene frekvencije koje ljudski vizualni sustav više primjećuje te treba izbjegavati označavanje u blizini tih frekvencija. Zbog toga je frekvencija vala sinusa također ograničena.

Početni kut vala sinusa bira se kao primarni nosilac informacije.

Kod ove metode riječ se ne pomiče samo horizontalno, nego se po potrebi sužava ili širi tako da se postigne traženi S_a.

Neka novi prosječni razmak nakon modifikacije razmaka između riječi u liniji teksta treba biti S_a'. Tada je promjena ukupne duljine razmaka između riječi u pikselima:

$$S_{tc} = (S_a' - S_a)(d-1) \tag{4.6}$$

gdje je d broj riječi, a S_a originalan prosječan razmak u liniji teksta opisan u jednadžbi (4.4).

Ako je $S_{tc} > 0$ tada će se ukupan razmak između riječi raširiti, a riječi u ovoj liniji će se skupiti. Ako je $S_{tc} < 0$ tada će se ukupan razmak između riječi u liniji teksta smanjiti, a riječi će se raširiti.

Neka je i-ta riječ ove linije ima širinu prije modifikacije Pxl_i u pikselima, tada se skupljanje ili širenje širine ove riječi u pikselima računa kao:

$$ES_i = \left\lfloor \frac{S_{tc}}{\sum_{i=1}^{d} Pxl_i} Pxl_i \right\rfloor, \text{ ako je } S_{tc} \geq 0 \tag{4.7}$$

$$ES_i = \left\lceil \frac{S_{tc}}{\sum_{i=1}^{d} Pxl_i} Pxl_i \right\rceil, \text{ ako je } S_{tc} < 0 \tag{4.8}$$

ES_i se zaokružuje na najbliži cijeli broj, s obzirom da predstavlja broj piksela. Dakle, može postojati razlika između S_{tc} i sume ES_i, koja se računa kao:

$$S_d = S_{tc} - \sum_{i=1}^{d} ES_i \tag{4.9}$$

U ovoj implementaciji razlika Sd se pribraja najvećem ESi.

Skupljanje ili širenje riječi ostvaruje se brisanjem ili kopiranjem svakog Iv_i–tog stupca. Interval Iv_i se računa kao:

$$Iv_i = \left\lfloor \frac{Pxl_i}{|ES_i|} \right\rfloor \tag{4.10}$$

Interval Iv_i zaokružuje se na cijeli broj. Nakon skupljanja ili širenja određene riječi, nova širina u pikselima računa se kao:

$$Pxl_i' = Pxl_i - ES_i \tag{4.11}$$

Dvije strane linije teksta ne mijenjaju se dok se linija skuplja ili širi. Za skupljanje ili širenje riječi, kod riječi s lijeve strane linije lijevi rub riječi je fiksan, dok se riječ skuplja ili širi. Ako su riječi s desne strane linije, desna strana riječi se drži fiksna dok se riječ širi ili skuplja.

Radno okruženje je jedna stranica teksta ili više stranica koje čine jedan dokument. Relevantne linije teksta u radnom okruženju su uzorci za sinus za označavanje. Početni kut može biti ili apsolutni početni kut ili relativan početni kut, ako se koristi više različitih valova.

Za ovu metodu označavanja razvijeni su privatni i javni algoritmi za označavanje.

4.2.4 Privatno označavanje

1. Računa se srednja vrijednost *Sa*

$$a_1 = \frac{\sum_{n=p}^{q} S_{an}}{q - p + 1}, \quad 0 \le p < q < N \tag{4.12}$$

   gdje su *p* i *q* indeksi prve i zadnje linije teksta u radnom okruženju u koje se označavaju sinusnim valom.

2. Za svaku liniju računa se komponenta vodenog žiga koja je određena sljedećim valom sinusa:

$$W_n = C_1 a_1 \sin(\omega_1 (n - p) + \phi_1) \tag{4.13}$$

   gdje je W_n željena komponenta vodenog žiga za privatno označavanje *n*-te linije teksta; ω_1 i ϕ_1 su frekvencija u radijanima i početni kut vala sinusa. C_1 je konstanta koja određuje amplitudu sinusa.

3. W_n se dodaje S_a za *n*-tu liniju te se generira novi prosječni razmak:

$$S_{an}' = S_{an} + W_n \tag{4.14}$$

4. Na kraju riječi svake od odabranih linija modificiraju se primjenom formula (4.6) do (4.11).

Privatna metoda može se shvatiti kao dodavanje konstantnog dijela originalnoj slučajnoj varijabli *X(n)*, te se tako kreira slučajna varijabla *Y(n)*

$$Y(n) = X(n) + W_n \tag{4.15}$$

gdje je *Y(n)* slučajna varijabla za privatno označavanje, a W_n vodeni žig za privatno označavanje.

4.2.5 Javno označavanje

Kod privatnog označavanja susjedne linije teksta imaju slučajne vrijednosti S_a, Kod javnog označavanja vrijednosti S_a linija koje se koriste kod javnog označavanja trebaju imati određenu vezu kako bi se mogle koristiti direktno kao uzorci za val sinusa.

Neprikladno je uzimati sve linije teksta tekstualnog dokumenta za javno označavanje zbog varijacija u S_a kod originalnih linija teksta. Promatranjem različitih profila S_a vidljivo je da linije s velikim brojem riječi imaju bliske vrijednosti S_a. Ovo je pogodno iz dva razloga. Prvo, u liniji teksta s velikim brojem riječi, prosječnoj riječi i odgovarajućem razmaku dodijeljen je i manji broj piksela. Dakle razlika između S_a susjednih linija je manja. Drugo, linija teksta s većim brojem riječi ima manju vjerojatnost da bude poravnata s obje strane ili je to poravnanje manje vidljivo.

1. S obzirom na prethodno opisana opažanja prvo se bira ključ tako da se linije čiji je broj riječi veći ili jednak ključu označavaju.
2. Nakon toga bira se skup linija S_w iz dokumenta tako da broj riječi svake linije nije manji od izabranog ključa.
3. Računa se srednja vrijednost S_a za svaku od linija iz skupa S_w:

$$a_2 = \frac{\sum_{m=u}^{v} S_{am}}{v-u+1}, \quad 0 \le u < v < N \tag{4.16}$$

   gdje u i v imaju slično značenje kao i p i q u jednadžbi, ali u i v su indeksi linija iz skupa S_w; m je indeks linije teksta iz skupa S_w, a S_{am} je S_a m-te linije.
4. Za svaku liniju teksta iz S_w računa se komponenta vodenog žiga određena valom sinusa:

$$W_m = C_2 a_2 \sin(\omega_2 (m-u) + \phi_2) \tag{4.17}$$

   W_m je željena komponenta vodenog žiga za javno označavanje m-te linije; ω_2 i ϕ_2 su frekvencija u radijanima i početni kut sinusa.
5. Za svaku liniju iz S_w, S_a zamjenjuje se sumom a_2 i W_m te se tako generira novi razmak:

$$S_{am}' = a_2 + W_m, \text{ ako je m-ta linija} \in S_w, \quad \text{inače nema izmjena} \tag{4.18}$$

6. Na kraju sve linije teksta mijenjaju se prema jednadžbama (4.6) do (4.11). Dakle za linije iz skupa S_w dobiva se nova slučajna varijabla za javno označavanje $Y(m)$:

$$Y(m) = a_2 + W_m \tag{4.19}$$

4.2.6 Detekcija i svojstva

Ako je tekst označen privatnom metodom, slučajna varijabla *Y(n)* dobiva se rekonstrukcijom S_a prema jednadžbi (4.4). S originalnim neoznačenim tekstom komponenta vodenog žiga W_n za privatno označavanje iz jednadžbe (4.15) računa se kao:

$$W_n = Y(n) - X(n) \tag{4.20}$$

Ako je tekst označen javnom metodom i ako se pretpostavi da je ključ poznat, moguća je rekonstrukcija skupa S_w kao i ponovno računanje a_2 iz jednadžbe (4.16). Komponenta vodenog žiga W_m za javno označavanje iz jednadžbe (4.19) računa se kao:

$$W_m = Y(m) - a_2 \text{, za linije teksta iz } S_w \tag{4.21}$$

Originalan početni kut detektira se računanjem unakrsne korelacije (eng. *cross-corelation*) detektirajućeg vala sinusa s W_n (vodeni žig za privatno označavanje) ili W_m (vodeni žig za javno označavanje):

$$r(j) = \frac{1}{T}\sum_{n=0}^{T-1} W(n) A_d \sin(\omega_d n + j), \tag{4.22}$$

gdje *W* predstavlja W_n ili W_m; ω_d je frekvencija u radijanima detektirajućeg sinusnog vala; a *j* predstavlja vremenski pomak u broju linija teksta i varira kako bi se detektirala označena informacija. Kroz *j* koji stvara ekstremnu vrijednost *r(j)* obnavlja se originalna označena informacija. A_d je amplituda detektirajućeg sinusnog vala. *T* je sumarni broj koji ovisi o broju stavki u W_n ili W_m kao i ω_d.

Jedan od parametara korišten u eksperimentima je broj uzoraka (eng. *half wave sampling points*), odnosno broj linija za označavanje *N* u jednadžbama (4.13) i (4.17) za koji vrijedi:

$$0 \le \omega N < \pi \text{, gdje } \omega \text{ predstavlja } \omega_1 \text{ ili } \omega_2 \tag{4.23}$$

Rezultati su prikazani u tablicama 1 i 2.

Iz eksperimenata je vidljivo da se razmak između riječi u tekstualnim dokumentima može označiti vodenim žigom bez većih vidljivih izmjena te se isti može ispravno detektirati.

Tablica 4.1 Rezultati detekcije za privatno označavanje

	Broj uzoraka			
	10	7	5	3
Točnost	20/20	20/20	20/20	20/20

Tablica 4.2 Rezultati detekcije za javno označavanje

	Broj uzoraka			
	7	6	5	3
Točnost	14/15	15/15	14/15	21/21

4.2.7 Zaključak

Razmak je jedinstvena karakteristika tekstualnog dokumenta. Prethodno je opisan novi algoritam za označavanje teksta digitalnim vodenim žigom korištenjem razmaka između riječi. Opisana metoda neznatno mijenja razmak između riječi tako da su različite linije iz teksta uzorci za val sinusa. Preliminarni testovi pokazali su obečavajuće rezultate. Ova metoda može se primijeniti i na javno i na privatno označavanje. Ugrađivanje informacije u horizontalnom i vertikalnom smjeru čini ovu metodu otporniju na vanjske utjecaje.

4.3 Označavanje elektroničkih tekstualnih dokumenata i slika teksta pomicanjem riječi ili linija

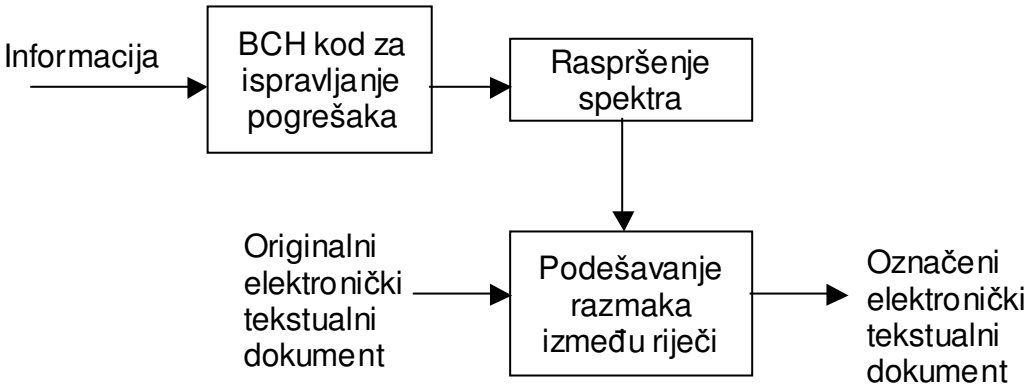
4.3.1 Uvod

Ovo poglavlje opisuje metodu [6] za označavanje elektroničkih tekstualnih dokumenata koja je slična postojećim metodama koje pomiču riječi i linije teksta. Ali za razliku od postojećih metoda za detekciju vodenog žiga nije potreban originalan dokument. Također metoda se može primijeniti na dokumente koji su poravnati na lijevi rub, desni rub ili na oba ruba te na pravilan i nepravilan razmak između linija teksta. Odlomci poravnati na obje strane vrlo su učestali u elektroničkim dokumentima. Kako bi "prisilili" krajeve posljednje riječi da bude točno na desnoj margini pojedini jezični procesori automatski i sistematično šire riječi unutar pojedine linije. Nepravilan razmak između linija rezultat je umetanja matematičkih simbola, slova koja su ispod ili iznad normalnih slova (eng. *super-* ili *sub-script*) ili drugih objekata. Kako bi se prilagodilo najvišem objektu svake linije jezični procesor automatski podešava razmak između linija koliko je potrebno.

U sljedećim poglavljima opisan je algoritam za označavanje i detekciju, eksperimentalni rezultati te zaključak.

4.3.2 Algoritam za označavanje

Kod ove metode koriste se tehnike raspršenja spektra i BCH tehnike kodiranja pogrešaka. Te se tehnike koriste kao odgovor na efekte koje uzrokuju nepravilan razmak između riječi ili linija tekstualnih dokumenata. Tehnika koja se predlaže za nepravilan razmak između linija vrlo je slična onoj za nepravilan razmak između riječi. Te se zbog izbjegavanja redundancije opisuje ona za razmak između riječi.



Slika 4.2 Ugrađivanje informacija u elektronički tekstualni dokument

Slika 4.2 prikazuje proces ugrađivanja informacije kod predloženog algoritma. Proces ugrađivanja započinje upotrebom BCH tehnike za ispravljanje pogrešaka radi zaštite informacije od šuma. Nakon toga koristi se jedinstvena m-sekvenca za raspršenje svakog od kodiranih bitova informacije. Tako raspršeni bitovi ugrađuju se u tekstualni dokument neznatno povećavajući ili smanjujući razmak između riječi. Dokument se pregledava od početka do kraja te se svaki razmak neznatno povećava ili smanjuje za malu delta vrijednost ovisno o vrijednosti odgovarajućeg bita vodenog žiga. Ako je bit nula razmak se smanjuje, ako je bit jedan razmak se povećava.

Delta određuje pomak, rastom delte raste i snaga vodenog žiga. Ali mora se voditi računa da delta bude dovoljno mali tako da unatoč povećanju ili smanjenju razmaka riječi ostanu odvojene jedna od druge.

Tipičan tekst s dvostrukim proredom pisan je u fontu *Times New Roman*, veličine 11 na stranici od 8.5x11 inča ima oko 25 linija. Svaka linija ima u prosjeku 13 riječi. Znači, svaka linija ima 12 razmaka između riječi, što znači da ima približno 300 razmaka između riječi po stranici.

Ako se koristi 16-bitna sekvenca za raspršenje svakog bita informacije, tada se 18-bitna informacija može pohraniti u svakoj stranici. Ova veličina informacije dovoljna je za 262144 različitih identifikacijskih brojeva, ali je dokument podložan šumu koji uzrokuje ispisivanje i skeniranje. Korištenjem BCH koda štiti se informacija od šuma, ali i značajno smanjuje broj dozvoljenih identifikacijskih brojeva.

Tablica 4.3 prikazuje veličinu dozvoljene informacije i broj grešaka koje se mogu ispraviti različitim BCH kodovima. Teoretski (15,5) BCH kod pruža najveću zaštitu jer može ispraviti najveći broj grešaka, ali ovaj kod također dozvoljava i najmanji broj bitova informacije. (15,5) BCH kod može ispraviti do tri greške, ali dozvoljava samo 32 različita identifikacijska broja.

S druge strane (15,11) BCH kod dozvoljava najveću veličinu informacije, ali ispravlja i najmanji broj pogrešaka. Ovaj kod može ispraviti samo jednu grešku, ali dozvoljava 2048 različitih identifikacijskih brojeva. (7,4) BCH kod predstavlja ravnotežu između zaštite i veličine informacije. Taj kod može ispraviti jednu pogrešku u svaka četiri bita.

Kada se spoji dva (7,4) koda za zaštitu osam bitova, mogu ispraviti dvije greške te se dozvoljava ugrađivanje 256 različitih identifikacijskih brojeva.

Tablica 4.3 Veličina informacije, broj dozvoljenih identifikacijskih brojeva i broj bitova koji se mogu ispraviti za različite BCH kodove

BCH kod	Duljina	Veličina informacije	Paritetni bitovi	Broj grešaka koje se mogu ispraviti	Broj identifikacijskih brojeva
(7,4)	7	4	3	1	16
(15,11)	15	11	4	1	2048
(15,7)	15	7	8	2	128
(15,5)	15	5	10	3	32

Stranica s jednostrukim proredom ima približno dva puta više razmaka nego ona s dvostrukim. Takva stranica dozvoljava ugrađivanje dvostruko većeg broja bitova, što znatno povećava broj različitih identifikacijskih brojeva.

Broj različitih identifikacijskih brojeva može se povećati korištenjem više stranica teksta za označavanje jednog identifikacijskog broja. Ali ovo poboljšanje komplicira proces dekodiranja.

16-bitni kod za raspršenje spektra može se generirati 4-bitnim posmačnim registrom. Taj kod generira m-sekvencu periode 16, koja se označava s *m(n)*. Ta m-sekvenca ima odgovarajuće korelacijske osobine za upotrebu s detektorom baziranim na korelaciji. Kod raspršenja, *c(n)*, generira se iz *m(n)* na sljedeći način:

$$c(n) = 2m(n) - 1 \tag{4.24}$$

Time se raspon m-sekvence mijenja iz {0,1} na {-1,1}. Ako se svaki bit kodirane informacije označi s $b \varepsilon \{-1,1\}$, tada se primjena tehnike raspršenja spektra na kodiranu informaciju opisuje s:

$$w(n) = b \times c(n) \tag{4.25}$$

gdje je *w(n)* 16-bitna sekvenca raspršenja spektra koja predstavlja bit *b*.

4.3.3 Označavanje elektroničkog dokumenta

Kod podešavanja razmaka između riječi u stvarnom vremenu i kod dokumenta poravnatog s obje strane često sam jezični procesor podešava razmake između riječi kako bi se očuvalo poravnanje. Ekstreman slučaj je kada ta automatska podešavanja pomaknu zadnju riječ trenutne linije u novu liniju. Ovaj slučaj moguće je izbjeći ako se podesi razmak između svake riječi u svakoj liniji.

Ako $ows_{i,j}$ predstavlja širinu *j*-tog originalnog razmaka između riječi i-te linije, a $nws_{i,j}$ predstavlja novu širinu nakon označavanja, tada je zbroj ovih širina prije, odnosno poslije označavanja:

$$ows_i = \sum_{j=1}^{N_i} ows_{i,j}$$
$$nws_i = \sum_{j=1}^{N_i} nws_{i,j} \tag{4.26}$$

gdje je N_i broj razmaka između riječi *i*-te linije. Ako $wl_{i,j}$ predstavlja širinu *j*-te riječi *i*-te linije, tada je zbroj širina svih riječi te linije:

$$swl_i = \sum_{j=1}^{N_i+1} wl_{i,j} \tag{4.27}$$

Kao kompenzacija razlike između nsw_i i osw_i širina svake riječi mora biti podešena na:

$$wl_{i,j}' = wl_{i,j} + \Delta_{i,j} \tag{4.28}$$

gdje $\Delta_{i,j}$ predstavlja vrlo mali broj dobiven sljedećom jednadžbom:

$$\Delta_{i,j} = \begin{cases} \left\lfloor (nsw_i - osw_i)\dfrac{wl_{i,j}}{swl_{i,j}} \right\rfloor & ako\ je\ (nsw_i - osw_i) \geq 0 \\ \left\lceil (nsw_i - osw_i)\dfrac{wl_{i,j}}{swl_{i,j}} \right\rceil & ako\ je\ (nsw_i - osw_i) < 0 \end{cases} \tag{4.29}$$

Opisani proces označavanja može se implementirati na razini upravljačkog programa (eng. *driver-a)* za *postscript* pisač. U tom slučaju, upravljački program za pisač stvara *postscript* dokument koji sadrži instrukcije koje opisuju stranicu. Pisač interpretira te instrukcije te ispravno ispisuje stranicu. Pisač ispisuje označeni dokument korištenjem izmijenjenih instrukcija u *postscript* dokumentu kao što je prethodno opisano.

4.3.4 Označavanje ispisanog dokumenta

Označavanje ispisanog dokumenta teže je nego označavanje elektroničkog dokumenta. Taj proces sličan je procesu opisanom u sljedećem poglavlju o detekciji vodenog žiga u ispisanom dokumentu. U tom procesu ispisani dokument prvo se skenira. Nakon skeniranja dokument se obrađuje procesorom za obradu slika kako bi se identificirale linije i razmaci između riječi. Kada su ti razmaci identificirani svaka riječ se neznatno pomiče. Isti proces koristi se i prilikom identifikacije i izmjene

razmaka između riječi smanjivanjem ili povećavanjem riječi, radi održavanja poravnanja. Prilikom ovog procesa posebno je važno ne unositi dodatan šum.

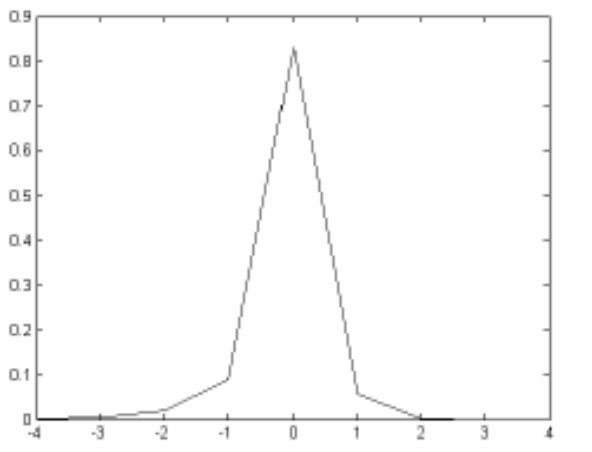
4.3.5 Detekcija vodenog žiga u elektroničkom dokumentu

Detekcija vodenog žiga jednostavan je proces. Detekcija počinje mjerenjem i bilježenjem razmaka, $nws_{i\,j}$ (širina razmaka između riječi nakon označavanja), između dvije uzastopne riječi. Prije označavanja, $ows_{i,j}$ (širina razmaka prije označavanja) svake linije približno je jednak. Jezični procesor neznatno podešava ove razmake u svrhu poravnanja. Dakle, srednja vrijednost $\overline{nws}_{i,j}$ dobra je procjena $ows_{i,j}$. Zbog toga se $\overline{nws}_{i,j}$ računa i oduzima od svakog od zabilježenih razmaka radi procjene *n*-tog uzorka, *w(n),* signala vodenog žiga. Rezultirajuće procjene vodenog žiga *w(n)'* segmentiraju se u segmente od 16 uzoraka svaka. Svaki od ovih segmenata korelira s originalnom m-sekvencom za dohvat bita informacije. Na kraju se izvodi BCH dekodiranje na bitovima informacije radi ispravljanja pogrešaka.

Procjena, *w(n)'*, *n*-tog uzorka signala vodenog žiga može se izraziti:

$$w(n)' = w(n) - (w(n+1) + w(n-1))/2 + \phi(n) \tag{4.30}$$

gdje $\phi(n)$ predstavlja slučajan šum. Za elektroničke dokumente $\phi(n)$ je šum koji nastaje zbog nepravilnog razmaka između riječi prije označavanja. Slika 4.3 prikazuje vjerojatnosnu razdiobu $\phi(n)$ za poravnani tekst veličine 11, *Times New Roman*. Iz slike je vidljivo da je šum $\phi(n)$ srednje vrijednosti nula, Gaussov šum s varijancom od 0.23. Potrebno je spomenuti da je $\phi(n)$ nula za ne poravnati tekst. Za skenirani dokumenti $\phi(n)$ također uključuje šum dobiven ispisom i skeniranjem.



Slika 4.3 Vjerojatnosna funkcija šuma ϕ(n), zbog poravnavanja linije

Izraz $(w(n+1) + w(n-1))/2$ još je jedan izvor šuma koji ne bi trebao imati utjecaja na detekciju. Zamjenom *w(n)* iz jednadžbe (4.25) u jednadžbu (4.30) dobiva se:

$$w(n)' = bc(n) - (bc(n+1) + bc(n-1))/2 + \phi(n) \tag{4.31}$$

Primjenom korelacijskog detektora na jednadžbu (4.31) dobiva se:

$$\sum_{n=1}^{N} w(n)'c(n) = b\sum_{n=1}^{N} c(n)c(n) - \frac{1}{2}\sum_{n=1}^{N} bc(n+1)c(n) + \frac{1}{2}\sum_{n=1}^{N} bc(n-1)c(n) + \sum_{n=1}^{N} \phi(n)c(n) \qquad (4.32)$$

gdje je *N* duljina koda za raspršenje *c(n)*. S obzirom da je *c(n)* m-sekvenca, drugi i treći dio desne strane jednadžbe (4.32) prelaze u nulu. Ovakav rezultat se dobiva jer je autokorelacija m-sekvence delta funkcija. Zadnji izraz s desne strane jednadžbe (4.32) reprezentira šum male magnitude *η(n)*. Time se jednadžba (4.32) pojednostavljuje na:

$$\sum_{n=1}^{N} w(n)'c(n) = b + \eta(n) \qquad (4.33)$$

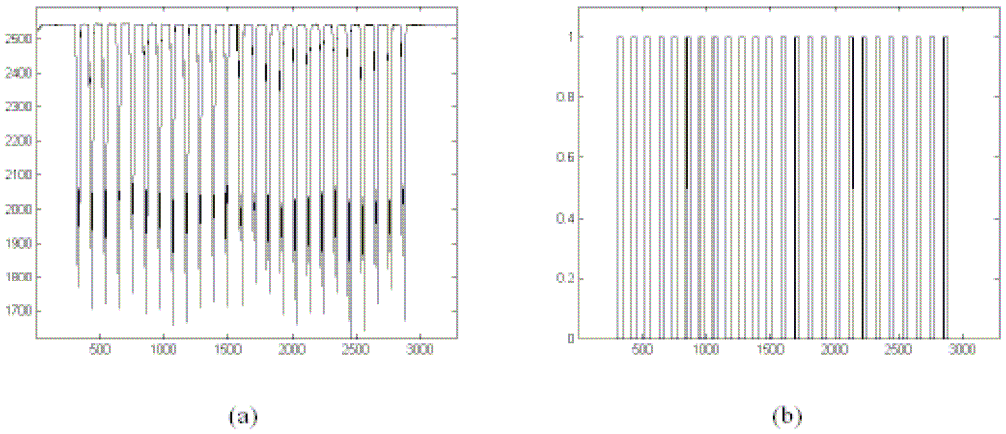
Vrijednost *b* vodenog žiga dobiva se primjenom sljedećeg praga na rezultate korelacije:

$$b = \begin{cases} 1 & \sum_{n=1}^{N} w(n)'c(n) \geq 0 \\ -1 & \sum_{n=1}^{N} w(n)'c(n) < 0 \end{cases} \qquad (4.34)$$

4.3.6 Detekcija vodenog žiga u ispisanom dokumentu

Detekcija vodenog žiga u ispisanom dokumentu malo je zahtjevnija. Proces se može opisati sljedećim koracima:

1. Skenirati dokument tako da bude prihvatljive kvalitete i rezolucije. Što je viša kvaliteta i rezolucija bolji su i rezultati detekcije.
2. Pretvoriti sliku u binarnu sliku korištenjem odgovarajućeg praga. Vrijednost praga može se jednostavno odrediti iz histograma slike, koji je bimodalan. Vrijednostima višima od praga dodjeljuje se vrijednost 1, a vrijednostima ispod praga 0. Dakle, tekst će imati vrijednost nula.
3. Ispraviti bilo kakvo odstupanje između orijentacije skeniranog dokumenta i elektroničkog dokumenta. Detektor može uzeti smjer linija skeniranog dokumenta kao početnu vrijednost. U željenoj aplikaciji korisnik treba paziti da ispravno postavi dokument u skener. Time se uzrokuju samo neznatna odstupanja u orijentaciji, koja se lagano ispravljaju.



Slika 4.4 (a) Vertikalni profil tipičnog tekstualnog dokumenta, (b) lokacije linija

4. Dohvatiti linije skeniranog dokumenta. To se može postići računanjem vertikalnog profila, gdje je vertikalni profil *v(i)*, slike *I(i,j)*:

$$v(i) = \sum_{j=1}^{W} I(i, j) \tag{4.35}$$

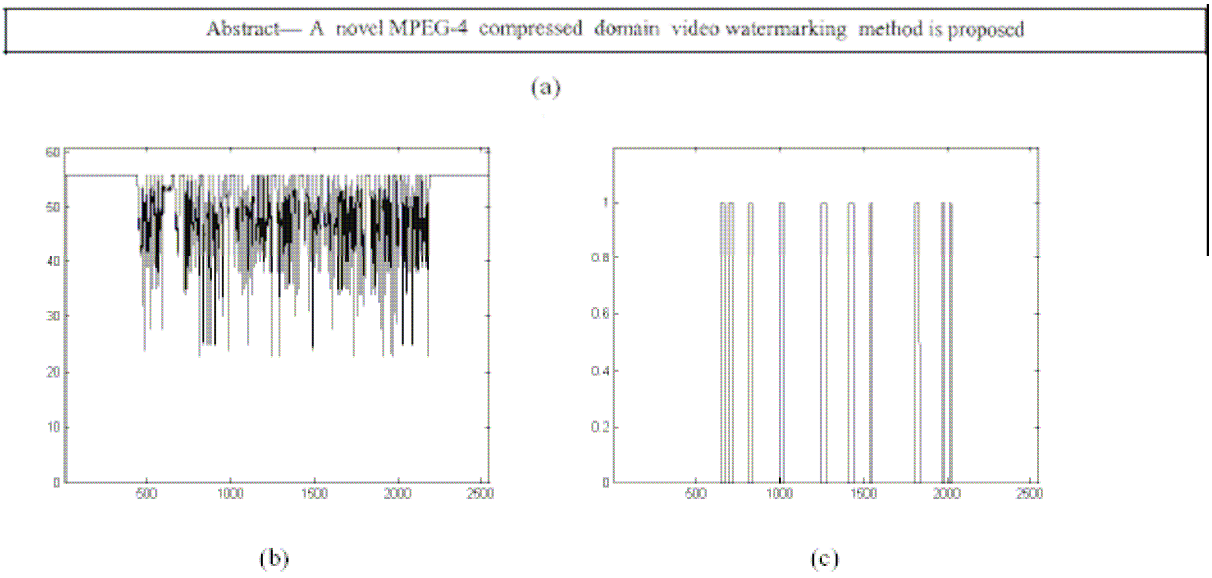
gdje je *W* širina slike *I(i,j)*. Slika 4.4(a) prikazuje vertikalni profil tipičnog tekstualnog dokumenta skeniranog s 300 DPI te pretvorenog u crno-bijelu sliku. Slika 4.4(b) prikazuje lokacije dohvaćenih linija istog dokumenta. Lokacije su određene uspoređujući profil *v(i)* sa zadanim pragom te bilježenjem lokacija dolina.

5. Detektirati i dohvatiti razmake između svake dvije uzastopne riječi. Ovaj korak postiže se računanjem horizontalnog profila *h(j)*, malog dijela slike *S(i,j)* oko svake linije:

$$h(j) = \sum_{i=1}^{H} S(i, j) \tag{4.36}$$

gdje *H* predstavlja visinu djelića slike *S(i,j)*. Slika 4.5 (a) prikazuje segment slike oko linije u tipičnom tekstualnom dokumentu. Slika 4.5 (b) prikazuje horizontalan profil iste. Iz tog horizontalnog profila mogu se izračunati razmaci između riječi, detektiranjem vršnih vrijednosti horizontalnog profila.

Vrlo široke i vrlo uske vršne vrijednosti ignoriraju se. Vrlo široke vršne vrijednosti predstavljaju margine stranice, a vrlo uske vršne vrijednosti razmak između slova u riječi. Slika 4.5 (c) prikazuje lokacije detektiranih razmaka između riječi prikazanih u (a) dijelu.



Slika 4.5 (a) Mali dio slike oko linije tipičnog tekstualnog dokumenta, (b) Horizontalni profil segmenta slike prikazanog u (a) dijelu, (c) lokacije dohvaćenih razmaka

6. Povezati sve izmjerene razmake iz svih linija te nastaviti kao u dijelu o detekciji u elektroničkom dokumentu; povezati svakih 16 segmenata sa sekvencom za raspršenje za dobivanje bita vodenog žiga; dekodirati bitove vodenog žiga s BCH dekoderom za ispravljanje eventualnih pogrešaka i dobivanje bitova informacije.

4.3.7 Eksperimentalni rezultati

Iz ostvarene implementacije algoritma opisanog u poglavljima o označavanju i detekciji te s različitim konfiguracijama iste ostvarena su sljedeća zapažanja.

U jednoj od konfiguracija korištena je 32-bitna m-sekvenca i 8-bitna informacija (256 različitih identifikacijskih brojeva), ali bez kodova za ispravljanje pogrešaka. 8-bitna informacija dovoljna je za praćenje dokumenta i identifikaciju originalnog primatelja za organizaciju s 256 zaposlenika.

Gore spomenuta konfiguracija testirana je ugrađivanjem informacije u tekstualni dokument veličine 11 *Times New Roman*, s 256 različitih identifikacijskih brojeva. Korištenjem detekcijskog algoritma za detekciju svakih od 256 identifikacijskih brojeva iz označenih, neizmijenjenih tekstualnih dokumenata ustanovljeno je da je uspješnost detekcije 98.8%. Svaki od pogrešno detektiranih identifikacijskih brojeva ima samo jedan bit greške. Tablica 4.4 prikazuje slučajeve pogrešne detekcije te uspoređuje originalno ugrađene vrijednosti s pogrešno detektiranima. Tablica također prikazuje bit pogreške u svakom od slučajeva. S obzirom da je greška bila u samo jednom bitu, upotrebom bilo kojeg od prije spomenutih BCH kodova za ispravljanje pogrešaka sve greške bile bi otklonjene.

Slična uspješnost detekcije postignuta je i kada je algoritam testiran i na izmijenjenom tekstu. Izmjene su uključivale promjenu fonta i veličine slova; izmjenu riječi, izmjenu poravnanja te mijenjanje lijeve i desne margine stranice.

Tablica 4.4 Pogrešno detektirane vrijednosti i njihove originalno ugrađene vrijednosti

Detektirana vrijednost		Ugrađena vrijednost	
Decimalno	Binarno	Decimalno	Binarno
9	00001001	41	00101001
25	00011001	57	00111001
145	10010001	209	11010001

Operacije izmjene teksta kao što su brisanje i umetanje riječi postigle su dvojake rezultate. U većem broju detekcija je bila uspješna jer je tehnika raspršenja spektra otporna na lokalne greške, pogotovo ako se greške pojavljuju pri kraju sekvence za raspršivanje. U ovom slučaju većina sekvence za raspršivanje ostala je nepromijenjena pa je detektor uspio detektirati ugrađeni bit. Ali, ako je greška bliže sredini niti jedan od dijelova nije dovoljan za ispravnu detekciju ugrađenog bita.

U drugoj konfiguraciji, korišten je (15,7) BCH kod za zaštitu dokumenata od dvostrukih grešaka. Konfiguracija zahtijeva smanjivanje sekvence za raspršivanje s 32 na 16 bitova za 15 bitni kod u stranici s dvostrukim proredom. Ovaj korak nije potreban kod stranice s jednostrukim proredom, jer takva stranica ima dovoljan broj razmaka. 7-bitna informacija dovoljna je za praćenje i identifikaciju dokumenata za organizaciju s 128 zaposlenika. Uspješnost detekcije bila je 100% kod neizmijenjenih, označenih dokumenta. Ipak bliže promatranje pokazuje da je 40% brojeva imalo greške koje su ispravili BCH kodovi za ispravljanje grešaka. Većina od ovih pogrešaka bile su jednostruke pogreške, ali bilo je i manji broj dvostrukih pogrešaka. Ovaj rezultat pokazuje da smanjivanje sekvence za raspršivanje s 32 na 16 bitova ima negativan učinak na mogućnost greške.

Veličina sekvence za raspršivanje i informacije može se povećati korištenjem dva spojena (7,4) BCH koda umjesto (15,7). Takav spojeni BCH kod omogućava povećavanje informacije na 8 bitova i sekvence na 20 bitova s istom veličinom dokumenta. 8-bitna informacija omogućava 256 različitih identifikacijskih brojeva. Iz prijašnjeg eksperimenta vidljivo je da je maksimalan broj bitova grešaka 2, a dva spojena (7,4) BCH koda ispravljaju 2 bita greške. Veći broj bitova sekvence smanjuje broj grešaka i prije samog ispravljanje grešaka od strane BCH koda.

Matlab je korišten za implementaciju detekcije kod ispisanih dokumenata, za detekciju linija i razmaka između riječi. Preliminarni rezultati ukazuju da je algoritam ispravno detektirao i izmjerio razmake u ispisanom dokumentu. Ipak potrebno je dodatno podešavanje kako bi detekcija bila pouzdanija.

4.3.8 Zaključak

U ovom poglavlju opisan je algoritam za označavanje povjerljivih dokumenata te detekciju originalnog primatelja bilo gdje. Algoritam je baziran na podešavanju razmaka između riječi ili između linija teksta. Ovaj algoritam daje dobre rezultate za

sva poravnanja teksta (lijevo, desno, te s obje strane), kao i za tekst s nepravilnim razmakom između linija. Algoritam koristi tehniku za raspršenje spektra te BCH kodove za ispravljanje pogrešaka i ne treba originalan dokument za detekciju. Tehnika raspršenja spektra otklanja greške uzrokovane šumom nastalim nepravilnim razmakom. BCH kodovi za ispravljanje grešaka pomažu kod grešaka uzrokovanih šumom zbog ispisa i skeniranja. Rezultati simulacije pokazali su da je algoritam otporan na određene oblike formatiranja teksta kao što je izmjena fonta i margina. Također preliminarni rezultati detekcije i mjerenja razmaka između riječi i linija kod ispisanog dokumenta su obećavajući. Dodatna podešavanja i istraživanja algoritama za detekciju vodenog žiga u ispisanim dokumentima su u tijeku.

4.4 Označavanje značajki teksta

4.4.1 Uvod

Kod aplikacija koje se bave identifikacijom, autentičnosti i zaštitom, izmjena skrivenih podataka znači da je i sam dokument bio izmijenjen. Dakle, potrebne su lomljive ili polu-lomljive metode [7]. Lomljive metode prihvatljive su za digitalne dokumente dok su polu-lomljive (otporne na nenamjerne napade, npr. šum nakon ispisa i skeniranja) prihvatljive za digitalne i ispisane dokumente.

Glavni zahtjevi za polu-lomljivu metodu skrivanja podataka trebali bi biti:

1. Funkcionira kod digitalnih i ispisanih oblika dokumenata
2. Treba biti nezavisna od formata dokumenta, s time da format podržava određen nivo opisa teksta. Neki od modernih formata koji zadovoljavaju ovaj uvjet su: *Microsoft Office Word* (*DOC*), *Rich Text Format* (*RTF*), *PostScript* (*PS*), *Portable Document Format* (*PDF*) i drugi.
3. Originalni tekstualni dokument mora se moći pretvoriti iz jednog formata u drugi tako da zadrži skrivenu informaciju.
4. Označeni dokumenti ne bi se trebali vidljivo razlikovati od originalnog teksta.
5. Potrebna je veća stopa označavanja. Tako da i pojedine stranice sadrže određene osnovne informacije (npr. ime autora, vrijeme i datum kreiranja, komentari, itd.)
6. Jednostavna za automatizaciju. Automatizacija i procesiranje bez nadzora su važne značajke koje čine rješenje zanimljivim za praktične primjene.

U nastavku poglavlja bit će opisane dvije polu-lomljive metode označavanja. Prva metoda, kvantizacija boje, može se koristiti za digitalne i ispisane dokumente. Druga metoda, *halftone* kvantizacija odnosi se na ispisane tekstualne dokumente.

4.4.2 Kvantizacija boje

U ovoj metodi značajka teksta u koju se ugrađuje informacija je boja teksta. Glavna ideja ove metode je kvantizirati boju svakog znaka tako da ljudski vizualni sustav ne može odrediti razliku između originalnih i kvantiziranih znakova, ali da specijalizirani

čitač može odrediti razliku, npr. skener s velikim dinamičkim rasponom u slučaju ispisanih dokumenata.

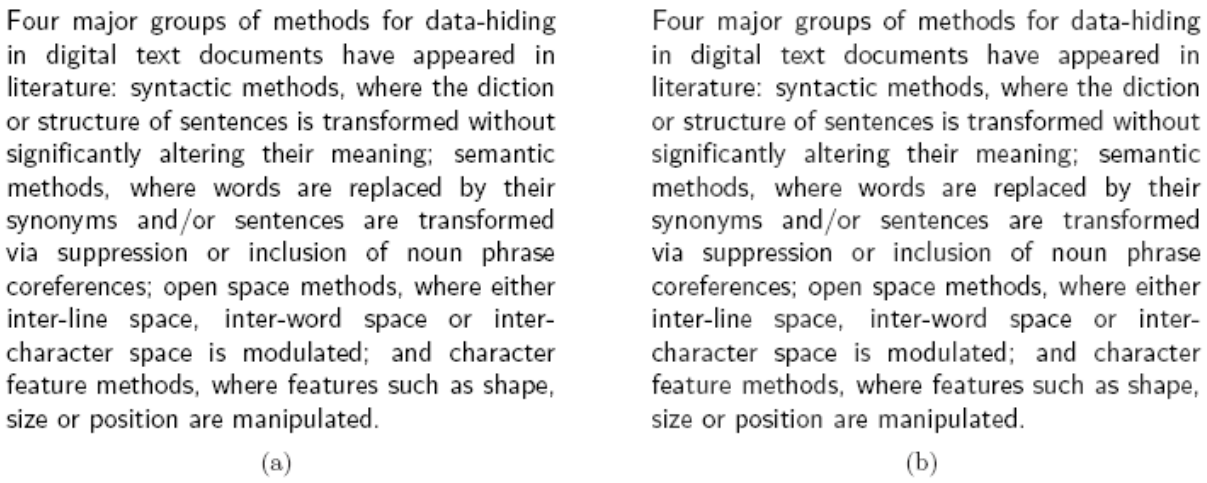
Slika 4.6 prikazuje primjer kvantizacije boje. Dakle, tamni znakovi se kodiraju kao 0, a svjetliji znakovi kao 1. Znači može se ugraditi binarna sekvenca. Također se može primijetiti da se po tekstu ugrađuje više informacija nego kod metoda koje mijenjaju razmak između riječi i linija. Kako bi dokument bio otporan na pretvorbu iz digitalnog u analogni pa opet u digitalni oblik određeni znakovi mogu se izuzeti iz označavanja. Manji znakovi, kao što su točka i zarez, nisu dobri nosioci informacije za ispisane dokumente. Kod digitalnih dokumenata ne bi trebalo biti ovih problema.

Slika 4.6 Kvantizacija boje: (a) originalan tekst, (b) označeni tekst

Ova metoda zadovoljava zahtjeve 1, 2, 3. Zahtjev 4 također je zadovoljen jer se zna da ljudski vizualni sustav ne detektira manje promjene u luminaciji. Također varijacije luminacije preko svijetlih ili tamnih podloga manje su vidljive nego kod sivih podloga. Na sreću većina dokumenata pisana je tamnim slovima preko svijetle podloge. Korištenjem modernih tekst procesora vidljivo je da se u digitalnom okruženju ovom metodom može ugraditi do 4 bita po znaku (korištenjem razina sive od 0-15), a da žig i dalje ne bude vidljiv ljudskom oku. Ako se skriven tekst ugrađuje u ispisane dokument onda će se ugraditi 1 do 2 bita informacije po znaku.

Dvo-razinski kvantizator

Najjednostavnija metoda ugrađivanja informacija je korištenje dvorazinskog kvantizatora. U ovom pristupu bira se referentna boja koja reprezentira 0. Dobar izbor je originalna boja teksta u dokumentu (većinom je crna). Zatim se bira svjetlija nijansa koja reprezentira 1. Slika 4.7 je primjer ove metode, gdje je 0 označena s crnom (luminacija je 0), a 1 sa svjetlijom nijansom crne (luminacija 46).



Slika 4.7 Dvo-razinski kvantizator (a) originalni tekst; (b) označeni tekst

Rezultati ove metode opisani su u poglavlju o eksperimentalnim rezultatima.

Višerazinska kvantizacija

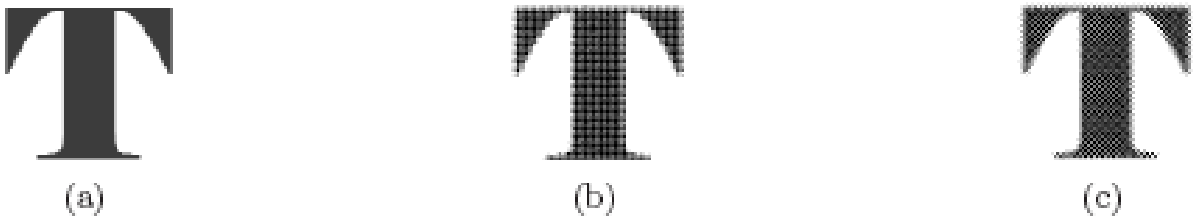
Prethodno opisana metoda može se proširiti na više razina. Umjesto korištenja dvije razine boje, može se koristiti četiri ili osam razina boje. Ova metoda dopušta označavanje više bitova po znaku. Što se tiče praktičnih primjena ova metoda ovisi o kvaliteti pisača i skenera. Zbog razvoja kvalitetnijih pisača i skenera ovu metodu ne treba odbaciti.

4.4.3 *Halftone* kvantizacija

Ova metoda oslanja se na *halftoning*, široku korištenu metodu ispisa koja omogućava da slike koje imaju kontinuirajuću boju mogu biti ispisane s tintom jedne boje (siva skala) ili nekoliko tinta u boji. Ovdje se ograničava na crno-bijele pisače.

Kako bi se simulirala dana nijansa sive, *halftone* pisač koristi *halfotne* zaslon. Ova metoda iskorištava činjenicu da može postojati više različitih *halftone* zaslona koji vode k istoj nijansi sive. Ta osobina može se koristiti kod skrivanja podataka korištenjem različitih *halftone* zaslona za označavanje pojedinih znakova, ovisno o poruci koja se želi ugraditi. Tipične značajke *halftone* zaslona koje su korisne za ugrađivanje podataka su: kut zaslona i oblik točki na zaslonu (eliptičan, okrugli, pravokutni).

Slika 4.8 prikazuje primjer primjene ove metode gdje se kut zaslona od 0° koristi za kodiranje 0, a kut od 45° za kodiranje 1. Jedna od većih prednosti ove metode jest da svi znakovi iste nijanse sive. Ako se ne kombinira s nekom od tehnika kvantizacije boje ova metoda može se koristiti samo za označavanje ispisanih dokumenata. Tako se npr. s dvije nijanse sive može ugraditi informacija u digitalnu verziju dokumenta, a korištenjem *halftone* zaslona s uzorcima u kombinaciji s dva kuta zaslona za ugrađivanje informacije u ispisanu verziju tekstualnog dokumenta.



Slika 4.8 Halftone kvantizacija: (a) originalan znak; (b) označeni znak za m=0; (c) označeni znak za m=1

4.4.4 Eksperimentalni rezultati

U ovom dijelu opisuje se praktična implementacija prethodno opisane kvantizacije boje. Kao što je prethodno spomenuto ova metoda može se koristiti za označavanje digitalnih i ispisanih tekstualnih dokumenata.

Implementacija ove metode u digitalnom okruženju prilično je jednostavna. U eksperimentima implementiran je prototip za *Microsoft Office Word* dokument sposoban za ugrađivanje i izdvajanje proizvoljne poruke. Ako se pretpostavi

savršena sinkronizacija kod čitanja označenih znakova, prototip je sposoban izdvojiti poruku bez grešaka. Dakle, za ovaj slučaj nije potreba uporaba kodova za otklanjanje pogrešaka za pouzdano izdavanje ugrađene poruke. Također je potvrđeno da se pretvorbom iz *DOC* formata u *PDF* ili *PS* format zadržava informacija o boji svakog znaka. Implementacija uspijeva izdvojiti ugrađenu informaciju iz dokumenata dobivenih pretvorbom iz *DOC* u *PDF* i *PS* formate.

Slijedi opis proširene implementacije metode kvantizacije boje za tekstualne dokumente koji su podložni ispisu i ponovnom skeniranju. Ova implementacija koristi samo dvorazinski kvantizator, ali može se proširiti tako da koristi i višerazinski kvantizator. Tablica 4.5 prikazuje opremu koja se koristila za potrebe eksperimenta. Uobičajene postavke printera (rezolucija, frekvencija zaslona, *halftone* algoritam) korištene su za ispis tekstualnih dokumenata. Za skeniranje tekstualnih dokumenata korištena je rezolucija r_s=600 ppi, siva kala, 8 bitova dubine, cijeli dinamički raspon, γ-korekcija je postavljena na 1, te *unsharp mask* filtar visoke razine ovisno o sučelju upravljačkog programa svakog od skenera.

Tablica 4.5 Korištena oprema za potrebe eksperimenta

Model	Tip
HP Color LaserJet 4600	Laserski pisač
Epson Perfection 3170 Photo	CCD skener
Epson Perfection 4990 Photo	CCD skener
Canon LiDE 50	CCD skener

Radi jednostavnosti, prvo su odabrani slučajni crni tekstovi koji koriste latinicu (A,B,...,Z,a,b,...,z), uobičajeni interpunkcijski znakovi, specijalni znakovi (zarez, točka, dvotočka, točka-zarez,-,?,!,",',(,),<,>,@,|), brojevi (0,1,...,9) i aritmetički znakovi (+,-,*,/,=). Ova implementacija može raditi i s drugim abecedama. Kao font korišten je *Arial*, veličine 10. Kako bi sustav bio što otporniji na skeniranje i ispis neki od znakova su isključeni iz označavanja. To su sljedeći znakovi: zarez, točka, dvotočka, točka-zarez, dvostruki i jednostruki navodnici i minus. Jednaki broj proizvoljnih poruka ugrađen je u digitalne tekstove korištenjem metode dvorazinskog kvantizatora. Označeni digitalni tekst nakon toga je ispisan i skeniran s opisanom opremom. Nakon toga skenirani dokument je procesiran kako bi se dohvatila ugrađena informacija.

Proces dohvaćanja informacije može se podijeliti u 3 dijela: segmentacija znakova, demodulacija značajki teksta (u ovom slučaju boje) te dekodiranje zasnovano na kvantizaciji.

Rezultati su bili slični za sve korištene skenere. Tablica 4.6 prikazuje rezultate za Epson Perfection 3170 Photo skener. Za označavanje 0 izabrana je crna boja luminacije 0. Za označavanje 1 izabrano je više različitih luminacija označenih koje predstavlja varijabla $Q_1(x)$.

Tablica 4.6 Svojstva metode dvo-razinske kvantizacije boje

$Q_1(x)$	Broj grešaka	Postotak greške
41	1342	32.7%
46	824	20.1%
51	315	7.7%
56	120	2.9%
61	62	1.5%
66	23	0.6%

4.4.5 Zaključak

U ovom poglavlju opisana je nova metoda za rješavanje problema skrivanja podataka u tekstualnim dokumentima. Glavna ideja bila je da se tekstualni znak smatra kao struktura koja se sastoji od više značajki kao što su oblik, pozicija, orijentacija, veličina, boja, itd. Od tih značajki odabrana je boja te je prikazana metoda kvantizacije boje kao nova metoda za polu-lomljivo skrivanje podataka u digitalnim i ispisanim dokumentima. Eksperimenti su pokazali da ljudsko oko vrlo malo zamjećuje ovu metodu. Može se veliki broj bitova informacije ugraditi u dokument te se može u potpunosti automatizirati. Iz prethodno spomenutog vidljivo je kako je ova metoda prihvatljiva za sljedeće primjene: identifikacija dokumenata, autentičnost i sprječavanje izmjena.

4.5 Jezično označavanje

4.5.1 Uvod

Iako se povećala količina slika te video podataka, tekstualni podaci i dalje čine većinu prometa na Internetu i drugih tipova podataka iz svakodnevnog života. Većina časopisa, novina i znanstvenih publikacija pruža članke u digitalnom formatu. Iako je to poboljšalo način na koji čitatelji mogu pretraživati i pristupati informacijama, također otvara različita pitanja autorima o tome kako se njihov rad distribuira i ponovno koristi. Prava vlasništva posebno su važna za tekstualne podatke jer ih je lakše skinuti s Interneta i modificirati od recimo slika ili filmova. Ovo poglavlje bavi se trenutnim stanjem jezičnog (eng. *natural language, NL)* označavanja, koje označava dokumente manipuliranjem semantičkom i/ili sintaksnom strukturom rečenice. Ovaj pristup razlikuje se od svih prethodno opisanih koji mijenjaju izgled elemenata teksta, kao što je izmjena formata ili veličine teksta, razmaka između riječi ili linija, itd. U usporedbi s tim metodama *NL* označavanje vrlo je mlado područje. Osim zaštite sadržaja, otporni *NL* algoritmi označavanja imat će i primjene kao što su revizija teksta, sprečavanje izmjena i traženje izdajica [8].

4.5.2 Tehnike jezičnog procesiranja i izvori informacija

Jezično procesiranje (eng. *Natural Language Processing, NLP*) bavi se algoritmima koji će analizirati, razumjeti i automatski generirati *natural language*. Ovo poglavlje ukratko opisuje *NLP* tehnike i izvore informacija koji su od interesa za skrivanje informacije u *natural language* tekstu.

Izvori podataka

Uspjeh skrivanja informacije ovisi o pribavljanju dobrih modela medija u koje će se informacija ugraditi, a to se može ostvariti velikim skupovima podataka. Statistički reprezentativan uzorak *natural language*-a zove se korpus. S obzirom da se većina *NLP* istraživanja zasniva na statističkoj analizi i sustavima strojnog učenja, potrebni su veliki korpusi u obliku čitljivom za strojeve. Zbog toga je stvoren veći broj korpusa u elektroničkom obliku koje se koriste u *NLP* istraživanju.

Osim korpusa postoje i elektronički rječnici koji su u stvari velike baze leksičkih veza između riječi. Najpoznatiji takav rječnik je *Wordnet*. U *Wordnet*-u engleske imenice, pridjevi, prilozi i prijedlozi su organizirani u setove sinonima. *VerbNet* je također još jedan elektronički rječnik koji je leksikon glagola sa sintaksnom i semantičkom informacijom o glagolima iz engleskog jezika.

Lingvističke transformacije

Kako bi se ugradila informacija u *NL* tekst potrebna je sistematična metoda za izmjenu ili transformiranje teksta. Ove transformacije trebale bi sačuvati gramatiku rečenica. Idealno je da se ne primjećuju ni promjene u značenju rečenice uzrokovane ovim transformacijama. Obično se koriste tri vrste transformacija za izmjene: supstitucija sinonima, sintaksne transformacije i semantičke transformacije.

Supstitucija sinonima najšire je korištena lingvistička transformacija za sustave skrivanja podataka jer je i najjednostavnija. Supstitucija sinonima uzima u obzir smisao riječi. Kako bi se sačuvao smisao rečenice riječ mora biti zamijenjena sinonimom istog smisla. Elektronički rječnik *Wordnet* klasificira sve riječi i fraze u skupove sinonima te time olakšava potragu za sinonimom tražene riječi. Ipak određivanje ispravnog smisla dane riječi veliki je problem jer je teško naći definiciju za smisao riječi.

Drugi tip transformacija su sintaksne transformacije. U njih se ubrajaju stvaranje pasivnog oblika rečenice, te stvaranje složenije rečenice spajanjem glavne i podređene rečenice kojima se ostvaruje smisao koji je mogao biti ostvaren jednostavnijom rečenicom (eng. *clefting)*.

Tablica 4.7 prikazuje neke od čestih sintaksnih transformacija u engleskom jeziku. Osim tih, postoji još jedan grupa sintaksnih transformacija koja se bazira samo na kategorizaciji glavnog glagola u rečenici. Glagoli se mogu klasificirati prema zajedničkom značenju i ponašanju. Različite klase glagola dozvoljavaju različite transformacije rečenice. Tablica 4.8 prikazuje primjer znan kao izmjena lokacije (eng. *locative alternation*).

Tablica 4.7 Česte sintaksne transformacije u engleskom jeziku

Transformacija	Originalna rečenica		Transformirana rečenica
Passivization	The slobbering dog kissed the big boy.	⇒	The big boy was kissed by the slobbering dog.
Topicalization	I like bagels.	⇒	Bagels, I like.
Clefting	He bought a brand new car.	⇒	It was a brand new car that he bought.
Extraposition	To believe that is difficult.	⇒	It is difficult to believe that.
Preposing	I like big bowls of beans.	⇒	Big bowls of beans are what i like.
There-construction	A unicorn is in the garden.	⇒	There is a unicorn in the garden.
Pronominalization	I put the letter in the mailbox.	⇒	I put it there.
Fronting	"What!" Alice cried.	⇒	"What!' cried Alice.

Treći tip lingvističkih transformacije su semantičke transformacije. Ova metoda generira semantičke transformacije koje čuvaju smisao korištenjem koreferenci imenica i glagola. Dvije imeničke fraze su koreferentne ako se odnose na isti entitet. Ovisno o konceptu koreference uvode se različite transformacije. Jedna takva informacija je *coreferent pruning* gdje se informacija o koreferenci koja se ponavlja briše. Suprotno od ove transformacije je *corefernt grafting* koja se također izvodi kada se informacija o koreferenci ponavlja ili se dodaje tekstu korištenjem baze činjenica. Na kraju može se izvesti i *coreferent substitution* koja se može gledati kao

Tablica 4.8 Primjer izmjene lokacije

Jack sprayed paint on the wall	⇒	Jack sprayed the wall with paint
'Henry clared the dishes from the table	⇒	Henry cleared the table of the dishes

kombinacija prijašnje dvije transformacije. Slika 4.9 prikazuje novinski članak na kojem će biti pokazane semantičke transformacije. Analiza se usredotočuje na referentni pojam "Bobby Fischer". *Prunning* se primjenjuje na prvu rečenicu, a izdvojena informacija se koristi za transformaciju druge rečenice. Slično tome, informacija dobivena iz treće rečenice koristi se za *gratfing* četvrte.

Yet Iceland has offered a residency visa to ex-chess champion Bobby Fischer in recognition of a 30-year-old match that put the country ''on the map''. His historic win over Russian Boris Spassky in Reykjavik in 1972 shone the international spotlight on Iceland as never before. Now Iceland is keen to repay the favour by offering sanctuary to Mr Fischer, an American citizen. He is being detained in Japan and is wanted in the US for violating international sanctions against the former Yugoslavia by playing there in 1992.

Slika 4.9 Novinski tekst prije transformacija

Yet Iceland has offered a residency visa to Bobby Fischer in recognition of a 30-year-old match that put the country ''on the map''. Ex-chess champion's historic win over Russian Boris Spassky in Reykjavik in 1972 shone the international spotlight on Iceland as never before. Now Iceland is keen to repay the favour by offering sanctuary to Mr Fischer, an American citizen. He, an American citizen, is being detained in Japan and is wanted in the US for violating international sanctions against the former Yugoslavia by playing there in 1992.

Slika 4.10 Novinski tekst nakon transformacija

Jedan od problema s iznad opisanim pristupom jest razrješavanje koreference, što je ujedno i jedna od najtežih zadaća u NLP-u. Nadalje nije preporučljivo zamijeniti dvije koreferentne fraze u određenim okolnostima. Slika 4.11 prikazuje jedan od dobro poznatih primjera ovog problema.

Spiderman just saved us from death.
Peter Parker just saved us from death.

Slika 4.11 Primjer problema razrješenja koreferenci

Iako se fraze Spiderman i Peter Parker odnose na istu osobu, nekome tko ne zna ovu činjenicu prva rečenica se može činiti točna a druga ne.

Parsiranje

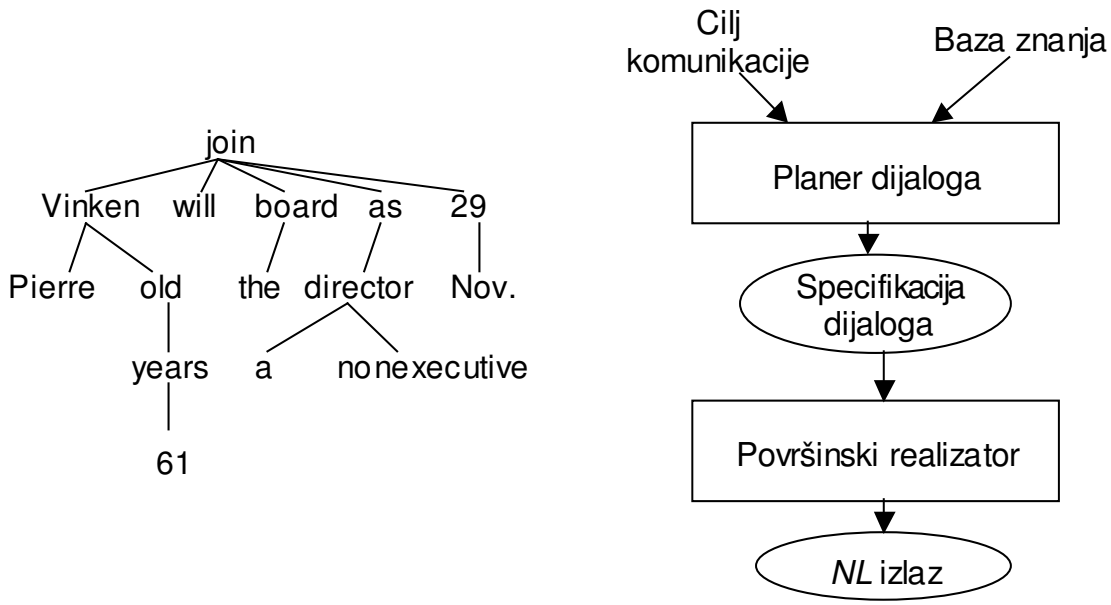
Parsiranje je proces koji od zadane rečenice stvara određenu vrste strukture. Izlaz parsiranja može biti morfološka, sintaksna, semantička struktura rečenice ili njihova kombinacija. Parsiranje je bitan proces jer se njime dobiva informacija o strukturi rečenice kao i o ulogama riječi koje ju čine. Većina parsera koriste *part-of-speech taggers* koji kategoriziraju riječi u predodređene razrede (kao što su imenice, pridjevi ili glagoli) i morfolške analizatore koji razbijaju riječi u morfeme kao jedan od koraka prije procesiranja.

Još uvijek nije dostupan u potpunosti implementiran semantički parser. Ipak postoje različiti alati koji pretvaraju strukture fraze generirane sinatksnim parserom u stabla zavisnosti, koja ilustriraju bit ili vezu između riječi u rečenici. Slika 4.12 (a) prikazuje primjer generiranog stabla zavisnosti za jednostavnu rečenicu.

Stvaranje *NL*-a

Stvaranje *NL*-a (eng. *Natural Language Generation - NLG*) definira se kao proces konstruiranja izlaza *NL*-a od ne lingvističke reprezentacije informacije prema određenim specifikacijama komunikacije. Slika 4.12 (b) prikazuje dijelove tipičnog *NLG* sustava. Dobar primjer *NLG* sustava je *Forecast Generator* (*FOG*), sustav za vremensku prognozu koju generira tekst na Engleskom i Francuskom. Ovaj sustav uzima meteorološke podatke i generira vremensku prognozu.

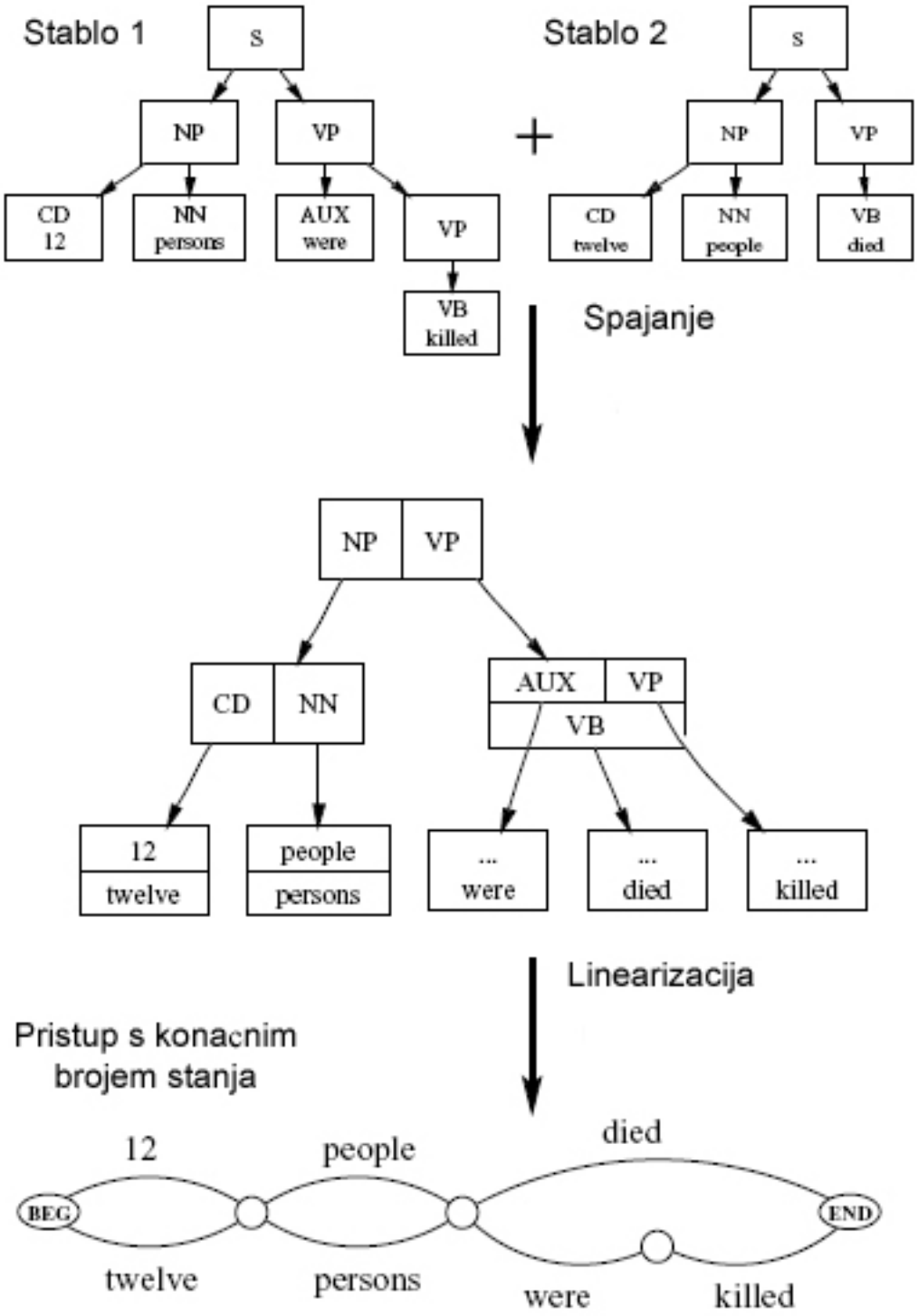
Što se tiče *NL* skrivanja informacija *NLG* je presudna komponenta. Nakon što je informacija dodana rečenici mijenjanjem strukturne reprezentacije, ova izmijenjena reprezentacija treba biti pretvorena natrag u *NL* korištenjem *NLG* sustava.



Slika 4.12 (a) Stablo zavisnosti za rečenicu: "Pierre Vinken, 61 years old, will join the board as a non executive director Nov. 29."; (b) Komponente tipičnog sustava za generiranje NL-a;

Parafraziranje teksta

Zadaća parafraziranja teksta uključuje mijenjanje parametara teksta kao što su duljina, čitljivost i stil za specifičnu namjernu bez gubitka temeljnog značenja teksta. Dakle parafraziranje teksta direktno je vezano uz *NL* označavanje. Parafraziranje teksta slično je strojnom prevođenju; samo što umjesto pretvaranja teksta iz jednog jezika u drugi, mijenja oblik teksta iz jednog oblika u drugi unutar istog jezika. Sustavi za parafraziranje teksta se većinom zasnivaju na kreiranju ili skupljanju skupova ili parova semantički ekvivalentnih riječi, fraza i uzoraka. Slika 4.13 prikazuje primjer parafraziranja teksta.



Slika 4.13 Primjer parafraziranja teksta korištenjem pristupa s konačnim brojem stanja

4.5.3 Dosadašnji rad na jezičnom označavanju

Kao što je prethodno spomenuto ovo je još vrlo mlado područje tako da za razliku od NL steganografije postoji manje dokumentacije.

Supstitucija sinonima temeljena na kvadratnim ostacima

Ideja o korištenju semantike i sintakse teksta za ubacivanje vodenog žiga je prvo predložena od strane Atallah et. al. [12] u 2000. gdje su se ASCII vrijednosti riječi koristene za ugrađivanje informacije u tekst izvršavanjem leksičke supstitucije u skupovima sinonima.

Neka je $m_{i \bmod k}$ bit vodenog žiga koji se želi ugraditi. w_i je trenutna riječ koja se razmatra je li pogodna za ugrađivanje. ASCII vrijednosti trenutne riječi (w_i) $A(w_i)$. Ako je:

$$m_{i \bmod k} = 1 \quad \text{i} \qquad x^2 = (A(w_i) + r_{i \bmod k})(\bmod p) \tag{4.37}$$

odnosno $(A(w_i) + r_{i \bmod k})$ je kvadratni ostatak modulo p tada se w_i ne mijenja. Inače se mijenja. p je 20-znamenkasti primarni ključ, k je broj bitova poruke vodenog žiga, a $r_0, r_1, \ldots, r_{k-1}$ sekvenca pseudoslučajnih brojeva generiranih korištenjem p kao početne vrijednosti (eng. *seed*).

Ugrađivanje informacije u stablo koje sadrži strukturu rečenice

U kasnijim radovima od Atallah et al [12] predlažu se dva algoritma ugrađivanja informacija u stablo sa strukturom rečenice umjesto korištenja leksičke supstitucije. Ove tehnike ugrađuju vodeni žig u parsirane reprezentacije rečenica umjesto u sam tekst kao kod leksičke supstitucije. Korištenje neposredne reprezentacije čini ove algoritme otpornijima na napade u usporedbi s leksičkim supstitucijskih sustavima.

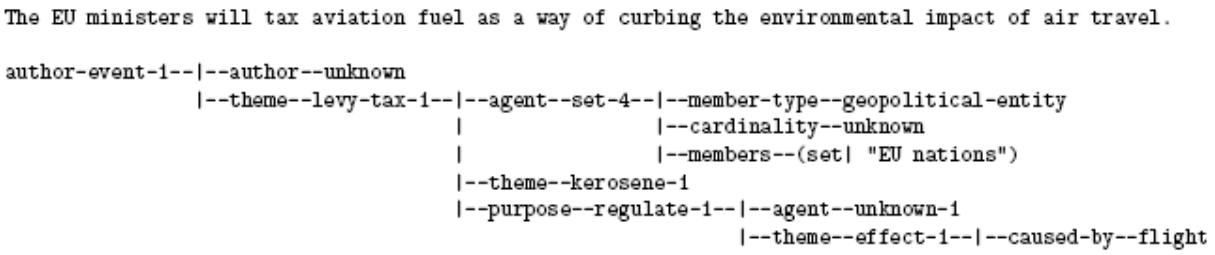
Razlike između dva predložena algoritma jest da prvi modificira sintaksno stablo dobiveno parsiranjem teksta u koji će se informacija ugrađivati, dok drugi ugrađuje u semantičko stablo. Sintaksno stablo je reprezentacija različitih dijelova rečenice koja je sintaksno parsirana. Slika 4.14 prikazuje primjere sintasknih stabala za dvije rečenice.

```
I took the book.
(S (NP I) (VP took (NP the book)) (. .))

The book was taken by me.
(S1 (S(NP (DT The) (NN book))(VP (VBD was) (VP (VBN taken) (PP (IN by) (NP (PRP me))))) (. .)))
```

Slika 4.14 Primjer sintaksnog stabla za dvije rečenice

Za razliku od sintaksnog, semantičko stablo koristi reprezentaciju teksta u obliku stabla koje se odnosi na reprezentaciju značenja riječi rečenice. Takve reprezentacije rečenice generiraju se korištenjem ontoloških semantičkih izvora. Slika 4.15 prikazuje primjer semantičkog stabla za zadanu rečenicu.



Slika 4.15 Primjer semantičkog stabla

Izbor rečenica koje će nositi informaciju vodenog žiga ovisi samo o strukturi stabla i vrši se na sljedeći način: Čvorovi stabla T_i rečenice s_i teksta su označeni prolasku s vrha prema dnu kroz stablo. Nakon toga, čvor s oznakom *j* se pretvara u 1 ako je

$$j + H(p) \tag{4.38}$$

kvadratni ostatak modulo *p*, a 0 inače, gdje je *p* tajni ključ a *H()* jednostrana *hash* funkcija. Nakon toga se generira nova sekvenca oznaka B_i prolaskom kroz drvo od dna prema vrhu. Rang d_i se tada dobiva za svaku rečenicu s_i korištenjem

$$d_i = H(B_i)\, XOR\, H(p) \tag{4.39}$$

te se rečenice sortiraju po rangu. Počevši od najmanje rangirane rečenice s_j, vodeni žig se umeće u nasljednika s_j. Rečenica s_j se naziva marker rečenica jer pokazuje na rečenicu koja nosi vodeni žig. Umetanje vodenog žiga nastavlja se sa sljedećom rečenicom u listi poredanoj po rangu. Kada su odabrane rečenice za označavanje bitovi se spremaju primjenom sintaksne ili semantičke transformacije.

4.5.4 Smjernice

Unatoč nekim poboljšanjima jezično označavanje i dalje je u povojima. Preporuča se suradnja zajednica koje se bave jezičnim označavanjem i označavanjem slika. Za neke aspekte jezičnog označavanja mogu se usvojiti neke od ideja označavanja teksta, dok se za druge aspekte moraju razviti u potpunosti novi pristupi koji mogu upravljati direktnom i rekurzivnom prirodom jezika.

Vjeruje se da su pristupi koji se oslanjanju na ugrađivanje informacija korištenjem sintaksne strukture rečenica obećavajući za jezično označavanje. Budući sustavi za jezično označavanje trebali bi voditi računa o koherentnoj semantici i retoričkoj strukturi rečenice.

Ocjena jezičnog sustava za označavanje predstavlja veći problem nego ocjenjivanje označavanja slika jer takvi sustavi moraju paziti na pitanja o značenju riječi ili rečenice, gramatici i stilu teksta. Trenutno ne postoje objektivne ocjene ljudske percepcije *NL* označenog teksta korištenjem različitih algoritama niti studije otpornost jezičnog označavanja na napade. Potrebno je uložiti još mnogo truda u ovo područje.

4.5.5 Zaključak

Jezično označavanje korištenjem lingvističkih tehnika novo je područje istraživanja s velikim potencijalom za mnoge primjene. Trenutno nema u potpunosti funkcionalnih sustava, iako je interes za ovo područje porastao. Došlo bi do velikog poboljšanja u jezičnom označavanju ako bi se iskustvo i znanje iz označavanja slika i audio zapisa moglo upotrijebiti uz pomoć istraživača iz ovog područja.

5. Opis praktičnog rada

5.1 Opis korištenog algoritma

Odabran je algoritam koji označava tekst klasificiranjem riječi i podešavanjem statistike razmaka između riječi [9]. Sve riječi u tekstu klasificiraju se prema nekoj značajci, zatim se od tih riječi stvaraju segmenti. U svaki se segment umeće ista količina informacije. Informacija se umeće mijenjanjem statistike razmaka između riječi određenog segmenta. Algoritam ima globalna svojstva, u smislu da skriva dio informacije u određeni segment čiji se elementi nalaze u cijelom dokumentu.

Pretpostavka je da je cijeli dokument već segmentiran u stranice, linije i riječi.

5.1.1 Klasifikacija riječi

Pretpostavlja se da linija teksta ima n riječi. i-ta riječi označava se s w_i, a širina w_i se označava s $l(w_i)$. Širina riječi, $l(w_i)$, mjeri se u pikselima. Neka je K broj klasa riječi, a $class(w_i)$ klasa riječi w_i. Širina riječi je značajka koja se koristi za klasifikaciju. Klasifikator koristi širine susjednih riječi kod određivanja klase. Tablica 5.1 prikazuje pravilo klasifikacije riječi koje generira dvije klase riječi, odnosno K=2. Određuje se klasa riječi w_i uspoređujući širine lijeve (w_{i-1}) i desne (w_{i+1}) riječi.

Tablica 5.2 prikazuje još jedno klasifikacijsko pravilo koje uspoređuje 5 uzastopnih riječi i stvara 4 klase, odnosno K=4. Za klasifikaciju prve i zadnje riječi smatra se da je lista cirkularna, odnosno lijeva riječ od w_1 je w_n, a desna riječ od w_n je w_1.

Tablica 5.1 Pravilo klasifikacije riječi (K=2)

Uvjeti	class(w_i)
$l(w_{i-1}) > l(w_{i+1})$	0
$l(w_{i-1}) \leq l(w_{i+1})$	1

5.1.2 Segmenti i klasifikacija

Segment se definira kao s uzastopnih riječi u liniji. Prvi segment je uređena lista (w_1, w_2,..., w_s). Sljedeći segment počinje od prve riječi prethodnog segmenta pa drugi segment izgleda (w_s, w_{s+1},..., w_{2s-1}). Slika 5.1 prikazuje primjer klasifikacije riječi i segmenata. Segment se uvodi kako pomaci riječi ne bi međusobno interferirali. S obzirom da su prva i posljednja zajedničke susjednim segmentima, njihove lokacije su fiksne. Dozvoljava se pomicanje samo unutarnjih riječi.

Segmenti dobivaju oznake iz oznaka pojedinih riječ u segmentu. Npr. riječi u drugom segmentu imaju oznake 1,0,0 tako da je oznaka segmenta 100. Broj klasa segmenata označava se s L, gdje se L računa prema formuli (5.5).

Tablica 5.2 Pravilo klasifikacije riječi (K=4)

Uvjeti	class(w_i)
a≥b i c≥d	00 (0)
a≥b i c<d	01 (1)
a<b i c≥d	10 (2)
a<b i c<d	11 (3)

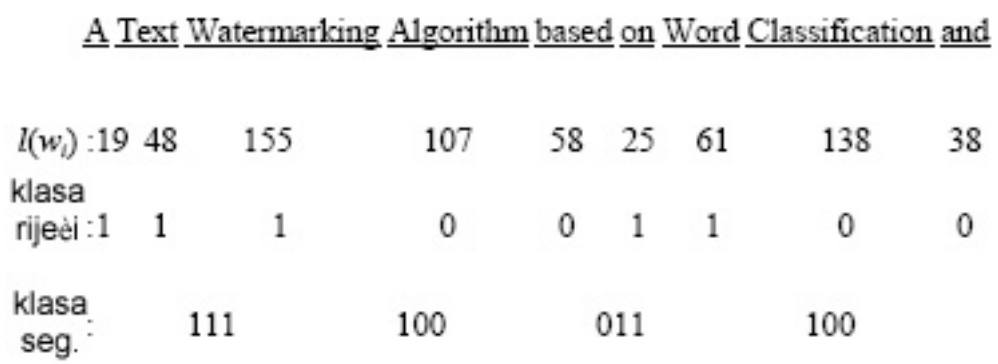
gdje su

$$a = l(w_{i-2}) + l(w_{i-1}) \tag{5.1}$$

$$b = l(w_{i+1}) + l(w_{i+2}) \tag{5.2}$$

$$c = l(w_{i-1}) + l(w_{i+1}) \tag{5.3}$$

$$d = l(w_{i-2}) + l(w_{i+2}) \tag{5.4}$$



Slika 5.1 Primjer klasifikacije riječi i klasifikacije segmenata

Broj klasa segmenata:

$$L = k^s \tag{5.5}$$

5.1.3 Umetanje i detekcija vodenog žiga

Strategija je definiranje određene statistike za klase segmenata te ugrađivanje signala vodenog žiga tako da statistika zadovoljava neke predodređene uvjete. Statistika se određuje za razmake između riječi za pojedine segmente.

Statistika razmaka između riječi

Prvo se odvija proces klasifikacije opisan u prethodnom poglavlju, s time da dokument može imati više od jedne stranice. Segmenti s istom oznakom nakon toga se grupiraju u skup segmenata *S(k)* gdje je $1 \le k \le L$. Pretpostavka je da dokument ima dovoljan broj segmenata i da je *L* skupova segmenata uravnoteženo što se tiče njihove veličine. To je realna pretpostavka s obzirom da tekstualni dokumenti obično imaju više tisuća riječi.

Korištenjem segmenata iz *S(k)*, formulira se statistika za *s-1* razmaka između riječi. Neka se statistika označava s Ω_i^k i neka vrijedi $1 \le k \le L$ i $1 \le i \le s-1$. Najjednostavniji slučaj statistike je srednja vrijednost $\Omega_i^k = \mu_i^k$, gdje je srednja vrijednost definirana u jednadžbi (5.6). Još jedna statistika je srednja vrijednosti i varijanca $\Omega_i^k = (\mu_i^k, \sigma_i^k)$ koje su definirane u jednadžbi (5.7).

$$\mu_{ii}^k = (1/m)\sum_{j=1}^{m} p_j^i, \quad 1 \le i \le s-1 \tag{5.6}$$

$$(\mu_{ii}^k, \sigma_i^k) = ((1/m)\sum_{j=1}^{m} p_j^i, (\sum_{j=1}^{m} (p_j^i - \mu_{ii}^k)^2 / m)^{1/2}), \quad 1 \le i \le s-1 \tag{5.7}$$

U ovim jednadžbama p_j^i predstavlja *j*-ti razmak *i*-tog segmenta, a *m* je broj segmenata u *S(k).*

Umetanje i detekcija signala digitalnog vodenog žiga

Fiksna količina informacije umeće se u svaku od klasa segmenata, Ako postoji *p* bitova informacije po klasi segmenta, veličina informacije koja se može ugraditi je *p*L* gdje je *L* broj klasa segmenata. Slijede jednostavna pravila kodiranja.

Pravilo 1: (s=3,L=64, Ω=μ, veličina informacije koja se može ugraditi je 64 bita)

```
Ako je (μ1≤μ2) signal je 1, inače 0
```

Pravilo 2: (s=3,L=64, Ω=(μ,σ), veličina informacije koja se može ugraditi je 128 bitova)

```
Ako je (μ1≤μ2) i (σ1≤σ2) signal je 00 (0)

Inače ako je (μ1≤μ2) i (σ1>σ2) signal je 01 (1)

Inače ako je (μ1>μ2) i (σ1≤σ2) signal je 10 (2)

Inače ako je (μ1>μ2) i (σ1>σ2) signal je 11 (3)
```

Kodiranje informacije zahtijeva specifičnu statistiku razmaka između riječi. Riječi u segmentu moraju se pomaknuti u lijevo ili u desno ovisno o traženoj distribuciji. To se jednostavno ostvaruje s obzirom da susjedni segmenti dijele rubne riječi. Lokacija zajedničkih riječi je fiksna, a pomiču se ostale *s-2* riječi.

Pravila se odnose na statistiku segmenta, a ne na pojedine riječi pa ako je statistika segmenta ispravna nema potrebe pomicati riječi tog segmenta. Inače se riječi pomiču jedan po jedan piksel dok se ne ostvare uvjeti. Tako se umeće informacija s najmanjom količinom pomaka riječi.

Detekcija signala slijedi sljedeće korake:

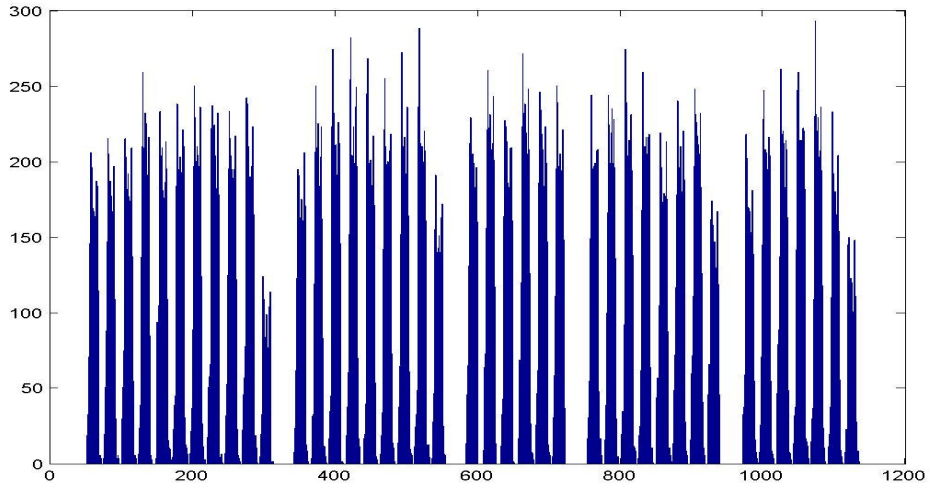
1. Primijeniti segmentaciju linije i riječi
2. Klasificirati riječi i segmente
3. Konstruirati skupove segmenata i izračunati razmake između riječi.
4. Izračunati statističke distribucije
5. Dekodirati signal iz distribucija

5.2 Praktična izvedba algoritma za pomicanje riječi

Algoritam je implementiran i izvršava se u Matlabu. Matlab je izabran jer ima izvrsnu podršku za obradu slika, a algoritam označava skenirane slike teksta. Algoritam je namijenjen tekstu poravnatom s obje strane (eng. *justified*).

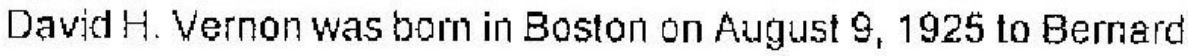
Prvo se učitana slika pretvara u binarnu sliku, crno bijelu sliku koja ima vrijednosti piksela 0 ili 1. S obzirom da vrijednost piksela 0 znači da je u pitanju tekst, odnosno crna boja, radi lakšeg daljnjeg rada ta binarna slika se invertira.

Nakon toga kreće se sa segmentiranjem dokumenta u linije. Osnova segmentiranja linija je vertikalni profil. Slike su u Matlabu reprezentirane matricama, pa je vertikalni profil zbroj vrijednosti piksela svakog retka matrice. Slika 5.2 prikazuje primjer jednog vertikalnog profila, u kojem su jasno vidljive linije, odnosno granice linija. S obzirom da je slika invertirana, minimumi predstavljaju razmak između linija, a maksimumi same linije. Iz takvog vertikalnog profila računaju se granice linija, početni i završni redak svake linije, i širine linija.

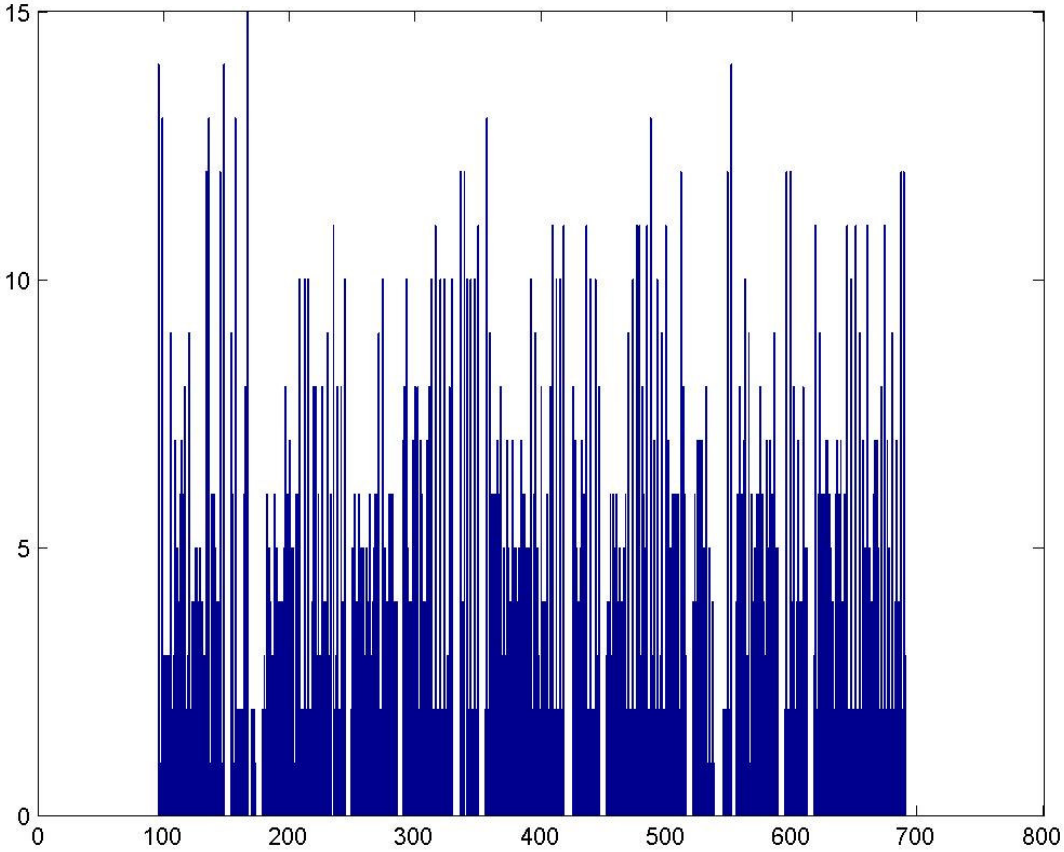


Slika 5.2 Primjer vertikalnog profila

Kad je stranica segmentirana u linije kreće se na segmentaciju riječi. Za svaku liniju određuje se širina riječi i razmaka između riječi u liniji te se zatim računaju klase riječi i klase segmenata riječi. Jedan od najvećih problema ovog programskog ostvarenja bilo je i ostaje upravo određivanje širine pojedinih riječi i razmaka između riječi. Za određivanje strukture linije koristi se horizontalni profil linije. Horizontalni profil je ustvari suma vrijednosti piksela pojedinog stupca linija. Kako bi se što više smanjio razmak između pojedinih slova u riječima, prvo se iscrtaju granice pojedinih slova, te se od tako izmijenjene linije računa horizontalni profil. Iscrtavanjem linija razmaci između riječi trebali bi postati izraženiji od razmak između slova, ali to nažalost nije uvijek slučaj. Slika 5.3 prikazuje primjer linije, a Slika 5.4 horizontalni profil te linije iz kojeg se vide granice linije.



Slika 5.3 Primjer linije



Slika 5.4 Primjer horizontalnog profila

Nakon računanja horizontalnog profila brišu se zaostali razmaci između slova. Računa se kolika je brojnost koje širine razmaka, jer se eksperimentom ustanovilo da će uvijek biti zaostalih razmaka između slova unutar riječi te da je to najbrojnija širina

razmaka. Nažalost uklanjanje najbrojnije širine razmaka ne spaja u potpunosti slova pojedine riječi. Problem su slova koja imaju jednu ili više okomitih stranica (npr. i,l,M,N, itd.), takve riječi unose dodatan razmak te se riječ koja ih sadrži može detektirati kao dvije riječi umjesto jedne. Također, neki interpunkcijski znakovi (npr. točka, zarez) mogu se detektirati kao zasebna riječ pa ih se treba spojiti s najbližom riječi. Detekcija tih interpunkcijskih znakova je jednostavna, jer je vrijednost zbroja iz horizontalnog profila za takve znakove manja u odnosu na ostala slova.

Poslije obrade linija, pristupa se određivanju klasa riječi i segmenata. Segment je velik 3 riječi. Ostvarene su dvije klase riječi, K=2, koja se računa prema Tablica 5.1. Kod druge klase riječi, K=4, klase se računaju prema Tablica 5.2 i jednadžbama (5.1), (5.2), (5.3) i (5.4). Za svaku liniju se računaju oznake segmenata te linije i spremaju se u novu matricu.

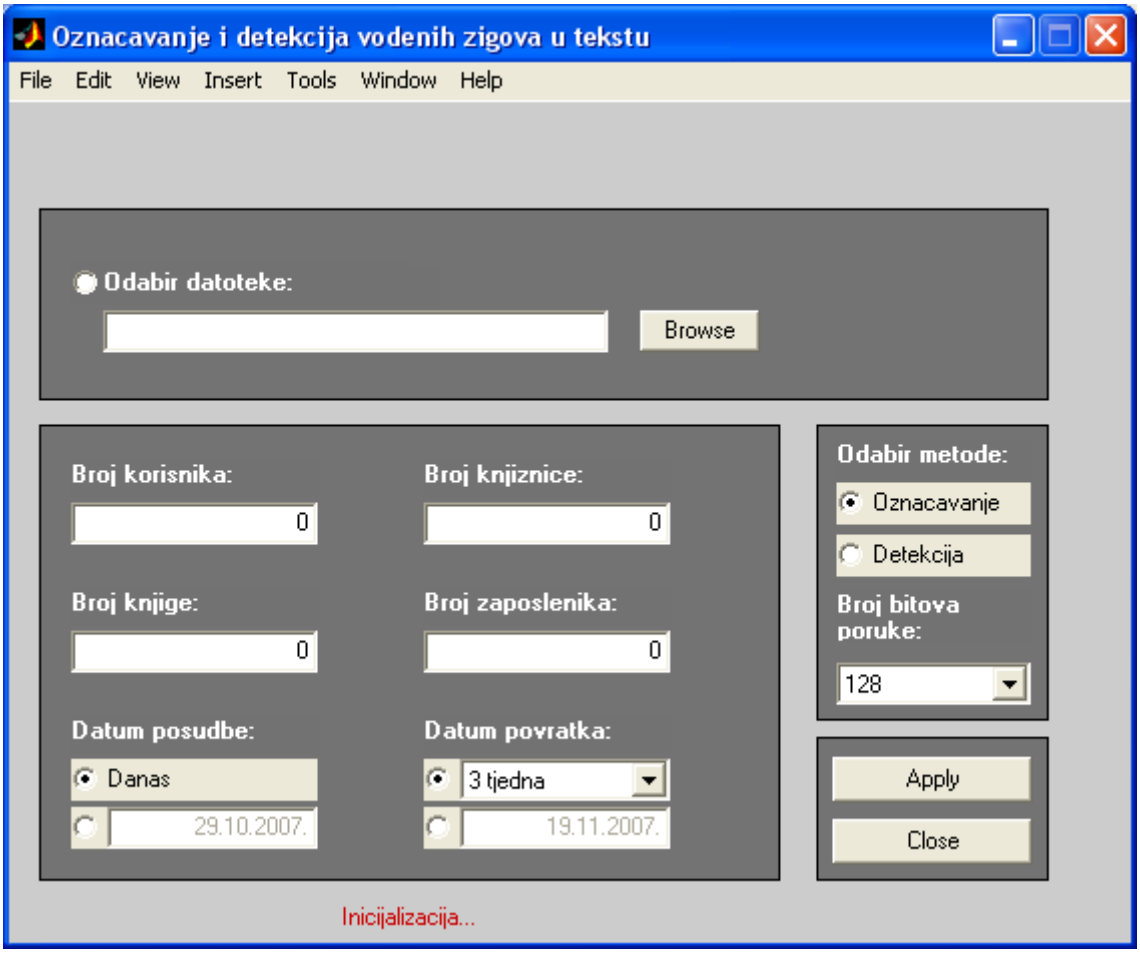
Iz matrice s oznakama segmenata za svaku liniju stvara se skup segmenata. Skup segmenata za svaku oznaku segmenta sadrži informaciju o broju linije u kojoj se pojedini segment s tom oznakom nalazi te rednom broju segmenta.

Kada je stvoren skup segmenata pristupa se računanju statistike pojedine oznake segmenta. Za statistiku su odabrane srednja vrijednost i standardna devijacija. Ako određena oznaka segmenta sadrži samo jedan segment, takva se oznaka briše, te se u taj segment ne ugrađuje informacija. Time se nažalost smanjuje veličina informacije koja se želi ugraditi. Oznake sa samo jednim segmentom brišu se jer nema smisla računati srednje vrijednosti i standardne devijacije samo jedne vrijednosti. Iz izračunatih vrijednosti računa se vrijednost vodenog žiga prema pravilu 1 odnosno 2. U slučaju detekcije iz tih vrijednosti računa se ugrađena poruka.

Ako se radi o označavanju tada se signal, odnosno bitovi originalnog teksta uspoređuju s bitovima poruke. Ako se bit poruke razlikuje od bita iz originalnog teksta za određenu oznaku segmenta, svi segmenti s tom oznakom se mijenjaju.

S obzirom da je ovo slijepi algoritam, za uklanjanje vodenog žiga iz teksta potreban je originalni tekst. Algoritam može ugraditi poruke veličine 8, 16, 64, 128 bitova. Uspješna detekcija najviše ovisi o tome koliko točno algoritam računa širinu riječi i razmaka između riječi prilikom označavanja i detekcije.

Slika 5.5 prikazuje korisničko sučelje programa. Korisnik može odabrati želi li označiti datoteku ili detektirati vodeni žig u željenoj datoteci. Osim toga može se odabrati veličina informacije koja se ugrađuje. Može se ugraditi 8, 16 ,64, 128 bitova informacije. Kod veličine informacije od 8 i 16 bitova, zbog malog raspona brojeva, označava se samo broj korisnika. Kod 128 bitova može se ugraditi željeni broj korisnika, broj knjižnice, broj zaposlenika koji se posudio dokument, broj samog dokumenta, te datum posudbe i povratka. Datum posudbe može se odabrati kao trenutni datum, ili se ručno unijeti. Datum povratka se može odabrati kao 3 tjedna, mjesec dana ili 3 mjeseca od dana posudbe ili se može ručno unijeti.



Slika 5.5 Korisničko sučelje programa

5.3 Eksperimentalni rezultati

Slika 5.6 Prikazuje originalni tekst korišten u eksperimentima. Ovaj tekst je odabran jer funkcija za obradu linije, odnosno funkcije koja računa širinu razmaka između linija i širinu riječi u većini slučajeva točno računa širinu razmaka i riječi. Nažalost ovaj tekst nema dovoljno oznaka segmenata za označavanje 64, odnosno 128 bitova poruke. U njega se mogu ugraditi poruke veličine 8, 16, 32 i 64 bita. Prilog 1 prikazuje rezultate ugrađivanja poruke veličine 8, 16, 32, 64 bita u originalan tekst.

Tablica 5.3 prikazuje primjer originalnog i označenog teksta kod ugrađivanja poruke veličine 8 bitova. U gornjem redu je originalni, a u donjem označeni tekst. Ubačena vrijednost je broj 20. U prvom stupcu se vidi neznatan pomak slova H udesno. U drugom stupcu se vidi da pomakom riječi *was* udesno jedan stupac razmaka prelazi preko slova b, pa je okomiti dio slova b tanji. Te na kraju u trećem stupcu se vidi pomak riječi *in* ulijevo. Kod neizmijenjene označene slike detekcijom se ispravno detektira svih 255 mogućih vrijednosti. Kod izmijenjene slike, kod 10 različitih vrijednosti, svih 10 unatoč izmjenama (brisanjem čitavih redaka) uspješno se detektira. Uspješnost detekcije ovisi o izbrisanom retku, makar se u ovom slučaja vrlo rijetko se detektira kriva vrijednost.

David H. Vernon was born in Boston on August 9, 1925 to Bernard and Ida Cohen Vernon. He and his wife Rhoda, who he met in grade school in their native Boston, were married June 1, 1947, following his freshman year at Harvard. He enrolled there after his return from the U.S. Navy service on a PT boat in the Pacific during World War II. Following graduation from Harvard College and the Harvard Law School, he began his distinguished teaching career as an instructor at New York University Law School where he received the LLM and SJD degrees. Later he was a professor at the University of Houston Law School and the University of New Mexico before going to the University of Washington Law School as professor and Associate Dean

Coming to Iowa as Dean of Law and professor in 1966-71, Dave Vernon was an imaginative leader in a time of growth when the law faculty more than doubled. He championed what he described as the "Graduate College approach" to legal education, characterized by smaller classes and a richer curriculum, enhanced by advanced courses and seminars, and emphasizing much greater individual attention to the development of students' basic professional skills, particularly their research and writing. He was instrumental in recruiting minority law students through summer orientation programs at both Iowa and New Mexico.

Dave Vernon was the Allan D. Vestal Professor of Law from 1986 to this date and before that he was Iowa Law School Foundation Professor. He served as president of the Association of American Law Schools (AALS) and delegate to the American Bar Association House of Delegates, representing AALS. He also was editor of the Journal of Legal Education and chair of the Board of Trustees of the Law School Admissions Council.

Dave Vernon was highly respected nationally in his scholarly fields -- contracts and conflict of laws -- and in transnational legal education. He was the author of six books, thirty articles, monographs and book reviews. He was invited as visiting professor at many American law schools as well as at the University of Durham, England, and Victoria University of Wellington, New Zealand as Fulbright Lecturer. In 1997 he received the Collegiate Teaching Award and the Regents Award for Faculty Excellence and in 2001 the Hancher-Finkbine Medallion Award.

Despite of his strong preference for full-time teaching, Dave Vernon twice agreed to serve as the University of Iowa's chief academic officer in I973-74 and 1988-89 and undertook special assignments for four University presidents. Known for his integrity, he worked assiduously and effectively to maintain the University's openness in time of campus unrest, to assure due process for faculty and students, and to promote inclusiveness throughout the University.

Slika 5.6 Originalni tekst

Tablica 5.3 Originalni i označeni tekst

David H. Vernon	was born	born in Boston
David H. Vernon	was born	born in Boston

Tablica 5.4, prvi redak prikazuje tekst označen vrijednosti 255. Drugi redak prikazuje izmijenjen tekst, dobiven brisanjem 3 retka. Zanimljivo je da se čak i s 3 izbrisana retka i dalje dobro detektira označena vrijednost. Točna vrijednost informacije može se očitati čak i samo iz 1. ili 2. ili 5. odlomka. Pogrešna vrijednost se detektira kad postoji samo 3. ili samo 4. odlomak.

Tablica 5.4 Označeni tekst i izmijenjeni označeni tekst

Coming to Iowa as Dean of Law and professor in 1966-71, Dave Vernon was an imaginative leader in a time of growth when the law faculty more than doubled. He championed what he described as the "Graduate College approach" to legal education, characterized by smaller classes and a richer curriculum, enhanced by advanced courses and seminars, and emphasizing much greater individual attention to the development of students' basic professional skills, particularly their research and writing. He was instrumental in recruiting minority law students through summer orientation programs at both Iowa and New Mexico. Dave Vernon was the Allan D. Vestal Professor of Law from 1986 to this date and before that he was Iowa Law School Foundation Professor. He served as president of the Association of American Law Schools (AALS) and delegate to the American Bar Association House of Delegates, representing AALS. He also was editor of the Journal of Legal Education and chair of the Board of Trustees of the Law School Admissions Council.
Vernon was an imaginative leader in a time of growth when the law faculty College approach" to legal education, characterized by smaller classes and a richer curriculum, enhanced by advanced courses and seminars, and emphasizing much greater individual attention to the development of students' basic professional skills, particularly their research and writing. He was instrumental in recruiting minority law students through summer orientation programs at both Iowa and New Mexico. Dave Vernon was the Allan D. Vestal Professor of Law from 1986 to this date and before that he was Iowa Law School Foundation Professor. (AALS) and delegate to the American Bar Association House of Delegates, representing AALS. He also was editor of the Journal of Legal Education and chair of the Board of Trustees of the Law School Admissions Council.

Kod označavanja informacije veličine 16 bitova rezultati su slični kao i kod 8 bitova. Od 10 slika označenih različitim vrijednostima te zatim izmijenjenih kod svih 10 je uspješno detektirana ugrađena vrijednost. Tablica 5.5 prikazuje razliku između originalnog i označenog teksta. U prvom i drugom stupcu vidljivo je da se riječ pomicala za više od 1 piksela, jer je ostao trag od okomitog dijela n odnosno m. U trećem stupcu se vidi pomak riječ of u lijevo. Također je ostao trag od slova f.

Tablica 5.5 Originalne i označene vrijednosti za veličinu informacije od 16 bitova

school in their	graduation from Harvard	Despite of his
school in their	graduation from Harvard	Despite of his

Tablica 5.6 Označeni tekst, gore je neizmijenjen, dolje je izmijenjen

Coming to Iowa as Dean of Law and professor in 1966-71, Dave Vernon was an imaginative leader in a time of growth when the law faculty more than doubled. He championed what he described as the "Graduate College approach" to legal education, characterized by smaller classes and a richer curriculum, enhanced by advanced courses and seminars, and emphasizing much greater individual attention to the development of students' basic professional skills, particularly their research and writing. He was instrumental in recruiting minority law students through summer orientation programs at both Iowa and New Mexico.

Dave Vernon was the Allan D. Vestal Professor of Law from 1986 to this date and before that he was Iowa Law School Foundation Professor. He served as president of the Association of American Law Schools (AALS) and delegate to the American Bar Association House of Delegates, representing AALS. He also was editor of the Journal of Legal Education and chair of the Board of Trustees of the Law School Admissions Council.

Vernon was an imaginative leader in a time of growth when the law faculty more than doubled. He championed what he described as the "Graduate College approach" to legal education, characterized by smaller classes and a richer curriculum, enhanced by advanced courses and seminars, and emphasizing much greater individual attention to the development of students' basic professional skills, particularly their research and writing. He was instrumental in recruiting minority law students through summer orientation programs at both Iowa and New Mexico.

Dave Vernon was the Allan D. Vestal Professor of Law from 1986

He served as president of the Association of American Law Schools (AALS) and delegate to the American Bar Association House of Delegates, representing AALS. He also was editor of the Journal of Legal Education and chair of the Board of Trustees of the Law School Admissions Council.

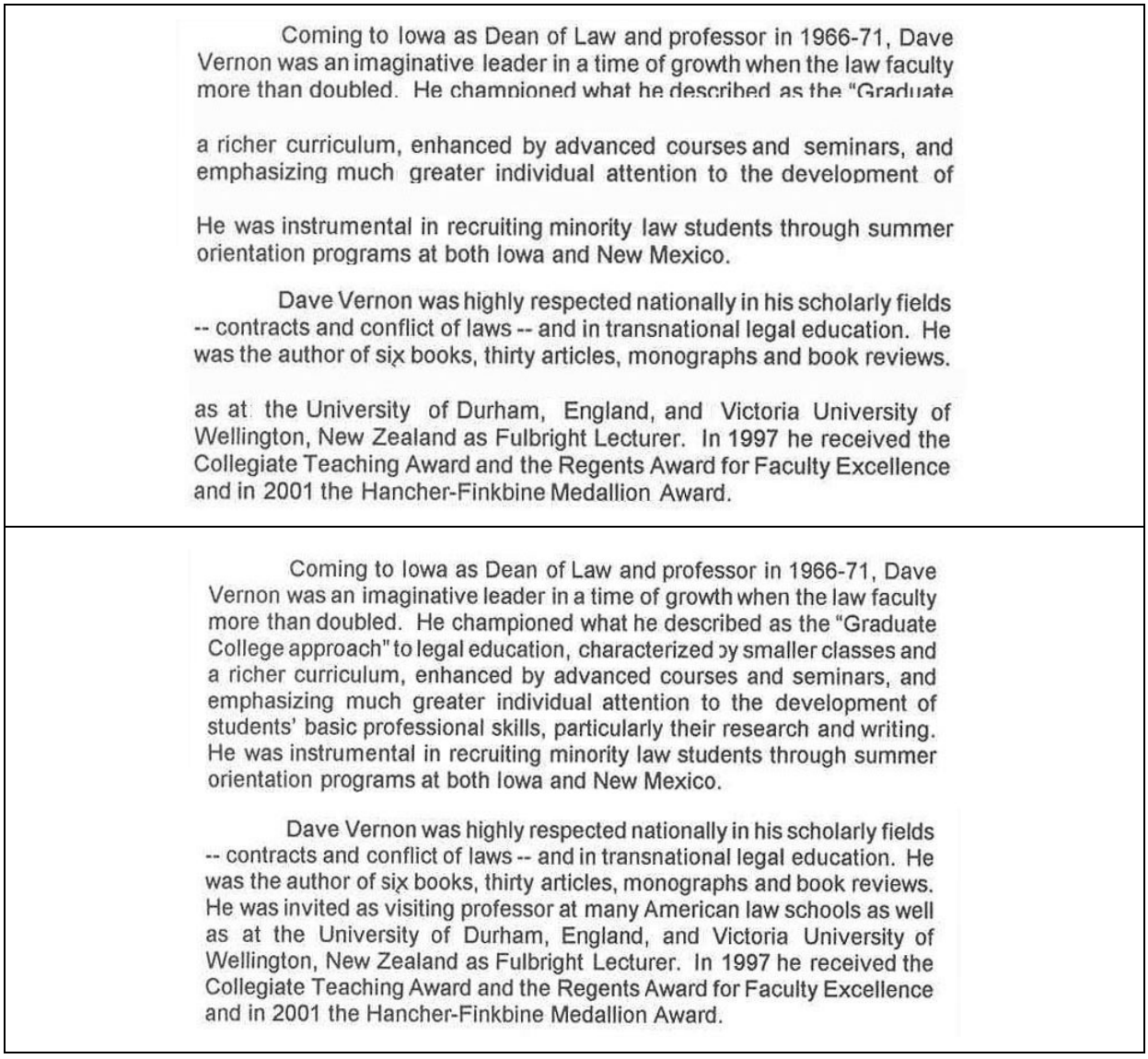
Tablica 5.6 prikazuje označeni tekst koji je u gornjem retku neizmijenjen, a u donjem izmijenjen. I ovdje se unatoč 2 izbrisana retka detektira ispravna vrijednost.

Tablica 5.7 Originalne i označene vrijednosti za veličinu informacije od 32 bitova

Professor of Law	characterized by smaller	Vernon was the
Professor ɔf Law	characterized ɔy smaller	Vernon was the

Kod slike označene porukom veličine 32 bita vidljiva su i veća izobličenja, uzrokovana većim pomakom riječi. Tablica 5.7 prikazuje te razlike.

Tablica 5.8 Označeni tekst, gore je neizmijenjen, dolje je izmijenjen



Coming to Iowa as Dean of Law and professor in 1966-71, Dave Vernon was an imaginative leader in a time of growth when the law faculty more than doubled. He championed what he described as the "Graduate

a richer curriculum, enhanced by advanced courses and seminars, and emphasizing much greater individual attention to the development of

He was instrumental in recruiting minority law students through summer orientation programs at both Iowa and New Mexico.

Dave Vernon was highly respected nationally in his scholarly fields -- contracts and conflict of laws -- and in transnational legal education. He was the author of six books, thirty articles, monographs and book reviews.

as at the University of Durham, England, and Victoria University of Wellington, New Zealand as Fulbright Lecturer. In 1997 he received the Collegiate Teaching Award and the Regents Award for Faculty Excellence and in 2001 the Hancher-Finkbine Medallion Award.

Coming to Iowa as Dean of Law and professor in 1966-71, Dave Vernon was an imaginative leader in a time of growth when the law faculty more than doubled. He championed what he described as the "Graduate College approach" to legal education, characterized ɔy smaller classes and a richer curriculum, enhanced by advanced courses and seminars, and emphasizing much greater individual attention to the development of students' basic professional skills, particularly their research and writing. He was instrumental in recruiting minority law students through summer orientation programs at both Iowa and New Mexico.

Dave Vernon was highly respected nationally in his scholarly fields -- contracts and conflict of laws -- and in transnational legal education. He was the author of six books, thirty articles, monographs and book reviews. He was invited as visiting professor at many American law schools as well as at the University of Durham, England, and Victoria University of Wellington, New Zealand as Fulbright Lecturer. In 1997 he received the Collegiate Teaching Award and the Regents Award for Faculty Excellence and in 2001 the Hancher-Finkbine Medallion Award.

Tablica 5.8 prikazuje označeni izmijenjeni i neizmijenjeni tekst. Opet se unatoč brisanju čak 3 retka detektirala točna vrijednost. I opet je bila točna detekcija kod svih 10 testnih primjera.

Tablica 5.9 Originalne i označene vrijednosti za veličinu informacije od 64 bita

as at the	undertook special assignments	to assure due
as at the	undertook special assignments	to assure due

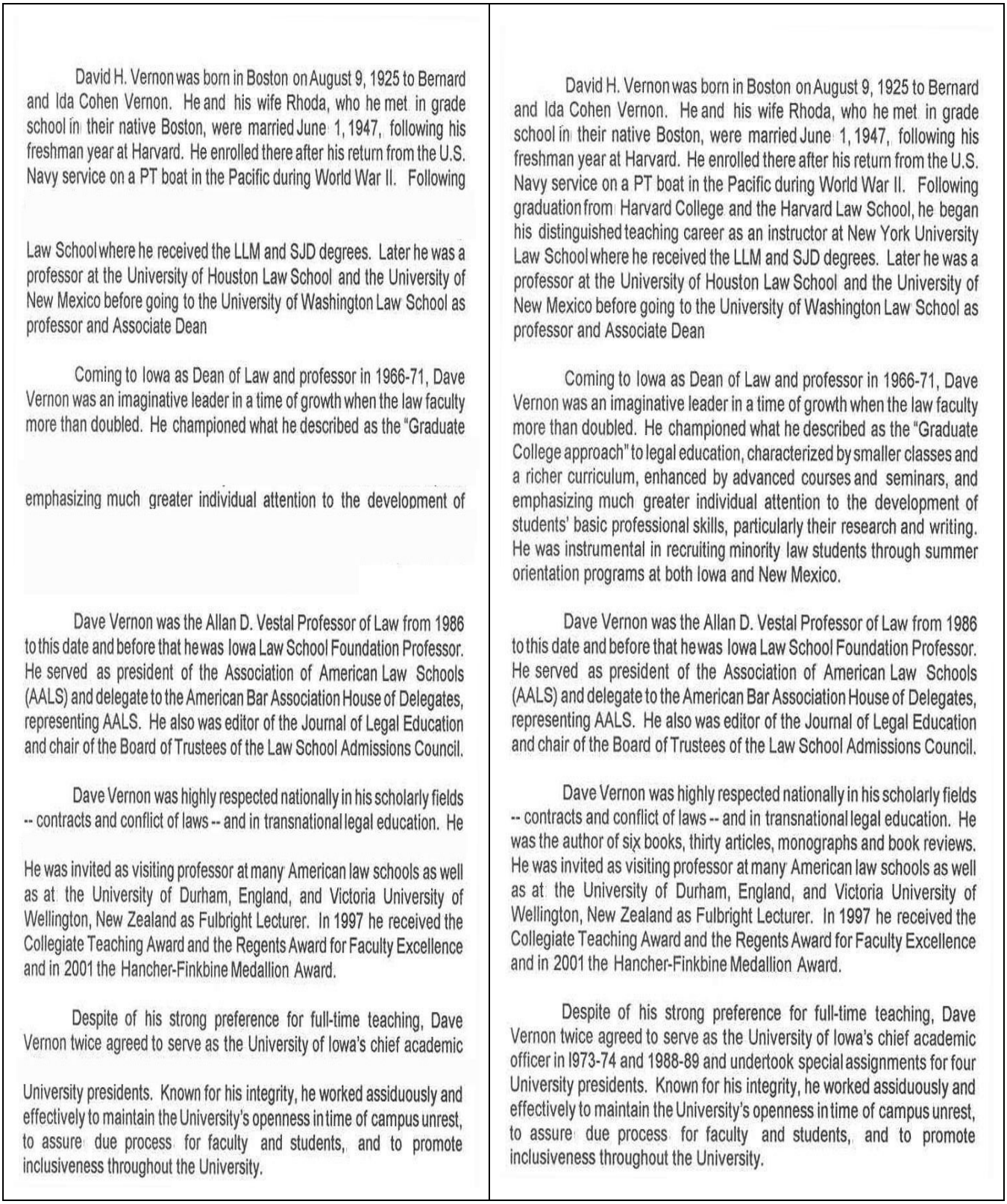
Tablica 5.9 prikazuje razlike između označenog i originalnog teksta. Iako je ugrađeno 64 bita informacija, nema prevelikih izobličenja, ali ona pak postoje.

Tablica 5.10 Označeni tekst, gore je neizmijenjen, dolje je izmijenjen

Coming to Iowa as Dean of Law and professor in 1966-71, Dave Vernon was an imaginative leader in a time of growth when the law faculty more than doubled. He championed what he described as the "Graduate College approach" to legal education, characterized by smaller classes and a richer curriculum, enhanced by advanced courses and seminars, and emphasizing much greater individual attention to the development of students' basic professional skills, particularly their research and writing. He was instrumental in recruiting minority law students through summer orientation programs at both Iowa and New Mexico.
Coming to Iowa as Dean of Law and professor in 1966-71, Dave Vernon was an imaginative leader in a time of growth when the law faculty a richer curriculum, enhanced by advanced courses and seminars, and emphasizing much greater individual attention to the development of students' basic professional skills, particularly their research and writing. He was instrumental in recruiting minority law students through summer

Tablica 5.10 prikazuje označeni izmijenjeni i neizmijenjeni tekst. Algoritam je opet uspio točno odrediti vrijednosti ugrađene informacije u svih 10 testnih primjera.

Tablica 5.11 Lijevo je označena i modificirana stranica teksta, desno je označena stranica bez modifikacija, veličina ubačene informacije je 64 bita



David H. Vernon was born in Boston on August 9, 1925 to Bernard and Ida Cohen Vernon. He and his wife Rhoda, who he met in grade school in their native Boston, were married June 1, 1947, following his freshman year at Harvard. He enrolled there after his return from the U.S. Navy service on a PT boat in the Pacific during World War II. Following

Law School where he received the LLM and SJD degrees. Later he was a professor at the University of Houston Law School and the University of New Mexico before going to the University of Washington Law School as professor and Associate Dean

Coming to Iowa as Dean of Law and professor in 1966-71, Dave Vernon was an imaginative leader in a time of growth when the law faculty more than doubled. He championed what he described as the "Graduate

emphasizing much greater individual attention to the development of

Dave Vernon was the Allan D. Vestal Professor of Law from 1986 to this date and before that he was Iowa Law School Foundation Professor. He served as president of the Association of American Law Schools (AALS) and delegate to the American Bar Association House of Delegates, representing AALS. He also was editor of the Journal of Legal Education and chair of the Board of Trustees of the Law School Admissions Council.

Dave Vernon was highly respected nationally in his scholarly fields -- contracts and conflict of laws -- and in transnational legal education. He

He was invited as visiting professor at many American law schools as well as at the University of Durham, England, and Victoria University of Wellington, New Zealand as Fulbright Lecturer. In 1997 he received the Collegiate Teaching Award and the Regents Award for Faculty Excellence and in 2001 the Hancher-Finkbine Medallion Award.

Despite of his strong preference for full-time teaching, Dave Vernon twice agreed to serve as the University of Iowa's chief academic

University presidents. Known for his integrity, he worked assiduously and effectively to maintain the University's openness in time of campus unrest, to assure due process for faculty and students, and to promote inclusiveness throughout the University.

David H. Vernon was born in Boston on August 9, 1925 to Bernard and Ida Cohen Vernon. He and his wife Rhoda, who he met in grade school in their native Boston, were married June 1, 1947, following his freshman year at Harvard. He enrolled there after his return from the U.S. Navy service on a PT boat in the Pacific during World War II. Following graduation from Harvard College and the Harvard Law School, he began his distinguished teaching career as an instructor at New York University Law School where he received the LLM and SJD degrees. Later he was a professor at the University of Houston Law School and the University of New Mexico before going to the University of Washington Law School as professor and Associate Dean

Coming to Iowa as Dean of Law and professor in 1966-71, Dave Vernon was an imaginative leader in a time of growth when the law faculty more than doubled. He championed what he described as the "Graduate College approach" to legal education, characterized by smaller classes and a richer curriculum, enhanced by advanced courses and seminars, and emphasizing much greater individual attention to the development of students' basic professional skills, particularly their research and writing. He was instrumental in recruiting minority law students through summer orientation programs at both Iowa and New Mexico.

Dave Vernon was the Allan D. Vestal Professor of Law from 1986 to this date and before that he was Iowa Law School Foundation Professor. He served as president of the Association of American Law Schools (AALS) and delegate to the American Bar Association House of Delegates, representing AALS. He also was editor of the Journal of Legal Education and chair of the Board of Trustees of the Law School Admissions Council.

Dave Vernon was highly respected nationally in his scholarly fields -- contracts and conflict of laws -- and in transnational legal education. He was the author of six books, thirty articles, monographs and book reviews. He was invited as visiting professor at many American law schools as well as at the University of Durham, England, and Victoria University of Wellington, New Zealand as Fulbright Lecturer. In 1997 he received the Collegiate Teaching Award and the Regents Award for Faculty Excellence and in 2001 the Hancher-Finkbine Medallion Award.

Despite of his strong preference for full-time teaching, Dave Vernon twice agreed to serve as the University of Iowa's chief academic officer in 1973-74 and 1988-89 and undertook special assignments for four University presidents. Known for his integrity, he worked assiduously and effectively to maintain the University's openness in time of campus unrest, to assure due process for faculty and students, and to promote inclusiveness throughout the University.

U ovom slučaju dolazi do pogrešne detekcije, ali ubačeni broj je 1234567, a detektirani 1234565, tako da čak uz ovoliko brisanja nije velika razlika.

6. Zaključak

Eksperimenti su pokazali da točnost detekcije u označenom izmijenjenom i označenom neizmijenjenom tekstu najviše ovisi o točnosti funkcije koja određuje širinu razmaka između riječi i širinu riječi. Što je ubačena informacija manja to su i manja izobličenja pa i funkcija za obradu linije točnije računa širine.

Tako za najmanju veličinu ugrađene informacije za sve moguće vrijednosti informacije kod neizmijenjenog označenog teksta detekcija detektira točne vrijednosti. Također kod najmanjeg broja bitova informacije napadač mora uložiti mnogo truda da slomi vodeni žig jer se zbog male količine informacije, informacija nalazi na cijelom dokumentu. Odnosno postoji mali broj oznaka segmenata (0-7 za 8 bitova informacije) ali zato za svaku oznaku postoji dovoljan broj segmenata za ispravnu detekciju. Kako veličina informacije raste tako se smanjuje broj segmenata za svaku oznaku, a time se povećava mogućnost greške. Tako se može dogoditi da napad izmijeni tekst na način da oznaka segmenta koja je imala samo 1 segment više ne postoji čime se može izmijeniti vrijednost očitanog vodenog žiga.

Rješenje za taj problem je dovoljno veliki tekstualni dokument, jer je time i veća vjerojatnost da će svaka od oznaka segmenata imati dovoljan broj segmenata da vodeni žig bude otporan na napade. Također je pitanje hoće li se isplatiti napadaču obraditi veliki tekstualni dokument, pogotovo ako je dokument u ispisanom obliku pa tada napadač treba izdvojiti dosta vremena dok skenira sve stranice dokumenta, a nakon toga još i vrši izmjene nad svim stranicama.

Implementirani algoritam dobar je za zaštitu tekstualnih dokumenata koji će se ispisivati. Ispisivanjem i ponovnim skeniranjem unosi se mali šum, koji ne smeta jer se slika pretvara u binarnu. Još jedna od prednosti implementiranog algoritma je što razlike između originalnog i označenog teksta nisu jako vidljive. Ako se riječ pomiče za samo jedan piksel razlike gotovo nisu vidljive. U slučaju pomaka za više piksela razlike mogu postati vidljivije, ali to nije uvijek slučaj.

Otpornost opisanog algoritma na napade ne treba ovisiti o veličini ugrađene informacije, ako se povećanjem informacije povećava i količina teksta za označavanje. Naravno preduvjet za dobru detekciju, neovisno o veličini ugrađene informacije, je dobro ugođena funkcija za obradu linije. Tako da bi cilj budućeg rada na ovom algoritmu bilo što bolje ugađanje funkcije za obradu linije, te proširivanje algoritma da ne označava samo tekst poravnat na obje strane, nego i druga poravnanja.

7. Literatura

[1] Edin Muharemagic, Borko Furht, “Multimedia Security: Watermarking Techniques“

[2] Fred Mintzer, Gordon W. Braudaway i Minerva M. Yeung, “Effective and ineffective Digital Watermarks“

[3] Fran Hartung, Matrin Kutter, “Multimedia Watermarking Techniques“

[4] http://www.dlib.org.ar/dlib/july98/gladney/07gladney.html#deployment_problem

[5] Ding Huang, Hong Yang, Interword Distance Changes Represented by Sine Waves For Watermarking Text Images“

[6] Adnan M. Alattar, Osama M. Alattar, “Watermarking Electronic Text Documents Containing Justified Paragraphs and Irregular Line Spacing“

[7] R. Vill´an, S. Voloshynovskiy, O. Koval, J. Vila, E. Topak, F. Deguillaume, Y. Rytsar, T. Pun, “Text Data-Hiding for Digital and Printed Documents:Theoretical and Practical Considerations“

[8] Mercan Topkara, Cuneyt M. Taskiran, Edward J. Delp, “Natural Language Watermarking“

[9] Young-Won Kim, Kyung-Ae Moon, Il-Seok Oh, “A Text Watermarking Algorithm based on Word Classification and Inter-word Space Statistics“

[10] J. T. Brassil, S. Low, N. F. Maxemchuk, and L. O’Gorman, “Electronic Marking and Identification Techniques to Discourage Document Copying,” *IEEE Journal on Selected Areas in Communications*, vol. 13, no. 8, October 1995, pp. 1495-1504.

[11] J. T. Brassil, S. Low, and N. F. Maxemchuk, "Copyright Protection for the Electronic Distribution of Text Documents," *Proceedings of the IEEE*, vol. 87, no. 7, July 1999, pp.1181-1196.

[12] M. Atallah, C. McDonough, S. Nirenburg, and V. Raskin, “Natural Language Processing for Information Assurance and Security: An Overview and Implementations,” Proceedings 9th ACM/SIGSAC New Security Paradigms Workshop, September, 2000, Cork, Ireland, pp. 51–65.

8. Prilog 1

Primjeri označavanja teksta porukom različite duljine

David H. Vernon was born in Boston on August 9, 1925 to Bernard and Ida Cohen Vernon. He and his wife Rhoda, who he met in grade school in their native Boston, were married June 1, 1947, following his freshman year at Harvard. He enrolled there after his return from the U.S. Navy service on a PT boat in the Pacific during World War II. Following graduation from Harvard College and the Harvard Law School, he began his distinguished teaching career as an instructor at New York University Law School where he received the LLM and SJD degrees. Later he was a professor at the University of Houston Law School and the University of New Mexico before going to the University of Washington Law School as professor and Associate Dean

Coming to Iowa as Dean of Law and professor in 1966-71, Dave Vernon was an imaginative leader in a time of growth when the law faculty more than doubled. He championed what he described as the "Graduate College approach" to legal education, characterized by smaller classes and a richer curriculum, enhanced by advanced courses and seminars, and emphasizing much greater individual attention to the development of students' basic professional skills, particularly their research and writing. He was instrumental in recruiting minority law students through summer orientation programs at both Iowa and New Mexico.

Dave Vernon was the Allan D. Vestal Professor of Law from 1986 to this date and before that he was Iowa Law School Foundation Professor. He served as president of the Association of American Law Schools (AALS) and delegate to the American Bar Association House of Delegates, representing AALS. He also was editor of the Journal of Legal Education and chair of the Board of Trustees of the Law School Admissions Council.

Dave Vernon was highly respected nationally in his scholarly fields -- contracts and conflict of laws -- and in transnational legal education. He was the author of six books, thirty articles, monographs and book reviews. He was invited as visiting professor at many American law schools as well as at the University of Durham, England, and Victoria University of Wellington, New Zealand as Fulbright Lecturer. In 1997 he received the Collegiate Teaching Award and the Regents Award for Faculty Excellence and in 2001 the Hancher-Finkbine Medallion Award.

Despite of his strong preference for full-time teaching, Dave Vernon twice agreed to serve as the University of Iowa's chief academic officer in I973-74 and 1988-89 and undertook special assignments for four University presidents. Known for his integrity, he worked assiduously and effectively to maintain the University's openness in time of campus unrest, to assure due process for faculty and students, and to promote inclusiveness throughout the University.

Slika 8.1 Tekst označen porukom veličine 8 bitova

David H. Vernon was born in Boston on August 9, 1925 to Bernard and Ida Cohen Vernon. He and his wife Rhoda, who he met in grade school in their native Boston, were married June 1, 1947, following his freshman year at Harvard. He enrolled there after his return from the U.S. Navy service on a PT boat in the Pacific during World War II. Following graduation from Harvard College and the Harvard Law School, he began his distinguished teaching career as an instructor at New York University Law School where he received the LLM and SJD degrees. Later he was a professor at the University of Houston Law School and the University of New Mexico before going to the University of Washington Law School as professor and Associate Dean

Coming to Iowa as Dean of Law and professor in 1966-71, Dave Vernon was an imaginative leader in a time of growth when the law faculty more than doubled. He championed what he described as the "Graduate College approach" to legal education, characterized by smaller classes and a richer curriculum, enhanced by advanced courses and seminars, and emphasizing much greater individual attention to the development of students' basic professional skills, particularly their research and writing. He was instrumental in recruiting minority law students through summer orientation programs at both Iowa and New Mexico.

Dave Vernon was the Allan D. Vestal Professor of Law from 1986 to this date and before that he was Iowa Law School Foundation Professor. He served as president of the Association of American Law Schools (AALS) and delegate to the American Bar Association House of Delegates, representing AALS. He also was editor of the Journal of Legal Education and chair of the Board of Trustees of the Law School Admissions Council.

Dave Vernon was highly respected nationally in his scholarly fields -- contracts and conflict of laws -- and in transnational legal education. He was the author of six books, thirty articles, monographs and book reviews. He was invited as visiting professor at many American law schools as well as at the University of Durham, England, and Victoria University of Wellington, New Zealand as Fulbright Lecturer. In 1997 he received the Collegiate Teaching Award and the Regents Award for Faculty Excellence and in 2001 the Hancher-Finkbine Medallion Award.

Despite of his strong preference for full-time teaching, Dave Vernon twice agreed to serve as the University of Iowa's chief academic officer in I973-74 and 1988-89 and undertook special assignments for four University presidents. Known for his integrity, he worked assiduously and effectively to maintain the University's openness in time of campus unrest, to assure due process for faculty and students, and to promote inclusiveness throughout the University.

Slika 8.2 Tekst označen porukom veličine 16 bitova

David H. Vernon was born in Boston on August 9, 1925 to Bernard and Ida Cohen Vernon. He and his wife Rhoda, who he met in grade school in their native Boston, were married June 1, 1947, following his freshman year at Harvard. He enrolled there after his return from the U.S. Navy service on a PT boat in the Pacific during World War II. Following graduation from Harvard College and the Harvard Law School, he began his distinguished teaching career as an instructor at New York University Law School where he received the LLM and SJD degrees. Later he was a professor at the University of Houston Law School and the University of New Mexico before going to the University of Washington Law School as professor and Associate Dean

Coming to Iowa as Dean of Law and professor in 1966-71, Dave Vernon was an imaginative leader in a time of growth when the law faculty more than doubled. He championed what he described as the "Graduate College approach" to legal education, characterized by smaller classes and a richer curriculum, enhanced by advanced courses and seminars, and emphasizing much greater individual attention to the development of students' basic professional skills, particularly their research and writing. He was instrumental in recruiting minority law students through summer orientation programs at both Iowa and New Mexico.

Dave Vernon was the Allan D. Vestal Professor of Law from 1986 to this date and before that he was Iowa Law School Foundation Professor. He served as president of the Association of American Law Schools (AALS) and delegate to the American Bar Association House of Delegates, representing AALS. He also was editor of the Journal of Legal Education and chair of the Board of Trustees of the Law School Admissions Council.

Dave Vernon was highly respected nationally in his scholarly fields -- contracts and conflict of laws -- and in transnational legal education. He was the author of six books, thirty articles, monographs and book reviews. He was invited as visiting professor at many American law schools as well as at the University of Durham, England, and Victoria University of Wellington, New Zealand as Fulbright Lecturer. In 1997 he received the Collegiate Teaching Award and the Regents Award for Faculty Excellence and in 2001 the Hancher-Finkbine Medallion Award.

Despite of his strong preference for full-time teaching, Dave Vernon twice agreed to serve as the University of Iowa's chief academic officer in 1973-74 and 1988-89 and undertook special assignments for four University presidents. Known for his integrity, he worked assiduously and effectively to maintain the University's openness in time of campus unrest, to assure due process for faculty and students, and to promote inclusiveness throughout the University.

Slika 8.3 Tekst označen porukom veličine 32 bita

David H. Vernon was born in Boston on August 9, 1925 to Bernard and Ida Cohen Vernon. He and his wife Rhoda, who he met in grade school in their native Boston, were married June 1, 1947, following his freshman year at Harvard. He enrolled there after his return from the U.S. Navy service on a PT boat in the Pacific during World War II. Following graduation from Harvard College and the Harvard Law School, he began his distinguished teaching career as an instructor at New York University Law School where he received the LLM and SJD degrees. Later he was a professor at the University of Houston Law School and the University of New Mexico before going to the University of Washington Law School as professor and Associate Dean

Coming to Iowa as Dean of Law and professor in 1966-71, Dave Vernon was an imaginative leader in a time of growth when the law faculty more than doubled. He championed what he described as the "Graduate College approach" to legal education, characterized by smaller classes and a richer curriculum, enhanced by advanced courses and seminars, and emphasizing much greater individual attention to the development of students' basic professional skills, particularly their research and writing. He was instrumental in recruiting minority law students through summer orientation programs at both Iowa and New Mexico.

Dave Vernon was the Allan D. Vestal Professor of Law from 1986 to this date and before that he was Iowa Law School Foundation Professor. He served as president of the Association of American Law Schools (AALS) and delegate to the American Bar Association House of Delegates, representing AALS. He also was editor of the Journal of Legal Education and chair of the Board of Trustees of the Law School Admissions Council.

Dave Vernon was highly respected nationally in his scholarly fields -- contracts and conflict of laws -- and in transnational legal education. He was the author of six books, thirty articles, monographs and book reviews. He was invited as visiting professor at many American law schools as well as at the University of Durham, England, and Victoria University of Wellington, New Zealand as Fulbright Lecturer. In 1997 he received the Collegiate Teaching Award and the Regents Award for Faculty Excellence and in 2001 the Hancher-Finkbine Medallion Award.

Despite of his strong preference for full-time teaching, Dave Vernon twice agreed to serve as the University of Iowa's chief academic officer in I973-74 and 1988-89 and undertook special assignments for four University presidents. Known for his integrity, he worked assiduously and effectively to maintain the University's openness in time of campus unrest, to assure due process for faculty and students, and to promote inclusiveness throughout the University.

Slika 8.4 Tekst označen porukom veličine 64 bita

Sažetak

U ovom radu ukratko su opisane osnove i primjene označavanja digitalnim vodenim žigom. Dan je pregled algoritama za označavanje teksta. Za svaku od četiri vrste algoritama detaljnije je opisan primjer. Također je opisana i implementacija algoritma za označavanje teksta klasificiranjem riječi i podešavanjem statistike razmaka između riječi te su opisani rezultati eksperimenata.

Abstract

In this paper, basic principles and applications of digital watermarking are described. In addition to that, an overview of current text watermarking algorithms is included. Each of the four types of text watermarking algorithms is described. An algorithm based on word classification and inter-word space statistics was implemented and the experimental results are discussed.

Sadržaj