OCR tehnologije za digitalizaciju sadržaja

Stjepanović, Nikolina

Sveučilište Jurja Dobrile u Puli
Odjel za informacijsko - komunikacijske tehnologije

NIKOLINA STJEPANOVIĆ

OCR TEHNOLOGIJE ZA DIGITALIZACIJU SADRŽAJA

Završni rad

Pula, rujan 2017. godine

Sveučilište Jurja Dobrile u Puli
Odjel za informacijsko - komunikacijske tehnologije

NIKOLINA STJEPANOVIĆ

OCR TEHNOLOGIJE ZA DIGITALIZACIJU SADRŽAJA

Završni rad

JMBAG: 0319001208, redoviti student

Studijski smjer: Informatika

Predmet: Informatizacija uredskog poslovanja

Znanstveno područje: Društvene znanosti
Znanstveno polje: Informacijske i komunikacijske znanosti
Znanstvena grana: Informacijski sustavi i informatologija
Mentor: doc. dr. sc. Darko Etinger

Pula, rujan 2017. godine

IZJAVA O AKADEMSKOJ ČESTITOSTI

Ja, dolje potpisana Nikolina Stjepanović, kandidatkinja za prvostupnicu informatike ovime izjavljujem da je ovaj Završni rad rezultat isključivo mojega vlastitog rada, da se temelji na mojim istraživanjima te da se oslanja na objavljenu literaturu kao što to pokazuju korištene bilješke i bibliografija. Izjavljujem da niti jedan dio Završnog rada nije napisan na nedozvoljen način, odnosno da je prepisan iz kojega necitiranog rada, te da ikoji dio rada krši bilo čija autorska prava. Izjavljujem, također, da nijedan dio rada nije iskorišten za koji drugi rad pri bilo kojoj drugoj visokoškolskoj, znanstvenoj ili radnoj ustanovi.

Student

_____

U Puli, _____, _____ godine

IZJAVA

o korištenju autorskog djela

Ja, Nikolina Stjepanović dajem odobrenje Sveučilištu Jurja Dobrile u Puli, kao nositelju prava iskorištavanja, da moj završni rad pod nazivom “OCR tehnologije za digitalizaciju sadržaja” koristi na način da gore navedeno autorsko djelo, kao cjeloviti tekst trajno objavi u javnoj internetskoj bazi Sveučilišne knjižnice Sveučilišta Jurja Dobrile u Puli te kopira u javnu internetsku bazu završnih radova Nacionalne i sveučilišne knjižnice (stavljanje na raspolaganje javnosti), sve u skladu s Zakonom o autorskom pravu i drugim srodnim pravima i dobrom akademskom praksom, a radi promicanja otvorenoga, slobodnoga pristupa znanstvenim informacijama.

Za korištenje autorskog djela na gore navedeni način ne potražujem naknadu.

U Puli, _____ (datum)

Potpis

_____

Sadržaj

Uvod

Sa stanjem u kojoj se tehnologija danas nalazi i činjenicom da svakodnevno napreduje sve više, sve je veća zainteresiranost za digitalizacijom dokumenata. Digitalna rješenja su prihvaćena od sve više organizacija i tvrtki jer su odličan način za ubrzanje poslovanja, a istovremeno smanjenje troškova.
Ručno pretipkavanje dokumenata je vrlo sporo i može rezultirati mnogim greškama u pretipkavanju. Danas postoji tehnologija koja je vodeća u digitalizaciji teksta. OCR tehnologija, odnosno optičko prepoznavanje znakova, postala je jedna od najuspješnijih primjena tehnologije koja uključuje prepoznavanje uzorka i umjetnu inteligenciju. Predstavlja tehnologiju koja nam omogućava pretvorbu papirnog dokumenta u računalnu datoteku. Razvitak te tehnologije započeo je još u dvadesetom stoljeću, ali kao pomoć slabovidnim i slijepim osobama u čitanju. Kroz godine počele su ga koristiti banke, pošte i ostale državne institucije. Do danas ta tehnologija je veoma uznapredovala te se koristi kao glavna medota za digitalizaciju tiskanih tekstova. Iako se OCR sustavi koriste u velikoj količini i u različitim primjenama, računala se i dalje nisu u mogućnosti natjecati sa ljudskim sposobnostima čitanja.

Ovaj rad je podijeljen u 6 poglavlja, u kojem nakon obrade teme slijedi zaključak, literatura te popis slika i tablica.

U prvom poglavlju pod nazivom „OCR tehnologija“ govori se o analizi slike dokumenta te o samoj definiciji i vrstama prepoznavanja teksta i dokumentacije. Također se OCR tehnologija sagledava u području umjetne inteligencije.

U drugom poglavlju s naslovom „Povijest OCR tehnologije“ prolazi se kroz sva razdoblja u kojima je OCR napredovao. Navedeni su značajniji OCR uređaji te fontovi.

Treće poglavlje nosi naslov „Komponente OCR sustava“. U tom je poglavlju cijelokupni OCR proces podijeljen na četiri dijela - obradu dokumenta, metode prepoznavanja znakova, klasifikaciju te dodatnu obradu.

U četvrtom poglavlju pod nazivom „Prednosti i nedostaci OCR tehnologije“ detaljnije su razjašnjene pozitivne i negativne strane OCR.

Peto poglavlje „Funkcionalnost OCR tehnologije“ govori o upotrebama OCR tehnologije, od početka njenog razvitka do danas.

Poslijednje poglavlje „Rješenja“ sadrži tri softverska rješenja OCR - najkvalitetniji, srednje kvalitetan, a pristupačniji te besplatno rješenje.

1. OCR TEHNOLOGIJA

1.1. Analiza slike dokumenta

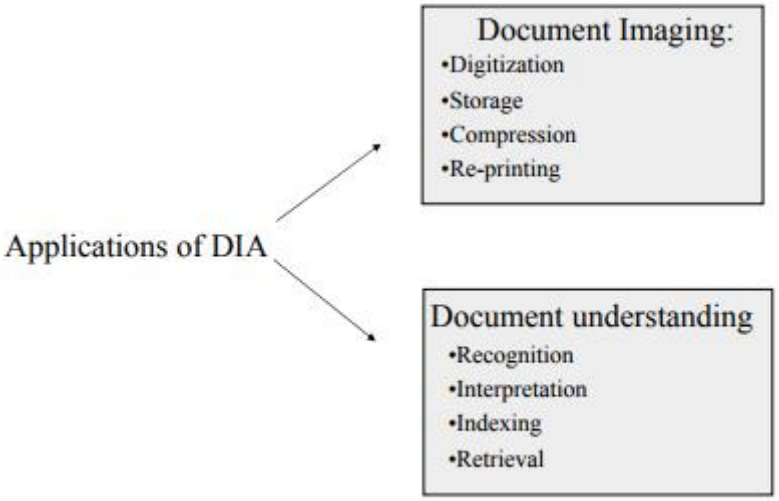
DIA (eng. *Document Image Analysis*) je područje digitalne obrade slika kojoj je cilj pretvorba slike dokumenta u oblik koji se može izmjeniti, pohraniti, preuzeti, prenostiti i ponovno upotrijebiti. Dokument predstavlja predmet koji je kreiran s ciljem da prenese informaciju, a može biti skenirana slika papirnog dokumenta, elektronički dokument, multimedijski dokument i slično.

Slike nastaju skeniranjem papirnih dokumenata, ali iako je slika u digitalnom formatu, tekstu i slikama se ne može lako pristupiti jer je dokument u suštini fotografija. To je područje u kojem analiza slike dokumenta postaje korisna jer je u mogućnosti izdvojiti tekst, rukopis, barkod i ostalo iz skeniranih slika i pretvoriti ih u tekstualnu datoteku koja se može pretraživati.

Iako ima mnogo različitih vrsta softvera za analizu dokumenata, najpoznatiji su softveri za optičko prepoznavanje znakova (OCR) koji se koristi za izdvajanje tiskanog teksta iz skeniranih slika. Osim teksa, on može obuhvatiti i druge atribute kao što su veličina slova, oblikovanje linija retka, riječi i slično.

Softveri za analizu slike dokumenta koji služi za izdvajanje teksta iz slika uglavnom uvijek koristi OCR tehnologiju. Postoje i naprednije tehnologije kao što su ICR (eng. *Intelligent Character Recognition*) koje se vrlo često dodaju OCR softverima, a koji služi za prepoznavanje ručno pisanog teksta na fotografijama. Uz OCR, često ide i OMR (eng. Optical Mark Recognition), tehnologija koja se koristi za dohvaćanje znakova iz kvadratića (eng. *checkboxes*). Ukoliko je potrebno i prepoznavanje barkodova, postoji i tehnologija OBR (eng. *Optical Barcode Recognition*).

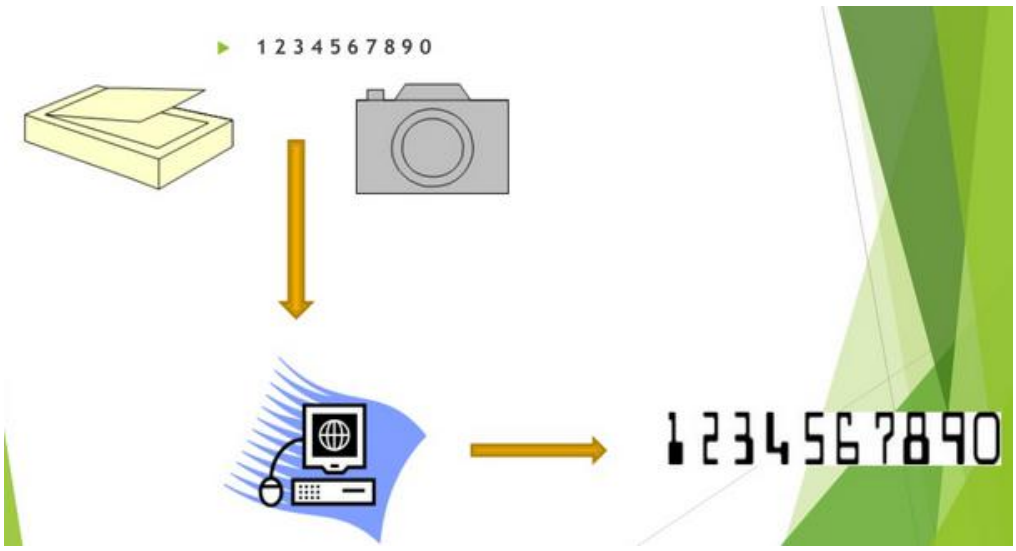
Ukoliko su u OCR softvere uključene sve tehnologije, brzina sustava, točnost i preciznost prilikom prepoznavanja se povećava uz što manju uključenost ljudskog faktora. (CVISION Technologies, 2017.)



Slika 1. Primjena analize slike dokumenta. Izvor: http://www.cvc.uab.es/~ernest/slides/ocr0607.pdf

1.2. Definicija OCR tehnologije

OCR (eng. *Optical Character Recognition*) je računalna tehnologija koja omogućuje pretvorbu skeniranih papirnatih dokumenata, PDF dokumenata i slika dokumenata snimljenih pomoću digitalnog fotoaparata u formate koji se mogu uređivati. Drugim riječima, elektronska ili mehanička pretvorba tipografskog ili tiskanog teksta u obliku slike u strojno kodirani tekst. (Panjwani, 2015)

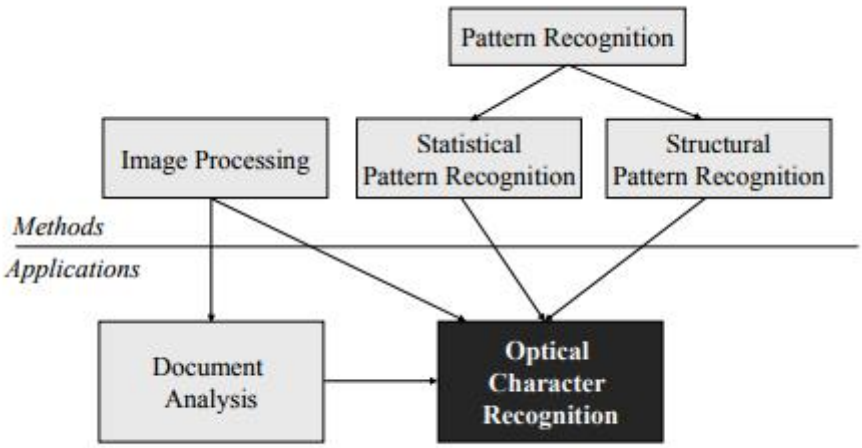


Slika 2. Proces OCR tehnologije. Izvor:

https://www.slideshare.net/karanpanjwani752/optical-character-recognition-ocr

OCR tehnologija se može koristiti u mnogim slučajevima, kao što su korekcije na papirnatom dokumentu, na primjer na tekstu ugovora, članku iz časopisa ili letka ili za upotrebu dijela teksta prilikom kreiranja drugog dokumenta. Također se koristi za digitalizaciju dokumenata kao što su računi, bankovni izvještaji, podsjetnica, pošta, statističkih podataka ili bilo kakve slične dokumentacije. OCR je tehnologija koja nam omogućava da se koristimo tiskani tekst u digitalnom obliku, a tako ga možemo uređivati, lakše pretraživati, pohraniti i sačuvati, prikazati na webu.

Kao program oponaša ljudski proces čitanja. Samim time se nalazi u domeni umjetne inteligencije (eng. *artificial intelligence*). Kao i prepoznavanje glasa, lica, otiska prsta i mrežnice, OCR također spada u pod-granu prepoznavanja uzorka (eng. *pattern recognition*). (Vynckier, 2017)



Slika 3. Metode korištene pri OCR. Izvor: http://www.cvc.uab.es/~ernest/slides/ocr0607.pdf

Prema Wallsu (2008) i Panjwani (2015), vrste prepoznavanja s obzirom na kakav font dokumentacija sadrži i na temelju čega se prepoznavanje vrši:

1. OCR (eng. *Optical Character Recognition*) - tipografski tekst, prepoznavanje na temelju znaka
2. OWR (eng. *Optical Word Recognition*) - tipografski tekst, prepoznavanje na temelju cijele riječi
3. ICR (eng. *Intelligent Character Recognition*) - rukopis ili kurzivni tekst, prepoznavanje na temelju znaka, uglavnom koristi strojno učenje
4. IWR (eng. *Intelligent Word Recognition*) - rukopis ili kurzivni tekst, prepoznavanje po riječi

Prema Wallsu (2008), dokumenti koje želimo digitalizirati mogu postojati u slijedeća tri oblika:

1. Strukturirani dokument - dokument sadrži predvidljive informacije na unaprijed određenim pozicijama na dokumentu (npr. računi, potvrde o dostavi, formulari i slično)

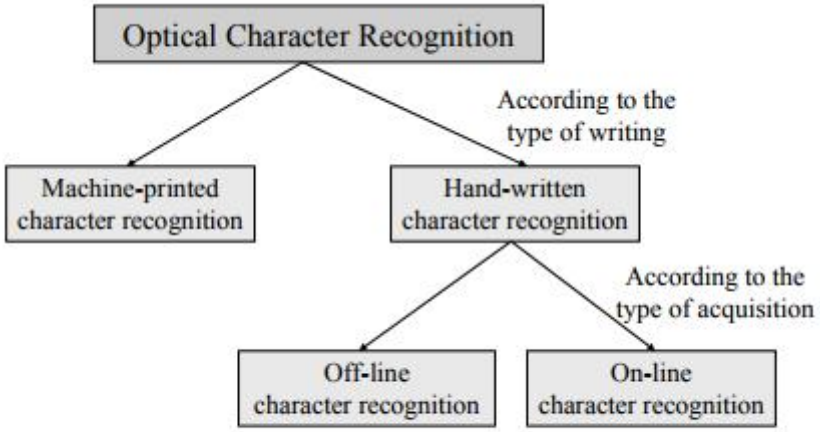
2. Polu-strukturirani dokument - dokument sadrži predvidljive informacije, ali pozicije gdje se informacije nalaze na dokumentu nisu određene (npr. računi za transport, medicinske usluge, transakcije i slično)

3. Nestrukturirani dokument - informacije ni pozicija na kojima se informacija nalazi na dokumentu nisu predvidljive (npr. medicinske evidencije, knjige, časopisi i slično)

Prema Valvenyju (2006-2007), po vrsti zapisivanja znakova na dokumentu, dokument možemo dijeliti na računalno ispisani dokument i na ručno ispisani dokument.

Prilikom prepoznavanja znakova ispisanih računalom, znakovi su definirani oblikom fonta pod koji spada dimenzija fonta (visina, širina fonta te razmak između znakova) te oblik fonta (tipografski efekti kao što su podebljanost, zakrivljenost, podcrtanost...)
OCR sustavi prilikom prepoznavanja znakova ispisanih računalom mogu naići na poteškoće zbog sličnih oblika različitih znakova, mnogo različitih fontova,

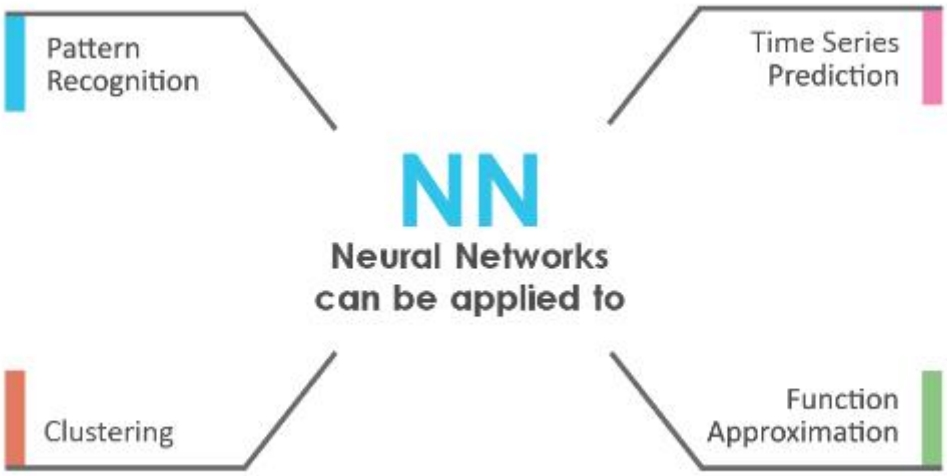
spojenih znakova, nedostataka na dokumentu kao što su prelomljene linije, mrlje, teški neprepoznatljivi znakovi, stari dokumenti, kopije kopija i slično.
Prilikom prepoznavanja fontova ispisanih rukom, OCR sustav se može susreći sa izazovima kao što su segmentacija znakova i varijabilnost oblika za isti znak.



Slika 4. Kategorizacija dokumenta prema načinu na koji je on zapisan. Izvor: http://www.cvc.uab.es/~ernest/slides/ocr0607.pdf

1.3. OCR kao koncept umjetne inteligencije

Prema Vynckieru, jedan od najvažnijih koncepata na kojem se OCR tehnologija bazira jesu neuronske mreže. Neuronske mreže su sofisticirani, fleksibilni algoritmi koji uče na temelju primjera, a kao takvi su odlični za OCR tehnologiju. One se mogu primjeniti na različite zadatke kao što su prepoznavanje uzorka, predviđanje vremenskih serija, usklađivanja funkcija, grupiranja i slično.



Slika 5. Primjena neuronskih mreža. Izvor: http://www.cvisiontech.com/resources/ocr-primer/ocr-neural-networks-and-other-machine-learning-techniques.html

Neuronske mreže su alati koji mogu pomoći u rješavanju problema s tipovima OCR. Inspirirane su načinom na koji funkcionira ljudski mozak i načinom na koji on obrađuje informacije. Sadrže zbirke matematičkih modela koji predstavljaju svojstva biološkog živčanog sustava. Neuronske mreže se sastoje od velikog broja međusobno povezanih elemenata za obradu (čvorovi) koji su vezani vezama (linkovima). Uče pomoću treninga, odnosno izlaganja skupu ulaznih i izlaznih podataka (obrazaca) te se znanja nužna za rješavanje problema pohranjuju.

Danas se koriste za rješavanje složenih problema. Najbolje su prilikom rješavanja problema koji su prekomplicirani za konvencionalne tehnologije (ukoliko problem nema algoritamsko rješenje ili problem za koje je algoritamsko rješenje previše komplicirano da bi se tražilo) i za probleme za koje ljudi imaju mogućnost rješavanja, ali tradicionalne metode nisu prikladne. Vrlo su dobre prilikom prepoznavanja uzoraka i robusnih klasifikatora, a imaju i mogućnost generalizacije u donošenju odluka na temelju nepreciznih ulaznih informacija. Nude rješenja za razne probleme klasifikacije kao što su prepoznavanje govora, signala, funkcionalna predviđanja, modeliranje sustava i slično. Glavne prednosti neuronskih mreža su njihova sposobnost učenja i otpornost na “iskrivljene” ulazne podatake. (CVISION Technologies, 2017)

OCR program svakodnevno nailazi na nove uzorke slova, razne fontove koji se ne koriste često, na dokumente koji mogu biti loše skenirani, neuredni, zamrljani, neuredno ručno pisani i slično. Neuronske mreže se tu izdvajaju kao rješenje jer imaju mogućnost analiziranja sličnosti (nepoznati uzorak uspoređuje sa već poznatim uzorcima) i generiraju izlaz koji je najsličniji već postojećim uzorcima.

Kao što je već navedeno, neuronske mreže usmjerene su na razvoj mreža koje bi trebale funkcionirati kao ljudski mozak u rješavanju problema. One služe kako bi se razumjele biološke neurološke mreže te kako bi riješili probleme u području umjetne inteligencije. Imaju veliku mogućnost prepoznavanja obrazaca, učenja iz iskustva, organizaciji i grupiranju podataka, sortiranju podataka kao relevantne i irelevantne i slično. Često se koriste u "data miningu"- ekstrakciji znanja i razumijevanju iz čistih, sirovih podataka. OCR kao sustav također pripada "data miningu" zbog mogućnosti određivanja koji oblici pripadaju kojem slovu (npr. razlika između slova "A" i slova "B"). Postoje razni uzorci određenog simbola (npr. slova "a", odnosno "A"). Osnovni uzorak slova unosi se u neuronsku mrežu koja razvija ideju o tome kako određeno slovo izgleda. Trebalo bi u mrežu unijeti i dovoljan broj uzoraka slova kako bi bio u mogućnosti prepoznati svaki oblik koji se unese na prepoznavanje.

Slika 6. Neki od mogućih oblika slova A. Izvor:

http://www.how-ocr-works.com/history/neural-networks.html

Zahvaljujući mogućnosti da program u omjer uzme visinu i širinu slova i da prepozna o kojem je slovu riječ pokazuje da neuronske mreže imaju "inteligentno" razmišljanje i dolaženje do rješenja - analizira segmentirane znakove kao što ljudi nesvjesno rade, analizira petlje, rupe, čvorove, kutove te dolaze do rješenja na temelju unaprijed definiranog izvora znanja.
Nakon što pomoću neuronskih mreža sustav prepozna sve znakove, na red dolaze ekspertni sustavi koji potvrđuju rezultat te daju konačan odgovor. Oni imaju mogućnost ispravka rezultata ili potvrde, a ukoliko su dostupne alternative, pomažu u određivanju koja je od solucija najtočnija. Pod ekspertne sustave spada automatsko učenje, uporaba tipografskih pravila i uporaba lingvistike.
Automatsko učenje daje mogućnost sustavu da uči oblike slova bez povratne informacije korisnika - dolazi do rezultata tako što proučava sve moguće pojave određenog oblika fonta. Ako jedna riječ ima neko značenje i pojavi se negdje drugdje u dokumentu, da li će imati isto značenje? OCR sustav se prilagođava tipografskim podacima koji se spominju u dokumentu te automatski nauči čitati font kojim je dokument napisan.
Najpopularniji pristup neuronskih mreža u rješavanju OCR problema se temelji na "backpropagation learning-u", odnosno pripremanju seta vježbi te učenju mreža da prepoznaju uzorke iz seta vježbi. Prilikom treninga, mreže se podučavaju kako bi one na određeni ulaz odgovorile željenim izlazom. Obuka se temelji na dva dijela: mogući ulaz i željeni izlaz sa obzirom na specifičan ulaz. Cilj treninga je mogućnost da neuronske mreže na proizvoljni ulaz kreiraju izlaz, na temelju kojeg možemo riješiti vrstu uzorka na predstavljenoj mreži. (CVISION Technologies, 2017)

1.3.1. Primjer učenja OCR fontova pomoću Bitmapa

Ukoliko želimo trenirati mrežu koja bi trebala biti u mogućnosti prepoznati 26 velikih slova koji su prikazani kao slike od 16x16 piksela, najjednostavniji način je stvaranje vektora veličine 256 (u ovom slučaju), a vektor će sadržavati "1" na svim pozicijama koje odgovaraju slovu, odnosno "0" za sve pozicije koje

predstavljaju pozadinu. Svaki uzorak se kodita kao vektor veličine 26 (jer u ovom slučaju prepoznajemo 26 slova), te stavljamo “1” na pozicije na kojima se slobo nalazi, a “0” na ostalim pozicijama. Nakon što postavimo uzorke za svako slovo, započinje učenje neuronskih mreža.
Za navedeni primjer koristi se jedan sloj neuronske mreže koji ima 246 ulaza koji odgovaraju veličini ulaznog vektora, i 26 neurona u sloju koji odgovara veličini izlaznog vektora. U svakom koraku učenja svi se uzorci iz seta za treniranje prikazuju u mreži te se izračunava sažetak kvadratne pogreške - kada je pogreška manja od navedenog ograničenja, treniranje mreže je gotovo te se ona može koristiti za prepoznavanje. (CVISION Technologies, 2017)

1.3.2. Primjer učenja OCR fontova pomoću obrazaca na temelju značajki

Pristup učenja pomoću Bitmapa funkcionira vrlo dobro, ali je istovremeno i vrlo ograničen ukoliko ga želimo proširiti i nadopuniti. Postoje slučajevi u kojima nam treba generaliziraniji pristup koji će uključivati i varijacije fontova.
Učenje pomoću obrazaca (eng. *Feature-based Classifiers*) se koristi ukoliko u definiciji fontova postoji određeni broj bitmap varijacija, bolji skup ulaza koji predstavljaju font bio bi odabran kao klasifikator (obrazac) koji bi bio nepromjenjiv na promjene veličine fonta ili točaka. Takvi klasifikatori bi trebali uključivati topološke karakteristike kao što su Eulerov broj, geometrijska svojstva (konkavnost, konveksnost i slično).
Kako danas postoji mnogo različitih vrsta fontova, prvo se treba pomoću OCR sustava odrediti koji je znak u pitanju. Kada znamo o kojem je znaku riječ, uspoređujemo ga sa svim definicijama toga znaka kako bismo odredili njegov font i njegovu veličinu. (CVISION Technologies, 2017)

2. POVIJEST OCR

Razvoj OCR tehnologije je vrlo opširan i dugotrajan, a kroz povijest se koristila u mnogobrojne svrhe. Prema Cheriet, Kharma, Liu, Suen (2007), najznačajnija povijest OCR seže još iz 1950ih godina kada su znanstvenici pokušavali zabilježiti slike znakova i teksta pomoću mehaničkih i optičkih sredstava. U početku skeniranje je bilo vrlo sporo te je moglo digitalizirati jedan redak znakova, a s vremenom se tehnologija uznapredovala na mogućnost skeniranja cijele stranice puno većom brzinom. Do 1960ih i 1970ih godina OCR se počeo širiti na poslovni svijet kao što su banke, pošte, avionske kompanije, bolnice i slično. Strojevi su u to vrijeme imali veliki postotak pogreške ukoliko je kvaliteta isprintanog dokumenta bila loša, ali postojala je sve veća potreba za napretkom OCR. Nastali su novi fontovi, OCR-A i OCR-B, koji su imali veću mogućnost prilagodbe i vrlo brzo su prihvaćeni od ISO (eng. *International Standards Organization*). Takva su postignuća omogućila vrlo dobre rezultate za mnogo razumnije cijene.

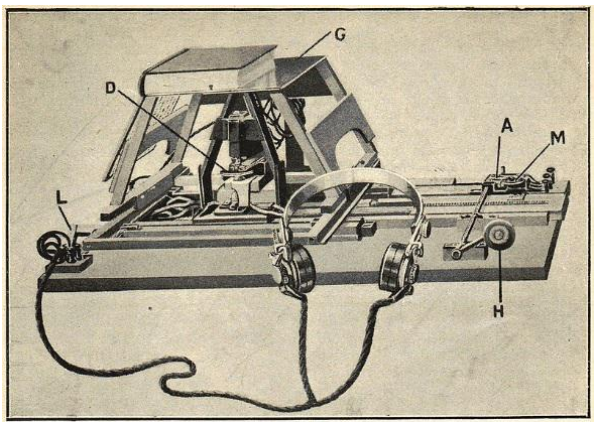
Kako je OCR razvoj napredovao, postojali su sve veći zahtjevi za prepoznavanje rukopisa jer je mnogo dokumenata, kao što su adrese na kovertama, imena, adrese, identifikacijski brojevi, cijene na čekovima, bile ispisane rukopisom. Kako su rani sustavi bili bazirani na prepoznavanju uzoraka koji se bazirao na jednostavnim geometrijskim linijama, vjerojatnost da se prepozna rukopis bio je vrlo minimalan. Prvo su dizajnirani modeli u kojima su postojala pravila kako bi OCR strojevi što lakše prepoznali rukopis, kao što su da se piše razumljivo, velikim tiskanim slovima, da se koriste jednostavni oblici znakova, da se pišu što odvojenije i slično. Kako razvoj i dalje napreduje ljudi su mogli početi pisati kako žele, a više ne postoje specifični modeli koje se treba poštovati, sve zbog algoritama kao što su obrada dokumenta, izvlačenje značajki, klasifikacija i ostale metode koje će biti opisane u nastavku.

Prema Schantzu (1982), navedena su okvirna razdoblja razvoja OCR, a neka prijelomna razdoblja su detaljnije objašnjena.

- Od 1870. do 1931. nastale su prve ideje o OCR tehnologiji koja bi služila kao pomoć slijepim i slabovidnim osobama. Nastali su strojevi “Optophone” i “Tauschek”. (3)
- Od 1931. do 1954. nastali su prvi OCR alati i strojevi koji su imali mogućnost interpretacije Morseovog koda i čitanje teksta na glas. Prva kompanija koja je uspjela izgraditi stroj i prodati ga je “Intelligent Machines Reasearch Corporation”. (4)
- Od 1954. do 1974. razvijen je “Optacon” - prvi prijenosni OCR uređaj. Taj i slični uređaji korišteni su za čitanje adresa u poštanskim uredima i za čitanje kupona iz “Reader’s Digest”. (3) (5)
- Od 1974. do 2000. osnovane su kompanije kao što su ABBYY, Caere Corporation i Kurzweil Computer Products. (6) Razvijaju se uređaji koji mogu prepoznati više fontova i koji mogu pročitati bilo kakav tekst, a najčešće se koriste pri čitanju putovnica i oznaka cijena.
- Od 2000. do danas napravljeni su softveri koji su besplatni i dostupni online, a najpoznatiji su Adobe Acrobat i Google Drive. (7)

2.1. Povijest OCR uređaja

Optophone je uređaj koji je nastao oko 1914. godine, a izumio ga je Dr, Edmund Rournier d’Albe. Uređaj je izumljen kako bi omogućio slijepima čitanje. Iako je bio nalik na današnji skener, on nije skeniranu sliku pretvarao u elektronički tekst nego u glazbeni ton. Koristeći foto senzore selenija koji je prepoznavao tamnije dijelove stranice, odnosno tintu, stroj je specifično slovo pretvarao u glazbeni ton koju je slijepa osoba mogla tumačiti. Nedostatak je bio u tome što je osoba morala naučiti i razumjeti koja melodija predstavlja koje slovo, što je vrlo dug i mukotrpan proces. U pokusu napravljenom 1918. godine, Mary Jameson je bila u mogućnosti pročitati, odnosno protumačiti jedno slovo po minuti. Uz napredak tehnologije. kroz sljedećih par godina uspjela je pročitati i do 60 slova u minuti te je time postala prva slijepa osoba koja je pročitala stranicu knjige. (Wikipedia, 2017)



Slika 7. Optophone. Izvor: https://en.wikipedia.org/wiki/Optophone

Optacon je elektromehanički uređaj koji pomaže slijepima u čitanju tiskanog materijala koji nije napisan u Brailleovom pismu. Izumio ga je John Linvill oko 1966. godine. Uređaj radi tako što pretvara tiskani materijal u taktilnu sliku koja se putem impulsa šalje slijepoj osobi. Osoba uređaj koristi tako što stavi prst indeks u poseban dio uređaja. Da bi osoba razumjela tekst koji je uređaj skenirao mora naučiti micati kameru uređaja točno po liniji tiskanog materijala, naučiti oblike slova i prepoznavati oblike slova koje su ujedinjene u riječi i rečenice. (Wikipedia, 2017)

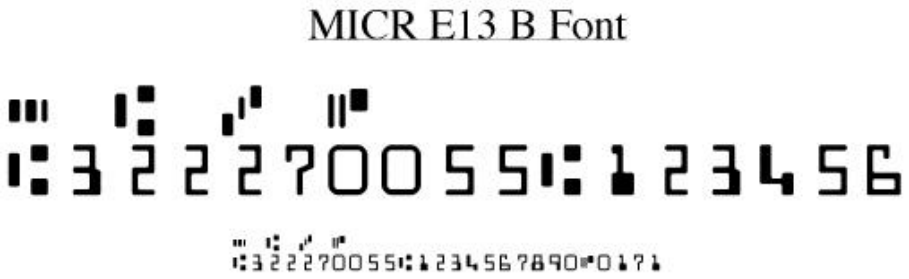
Slika 8. Optacon. Izvor: https://en.wikipedia.org/wiki/Optacon

”The Intelligent Machine Corporation” izumila je 1959. stroj koji je mogao čitati

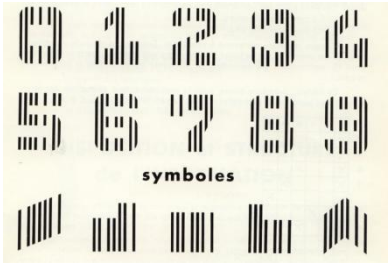
samo jedan font u jednoj veličini. S vremenom su nastali sustavi koji su imali mogućnost čitanja više fontova istovremeno. Takvi prvi sustavi su bili izuzetno spori, a uglavnom su bili limitirani na prepoznavanje specifičnih fontova: E13B, OCR-A i OCR-B. Ti fontovi su nastali kako bi ih računalo moglo pročitati, a nalazili su se na čekovima. (Wikipedia, 2017)

2.2. Povijest OCR fontova

Prema Smithu (2017), OCR-A je font koji je kreiran 1966. godine kojega strojevi mogu pročitati. Strpjevi toga vremena nisu bili u mogućnosti raspoznavati između različitih vrsta fontova pa je kreiranje jednoga specifičnoga bilo vrlo potrebno. OCR-A nije prvi font koji je stroj mogao pročitati - 1950ih u bankovim industrijama su razvijena dva fonta, E-13B (Sjeverna Amerika i Ujedinjeno Kraljevstvo) te CMC-7 (Francuska, Španjolska, Južna Amerika) koji su sadržavali brojeve od 0-9 i par specifičnih kodova koji su se koristili za komunikaciju računala.
Ideja OCR-A fonta je kreiranje skupa znakova koje će računalo, ali i ljudi, biti u mogućnosti pročitati. Problem je bio u tome što je dizajn tog fonta bio više prilagođen računalima. S vremenom je nastao font OCR-B čiji je dizajn bio mnogo privlačniji i mnogo sličniji standardiziranim europskim fontovima.



Slika 9. E13-B. Izvor: http://www.matchfonts.com/MICR-Font/



Slika 10. CMC-7. Izvor:

http://www.bytewaytech.com/recogniform-sdk/micr-cmc7/

Slika 11. OCR – A. Izvor: https://en.wikipedia.org/wiki/OCR-A

Slika 12. OCR – B. Izvor: https://en.wikipedia.org/wiki/OCR-B

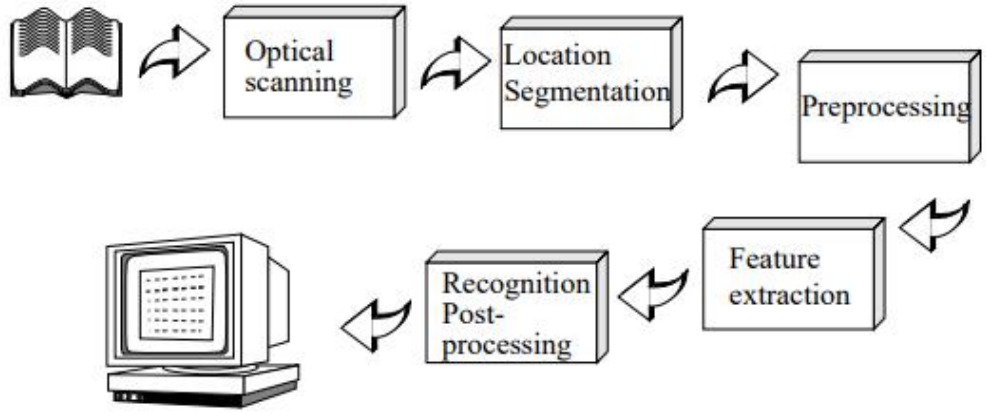
Također poznati font je Farrington B (7B-OCR), brojčani font koji se i dan danas koristi na kreditnim i debitnim karticama.

Slika 13. 7B-OCR. Izvor: http://www.barcodesoft.org/farrington7b-font

3. KOMPONENTE OCR SUSTAVA

Kako je već spomenuto, cilj OCR tehnologije je da tiskani, papirni dokument pretvori u računalnu datoteku koja se može uređivati. Tekstualnu datoteku možemo dobiti i pretipkavanjem dokumenata što može biti vrlo dug i mukotrpan proces. Prema Vynckieru, čovjek može u minuti utipkati oko 200 znakova, dok OCR programi imaju mogućnost prepoznavanja oko minimalno 1600 znakova u sekundi (kod boljih računala i do tri puta više). Ono što nam je potrebno kod OCR tehnologije je digitalni fotoaparat ili skener koji služi kao “oko računala”. Fotoaparat ili skener generiraju sliku dokumenta koju ne možemo obrađivati u nekom od programa za uređivanje teksta (npr. Word) - ti uređaji ne prepoznaju znakove na dokumentu nego spremaju cijeli dokument kao sliku. Pomoću OCR tehnologije, svaki znak na dokumentu se skenira zasebno te se dokument sprema na računalo kao tekstualni dokument. Ali kako to zapravo funkcionira?

Ljudsko oko prepoznaje znakove po obrascima od kojih se oni sastoje (tamnijih i svjetlijih dijelova, krivulja i slično), a mozak koristi te značajke kako bi zaključio o kojem je znaku riječ (mozak najčešće skenira cijelu riječ ili čak i skupinu riječi istovremeno). Nismo niti svjesni koliko truda je zapravo uloženo u čitanje samo jedne riječi, ali za računalo to nije tako lagan zadatak. Računalu moramo predočiti sliku onoga što želimo da pročita koju smo generirali putem skenera ili fotoaparata. Računalo, iako sada ima sliku, ono i dalje ne shvaća što se na slici nalazi - za njega je to samo skup piksela bez ikakvog značenja. Da bi računalo moglo shvatiti što se na slici nalazi, osmišljena je OCR tehnologija. Zadatak OCR je zapravo vrlo kompleksan. Razlog tome je što postoji veliki broj znakova, veliki broj fontova, a u obzir moramo uzeti i rukopis. Kao što je logično, svaka osoba piše pojedinačno slovo na sličan, ali opet dovoljno različit način da ga računalo teže prepozna.



Slika 14. Komponente OCR sustava. Izvor: https://www.nr.no/~eikvil/OCR.pdf

Na slici su prikazani koraci koji se izvršavaju prilikom prepoznavanja teksta. U nastavku su navedeni svi koraci prilikom digitalizacije dokumenta te su detaljnije objašnjeni.

3.1. Obrada dokumenta

Obrada dokumenta uključuje poboljšavanje kvalitete slike kako bi prepoznavanje teksta bilo što uspješnije i optimalnije te kako bi korak ekstrakcije (izvlačenja) značajki moglo funkcionirati što ispravnije i učinkovitije. Uključuje skeniranje, prikaz fotografije te obradu slike (poravnanje dokumenta, uklanjanje nedostataka kao što su prašina, ogrebotine, mrlje i slično, binarizaciju, određuje se zona interesa). (Vynckier, 2017)
Nakon što se raspoznaje razlika između znaka i njegove pozadine, slijedi korak u kojem se uklanjanju nedostatci i uklanjaju se suvišne informacije koje sustav ne treba obrađivati kako bi svi daljnji koraci bili što uspješniji. Nakon ove faze, podaci koji se trebaju obrađivati biti će smanjeni na male binarizirane slike sa jednostavnim obrisima. (Cheriet i sur, 2017.)

3.1.1. Ispis

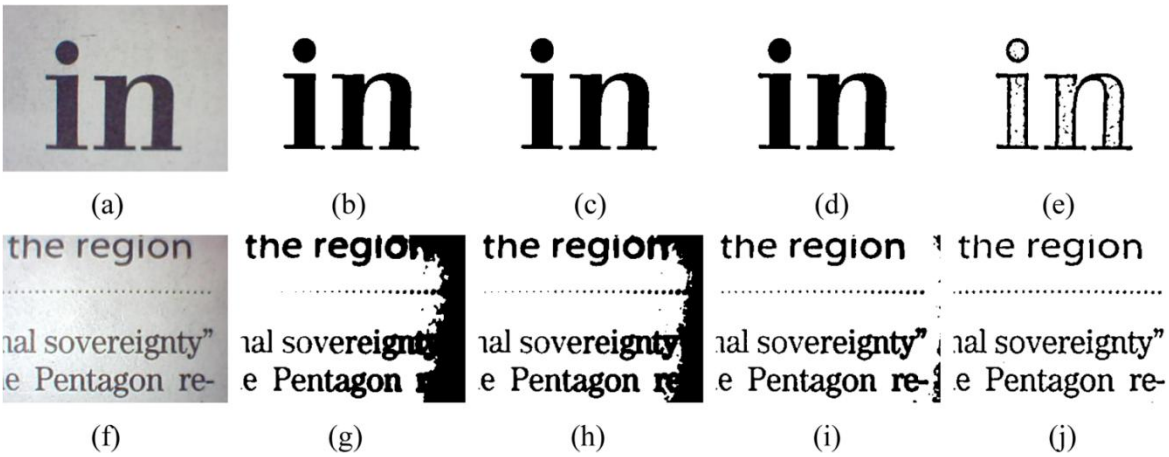
Kvaliteta originalnog ispisa dokumenta je vrlo bitna stavka u točnosti OCR procesa. Prljave oznake, mrlje kave, mrlje tinte, nabori i slično smanjuju vjerojatnost da OCR prepozna znakove, slova i riječi ispravno.

3.1.2. Skeniranje

Ispis dokumenta se provlači kroz optički skener. Skeniranje kao proces predstavlja trenutak u kojem nastaje slika originalnog dokumenta. Printani dokumenti uglavnom su spoj crne tinte na bijelom dokumentu, mnogi skeneri danas imaju mogućnost pretvaranja dokumenta u binarni prije izvođenja OCR procesa. Proces “thresholding” je vrsta obrade fotografije kojoj je zadatak segmentacija fotografije. Taj korak je potreban jer prepoznavanje koje slijedu u nastavku isključivo ovisi o kvaliteti binarne fotografije. (Eikvil, 1993)
Postoji mnogo vrsta skenera, a za OCR su najbolji oni koji imaju mogućnost povlačenja stranica automatski, jednu za drugom. Nakon što skenira jednu stranicu, OCR program prepoznaje tekst na njemu, a kada je gotov, automatski prelazi na iduću. Kod nekih drugih vrsta skenera moramo sami umetati stranicu po stranicu u skener što je mnogo dugotrajniji proces. Ukoliko koristimo digitalni aparat, trebao bi se koristiti makro fokus kako bi slika bila što oštrija. (Vynckier, 2017.)

3.1.3. Binarizacija dokumenta

Ukoliko skener nema mogućnost binarizacije, to je prvi korak OCR sustava. Dokument bi trebao biti jednobitni, odnosno OCR tehnologija se zasniva na binarnom procesu - prepoznaje stvari koje jesu ili koje nisu. Ako je originalna slika bez mrlja i grešaka, sve crne boje biti će dio znaka koji OCR treba prepoznati, a sve bijelo je pozadina. Ukoliko je tekst isprintan na šarenoj pozadini, program neće biti u mogućnost raspoznati pozadinu od samog znaka. Ako je pozadina slabijih boja, a sam tekst u kontrastu, ne bi trebalo biti problema u prepoznavanju (npr. svijetloplava pozadina sa tamnoplavim tekstom). U većini slučajeva možemo namjestiti kontrast i svjetlinu do trenutka gdje se raspoznaje tekst od pozadine, ali u situacijama gdje je tamni tekst ispisan na tamnoj pozadini to nije tako jednostavno. Isto vrijedi i ukoliko se svijetli tekst nalazi na svijetloj pozadini. Naravno, ukoliko se koristi kvalitetniji i sofisticiraniji OCR program, ne bi trebali brinuti o binarizaciji. (Vynckier, 2017.)



Slika 15. Proces binarizacije. Izvor:
http://electronicimaging.spiedigitallibrary.org/article.aspx?articleid=1352458

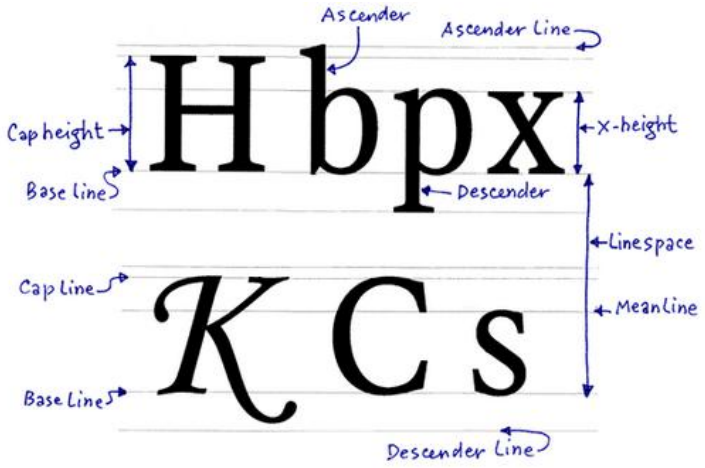
3.1.4. Segmentacija

Nakon što je obrada dokumenta gotova, trebali bismo imati crno - bijelu sliku na kojoj je tekst lako čitljiv, a pozadina vrlo neutralna.

Segmentacija predstavlja proces koji određuje sastavnice slike. Uključuje lociranje područja dokumenta te razlikuje tekst od fotografija, grafova i ostalih komponenti koje se mogu nalaziti na dokumentu. Primjerice, prilikom automatskog razvrstavanja pošte, adresa mora biti locirana i izdvojena od ostalih komponenti koje se nalaze na koverti, kao što su markice, logotipovi i ostalo. Primjenjena na sam tekst koji se nalazi na dokumentu, segmentacija predstavlja izolaciju znakova ili riječi. Većina sustava za prepoznavanje segmentira riječi u izolirane znakove koji onda prepoznaju pojedinačno. (Eikvil, 1993)

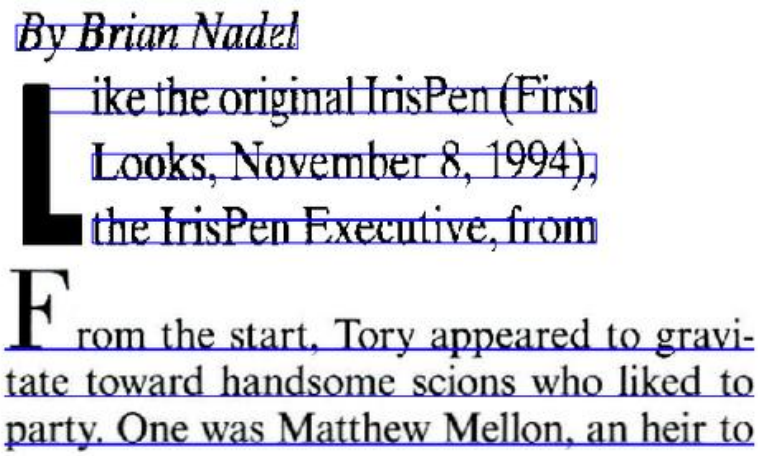
Prema Vynckieru (2017), segmentirati možemo linije (“line segmentation”), riječi (“word segmentation”) i znakove (“character segmentation”).

Kod segmentacije linije zone se interesa dijele u različite linije. Veličina maloga slova se zove “x-visina”, odnosno “x-height”, a predstavlja bazu znaka. Dio znaka koji se nalazi iznad “x-visine” predstavlja “ascender”, a ispod “descender”. Ti dijelovi znakova se ignoriraju prilikom određivanja veličine fonta.



Slika 16. Linijska segmentacija teksta. Izvor: http://www.how-ocr-works.com

Neka slova predstavljaju značajan problem segmentacije teksta, kao veliko početno slovo na početku članka iz novina ili knjiga. Na slici 10. prikazane su neke od mogućnosti:



Slika 17. Moguće komplikacije prilikom segmentacije. Izvor: http://www.how-ocr-works.com

Segmentacija riječi i znakova predstavlja odvajanje jedne riječi od druge, ali i odvajanje znakova unutar jedne riječi.

Slika 18. Segmentacija riječi i znakova. Izvor: http://www.how-ocr-works.com

Prema Eikvilu (1993), glavni problemi prilikom segmentacije mogu biti:

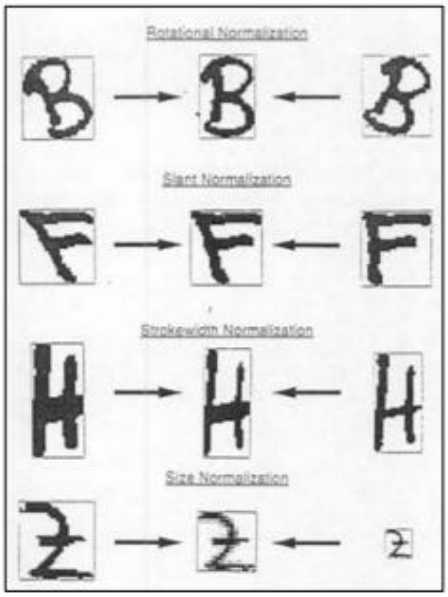
- Izdvajanje znakova koji se dodiruju - takvi nedostaci mogu dovesti do toga da se nekoliko različitih znakova tumači kao jedan znak. Takva spajanja će se na dokumentu najčešće dogoditi ukoliko je originalni dokument vrlo taman ili ako se skenira na loš način.
- Razlikovanje buke teksta - točke i naglasci se mogu pogrešno shvatiti.
- Zamjena slike, grafova, matematičkih simbola i ostaloga za tekst - dovodi do pogrešnog teksta.
- Zamjena teksta za slike, grafove, matematičke simbole i ostalo - tekst neće biti proslijeđen u fazu prepoznavanja.

3.1.5. Uklanjanje nedostataka

Fotografije nastale skeniranjem vrlo često imaju mnogo nedostataka te je najčešće potrebno smanjiti buku na fotografijama ili izoštriti rubove znakova, popuniti male rupe ili ukloniti dodatnu tintu na rubovima znakova i slično. Izglađivanje slike i uklanjanje buke fotografije vrši se pomoću filtriranja fotografije.

Normalizacija znakova se smatra jednim od najbitnijih koraka prilikom obrade dokumenta. Slika znaka se preslikava na ravninu čija je veličina unaprijed definirana. Cilj normalizacije znaka je smanjenje varijacije znaka i pretvaranje znaka u standardni oblik poboljšavajući njegovu orijentaciju, veličinu, nagib i

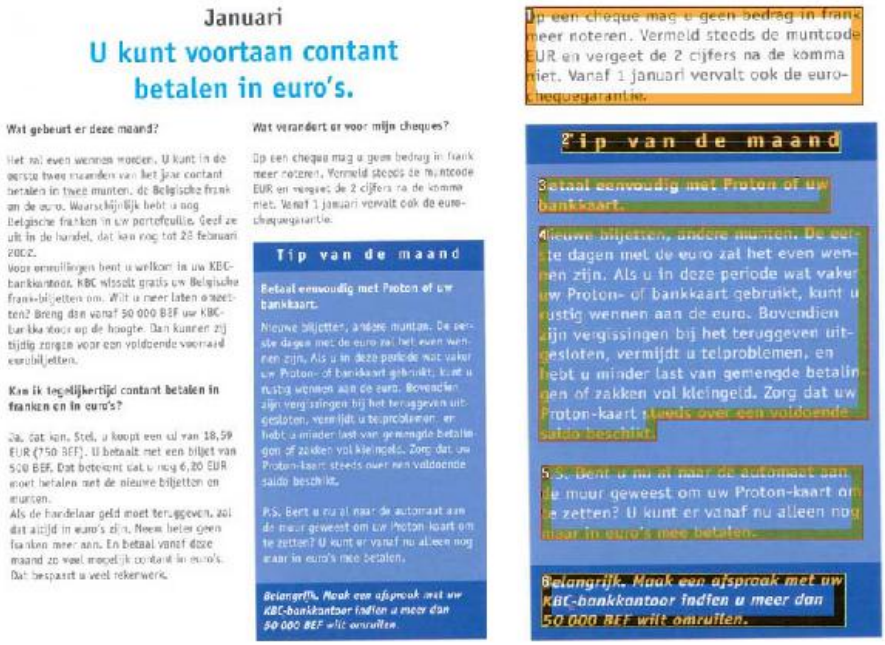
debljinu poteza, kako bi se olakšala ekstrakcija (izvlačenje) značajki i kako bi se poboljšala točnost klasifikacije.



Slika 19. Normalizacija znaka. Izvor: http://www.cvc.uab.es/~ernest/slides/ocr0607.pdf

3.1.6. Zona interesa

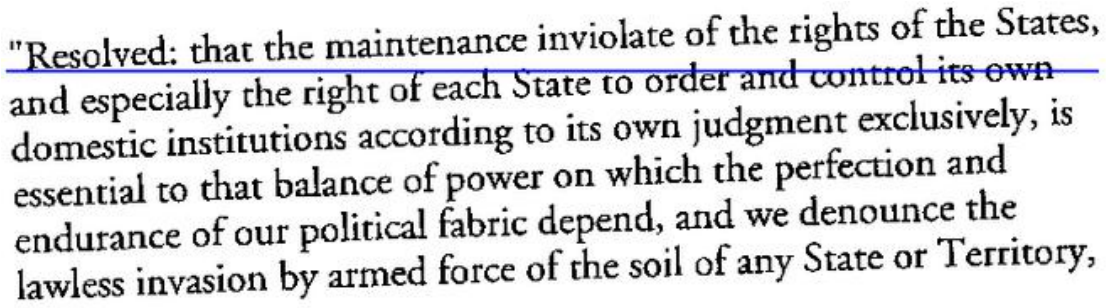
Sustavi OCR nude mogućnost odabira zone interesa - slika dokumenta može sadržavati veliki naslov, tekstualne stupce, fotografije, tablice i slično, a možda nemamo potrebe provući sve te podatke kroz program. Postoji više načina kako možemo odabrati zone interesa, a to su pomoću automatske analize stranice, ručno ili koristeći fiksnih, već postojećih prozora. Većina OCR programa danas obrađuju stranicu slovo po slovo, riječ po riječ, liniju po liniju.



Slika 20. Odabir zone interesa. Izvor: http://www.how-ocr-works.com

3.1.7. Ispravljanje nagiba dokumenta

Prilikom skeniranja dokumenta, papir može biti umenut nakošeno ili originalni dokument može sadržavati tekst koji je ispisan nakošeno. Takav efekt na fotografiji drastično smanjuje točnost cjelokupnog procesa te je on jedan od primarnih zadataka koji treba biti izvršen ukoliko očekujemo dobar rezultat OCR programa. Algoritmi za analizu izgleda stranice i prepoznavanje znakova vrlo su osjetljivi na zakretanje stranica.



Slika 21. Ispravljanje nagiba teksta. Izvor: http://www.how-ocr-works.com

3.2. Metode prepoznavanja znakova

Prema Woodfordu (2010/2014), dvije najpoznatije metode na temelju kojih su softveri bazirani su prepoznavanje uzorka (eng. *pattern matching*) i ekstrakcija značajki (eng. *feature extraction*).

3.2.1. Prepoznavanje uzorka

Princip prepoznavanja uzorka je specifičan jer se kod tog pristupa značajke ne izdvajaju nego se znak sagledava cijelosti. Matrica koja sadržava sliku ulaznoga znaka direktno se uspoređuje sa skupom prototipnih znakova koji predstavljaju klasu. Izračunava se udaljenost između uzoraka te se dodjeljuje klasa koja najviše odgovara ulaznom znaku. Takva tehnika je vrlo jednostavna za korištenje i implementaciju te se koristi u mnogim komercijalnim OCR sustavima. Dakako, takva tehnika je vrlo osjetljiva na nedostatke i na varijacije fontova i rukopisa. (Eikvil, 1993)

Kako nije moguće očekivati da svaki čovjek na svijetu znak piše na identičan način te kako nije moguće očekivati da će svaka dokumentacija biti ispisana u istom fontu, u 1960im godinama smišljen je specifičan font koji se zvao “OCR-A”. Font se nije koristio prilikom ispisivanja svih dokumenta, ali se koristio na čekovima i službenim dokumentima. Printeri su printali samo u tom fontu, a svaka OCR tehnologija tog vremena prepoznavala je taj font. Svi znakovi su bili jednako visoki, a svaka krivulja znaka je bila osmišljena tako da se u potpunosti razlikuje o drugih znakova i da program nema mogućnost zabune prilikom prepoznavanja znaka. Problem je bio u činjenici što nisu svi printeri printali u OCR-A fontu, a niti ljudski rukopis mu nije sličan. Trebalo je naučiti programe da prepoznaju znakove u mnogim različitim fontovima i iako su mogli prepoznavati mnogo znakova, nije bilo garancije da će prepoznati svaki novi font. Osim OCR-A, postojao je i OCR-B font.

3.2.2. Ekstrakcija značajki

Cilj ekstrakcije značajki je dohvatiti bitne karakteristike pojedinačnog simbola. Taj korak je jedan od najtežih koraka prilikom prepoznavanja znakova. U

takvoj se metodi ekstrahiraju jedinstvene značajke simbola te se uspoređuju sa opisima razreda dobivenih tijekom faze treninga. Dodjeljuje mu se opis koji najbolje odgovara ulaznom znaku. (Eikvil, 1993)

Znak se sagledava kao kombinacija individualnih linija i poteza. Ta metoda je poznata i kao ICR (eng. *Intelligent Character Recognition*). Znakovi se razlikuju jedan od drugoga na temelju principa kojima su zapisani. Svaki znak ima specifične komponente kao što su linije koje mogu biti nakošene ili prekrižene, kutovi između linija i slično. Kao primjer možemo uzeti slovo A. Ono je u različitim fontovima dovoljno različito zapisano da ga program ne može prepoznati, ali uvijek se sastoji od dvije nakošene linije koje se dodiruju na vrhu, a po sredini ima horizontalnu liniju koja ih spaja. Na temelju toga, program će prepoznati skoro svako slovo A, bez obzira kojim je fontom zapisano. Ova metoda je mnogo popularnija u OCR rješenjima jer ima mogućnosti prepoznavanja mnogo više fontova, a što je najvažnije, prepoznaje i ljudski rukopis.

Slika 22. Prepoznavanje značajki – učenje slova A. Izvor: http://www.explainthatstuff.com/how-ocr-works.html

3.3. Klasifikacija

Klasifikacija je proces prepoznavanja svakog znaka te dodjela odgovarajuće klase svakom znaku. Prema Eikvilu (1993), postoje dvije vrste klasifikacije za prepoznavanje znakova, a to su teorija odluke i strukturalna metoda.

3.3.1. Teorija odluke

Glavni principi teorije odluke su princip podudaranja, statistička klasifikacija i neuronske mreže.

Princip podudaranja obuhvaća tehnike koje se temelje na mjerama sličnosti, odnosno izračunava se udaljenost između vektora značajki koje opisuju karakter i opis svakog znaka. Najčešća mjera koja se koristi je Euklidska udaljenost.

Ukoliko koristimo statističku klasifikaciju primjenjujemo “propablistički” pristup prilikom prepoznavanja. Koristi se klasifikacija koja je optimalna, odnosno odabire se klasa koja je dala najmanju vjerojatnost pogreške.

Neuronske mreže su mreže koje se sastoje od nekoliko međusobno povezanih slojeva. Svaki element sloja izračunava procjenjeni zbroj ulaznih vrijednosti i pretvara ih u izlazni dokle god ne dobije kao rezultat željeni izlaz.

3.3.2. Strukturalna metoda

Unutar strukturalnih metoda, najpoznatije su sintaksne metode. Mjere sličnosti između znakova se formuliraju korištenjem gramatičkih pojmova. Svaka klasa je definirana svojim gramatičkim pojmovima.

3.4. Dodatna obrada

Dodatna obrada OCR uključuje (Eikvil, 1993):

3.4.1. Grupiranje

Kao rezultat svih navedenih koraka, dokument predstavlja samo skup pojedinačnih znakova. Znakovi sami po sebi ne sadržavaju dovoljno informacija te je cilj povezati pojedinačne znakove koji pripadaju istom nizu čineći od njih riječi i brojeve. Grupiranje znakova u nizove se temelji na mjestu tih znakova na dokumentu. Znakove za koje je utvrđeno da su povezani se grupiraju zajedno. Udaljenost između znakova unutar riječi i rečenice uglavnom je dovoljno velika da OCR sustav to može raspoznati, ali ukoliko je riječ o nagnutim fontovima ili rukopisu, sustav može učiniti pogreške prilikom grupiranja znakova. Mnogi OCR sustavi koriste riječnike pojedinih jezika te mogu grupirati i prepoznati riječi pomoću baze podataka koju sadržavaju.

Ukoliko se riječ koju je sustav dobio grupiranjem ne nalazi u riječniku, on javlja grešku te ju promjeni u što sličniju riječ.

3.4.2. Prepoznavanje grešaka i korekcija

Većina programa daje mogućnost da pregledamo i ispravimo svaku stranicu tokom samog procesa. Program skenira i pregleda stranicu, zatim koristi provjeru pravopisa kako bi istaknuo na neke krivo napisane riječi te daje mogućnost da ih automatski ispravimo. Tu mogućnost se može i isključiti jer nije pogodna ukoliko se radi o velikoj količini stranica. Kvalitetniji i sofisticiraniji programi imaju mogućnost analize riječi koje se nalaze oko riječi koja je eventualno pogrešna. Na primjer, ukoliko je program prepoznao riječ kao “pos laje”, pretpostavlja da uz “laje” ide “pas”, a ne “pos”.
Iako je OCR tehnologija do danas vrlo napredovala, niti jedan program nije savršen i uvijek postoje greške, pogotovo ako je ispis dokumenta vrlo star ili u lošem stanju. Kako bi skenirani dokument bio pretvoren u tekstualni dokument sa stopostotnom točnošću, zadnji korak procesa je da sami prekontroliramo, pronađemo i ispravimo greške.

3.5. Točnost OCR

Točnost OCR ovisi o kvaliteti dokumenta, odnosno slike koju obrađujemo, ali i o fontu koji se nalazi na dokumentu.
Ukoliko je riječ o specifičnom isprintanom tekstu, kao što su OCR-A, OCR-B i ostali dizajnirani fontovi, većina OCR sustava znakove vrlo lako raspoznaje te je točnost oko 99.99% uz vrlo brzo čitanje znakova.
Ukoliko je riječ o OCR strojevima koji moraju prepoznavati više različitih fontova istovremeno (bilo stiliziranih ili nestiliziranih), koristi se ekstrakcija značajki. Baza podataka sadrži opise svakog znaka, a ne znak sam za sebe. Zbog toga u mogućnosti je pročitati mnogo različitih fontova, ali nije u mogućnosti pročitati sve fontove koji postoje. Iako mnogi sustavi tvrde da mogu prepoznavati više različitih fontova istovremeno, nijedan sustav nije u

mogućnosti prepoznati 100% znakova. (Eikvil, 1993)

Kako bi preciznost i točnost bila što veća, možemo poduzeti neke od koraka prilikom procesa. Optimizacija se općenito dijeli na dvije faze - prije skeniranja i tijekom skeniranja dokumenta.

Na kvalitetu i točnost rezultata utječe oblik dokumenta, odabir fonta i odabir boja. Tekst bi trebao biti ispisan na bijeloj površini, sa ograničenim linijama, bojama i fontovima kao što su Courier i Sans Serif, sa veličinom slova 10-13. Tijekom skeniranja na točnost najviše utječu razlučivost i čistoća skenirane slike. Slike bi trebale biti skenirane na najmanje 300 točaka po inču. Čak i nakon optimizacije dokumenata, softver i dalje može izbaciti greške. Ispravljanje grešaka se radi ručno na kraju skeniranja dokumenta. (DocuFi, 2017.)

Ukoliko je riječ o latinskom pismu, točnost i dalje nije stopostotna, bez obzira da li se radi o čistoj ili oštećenoj slici dokumenta. Pretpostavlja se da je točnost između 81% - 99%. Što se tiče rukopisa, teksta drugih pisama i kurzivnog teksta, točnost još nije određena jer ih većina OCR sustava danas i dalje nije u mogućnosti pročitati. Kao takva, ta pisma su još uvijek predmet istraživanja. Točnost bila mnogo veća kada bi se prepoznavanje vršilo pomoću cijele riječi i rječnika nego analiza pojedinačnih značajki znakova. (Panjwani, 2015.)

Postoji mogućnost kvalitetnog čitanja rukopisa, ali se moraju poštovati određeni standardi (znakovi moraju biti tiskani što veće kako bi zadržali dobru razlučivost, trebali bi biti tiskani u određene kućice, trebale bi se izbjegavati praznine u pisanju i dodatne stilske varijacije).

4. PREDNOSTI I NEDOSTACI OCR TEHNOLOGIJE

OCR sustavi imaju mnogo više prednosti nego nedostataka. Ukoliko želite pretvoriti dokument u digitalni format za uređivanje, OCR je najbolji izbor. Može uštediti vrijeme i trud jer pruža brzu alternativu pretipkavanja. Također pruža pretvorbu dokumenta u razne formate kao što su Microsoft Word, Text, Excel, PDF i mnoge druge. (Investintech, 2000-2017.)
Dokumenti se mogu pohraniti, uređivati, distribuirati. Prednosti su i lagana, praktična i besplatna dostupnost on-line usluga (npr. Google Docs), bitno je samo da postoji internetska mreža. Danas je OCR neophodan sustav u većini velikih tvrtki kao što su pravne, financijske institucije ili vladine agencije. (CVISION Technologies, 2017.)

Iako je OCR tehnologija mnogo poboljšana od svojih početaka, greške se i dalje pojavljuju. Sustavi su većinom omnifont, to jest prepoznaju skoro sve dostupne fontove, ali kurzivni tekst je još uvijek vrlo teško čitljiv i prepoznatljiv. Najveći nedostatak se nalazi činjenici da cijeli proces ovisi o kvaliteti dokumenta koji želimo digitalizirati. Ukoliko je tekst pisan rukom, ako je dokument star, razderan, ima mrlje ili neke druge oznake koje sprječavaju lako prepoznavanje sadržaja, stroj će imati poteškoća sa čitanjem i prevođenjem sadržaja u digitalni oblik. (Investintech, 2000-2017.)

Prema Eikvilu (1993), najčešći nedostaci koji uzrokuju greške:

- Varijacije u obliku znakova,
- Deformacije kao što su slomljeni znakovi, zamrljani znakovi ili mrlje na dokumentu,
- Neprikladne varijacije razmaka između znakova,
- Mješavina teksta, slika, grafikona, matematičkih simbola...

Vjeruje se da će čitanje kurzivnog teksta biti moguće jedino koristeći kontekstualne i gramatičke informacije što može biti vrlo zahtjevno i sporo prilikom učenja. (Panjwani, 2015.)

5. FUNKCIONALNOSTI OCR

Tokom godina OCR sustavi su se koristili u razne svrhe, a nastalo je mnogo komercijalnih proizvoda koji zadovoljavaju zahtjeve korisnika. U ovom poglavlju će biti navedeni neki od glavnih područja primjene OCR tehnologije. OCR tehnologija i dalje je najpopularnija je prilikom digitalizacije sadržaja. (Eikvil, 1993)

5.1. Unos podataka

OCR se koristi u tehnologijama kojima je potreban unos velikih količina ograničenih podataka. U samom početku razvoja OCR, ta se tehnologija koristila za čitanje bankovnih dokumenata. U bankarskoj industriji se koristi još od 1950ih, a tada se koristila u obradi čekova. Svi bankovni čekovi bili tiskani u posebnom fontu, OCR - A, a ti brojevi su bili otisnuti na dnu čeka. “Bank of America” je prva banka koja je koristila OCR alate.

Slika 23. Ček iz američke banke. Na dnu se nalazi broj ispisan u OCR – A fontu. Izvor: http://www.how-ocr-works.com/history/history.html

Bili su ograničeni čitati i prepoznavati samo ograničeni skup tiskanih znakova, uglavnom brojki i nekoliko osnovnih simbola. Dizajnirani su za čitanje podataka kao što su brojevi računa, identifikacija kupaca, brojevi članaka ili količina novca. Formati dokumenata su također bili optimizirani i prilagođeni.

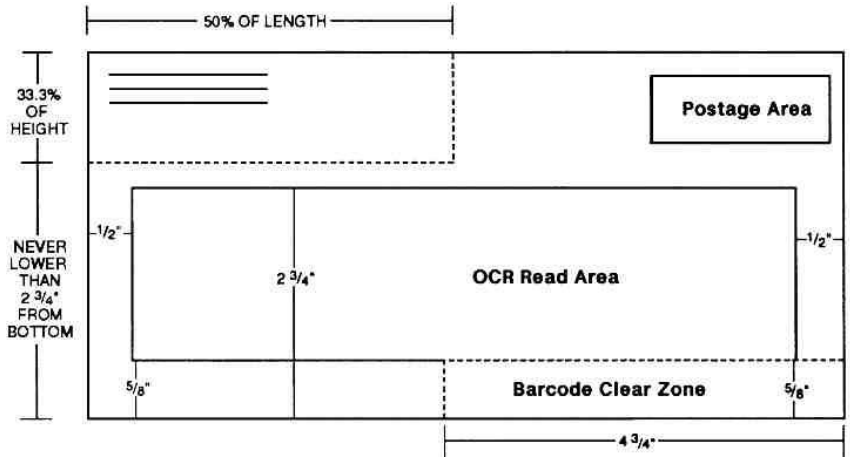
Prilikom prepoznavanja takve dokumentacije brzina je izuzetno velika (moguće je prepoznavanje i do 150.000 dokumenata po satu), stopa pogreške jednog znaka je 0.0001%, a odbijanja 0,01%. Zbog ograničenog broja znakova, čitači su vrlo tolerantni na kvalitetu dokumenta, ali takvi sustavi su vrlo skupi jer su dizajnirani za specifičnu namjenu.

5.2. Digitalizacija

OCR tehnologija danas se najčešće koristi u svrhu digitalizacije tekstualnog i ručno pisanog materijala (knjige, časopisi, magazini i slično). Većina dokumenata i dalje je sadržana u papirnatom obliku. Prema Stančiću, digitalizacija se gleda “kao složeni proces koji je potrebno ugraditi u institucijsku politiku te da je potrebno uspostaviti proces digitalizacije tako da on stvori novu, dodanu vijednost, a ne opterećenje”.
Prema Stolinskom i Bienieckom, potrebna je prvenstveno radi zaštite originalnog dokumenta. Original je dostupan na jednom mjestu, a kod digitalnog formata dostupnost je neograničena. Neki dokumenti se kopiraju u krug, mijenjaju između kopiranja, a kvaliteta dokumenta svakim kopiranjem sve više opada. Takav proces je dugotrajan i neefikasan. Sve više tvrtki, ali i pojedinaca ima potrebu za digitalizacijom dokumenata. Rad sa digitaliziranim materijalom jer mnogo efikasniji jer ne trebamo prostora za pohranu materijala. Također, informacije u elektroničkom obliku su lakše za oblikovanje, ispravljanje, dopunjavanje i pohranjivanje.
OCR se također koristi prilikom automatizacije ureda. Pri korištenju OCR tehnologije u ovom slučaju postoje ograničenja na format papira, font znakova i kvalitetu ispisa dokumenta.

5.3. Automatizacija procesa

Osim što OCR ima mogućnost čitanja i prepoznavanja onoga što je tiskano, koristi se i kao proces automatskog čitanja, na primjer automatskog sortiranja pošte.



Slika 24. Izgled koverti. Izvor:
http://www.omegaonline.com/envelope_design.htm

“OCR Read Area” predstavlja mjesto gdje pošiljatelj upisuje adresu primatelja, odnosno područje na koverti koje prolazi kroz OCR čitač. Adresa bi trebala biti zapisana koristeći samo velika slova, slova se ne bi trebala dodirivati i trebala bi biti otprilike jednako udaljena.

5.4. Ostale primjene OCR tehnologije

5.4.1. OCR kao pomoćna tehnologija osobama s oštećenjem vida

OCR tehnologija se u samim početcima koristila za razvoj uređaja koji će pomoći slabovidnim i slijepim ljudima u čitanju. Iako je kasnije napredak tehnologije otišao u drugom smjeru i danas se koristi u mnoge druge svrhe, veoma je poznata kao pomoćna tehnologija osobama sa oštećenjem vida.
Pomoću OCR programa možemo skenirati tiskani tekst, a on ga pretvara u sintetički govor koji čita na glas. Prema Budelliju (2010), tekst se može pohraniti kao audio format ili kao elektronički tekstualni format koji se dalje može koristiti za pretvorbu u Brailleovo pismo. Za razliku od Brailleovog pisma, koji je sustav pisanja ili tiskanja u kojem se kombinacije opipljivih točkica ili točaka koriste za čitanje slova pomoću dodira, OCR je mehanički ili elektronički prijevod skeniranih slika tiskanog ili rukom pisanog teksta u strojno kodirani tekst. Brailleovo pismo je tradicionalno sredstvo za čitanje i pisanje osoba sa oštećenjem vida, a OCR tehnologija slabovidnim i slijepim osobama

daje mogućnost da pristupe knjigama, časopisima i dokumentima kojima prije nisu imali pristup. (American Foundation for the Blind, 2017.)

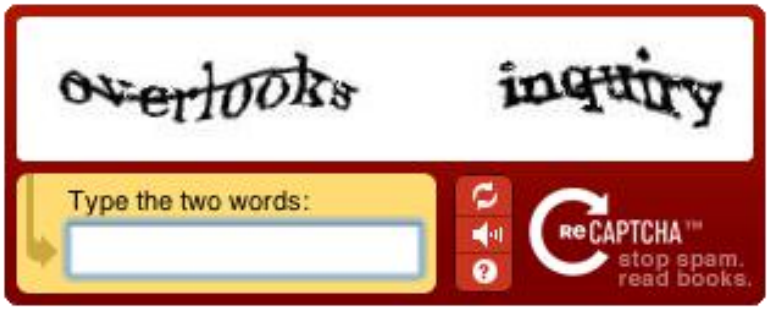
5.4.2. OCR prilikom prepoznavanja automobilskih tablica

OCR tehnologija se koristi i prilikom prepoznavanja automobilskih tablica. Prema web-stranici License Plate Recogniton (2017), sustav slika registarsku tablicu i pretvara sliku tablice u tekst. Ta tehnologija pomaže policiji tako što fotografira tablice automobila u prolasku te ih kasnije može iskoristiti ukoliko je vozilo napravilo prekršaj, provjerava vozila na graničnim prilazima i utvrđuje da li je vozilo sumnjivo. Može se koristiti i za automatsko otvaranje vrata garaže ukoliko vozilo pripada stanaru zgrade i slično.

5.4.3. OCR protiv CAPTCHA-e

CAPTCHA (eng. *Completely Automated Public Turing test to tell Computers and Humans Apart*) je program koji stvara test kojeg ljudi mogu pročitati, a računala ne mogu. Stvara se jednostavan test za koji CAPTCHA program može provjeriti točnost unesenog rješenja. (CARNet, 2010.) CAPTCHA testovi se koriste kao obrana za zaštitu web resursa. Vjeruje se da računala ne mogu točno riješiti taj test te se pretpostavlja da, ukoliko je test točan, rješenje je unio čovjek.

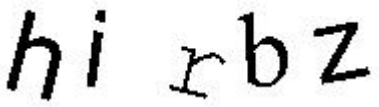
Prema Azad i Jain (2013), razvojem OCR tehnologije, jaz između ljudi i robota pri prepoznavanju znakova je sve manji, a taj trend bi mogao CAPTCHU učiniti vrlo neučinkovitim testom.



Slika 25. Primjer CAPTCHA testa. Izvor: http://www.captcha.net/

Ukoliko CAPTCHA slike obradimo pomoću OCR tehnologije, velika je vjerojatnost da će biti u mogućnosti prepoznati znakove koji se nalaze na fotografiji i time pobiti teoriju da računalo ne može točno riješiti test.

Slika 26. CAPTCHA test.



Slika 27. CAPTCHA nakon OCR

Izvor: https://webscraping.com/blog/Solving-CAPTCHA/

Rješenje testa koji je provučen kroz tri različita OCR programa:

	Captcha 1	Captcha 2	Captcha 3	Result
	7rrg5	hirbZ	izi3b	
Tesseract	7rrq5	hirbZ	izi3b	2 / 3
Gocr	7rr95	_i_bz	izi3b	1 / 3
Ocrad	7rrgS	hi_bL	iLi3b	0 / 3

Tablica 1. Rješenje CAPTCHA testa u različitim OCR programima. Izvor: https://webscraping.com/blog/Solving-CAPTCHA/

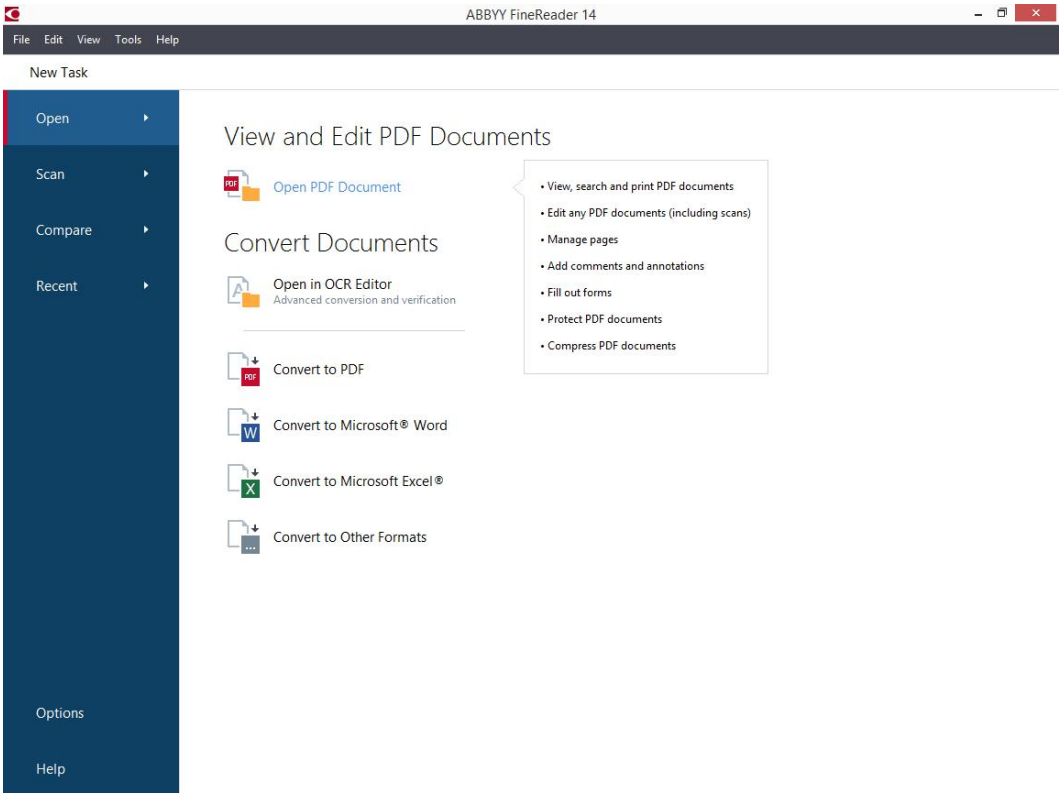
Ukoliko čovjek CAPTCHA test riješi pogrešno, stranica vrati odgovor sa novom slikom i novom prilikom za rješavanje testa, tako da činjenica da OCR programi nisu riješili test sa 100% točnošću nije problem.

6. RJEŠENJA

OCR tehnologija kroz godine je vrlo napredovala. Ranije su verzije radile istovremeno samo sa jednim fontom, a sustavu smo morali točno odrediti koja slika je za koji znak. Danas, sustavi su mnogo napredniji, podržavaju skoro sve formate slika, a barataju sa većinom postojećih fontova. Neki sustavi su toliko napredni da mogu dokument pretvoriti vizualno identični dokument (uključujući tablice, stupce, slike i ostalo), samo u digitalnom obliku. Kriteriji za kvalitetan sustav su prepoznavanje fontova, formatiranje fonta koji su podebljani, nakošeni ili podcrtani, mogućnost pretvorbe tablica u Excel. U nastavku će biti navedena neka od najpoznatijih OCR rješenja.

6.1. ABBYY FineReader

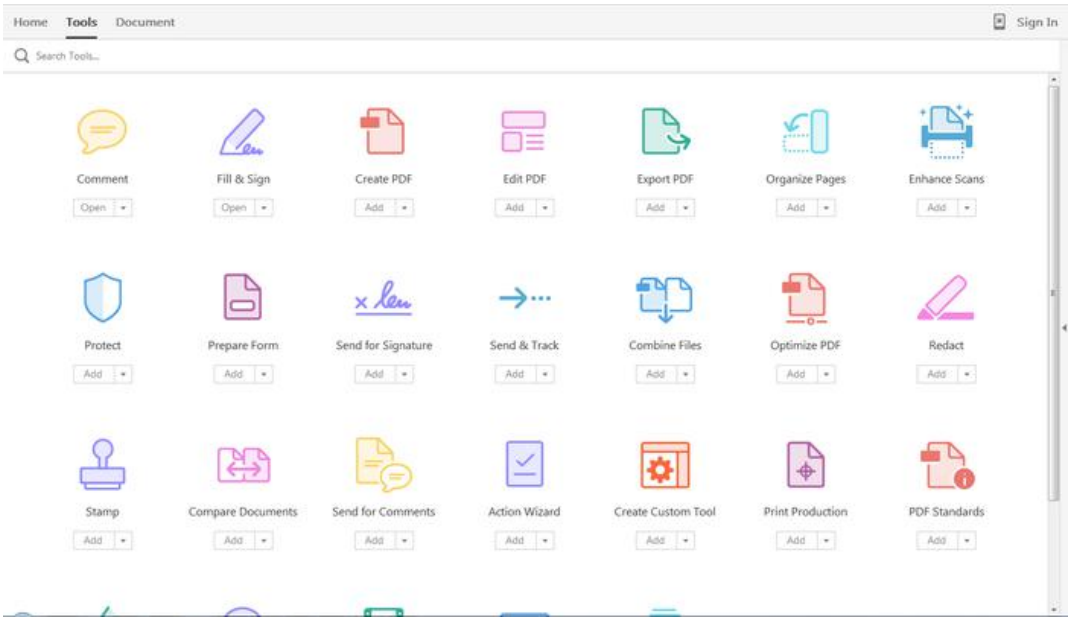
Prema web-stranici Digital Media (2017), ABBYY FineReader je OCR i PDF softverska aplikacija za povećanje poslovne produktivnosti u radu s dokumentima. Ima mogućnost kreiranja digitalne kopije dokumenata koju može uređivati bez pretipkavanja, pretraživati, pregledavati, komentirati, ispunjavati forme, pretvoriti u PDF, Word i mnoge druge formate. Uz skenirane slike, prepoznaje i slike dobivene pomoću digitalnog fotoaparata, a u sebi ima ugrađene alate za pred procesiranje slika i ručnu obradu fotografija. Točnost mu je prilično visoka, a prepoznaje tekstove na 190 jezika. Ima i mogućnost zadržavanja kompletnog izgleda dokumenta, uključujući naslove, tablice, fontove koji su korišteni u originalu, fusnote, brojeve stranica. Ima mogućnost auto detekcije jezika što štedi vrijeme, a ima i ugrađen hrvatski rječnik. Procjenjuje se da radi 100 puta brže od profesionalnog daktilografa, radi manje grešaka i nudi provjeru pravopisa. Cijena je ovisna o paketu, standardni paketi koštaju između 1000 i 2000 kn, a napredniji i do 6000 kn.



Slika 28. Izgled sučelja ABBYY FineReadera. Izvor: https://www.abbyy.com/en-eu/finereader/

6.2. Adobe Acrobat Pro

Adobe Acrobat Pro je OCR sustav koji ima mnogo manje mogućnosti od ABBYY FineReadera, vrlo je pouzdano i ekonomično rješenje. Ima mogućnost pretvorbe PDF datoteke u Word ili RTF. Podržava puno manje jezika, 25, a jezik se mora odabrati u dijaloškom okviru. Odabir jezika je izuzetno bitan jer koristi rječnike koji su specifični za jezik dokumenta. Dostupan je za Windows, Mac i browser. Cijena je oko 2000 kuna godišnje. (University of Illinois at Urbana-Champaign, 2017).

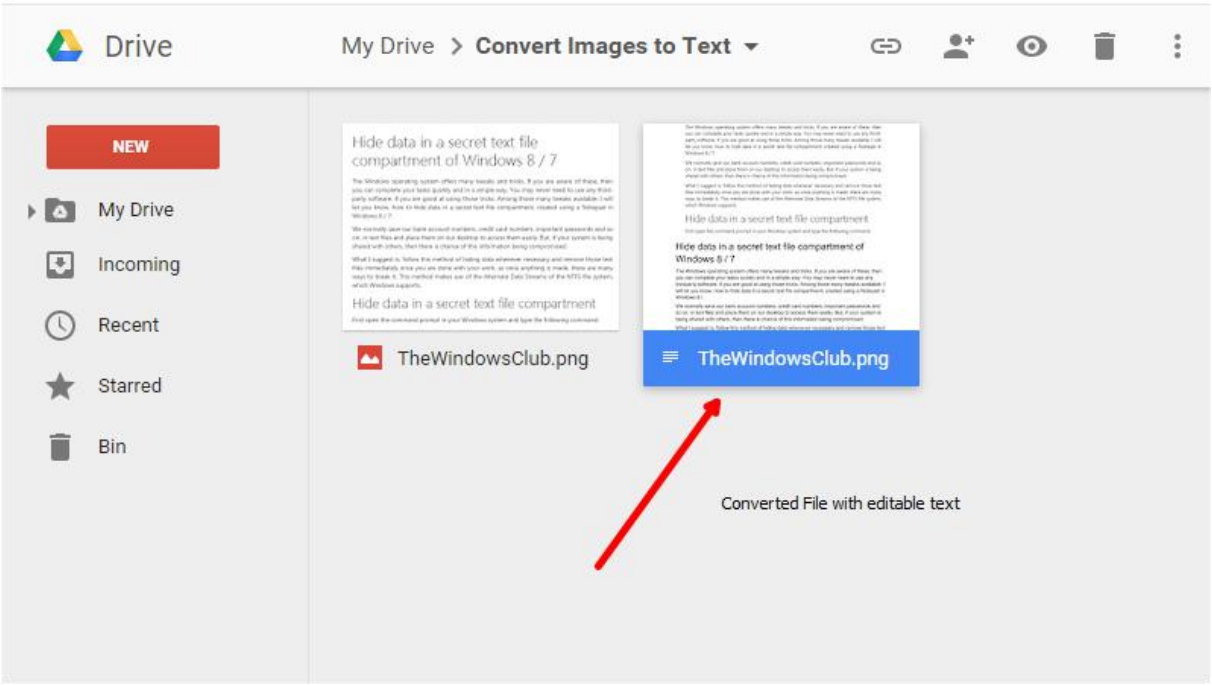


Slika 29. Izgled sučelja Adobe Acrobat Pro. Izvor: https://adobe-acrobat-professional.en.softonic.com/

6.3. Google Docs

Google Docs je brzo, jednostavno i besplatno rješenje. Podržava JPEG, PNG, GIF i PDF formate. Google Docs nije savršeno i izuzetno kvalitetno rješenje, ali može poslužiti svrsi. Kako bi rezultat bio bolji, Google Drive nudi savjete kako osigurati točnost (Google Support, 2017): tekst mora biti minimalno 10 piksela, dokument bi trebao biti točno orijentiran, a font bi trebao biti jednostavan (npr. Arial ili Times New Roman). Ima mogućnost prepoznavanja jezika na kojem je tekst napisan.

Ukoliko je font podebljan, kurzivan, font manji od 10 piksela, moglo bi biti mnogo krivih rezultata. Također vrlo vjerojatno neće prepoznati liste, tablice, stupce, fusnote i slično.



Slika 30. Izgled sučelja u Google Docsu. Izvor: http://www.thewindowsclub.com/google-drive-convert-image-to-text

ABBYY FineReader je najpouzdaniji sustav koji se preporuča onima koji često imaju potrebu izvlačiti tekst iz dokumenata, slika ili PDF-a, kojima je točnost i struktura dokumenta jako bitna. Google Docs se preporuča onima kojima usluga treba povremeno, nije im problem ispravljati greške ručno te im nije bitna struktura dokumenta. (Prevoditelj-teksta, 2012-2017.)

Zaključak

Iako prepoznavanje znakova u različitim računalnim fontovima nije jednostavno, prepoznavanje znakova napisanih rukopisom prilično je težak zadatak. Razlika između računala i čovjeka je činjenica da ljudski mozak ima mogućnost da se osloni na značenje napisanih riječi, na značenje već pročitanih riječi, ali i na činjenicu da vjerojatno zna o čemu osoba koja je napisala riječi razmišlja. To su područja u kojima se računalo ne snalazi jako dobro.

Postoje solucije koje bi pomogle računalu da lakše dođe do što točnijeg rješenja. Programi koji se koriste u područjima kao što je pošta ili banka, u kojima su pisma ili uplatnice češće pisane rukopisom nego tiskanim tekstom, imaju mogućnost da skeniraju samo poštanske brojeve, a ne cijele adrese. Također, očekuje se da ljudi paze na rukopis na uplatnicama i pismima, da imena i adrese pišu razumljivije, pazeći na razmake između slova, da koriste samo velika slova i slično. Uplatnice na sebi sadrže male odvojene kućice u koje bi se trebala upisivati samo jedna brojka ili slovo, a najčešće su u specifičnim, svjetlijim bojama kako bi programi lakše prepoznali tintu kojom su znakovi napisani od samih linija kvadratića. Današnji tableti i pametni telefoni sa „touchscreenom“ imaju ugrađenu tehnologiju koja prepoznaje znakove kako ih pišemo. Ukoliko povučemo jednu nakošenu liniju, pa drugu, pa ih spojimo horizontalno, spajajući sve te komponente, program zaključuje da smo najvjerojatnije napisali slovo A.

Cijene skenera i OCR tehnologije su značajnije niže nego unazad 20 godina, a kvaliteta je mnogo bolja. Napredni OCR danas minimiziraju stopu pojavljivanja grešaka, vrlo su pouzdani, učinkoviti, a cijena im nije jako visoka. Ručni unos podataka je dugotrajan proces, a OCR to može napraviti u nekoliko sekunda ili minuta, ovisno o broju znakova koje treba pretvoriti. Iako OCR nije savršen i postoje situacije u kojima nije popularan za korištenje (npr. pretvaranje audio datoteka), za tekstualne i slikovne dokumentacije je daleko najbolja solucija jer pruža fleksibilnost, brzinu i kontrolu nad dokumentima koja je bitna u svim profesionalnim radnim okruženjima.

Literatura

1. Schantz, H. F. (1982) The history of OCR: optical character recognition, Recognition Technologies Users Association.
2. Kurzweil Technologies (2017) Kurzweil Computer Products [online]. Kurzweil Technologies, Inc. Dostupno na: http://www.kurzweiltech.com/kcp.html [30.06.2017.]
3. Wikipedia (2017) Optophone [online]. Wikimedia Foundation, Inc. Dostupno na: https://en.wikipedia.org/wiki/Optophone [30.06.2017.]
4. Wikipedia (2017) Optacon [online]. Wikimedia Foundation, Inc. Dostupno na: https://en.wikipedia.org/wiki/Optacon [30.06.2017.]
5. ABBYY (2017) ABBYY FineReader 14 [online]. ABBYY. Dostupno na: https://www.abbyy.com/en-ee/finereader [30.06.2017.]
6. Vynckier, Ivo (2017) How OCR Works [online]. Vynckier Ivo. Dostupno na: http://www.how-ocr-works.com [02.07.2017.]
7. Woodford Chris (2010/2014) OCR (optical character recognition) [online]. Dostupno na: http://www.explainthatstuff.com/how-ocr-works.html [02.07.2017.]
8. Walls John (2008) OCR and Content Management with SAP and Imaging [online]. VerbellaCMG. Dostupno na: https://www.slideshare.net/verbella/ocr-and-content-management-with-sap-and-imaging [02.07.2017.]
9. Panjwani, Karan (2015) Optical Character Recognition (OCR) [online]. Dostupno na: https://www.slideshare.net/karanpanjwani752/optical-character-recognition-ocr [02.07.2017.]
10. CVISION Technologies (1998-2017) OCR Recognition [online]. CVISION Technologies, Inc. Dostupno na: http://www.cvisiontech.com/library/ocr/fast-ocr/ocr-recognition.html [02.07.2017.]
11. Investintech (2000-2017) OCR PROGRAM SOFTWARE [online]. Inevstintech.com Inc. Dostupno na: http://www.investintech.com/resources/articles/ocrprogram/ [02.07.2017.]

12. Stolinski, Sebastian i Bieniecki, Wojciech (nepoznato) Application of OCR systems to processing and digitalization of paper documents [online]. Computer Engineering Department, Technical University of Lodz, Poland. Dostupno na: http://sstolin.kis.p.lodz.pl/dane/pub/10_isim_ocr.pdf [02.07.2017.]
13. DocuFi (2017) How to improve OCR accuracy with document image cleanup [online]. DocuFi. Dostupno na: http://www.docufi.com/improve-ocr-accuracy [02.07.2017.]
14. American Foundation for the Blind (2017) Optical Character Recognition Systems [online]. 2017 American Foundation for the Blind. Dostupno na: http://www.afb.org/info/assistive-technology/optical-character-recognition-systems/35 [02.07.2017.]
15. Budelli, Joe (2010) OCR helping the Visually Impaired [online]. AIIM 2017. Dostupno na: http://community.aiim.org/blogs/joe-budelli/2010/09/08/ocr-helping-the-visually-impaired- [02.07.2017.]
16. CVISION Technologies (1998-2017) OCR for Banking Industry [online]. CVISION Technologies, Inc. Dostupno na: http://www.cvisiontech.com/library/ocr/file-ocr/ocr-for-banking-industry.html [02.07.2017.]
17. License Plate Recogniton (2010) Introduction to License Plate Recognition [online]. License Plate Recogniton. Dostupno na: http://www.licenseplatesrecognition.com/ [02.07.2017.]
18. WebScraping (2012) Solving CAPTCHA with OCR [online]. Dostupno na: https://webscraping.com/blog/Solving-CAPTCHA/ [02.07.2017.]
19. Azad, Silky i Jain, Kiran (2013) CAPTCHA: Attacks and Weaknesses against OCR Technology [online]. Global Journal of Computer Science and Technology, Neural and Artificial Intelligence. Dostupno na: https://globaljournals.org/GJCST_Volume13/3-CAPTCHA-Attacks-and-Weaknesses.pdf [02.07.2017.]
20. CARNet (2010) CAPTCHA [online]. Nacionalni CERT. Dostupno na: http://www.cert.hr/sites/default/files/NCERT-PUBDOC-2010-06-302.pdf [02.07.2017.]
21. Jaszczak, Kaz (2011) Optical Character Recognition: A Backbone for

Postal and Mail Sorting Applications [online]. Mailing Systems Technology. Dostupno na: http://mailingsystemstechnology.com/article-2813-Optical-Character-Recognition-A-Backbone-for-Postal-and-Mail-Sorting-Applications.html [02.07.2017.]

22. Prevoditelj teksta (2012-2017) Najbolja OCR rješenja za prepoznavanje teksta [online]. Prevoditelj-teksa.com. Dostupno na: http://www.prevoditelj-teksta.com/2014/07/najbolja-ocr-rjesenja-za-prepoznavanje-teksta.html [02.07.2017.]
23. Digital Media (2017) ABBYY FineReader [online]. Digital Media d.o.o. Dostupno na: http://www.digitalmedia.hr/ocrscanpdf/abbyy-finereader/ [03.07.2017.]
24. University of Illinois at Urbana-Champaign (2017) An Introduction to OCR and Searchable PDFs: Adobe Acrobat Pro [online]. University Library, University of Illinois at Urbana-Champaign. Dostupno na: http://guides.library.illinois.edu/c.php?g=347520&p=4116755 [03.07.2017.]
25. Google Support (2017) Covert PDF and photo files to text [online]. 2017 Google. Dostupno na: https://support.google.com/drive/answer/176692?co=GENIE.Platform%3DDesktop&hl=en&oco=1 [03.07.2017.]
26. Stančić, Hrvoje (nepoznato) Digitalizacija kao dio složenog informatično-informacijskom ekosustava [online]. Odsjek za informacijske i komunikacijske znanosti, Filozofski fakultet Sveučilišta u Zagrebu. Dostupno na: http://dfest.nsk.hr/2013/pdf/Stancic_Hrvoje.pdf [04.07.2017.]
27. CVISION Technologies (1998-2017) OCR, Neural Networks and other Machine Learning Techniques [online]. CVISION Technologies, Inc. Dostupno na: http://www.cvisiontech.com/resources/ocr-primer/ocr-neural-networks-and-other-machine-learning-techniques.html [25.08.2017.]
28. Valveny, Ernest (2006-2007) Optical Character Recognition [online]. Computer Vision Center (CVC). Dostupno na: http://www.cvc.uab.es/~ernest/slides/ocr0607.pdf

29. Cheriet, Kharma, Liu, Suen (2007) Character Recognition Systems, John Wiley & Sons, Inc.
30. Smith, Ernie (2017) Seek And Spell [online]. Tendium: The Dull Side of the Internet (2015-2017). Dostupno na: http://tedium.co/2017/03/22/ocr-typography-optical-character-recognition-history/ [25.08.2017.]
31. Eikvil, Line (1993) OCR - Optical Character Recognition [online]. Dostupno na: https://www.nr.no/~eikvil/OCR.pdf

Popis slika

Popis tablica

Sažetak

OCR tehnologija (Optičko prepoznavanje znakova) je mehanička ili elektronska pretvorba dokumenata u strojno kodirani tekst. To je najpoznatija metoda prilikom digitalizacije papirnog izvora podataka kao što su računi, pisma, dokumenti i bilo koji drugi ispisani zapis. Takav digitalizirani tekst kasnije se može lakše pretraživati, kompaktnije pohraniti, prikazati na mreži i slično. Najpoznatije metode su “prepoznavanje uzoraka” i “prepoznavanje značajki”, gdje se prva metoda najčešće koristi kod tiskanog teksta, a ne rukom pisanog ili više stiliziranog fonta. Druga metoda rastavlja znakove na dijelove te ujedinjujući njihove značajke, dolazi do rješenja za pojedini znak. Zbog te mogućnosti OCR tehnologija spada pod umjetnu inteligenciju i strojno učenje.

Ključne riječi: OCR tehnologija, Optičko prepoznavanje znakova, digitalizacija, digitalizacija teksta, prepoznavanje uzoraka, prepoznavanje značajki, umjetna inteligencija, strojno učenje

Summary

OCR technology is mechancial or electronic conversion of documents into machine coded text. That is the most common method of digitizing paper data sources such as invoices, letters, documents and any other printed data. Digitized text can be easily searched, stored, showed online, etc. The most common methods are “pattern matching” and “feature extraction”, where the first method is used in printed texts, not printed or stylized fonts. The second method separates characters into their features and joining them together to find a solution for particular character. OCR technology is a part of artificial intelligence and machine learing.

Key words: OCR technology, Optical character recognition, digitalization, text digitalization, pattern matching, feature extraction, artificial intelligence, machine learning