

Primjena računarske grafike.
OCR-Optičko znakovno prepoznavanje.
6362

Rezime: U ovom seminarskom radu je razrađena jedna od primjena računarske grafike a to je optičko prepoznavanje znakova. Razrađena je aplikacija OCR te sami koraci kroz koje prolazi taj proces prepoznavanja znakova. Također je razrađen prvi Open Source OCR tj. Tesseract, te njegova arhitektura. A to sve povlači i priču o pretvaranju teksta u govor. A da bi priča bila upotpunjena navedena je i sama primjena OCR-a u nekoliko industrija i u stvarnom životu.

Ključne riječi: Optičko prepoznavanje znakova (OCR), Primjena računarske grafike, Pretvaranje teksta u govor (TTS), Tesseract, Ekstrakcija teksta, Umjetničko stvaranje govora.

Uvod

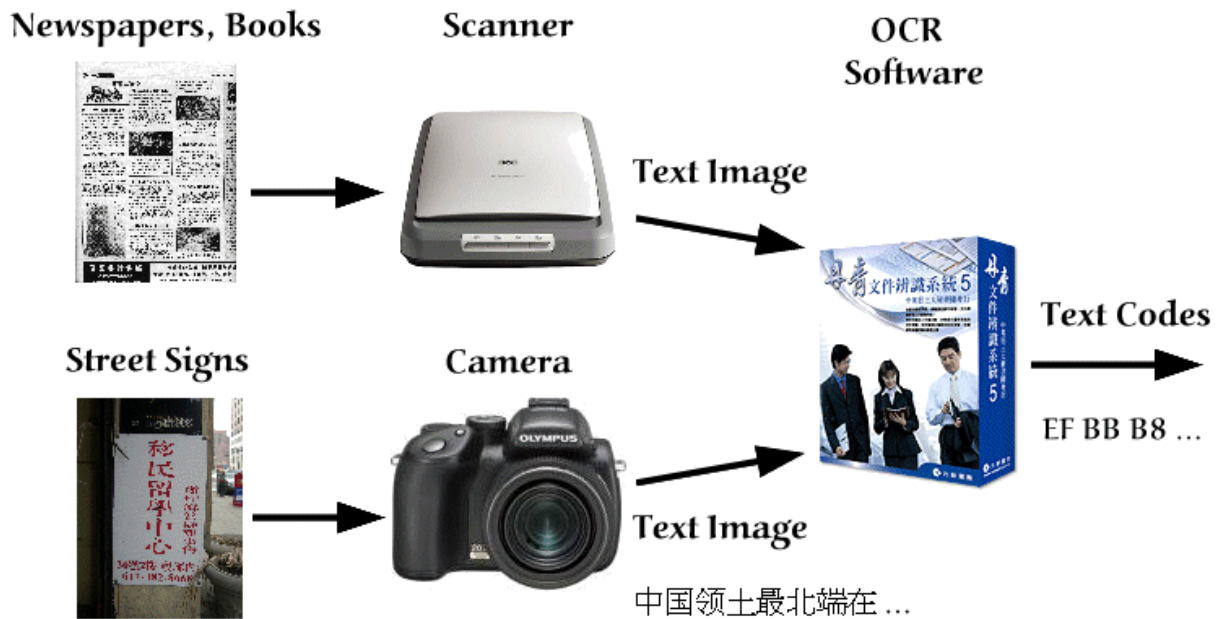


OCR je akronim za Optical Character Recognition odnosno optičko prepoznavanje znakova. OCR je ustvari mobilna aplikacija. OCR koristi android platforme pametnih mobilnih telefona. Suština je kombinirati optičko prepoznavanje znakova i umjetno stvaranje govora. Dakle, krajnji cilj programera bio je stvoriti aplikaciju prilagođenu korisnicima koja pretvara sliku u govor koristeći android telefone.

Slika 1. OCR

OCR zapravo uzima sliku kao input odnosno ulazne podatke, od te slike izdvaja tekst i to pretvara u govor. Ovaj sistem se koristi u nekoliko područja kao što su: bankarstvo, pravna industrija te ostale industrije, također za kućne potrebe te za uredske potrebe. OCR je uglavnom dizajniran za ljude koji nisu sposobni da čitaju bilo koji tip tekstualnog dokumenta. U ovom radu, metoda znakovnog prepoznavanja je prezentirana koristeći OCR tehnologiju i android telefone sa kvalitetnijom kamerom.

OCR



Slika 2. Kako čitati kineski jezik bez da ga poznajete? OCR

Kao što je već rečeno OCR je akronim za Optical Character Recognition odnosno optičko prepoznavanje znakova. Ova tehnologija dozvoljava automatsko prepoznavanje znakova kroz optički mehanizam. Kako bi vam bolje približili temu uporedit ćemo ovu tehnologiju sa čovjekovim tijelom, pa što se tiče ljudskog tijela, naše oči su optički mehanizmi. Slike koje vidimo očima su input, odnosno ulazni podaci za naš mozak. Sposobnost razumijevanja ovih podataka ovisi o mnogim faktorima svake osobe. OCR je tehnologija koja funkcioniše kao ljudska sposobnost čitanja. Iako, naravno OCR nije u mogućnosti da se natječe sa ljudskim sposobnostima čitanja.

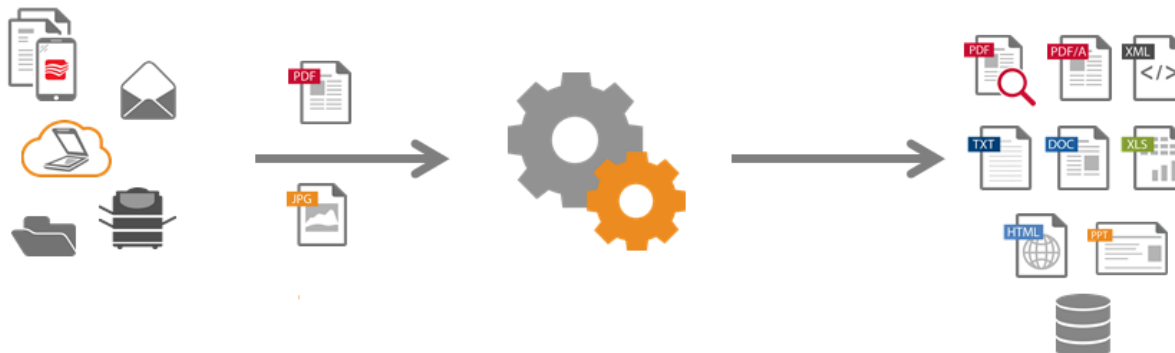


Slika 3. Blok dijagram OCR-a

OCR može da prepozna kako printani tekst tako i tekst pisan rukom. Bitno je napomenuti da performansa OCR-a direktno zavisi od kvaliteta ulaznog dokumenta. OCR je dizajniran za obradu slike koja se gotovo sva sastoji od teksta odnosno u sebi ima vrlo malo netekstualnih elemenata, naravno slika je napravljena sa mobilnom kamerom. Ova aplikacija je za mobilne telefone sa android operativnim sistemom koja kombinuje Google open-source OCR, Tesseract i OCR za prepoznavanje teksta. Također kombinuje Google prevoditelj i umjetno pretvaranje teksta u govor android operativnog sistema i tako korisnici mogu uslikati pomoću mobilnih telefona fotografiju koja sadrži tekst zatim slušaju šta zapravo piše na određenom dijelu fotografije. Većina programa za prepoznavanje znakova prepoznat će ulaznu sliku uslikanu sa skenerom ili digitalnom kamerom ili kompjuterskim softverom. Postoji problem sa razlikom velične kod kompjutera i skenera. Ako nemate skener ili digitalnu kameru onda se naravno pojavljuju problemi hardverskog tipa. OCR je tehnologija koja omogućava konvertovanje različitih tipova dokumenata kao što su naprimjer skenirani papirni dokumenti, PDF fajlovi ili

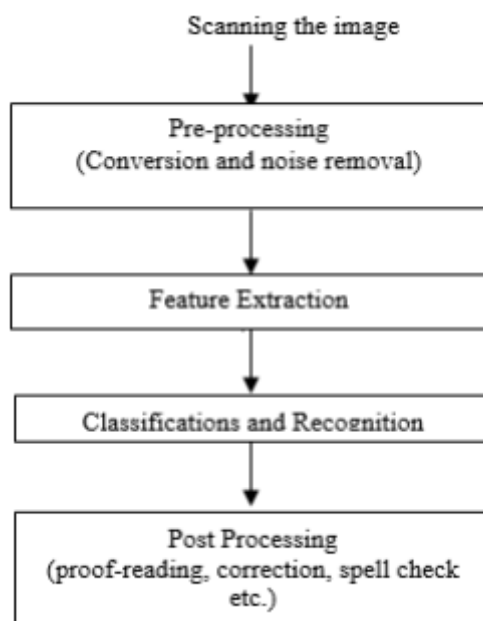
fotografije fotografisane digitalnom kamerom i to sve konvertuje u podatke koje se mogu uređivati i pretraživati.

What is Optical Character Recognition (OCR)?



Slika 4. OCR-Konvertovanje različitih tipova podataka.

Slike koje napravimo digitalnom kamerom su različite od dokumenata ili slika koje se skeniraju. One ponekad imaju nedostatke odnosno defekte poput izobličenja na krajevima i prigušene svjetlosti te tako postaju teške za očitati OCR aplikaciji jer u tom slučaju OCR ne može pravilno prepoznati tekst. Za takve situacije izabran je Tesseract zbog rasprostranjenosti, te mogućnosti nadogradnje, fleksibilnosti te cijele zajednice programera koja radi na tom programu te same činjenice da predobro radi. Da bi se izvela sama radnja prepoznavanja znakova aplikacija mora da prođe kroz faze. A te faze odnosno komponente OCR-a su obrađene u nastavku.

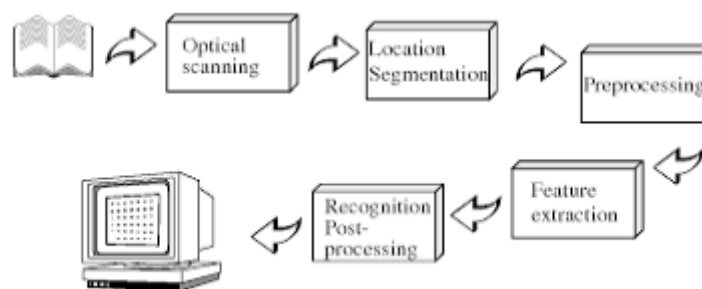


Slika 5. Faze OCR-a

Komponente OCR-a

Slike 5. i 6. prikazuju ukupno funkcionisanje optičkog prepoznavanja karaktera (OCR). Sastoji se od nekoliko koraka za prepoznavanje teksta. Ti koraci su: skeniranje, segmentacija, predradnja, ekstrakcija značajki, prepoznavanje. Ulazna slika za OCR može biti napisana rukom ili može biti tiskani tekst poput knjiga, časopisa, novina, itd. Takav se unos daje OCR-u. Prvo se skenira pomoću android mobilne kamere. To znači da digitalizira analogni dokument.

Tekstualne regije unutar slike se nalaze, zatim se simboli izdvajaju kroz segmentaciju, unaprijed obradite svaki simbol, izdvajaju se značajke zatim prepoznaju. Koraci su detaljno objašnjeni u nastavku.



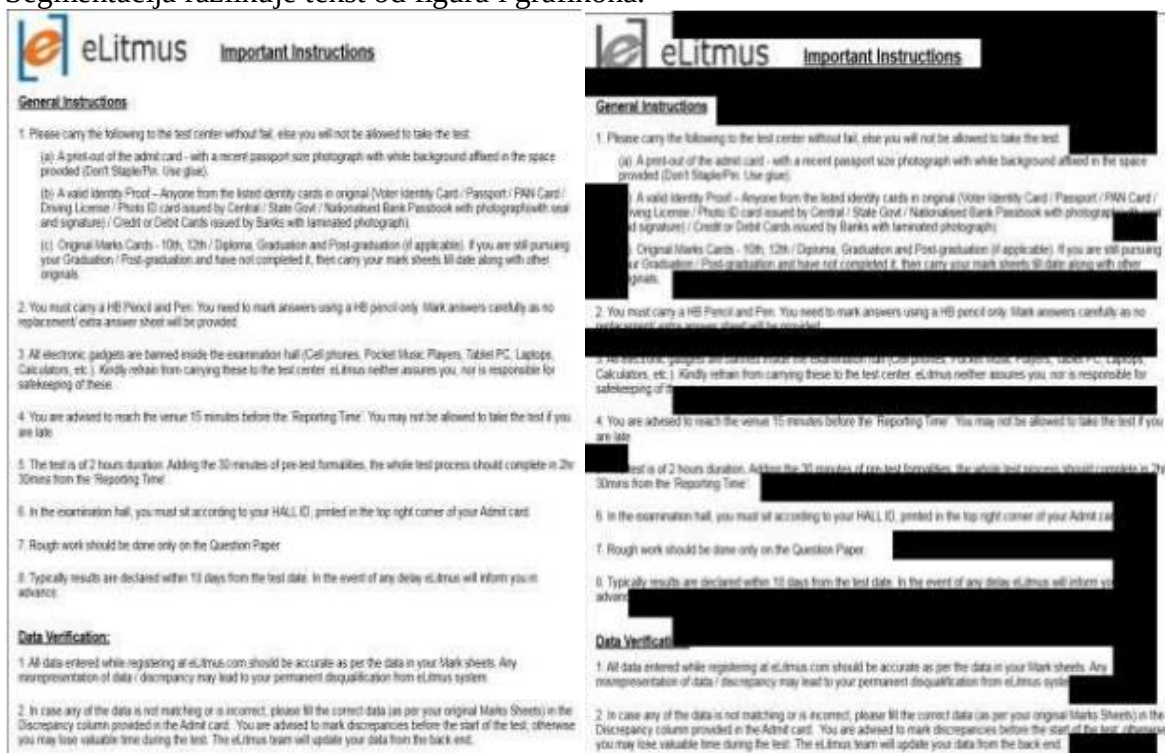
Slika 6. Funkcionisanje OCR-a

a) Skeniranje

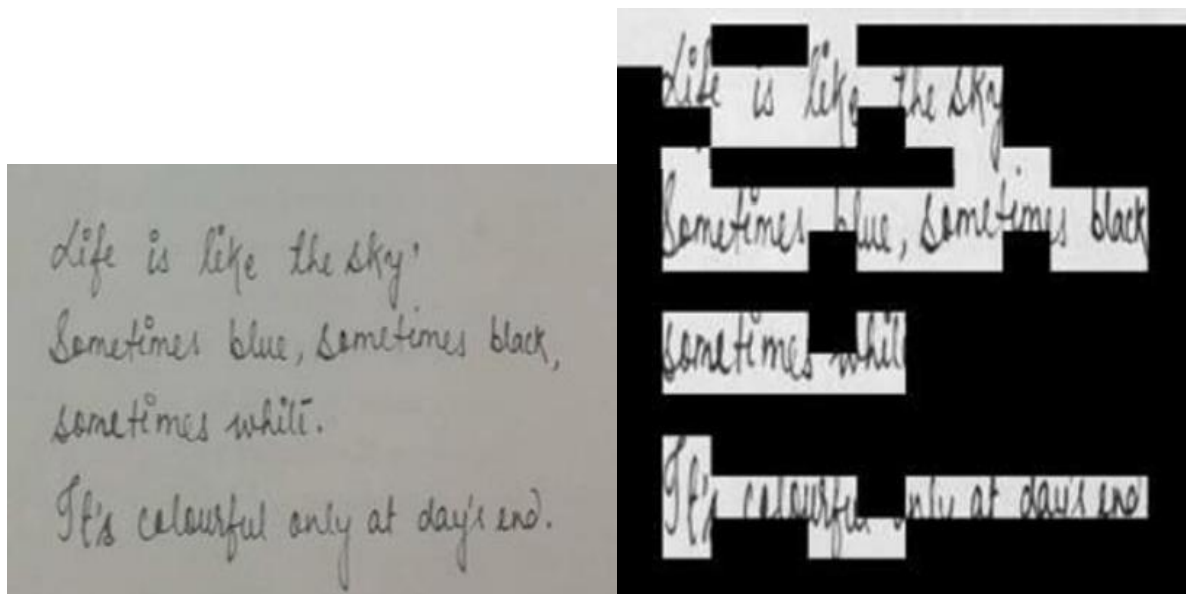
Ova aplikacija, tj. OCR sistem koristi kamere Android uređaja. Kamera snima sliku dokumenta. Ovo nije ništa drugo već proces skeniranja. Ukratko možemo reći da skeniranje čini originalni dokument kao digitalnu sliku. Generalno, originalno dodaci se sastoje od crno obojenog teksta na bijeloj pozadini, odnosno pozadini u boji. Uz skeniranje dolazi prag koji digitalnu sliku čini sivom. Praćenje praga je proces koji pretvara sliku sa više boja u sliku sa samo dvije boje obično je to crno-bijela slika. Razina praga je fiksna. Ako su nivoi sive boje ispod nivoa praga, onda je boja crna. A ako je nivo sive boje iznad nivoa praga, onda je boja identifikovana kao bijela. Ovo rezultira uštedom memorijskog prostora te smanjuje računarski napor.

b) Segmentacija

Proces lociranja regija na papiru štampanog ili rukom pisanog teksta je segmentacija. Segmentacija razlikuje tekst od figura i grafikona.



Slika 7. Prikaz printanog teksta prije i poslije procesa segmentacije.



Slika 8. Prikaz teksta na engleskom jeziku pisanog rukom prije i poslije procesa segmentacije.

Kada se segmentacija primijeni na tekst, znakovi ili riječi se izoliraju. Problem koji se najčešće javlja u segmentaciji je to što se pojavljuje zbrka između teksta i grafikona u slučaju sastavljenih i rastavljenih znakova. Obično se to sastavljanje i rastavljanje znakova javlja zbog skeniranja. Ako je dokument tamna fotokopija ili ako je skenirana na niskom pragu, nastaju sastavljanje znakova. A rastavljanje znakova nastaje ako je dokument svijetlija fotokopija ili skeniran na niskom pragu. OCR sistem se također zbunjuje tokom segmentacije kada su znakovi povezani s grafikonom.

c) Predobrada

Kao što smo vidjeli gore, tokom skeniranja može se javiti neki šum procesa. Zbog toga dolazi do lošeg prepoznavanja znakova. Ovaj problem se obično predviđi u predprocesu. To se sastoji od izgladivanja i normalizacije. U izravnavanju, na sadržaj slike se primjenjuju određena pravila i tehnike punjenja i stapanja.

d) Ekstrakcija karakteristika

Ekstrakcija izdvaja karakteristike simbola. U ovome su simboli karakterizirani tako da su nevažni atributi izostavljeni. Ekstrakcija osobina tehnika se ne podudara sa konkretnim uzorcima znakova, ali radije bilježi apstraktne osobine prisutne u znakovima poput praznih prostora, linija, itd. Tesseract algoritam (objašnjen u nastavku) se koristi za implementaciju ekstrakcije značajki. Značajka ekstrakcija se bavi reprezentacijom simbola. Slika znaka je preslikana na višu razinu pomoću izdvajanje posebnih karakteristika slike u elementu faza ekstrakcije.

e) Prepoznavanje

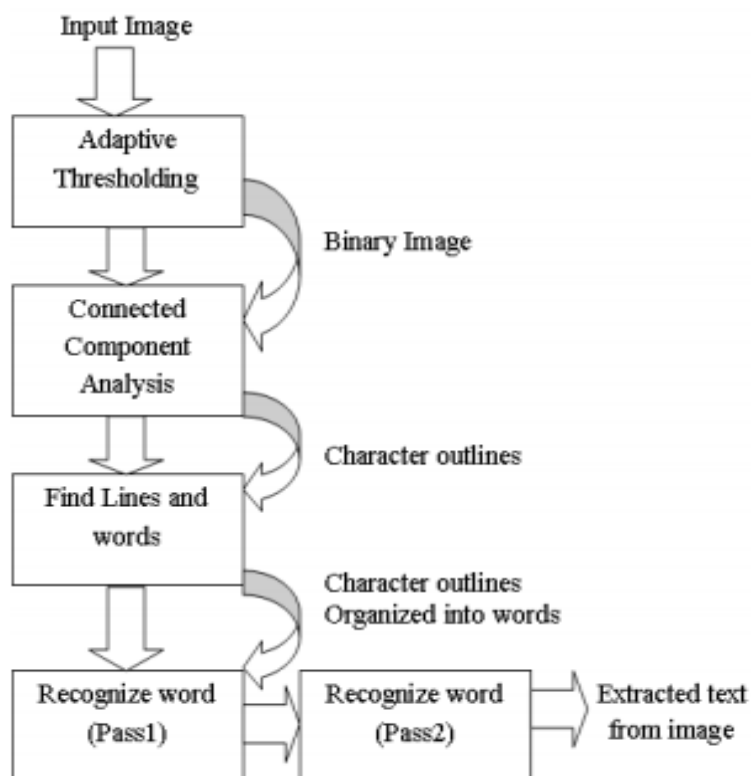
OCR sistem radi s Tesseract algoritmom koji prepoznaje likove. Tesseract identificira znakove u pikselu prednjeg plana, nazivaju se mrljice, a zatim pronalazi crte. Prepoznavanje znakova od riječi do riječi vrši se cijelim postupkom linije. Prepoznavanje uključuje pretvaranje ovih slika u tokove znakova koji predstavljaju slova prepoznatih riječi. Ukratko, prepoznavanje izdvaja tekst iz slika dokumenata. Tesseract je objašnjen u nastavku.

Tesseract

Tesseract je Open source OCR. Tesseract je prvi put razvijen između 1984 i 1994. I napokon je pušten kao Open source 2005. godine.

a) Arhitektura Tesseract-a

Tesseract može da prepozna i obavlja radnju i sa tekstom crnog fonta na bijeloj podlozi a i obrnuto. Tesseract obavlja aktivnost prepoznavanja riječi. Ta radnja se sastoji od dva prolaza. U prvom prolazu se pokušavaju prepoznati riječi. Onda te riječi prolaze kroz klasifikator koji prepoznaje tekst puno tačnije. Tokom drugog prolaza riječi koje nisu dobro prepoznate u prvom prolazu se prepoznavaju tokom ponovnog prolaska kroz stranicu. I u konačnici Tesseract rješava sumnjive prostore.



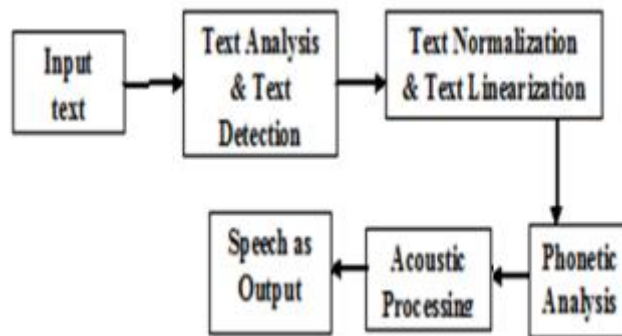
Slika 9. Arhitektura Tesseracta

Pretvaranje teksta u govor (TTS)

Pretvaranje teksta u govor, odnosno TTS (A text to speech), je sistem koji može automatski čitati tekst naglas koji je izdvojen iz OCR-a. Sintetizer umjetnog stvaranja govora može biti implementiran i od strane hardvera i od strane softvera. Sintetizer govora je umjetna proizvodnja ljudskog govora. Sistem koji se koristi za ovu svrhu se naziva sintetizer govora. TTS sistem konvertuje normalni tekst u govor. Sintetizer može ugraditi model glasnica i ostale karakteristike ljudskog glasa za stvaranje "sintetičkog" glasa. Koraci u procesu pretvaranja teksta u govor koji prate sliku 6. su:

a) Analiza teksta i otkrivanje:

Analiza teksta dio je predobrade. Analizira unosi tekst i organizuje se u obradiv popis riječi. Onda to pretvara u cjelovit tekst. Otkrivanje teksta lokalizira tekst iz tiskanih dokumenata.



Slika 10. Pretvaranje teksta u govor

b) Normalizacija i linearna obrada teksta:

Normalizacija teksta je transformacija teksta u izgovorni oblik. Često se vrši normalizacija teksta prije nego što se tekst obradi na neki način, kao što je generiranje sintetizirani govor ili automatizirano prevođenje jezika. Glavni cilj ovog postupka je prepoznati interpunkcijske znakove i stanke između riječi. Obično proces normalizacije teksta podrazumjeva pretvaranje svih malih i velikih slova, te podrazumjeva uklanjanje interpunkcijskih znakova, znakova akcenta, zaustavnih riječi ili "Previše uobičajene riječi" i drugih.

c) Fonetska analiza:

Pružaju fonetske abecede. Grafičar za fonemu vrši se konverzija. Zapravo je to pretvorba ortografski simboli u fonološke simbole

d) Akustička obrada:

Izvodi formantnu sintezu. Djeluje inteligentno i stoga ne zahtijeva nikakvu bazu uzoraka govora. Za izgovaranje teksta koristi se govornim osobinama osobe.

Aplikacije u kojima se koristi OCR

1. Bankarstvo

Jedna nadaleko poznata aplikacija je u bankarstvu, gdje se OCR koristi za obradu čekova bez ljudskog učešća. Slika čeka može se uslikati mobilnom kamerom, sve što je na njemu napisano skenira se odmah, a tačna suma novca se odmah prebacuje. Ova tehnologija je zamalo usavršena za tiskane čekove i prilično je precizna za rukopise također, iako povremeno zahtijeva i ručnu potvrdu. Sve u svemu, to smanjuje vrijeme čekanja u mnogim bankama.



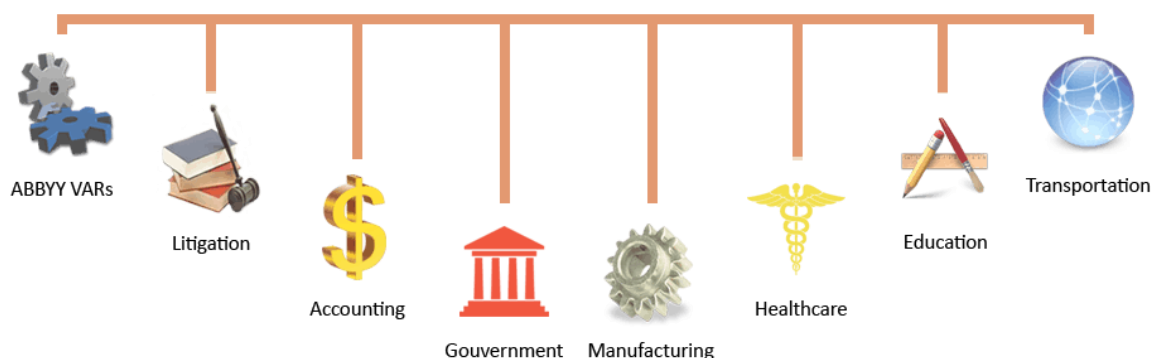
Slika 11. Primjena OCR-a u bankarstvu (Bankovni ček)

2. Pravna industrija

U pravnoj industriji također je bilo značajno pokret za digitalizaciju papirnih dokumenata. Kako bi se uštedio prostor i eliminirala potreba za pretraživanjem svake kutije sa papirima, dokumenti se jednostavno samo skeniraju. OCR dodatno pojednostavljuje proces tako što dokumenti pretražuju tekst, tako ih je lakše pronaći i raditi s jednom bazom podataka. Pravni profesionalci sada imaju brz i lak pristup ogromnoj biblioteci dokumenata u elektronskom obliku, koje jednostavno mogu pronaći kucanje nekoliko ključnih riječi.

3. Ostale industrije

OCR se široko koristi u mnogim drugim oblastima, uključujući obrazovanja, finansija i vladinih agencija. OCR je napravio bezbroj tekstova dostupnih na mreži, štedeći novac studentima i omogućavanje da se znanje dijeli.



Slika 12. Primjena OCR-a

4. Praćenje glasa

U nekim je slučajevima usmena informacija efikasnija od pisane poruke. Apel je jača, dok je pažnja možda se još uvijek usredotočuju na druge vizualne izvore informacija. Otuda ideja o uključivanju govora iz dokumenata.

ZAKLJUČAK

Optičko prepoznavanje znakova bavi se prepoznavanjem optičko obrađenih znakova. Pouzdano tumačenje teksta fotografija iz stvarnog svijeta je izazovan problem zbog varijacije u faktorima okoline ali to postaje jako lakše koristeći open source OCR.

PREDNOSTI

1. Ovaj pristup koristi operativni sistem Android.
2. Android je besplatan i otvoren softver. Tako da je OCR mobilna aplikacija znatno niža.
3. Rani OCR zahtijeva skupe skenere i posebne namjenski elektronski ili optički hardver, ali ovaj pristup je mobilna aplikacija koja ima ugrađenu kameru.
4. Pretvara rastersku sliku u tekst i tekst u glas koristeći OCR tehnike
5. Kao Android zasnovan na Linux kernelu, on ima svojstvo sigurnosti od virusne infekcije.

OGRANIČENJA

1. Tačnost OCR sistema direktno ovisi o kvalitet ulaznog dokumenta.
2. Izlaz iz OCR sistema često je prilično „bučan“ i garniran. Da bi se to popravilo, aplikacija će nakon obrade teksta poslati neke povratne informacije OCR paketu.

Literatura

- [1] <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.673.8061&rep=rep1&type=pdf> (dostupno 20.12.2019)
- [2] https://www.researchgate.net/profile/Chirag_Patel27/publication/235956427_Optical_Character_Recognition_by_Open_source_OCR_Tool_Tesseract_A_Case_Study/links/00463516fa43a64739000000.pdf (dostupno 20.12.2019)
- [3] <https://pdfs.semanticscholar.org/8a43/2184c311347a3851b0d3811d333662feb750.pdf> (dostupno 23.12.2019)
- [4] <http://ijsrcseit.com/paper/CSEIT1952175.pdf> (dostupno 24.12.2019)
- [5] <http://www.georgekung.com/ocr.html> (dostupno 28.12.2019)
- [6] <https://nitforall.blogspot.com/2018/08/ocr-optical-character-recognition-api.html> (dostupno 29.12.2019)
- [7] <https://www.wisetrend.com/solutions/ocr-legal-contracts-accounting-healthcare-education-industry/> (dostupno 29.12.2019)
- [8] <https://www.scribd.com/doc/48024585/opticko-prepoznavanje-znakova> (dostupno 25.12.2019)
- [9] <http://hr.android-pssystem.com/info/the-definition-of-ocr-function-31797553.html> (dostupno 30.12.2019)