hyväksymispäivä arvosana

arvostelija

Koneoppimisen sovellukset merkintunnistuksessa

Tuomo Salmenkivi

Helsinki 5.3.2018

HELSINGIN YLIOPISTO
Tietojenkäsittelytieteen osasto

HELSINGIN YLIOPISTO — HELSINGFORS UNIVERSITET — UNIVERSITY OF HELSINKI

Tiedekunta — Fakultet — Faculty	Laitos — Institution — Department
Matemaattis-luonnontieteellinen tiedekunta	Tietojenkäsittelytieteen osasto

Tekijä — Författare — Author
Tuomo Salmenkivi

Työn nimi — Arbetets titel — Title

Koneoppimisen sovellukset merkintunnistuksessa

Oppiaine — Läroämne — Subject
Tietojenkäsittelytiede

Työn laji — Arbetets art — Level	Aika — Datum — Month and year	Sivumäärä — Sidoantal — Number of pages
	5.3.2018	3 sivua + 0 liitesivua

Tiivistelmä — Referat — Abstract

Todo abstract

ACM Computing Classification System (CCS):
General and reference → Document types → Surveys and overviews
Applied computing → Document management and text processing → Document management → Text editing

Avainsanat — Nyckelord — Keywords
merkintunnistus, koneoppiminen

Säilytyspaikka — Förvaringsställe — Where deposited

Muita tietoja — övriga uppgifter — Additional information

Sisältö

1 Johdanto

Optista merkintunnistusta (OCR) käytetään muuttamaan skannattujen ja tyypillisesti konekirjoitettujen dokumenttien sisältö tietokoneen luettavaan ASCII merkistöön. Tämä mahdollistaa tekstin muokkaamisen, lukemisen ja säilömisen huomattavasti aiempaa tehokkaammalla tavalla.

Hyvin suuri osa arkipäiväisesti informaatiosta ei kuitenkaan ilmene dokumenteissa vaan esimerkiksi liikkeiden julkisivuissa tai tiekylteissä. Useat tavallisimmat OCR-ratkaisut eivät toimi tällaisissa luonnollisissa tilanteissa kovinkaan hyvin, sillä tyypillisesti nämä ratkaisut vaativat toimiakseen mustavalkoisen, tasaisen, linjoihin perustuvan tekstiympäristön.

Vaihtoehtoisia ratkaisuja perinteisemmille merkintunnistusratkaisulle on viime aikoina pyritty etsimään koneoppimisen kautta. Keskityn tässä tutkielmassa käsittelemään aluksi optisen tekstintunnistuksen käyttöä ja ongelmia yleisellä tasolla. Seuraavissa kappaleissa käsittelen optisen merkkitunnistuksen sovelluksia koneoppimisessa.

2 Merkintunnistus yleisesti

2.1 Merkintunnistuksen pääpiirteet

Merkintunnistus tarkoituksena on siirtää dataa tietokoneen ulkopuolisesta maailmasta muotoon, jossa sitä voidaan tutkia, editoida ja säilöä tehokkaasti. Hakusanojen avulle tiedon etsiminen suuresta määrästä dataa jota tietokone ei osaa käsitellä on ilman merkintunnistusta mahdotonta. Merkintunnistus-ohjelmistot sisältävät tyypillisesti samankaltaisia vaiheita

Ensimmäinen vaihe tekstin saamisessa tietokoneen käsiteltävään ASCII-merkistömuotoon on dokumentin siirtäminen tietokoneen muistiin. Nykyään digitaalisilla kameroilla kuvatiedosto saadaan kuvanottohetkellä automaattisesti digitaaliseen muotoon ja se voidaan myöhemmin ladata helposti kovalevylle tai flash-muistiin, mutta merkintunnistuksen alkuaikoina pakollinen ensimmäinen vaihe oli dokumentin skannaaminen. Ennen yleiskäytössä olevan skannerin yleistymistä, merkintunnistusta kehittävien tahojen ongelmana oli skannerilaitteiden kehitys, mutta yleisskannerien yleistyttyä ei tämä ollut enää suuri ongelma.

Ensimmäisen vaiheen jälkeen koneella on nyt käytössään digitaalisessa muodossa oleva kuvatiedosto. Kuvatiedosto ei vielä itsessään sisällä mitään tietoa merkeistä tai kuvassa olevasta kirjoituksesta vaan pelkästään pikseleistä, niiden väreistä ja sijainneista suhteessa kuvaan. Mikäli teksti on kirjoitus- tai tietokoneella ja samalla fontilla kirjoitettua on tässä vaiheessa tekstin tunnistaminen nukuteknologialla kohtalaisen helppoa. Vaikeudet alkavat siinä vaiheessa kun haluamme tutkia käsinkirjoitettua tai poikkeuksellisessa ympäristössä olevaa tekstiä. Ulkomaailmasta otetussa valokuvassa on niin paljon häiriötekijöitä, että tekstin löytäminen on vaikeaa ja käsinkirjoitetussa tekstissä merkkien tarkkuus suhteessa vertailuarvoihin heittää niin paljon, että tekstin tulkitseminen on haastavaa.

Seuraava vaihe vaihtelee riippuen merkintunnistusjärjestelmän implementaatiosta. Kun tarkastellaan dokumentteja, joissa kirjaisin on etukäteen tiedossa, voidaan yksinkertaisesti verrata jokaista merkkiä etukäteen tiedossa oleviin saman kirjaisimen merkkeihin. Kun saadaan osuma yhteensopivuudesta, tiedetään merkin olevan sama. Tämä vaihe ei kuitenkaan onnistu kun ei etukäteen tiedetä kirjaisinta, sillä eri kirjaisimien samat kirjaimet saattavat silti poiketa huomattavasti.

Tällaiseen tilanteeseen eräs kehitetty ratkaisu on ominaisuuksien tunnistaminen. Kun tunnistetaan merkin ominaisuuksia koko merkin sijaan ei merkin kirjaisimella ole välttämättä merkitystä. Esimkeriksi ison a-kirjaimen ominaisuuksina on käytännössä poikkeuksetta kaksi pystysuuntaista viivaa, jotka alkavat omista alakulmistaan ja kohtaavat keskellä sekä horisontaalinen viiva noin keskellä merkkiä joka yhdistää pystysuuntaiset viivat. Nyt kun etsimme tekstistä merkkejä, jotka täyttävät nämä ominaisuudet voimme suurella todennäköisyydellä löytää isot a-kirjaimet.

2.2 Merkintunnistuksen historiaa

Varsinainen tekstintunnistustamisen tutkimus sellaisenaan kun se tänä päivänä tunnetaan alkoi 1950-luvulla tarpeesta tunnistaa tekstiä pankkisekeissä. Alustavassa sekkien merkintunnistuksessa käytettiin apuna rautaoksidia sisältävää mustetta, jolloin erityisen lukulaitteen lukupää kykeni tunnistamaan merkit niiden magneettisuuden perusteella. Tätä kutsittiin MICR-teknologiaksi (Magnetic Ink Character Recognition). MICR-teknologiassa käytettiin tyypillisesti kirjaisimia E-13B sekä CMC-7. Vakiokirjaisimien käyttäminen oli etenkin varhaisvaiheen merkintunnistuksessa äärimmäisen tärkeää, sillä kirjaisimen pitäminen standardina paransi tarkkuutta ja sitä kautta luotettavuutta sekä MICR- että myöhemmin myös OCR-teknologioissa.

Varsinainen OCR-teknologia yleistyi vasta 1960-luvulla jota varten kehitettiin omat erityiset kirjaisimet, OCR-A sekä OCR-B.

2.3 OCR

3 Koneoppimisen sovelluksia merkintunnistuksessa