hyväksymispäivä arvosana

arvostelija

Koneoppimisen sovellukset merkintunnistuksessa

Tuomo Salmenkivi

Helsinki 4.3.2018

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
	4.3.2018	1 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 Merkintunnistuksen historia

2.1 Historia 1

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 MICR-teknologiassa ja myöhemmin myös OCR-ratkaisuissa.

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

3 Koneoppimisen sovelluksia merkintunnistuksessa