Mediasisältöjen tunnistaminen ja haku

Jorma Laaksonen

Aalto-yliopiston perustieteiden korkeakoulu
Tietotekniikan laitos
Espoo

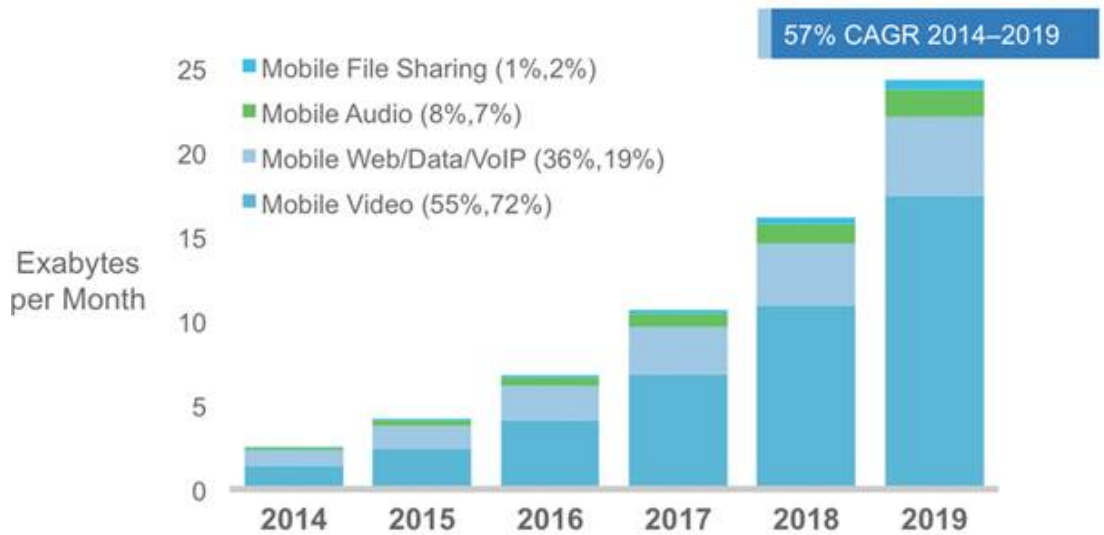
15.9.2016 @ Kynä ja kone

Sisältö

Mediadata on isoa dataa

. . . etenkin videodata

- Cisco Visual Networking Index -ennuste vuodelle 2019:
  - datasiirron yhteismäärä internetissä on 2,0 zettatavua (10^{21})
  - kaikesta internetdatasta videota on 80–90%
  - mobiilin videodatan määrä kasvaa 13 kertaiseksi vuodesta 2014
  - mobiilista datankäytöstä 72% on videota



Mobiilidatan käytön kasvu Ciscon mukaan

Mediadata on isoa dataa

. . . etenkin videodata

- Newstexin mukaan keväällä 2014 yhden minuutin aikana:
  - Instagrammin käyttäjät julkaisivat 220.000 uutta valokuvaa
    - 0,27 ms välein
  - YouTuben käyttäjät lisäsivät palveluun 72 tuntia uutta videota
    - 4,320 kertaa reaaliaika

© jdbaskin@flickr

Sisältö

Kuinka etsimänsä mediasisällön voi löytää?

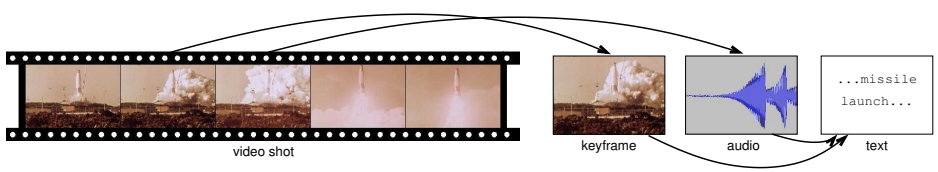
. . . kuten neulan heinäsuovasta

- Google ratkaissut kuvahakuongelman:
  - metadatan ja ympäröivän tekstin käyttö
  - keskenään samankaltaisten kuvien löytäminen
  - teksti- ja kuvainformaation yhdistäminen
  - lähes aina löytyy riittävän hyvä kuva
- Sama ei päde video- ja mediasisältöihin yleensä
- YouTubessa esimerkiksi:
  - metadatan ja tekstin käyttö
  - lataajan ja katsojien toimintahistoriat
  - suosittelut ja jakamiset
  - **ei sisältöanalyysiä**
- Aina tarvitaan etukäinen **indeksointi**

Mitä mediatiedostoista haluttaisiin löytää?

. . . eli mitä pitäisi indeksoida

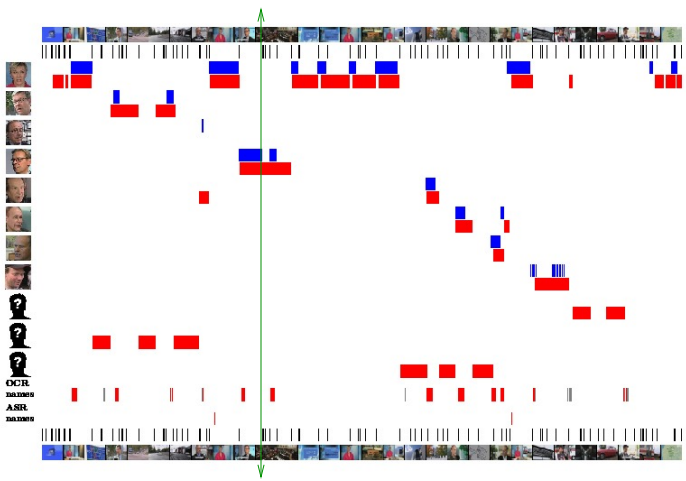
- Pitäisi löytää **tarkka kohta** videosta, jossa. . .
- kuvan perusteella
  - on tietty tai tietynlainen henkilö, paikka, rakennus, esine . . .
  - tapahtuu tiettyjä asioita tietyssä järjestyksessä
  - ihmiset/eläimet reagoivat tapahtumiin tietyllä tavalla
- äänen perusteella
  - on haluttu tai halutunlainen puhuja
  - on haluttu puhesisältö
  - on haluttu musiikki tai muut äänet
- metadatan perusteella
  - on tietty kuvaaja, kuvausaika tai -paikka
  - on tietyt kopiointioikeudet
- Tarvitaan kuva- ja äänianalyysien yhdistämistä



Mitä mediatiedostoista haluttaisiin löytää?

. . . eli mitä pitäisi indeksoida

- Esimerkki 1: Ylen alueellisten uutislähetysten indeksointi
  - visuaalisen sisällön tunnistaminen
  - henkilöiden nimien tunnistaminen kuvasta
  - kasvojen ja nimien yhteensovittaminen
  - puheäänien ja nimien yhteensovittaminen
  - puheen tunnistaminen

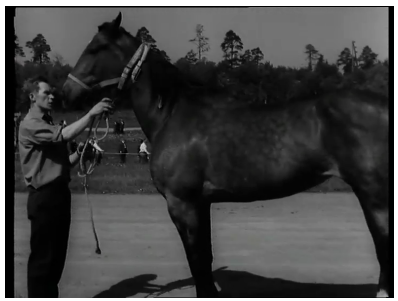


Lauri Lyly puhuu vaalirahoituksesta ja sanoo "seminaarilippuja ostetaan"

Mitä mediatiedostoista haluttaisiin löytää?

. . . eli mitä pitäisi indeksoida

- Esimerkki 2: Finlandia-katsaukset
  - 700 n. 5 minuutin mittaista uutiskatsausta vuosilta 1943–1962
  - Kansallinen audiovisuaalinen instituutti KAVI omistaa
  - kaikkien katsottavissa osoitteessa `http://www.elonet.fi/`
  - ei kunnollista indeksiä ja hakutoimintoa
  - hakulause “finlandia-katsaus hevonen” ei tuota yhtään osumaa
  - kuitenkin materiaalista löytyy *ainakin* seuraavat:



ravit Käpylässä

ratsastuskilpailu

promootio HY:ssä

Linnanmäki 10 v.

Mitä mediatiedostoista haluttaisiin löytää?

. . . eli mitä pitäisi indeksoida

- Esimerkki 2: Finlandia-katsaukset
  - 700 n. 5 minuutin mittaista uutiskatsausta vuosilta 1943–1962
  - Kansallinen audiovisuaalinen instituutti KAVI omistaa
  - kaikkien katsottavissa osoitteessa `http://www.elonet.fi/`
  - ei kunnollista indeksiä ja hakutoimintoa
  - hakulause “finlandia-katsaus hevonen” ei tuota yhtään osumaa
  - kuitenkin materiaalista löytyy *ainakin* seuraavat:

joulukadun avaus

Elanto 50 v.

Egyptin hallitsija

rauha alkanut

Mitä mediatiedostoista haluttaisiin löytää?

. . . eli mitä pitäisi indeksoida

- Esimerkki 3: Dokumenttielokuva *Helsinki ikuisesti*
  - voidaanko elokuvasta löytää kiinnostavia videosisältöjä?
  - elokuvan jokainen ruutu on analysoitu ja sisältö tunnistettu
  - sisältö voidan kuvailla avainsanoin ja hakea niiden avulla

Sisältö

Pikseleistä piirteiksi

. . . oleellinen irti epäoleellisesta

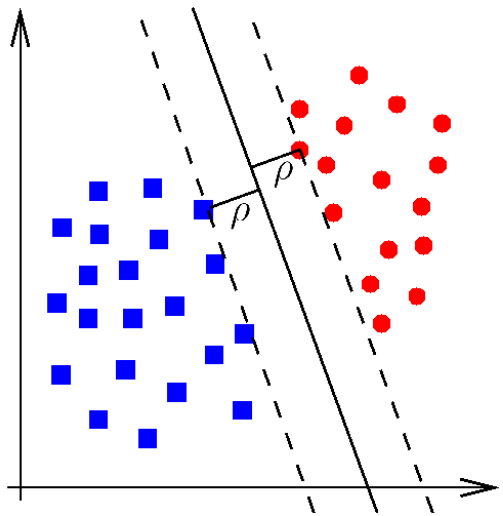
- Piirreirrotuksella pyritään
  - tiivistämään informaatio alkuperäistä pienempään datamäärään
  - korostamaan hakutehtävän kannalta tärkeää informaatiota
- Kuvista tai videon yksittäisistä ruuduista
  - lasketaan kuvailevia tunnuslukuja
  - usein korkeadimensioisia (5–20.000) vektoreita
  - keskiarvoja, histogrammeja, tietoa yhteisesiintymisestä
  - paikallisen tekstuurin kuvaus $\Rightarrow$ *bag of visual words*
  - syvillä neuroverkoilla lasketut aktivaatiopiirteet
- Videosta ruutujen välistä muutosta kuvaavat piirteet

© Siddhartha Chandra, INRIA

Piirteistä luokitteluiksi

. . . mikä erottaa kissat koirista?

- Muodostetaan koneoppimisella päätössääntö, joka kertoo:
  - mihin luokkaan kuva tai video kuuluu *tai*
  - kuinka todennäköisesti se kuuluu tiettyyn luokkaan
- Erlaisia luokittelumenetelmiä käytetty:
  - tukivektorikone (SVM)
  - luokittelupuut
  - neuroverkot
- Hyvien luokittelijoiden opettamiseen tarvitaan **paljon** dataa



Sisältö

Suuria aineistoja luokittelijoiden opetukseen

. . . mitä enemmän sitä parempi

- Data-aineistot ovat kasvaneet moneen suuntaan:
  - enemmän kuvaluokkia (20 $\longrightarrow$ 20.000)
  - enemmän näytteitä kustakin luokasta (100 $\longrightarrow$ 2.000)
  - vaikeampia esimerkkejä
- Data-aineistojen käyttö vakiintunut tietyksi *benchmark*iksi
  - tulokset keskenään vertailukelpoisia
  - toisten tutkijoiden koejärjestelyjä ei tarvitse toistaa
- Käytettävissä olevien data-aineistojen kasvu on
  - johtanut parempiin piirreirrotus- ja luokittelumenetelmiin
  - tehnyt uusia asioita mahdolliseksi

IMAGENET

`http://www.image-net.org`

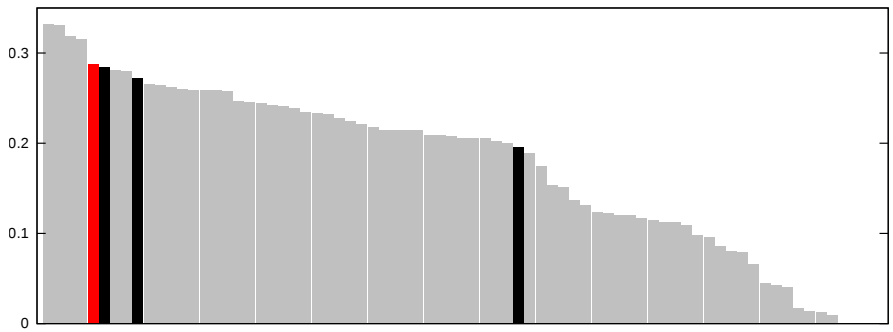
14.197.122 kuvaa

21.841 luokkaa

TRECVID

. . . NISTin vuosittainen videohakumenetelmien arviointi

- *Semantic Indexing* -tehtävä:
  - 1.400 tuntia videota, 600 GB dataa
  - 35.300 videota, n. 5 min/video, 880.000 otosta
  - 346 semanttista luokkaa
  - *auto*, *hevonen*, *tietokone*, *ihminen*, *näyttelijä*, *laulaminen*, . . .
- *Multimedia Event Detection* -tehtävä:
  - 5.151 tuntia videota, 2 TB dataa
  - 144.000 videota, n. 2 min/video
  - joka vuosi uudet “tapahtumat”
  - *syntymäpäiväjuhla*, *renkaanvaihto*, *eläimenhoito*, *paraati*, . . .



PicSOM-ryhmän sijoitus v. 2014 SIN-tehtävässä

Sisältö

Kuvien ja videoiden kuvailu lauseilla

. . . UUTTA v. 2015 !!!

- Kuvien ja videoiden kuvailu käyttää usein vain substantiiveja
- Tilanne on muuttumassa uusien tietokantojen myötä
- Microsoft Common Objects in Context (COCO)
  - `http://mscoco.org/explore/`
  - 80 kuvaluokkaa, 160.000 kuvaa, jokaisella viisi kuvailulausetta
  - arvioinnissa sekä automaattinen että ihmisen arviointi

“a man with a red helmet on a small moped on a dirt road” / “man riding a motor bike on a dirt road on the countryside” / “a man riding on the back of a motorcycle” / “a dirt path with a young person on a motor bike rests to the foreground of a verdant area with a bridge and a background of cloud wreathed mountains” / “a man in a red shirt and a red hat is on a motorcycle on a hill side”

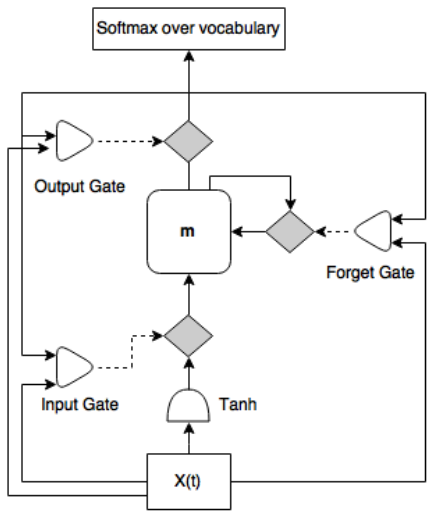
Lausekuvailun tuottaminen

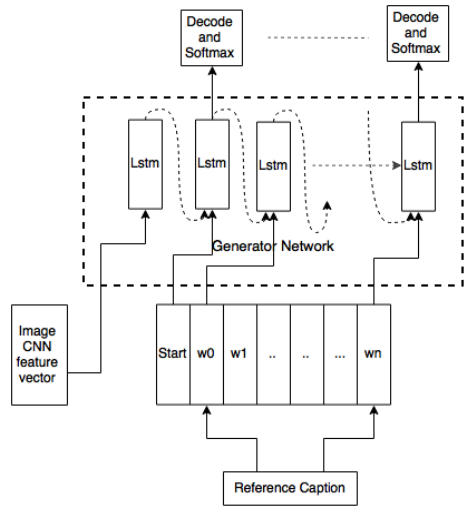
. . . vanha menetelmä uudessa tehtävässä

- Opitaan tunnettuista esimerkeistä malli, joka tuottaa todennäkäisyyden P lauseelle S, kun kuva I on annettu:

$$\begin{aligned} P(S|I) &= P(w_0 w_1 w_2 \cdots w_{n-1}|I) \\ &= P(w_0|I)\, P(w_1|w_0, I)\, P(w_2|w_1, I) \cdots P(w_{n-1}|w_{n-2}, I) \end{aligned}$$

- Kun malli on muodostettu, sillä tuotetaan uusia lausekuvailuja
- Long Short-Term Memory (LSTM) -neuroverkko:





Lausekuvailu videoille

. . . uusi tehtävä, oletettavasti vaikea

- *Large-Scale Movie Description Challenge* (LSMDC)
  - noin 100 elokuvaa
  - noin 100.000 videoleikettä lausekuvailuilla
  - kuvailut tarkoitettu alunperin näkövammaisille
  - tehtävä: tuotettava lausekuvailu 3–10 sekunnin leikkeille
  - arvioinnissa sekä automaattinen että ihmisen arviointi

“she flips open the lid of a pastry box and kisses his cheek”

Tuloksia kuvien ja videoiden lausekuvailusta

MS COCO Image Captioning Challenge 2015

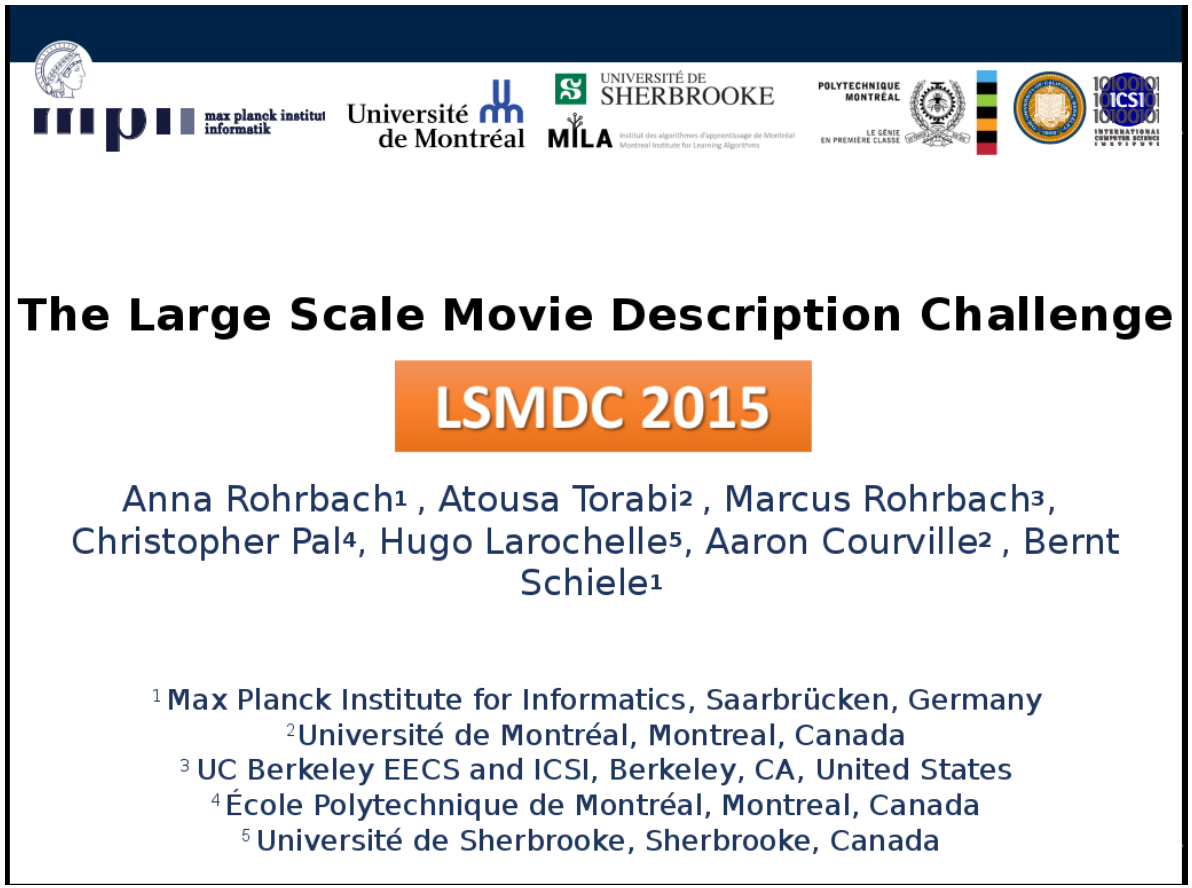
- a man riding a bike down a dirt road
- a man riding a bike down the street
- a man is riding a bike with a dog
- a man riding a bike down a road next to a forest
- a man riding a bike down a street next to trees

- Kuvailut ovat useimmiten tosia, mutta varsin yksinkertaisia, oleelliset yksityiskohdat jäävät usein kuvailematta
- Aallon ryhmä tulosten ykkösenä 5/2016 useimmilla mittareilla

Tuloksia kuvien ja videoiden lausekuvailusta

Large-Scale Movie Description Challenge 2015

- Aallon ryhmä kilpailun voittaja
- Kuvailut ovat useimmiten tosia, mutta varsin yksinkertaisia, oleelliset yksityiskohdat jäävät usein kuvailematta



The Large Scale Movie Description Challenge

LSMDC 2015

Anna Rohrbach[1], Atousa Torabi[2], Marcus Rohrbach[3], Christopher Pal[4], Hugo Larochelle[5], Aaron Courville[2], Bernt Schiele[1]

[1] Max Planck Institute for Informatics, Saarbrücken, Germany
[2] Université de Montréal, Montreal, Canada
[3] UC Berkeley EECS and ICSI, Berkeley, CA, United States
[4] École Polytechnique de Montréal, Montreal, Canada
[5] Université de Sherbrooke, Sherbrooke, Canada

Tuloksia kuvien ja videoiden lausekuvailusta

Microsoft Multimedia Video to Language Challenge 2016

- 10,000 videota, kukin 15-30 s, 20 lausetta kustakin
- Osa Microsoft Researchin laajempaa Multimedia Challengea
- Aallon ryhmä kilpailun voittaja ihmisarvioiden perusteella ja toinen automaattisten arvioiden perusteella

Automatic Multimodal Content Analysis

. . . pilot studies

- Esittely HY:n prof. Liisa Tiittulan ryhmän kanssa tehdystä hankesuunnitelmasta Suomen Akatemian ja EU:n hakuihin
  - Helsinki ikuisesti -elokuva
  - suomenkielinen AD-sisältökuvailu
  - visuaalinen sisällöntunnistus avainsanoille
  - visuaalinen sisällönkuvailu lauseilla
  - puheen- ja puhujantunnistus
  - katseen- ja kohteenseuranta