Jorma Laaksonen
vanhempi yliopistonlehtori, Tietotekniikan laitos, Aalto-yliopisto

Objektintunnistus ja suomalainen elokuva

Elokuvien kuvasisällön tutkiminen voi antaa meille kiinnostavaa tietoa siitä, millaisia esineitä suomalaisissa elokuvissa on esiintynyt ja miten esiintymisten lukumäärät ovat kehittyneet vuosien kuluessa. Tällaisen analyysin perusteella voitaisiin esimerkiksi vertailla eri aikakausien elokuvien esineistöä ja havaita, milloin jotkut esineet ovat näkyneet ensimmäisen tai viimeisen kerran suomalaisilla valkokankailla. Erityisesti näytelmäelokuvat ovat tässä suhteessa kiinnostavia, koska ne voivat sisältää myös fantasioituja asioita ja esineitä. Tässä katsauksessa esittelen MoMaF-projektissa tehtyjä kvalitatiivisia ja kvantitatiivisia kokeita, joilla halusimme selvittää, miten koneellinen objektintunnistus soveltuu elokuvien visuaalisen sisällön tarkasteluun.

Konenäkö ja kuvatietokannat

Esineiden tai yleisemmällä tasolla *objektien* tunnistaminen kuvista ja videoista on konenäön ja kuva-analyysin yksi klassinen ja keskeinen tutkimuskohde. Mielenkiinto tätä tutkimuskysymystä kohtaan on viime vuosina lisääntynyt samalla, kun yleensäkin tekoälyn ja koneoppimisen, etenkin *syvien neuroverkkojen*, tarjoamat mahdollisuudet on huomattu. Syvillä neuroverkoilla tarkoitetaan matemaattisia malleja, jotka voidaan opettaa tuottamaan haluttuja vastearvoja annetuille syötearvoille. Tähän käytetään hyvin suurta määrää syöte–vaste-pareja, joiden perusteella verkkojen sisältämiä – tyypillisesti miljoonia – parametriarvoja muokataan vaiheittain tuottamaan halutut vasteet.

Tärkeä osatekijä objektintunnistuksen kehityksessä on ollut se, että konenäkötutkijoiden käyttöön on tullut suuria kuvatietokantoja, joissa olevat objektit on ihmistyönä merkitty eli *annotoitu* suorakaiteen muotoisin rajauksin tai siluettimuodoin. Näiden kuvatietokantojen ansiosta on ollut mahdollista koneoppimisen avulla opettaa automaattisia tunnistimia, jotka kykenevät löytämään kyseisiä objekteja kuvista ja videoista. Esineiden tunnistamisen lisäksi kuvista ja videoista on visuaalisesti mahdollista tunnistaa muun muassa henkilöitä kasvojen perusteella, erilaisia näkymiä tai paikkoja (kuten *keittiö*, *urheilukenttä* tai *kauppa*) sekä toimintoja ja aktiviteetteja (kuten *hyppääminen*, *tanssiminen* tai *ruuanlaitto*).

Tunnetuin konenäkötutkimuksessa käytetty kuvatietokanta on Microsoft Researchin julkaisema *Common objects in context* eli COCO (Lin et al. 2014). Julkaisuaikanaan vuonna 2014 se oli poikkeuksellisen suuri kooltaan sisältäen yli 200 000 annotoitua kuvaa. Annotoituja objektiluokkia siinä on 80, mitä voidaan nykyisin pitää jo kovin pienenä määränä, mutta oleellista on, että objektien siluetit on COCOssa ihmistyönä merkitty niin, että on mahdollista tietää tarkasti, mitkä pikselit kuhunkin objektiin kuuluvat. Kuvassa 1 on esimerkki COCO-kuvatietokannan kuvasta ja siihen annotoiduista objekteista. Voidaan nähdä, että kuvaan on merkitty yksi tai useampia koiria,

ihmisiä, frisbeekiekkoja, käsilaukkuja ja matkapuhelimia. Annotointien tarkkuus jättää toivomisen varaa, sillä esimerkiksi kaikkia selvästikään erottuvia ihmisiä ei väkijoukosta ole merkitty.

COCOn objektiluokat voidaan karkeasti jakaa seuraaviin kategorioihin: ihminen, henkilökohtaiset tavarat (esim. sateenvarjo, reppu), kulkuneuvot (auto, polkupyörä), kadun esineet (liikennevalot, puistonpenkki), eläimet (kissa, kirahvi), urheiluvälineet (baseballmaila, rullalauta), ruokailuvälineet (kulho, haarukka), ruuat (pizza, porkkana), huonekalut (sohva, pöytä), elektroniikka (näppäimistö, TV), keittiökalusteet (mikroaaltouuni, lavuaari), muut (hammasharja, kirja). Kuinka ”tavallisia” kyseiset esineet konteksteineen ovat, on luonnollisesti hyvin aika-, paikka- ja kulttuurisidonnaista. COCO-tietokantaa suurempiakin annotoituja kuvatietokantoja on olemassa, tärkeimpänä Princetonin yliopiston ImageNet (Deng et al. 2009), jossa on yli 14 miljoonaa kuvaa ja lähes 22 tuhatta objektiluokkaa. ImageNetin ratkaiseva heikkous kuitenkin on, että siinä ei ole merkitty objektien siluetteja tai rajauksia. Vaikka kuvatietokannat lähes poikkeuksetta on annotoitu ainoastaan englanninkielisin termein, tämä ei aiheuta merkittävää kielellistä tai kulttuurista ongelmaa, koska termisanat voidaan ongelmitta kääntää toisille kielille.

Kuva 1. Esimerkki COCO-kuvatietokannan kuvasta ja siinä olevien objektien annotoinneista. Lähde: COCO-tietokanta: https://cocodataset.org/#explore?id=444350.

Objektien tunnistaminen kuvista perustuu siis koneoppimisen menetelmiin, jotka tarvitsevat opetusmateriaalikseen kuvia, joihin objektiluokkien nimet ja objektien sijainnit on merkitty. Yleistäen voidaan sanoa, että mitä enemmän opetusmateriaalia on käytettävissä, sitä paremmin menetelmät oppivat tunnistamaan haluttuja esineitä. Toisaalta mitä enemmän eri objektiluokkia halutaan tunnistaa, sitä epäluotettavampia tunnistukset oletettavasti ovat, koska mahdollisuuksia erilaisiin erehdyksiinkin tulee enemmän. Voidaan sanoa, että objektintunnistusta toteutettaessa joudutaan tasapainoilemaan mahdollisten tunnistusten rikkauden ja tarkkuuden välillä.

Visuaalisen objektintunnistuksen menetelmistä

2000-luvun kuluessa visuaalisen objektintunnistuksen menetelmät ovat kehittyneet huikeasti. Ennen niin sanottua syväoppimiseen perustuvien neuroverkkojen nykyistä vuonna 2012 alkanutta valtakautta parhaat menetelmät perustuivat kohteiden

kuvaamiseen osiensa muodostamana kokonaisuutena niin sanotun *deformable part model* -periaatteen mukaisesti (Felzenszwalb et al. 2010). Uudemmat menetelmät ovat periaatteeltaan suoraviivaisempia ja välttävät objektien osien etsimisen. Nykyiset menetelmät voidaan karkeasti jakaa yhden ja kahden vaiheen tekniikoihin, joista jälkimmäiset pyrkivät ensin löytämään tunnistettavan objektin ja sitten erikseen luokittelemaan sen, kun taas edelliset tekevät nämä kaksi vaihetta samanaikaisesti. Kahden vaiheen menetelmät ovat näistä vanhempia, ja niistä yleisimmin käytetty lienee Faster R-CNN (Ren et al. 2015). Vastaavasti tunnetuin yhden vaiheen menetelmä on *You only look once* eli YOLO (Redmon et al. 2018). Molemmat menetelmät tunnistavat kuvassa näkyviä objekteja tuntemiinsa objektiluokkiin ja osoittavat kuvasta mahdollisimman pienen suorakaiteen muotoisen alueen (engl. *bounding box*), jonka sisällä objekti sijaitsee. Lisäksi menetelmät palauttavat välillä 0–1 olevan lukuarvon, joka kertoo tunnistimen oman käsityksen tunnistuksensa tarkkuudesta. Mitä lähempänä arvoa yksi tarkkuusarvio on, sitä vakuuttuneempi tunnistin on omasta toiminnastaan.

Objektintunnistusmenetelmät ovat lähtökohtaisesti kuva- eikä videopohjaisia, eli ne tuottavat tunnistuksensa aina yksittäisiin videon ruutuihin. Siksi on mahdollista ja välttämätöntäkin jälkikäsitellä tunnistuksia ajallisesti yli peräkkäisten ruutujen. Tällä jälkikäsittelyllä voidaan paikata aukkoja, joita syntyy, kun tunnistin ei havaitse objektia aivan joka ruudussa, ja poistaa satunnaisia vääriä tunnistuksia, jotka usein esiintyvät vain yksittäisissä ruuduissa tai muutoin hyvin lyhyen aikaa. Tunnistimien arvioita omien tunnistustensa tarkkuudesta voidaan lisäksi hyödyntää osana ajallista käsittelyä ja poistaa lyhytaikaisia oletettavasti epäluotettavia tunnistuksia.

Kokeellisia tuloksia suomalaisilla elokuvilla

Aalto-yliopistossa tehdyssä tutkimuksessa (Xiang 2022) verrattiin Faster R-CNN- ja YOLO-menetelmiä objektintunnistuksessa yhden suomalaisen elokuvan tapauksessa. Kyseessä oli Valentin Vaalan vuonna 1958 ensi-iltaan tullut värielokuva *Nuori mylläri*, jonka tapahtumat sijoittuvat 1800- ja 1900-luvun vaihteeseen. Tutkimuksen tarkoituksena oli selvittää MoMaF-projektin jatkoa varten, kumpi menetelmistä toimii tarkemmin muutamille kyseisessä elokuvassa usein toistuville esineluokille, ja yleisemmin, ovatko yhden vai kahden vaiheen menetelmät tässä tehtävässä parempia. Tarkempaan analyysiin työssä valikoituivat objektiluokat *pullo* ja *hevonen*. Näistä pullot ovat pienempiä ja useimmiten näkyvät pienempikokoisina kuin hevoset. Toisaalta pullot ovat visuaalisesti keskenään ja katselukulmasta riippumatta varsin samannäköisiä toisin kuin hevoset, jotka eri asennoissa ja eri katselukulmista nähtyinä voivat olla hyvinkin erinäköisiä. Kuvassa 2 on esimerkkejä pullojen ja hevosten tunnistuksista *Nuoren myllärin* yksittäisissä ruuduissa.

Näemme, että Faster R-CNN- ja YOLO-menetelmät ovat tunnistaneet hieman erilaiset rajaukset objekteille ja että myös niiden arvioidut tunnistustarkkuudet voivat olla hyvin erilaisia. Yleisellä tasolla johtopäätös tästä tutkimuksesta oli, että kaksivaiheinen Faster R-CNN tunnisti yksivaiheista YOLOa herkemmin myös hyvin pieninä näkyviä objekteja, kuten pulloja. Suurempien objektien kohdalla puolestaan YOLOn tuottamat tunnistukset olivat luotettavampia, sillä Faster R-CNN tunnisti YOLOa useammin myös vääriä kohteita.

Osana MoMaF-hankkeen tutkimusta (Grósz et al. 2022) pyrittiin tunnistamaan Faster R-CNN -menetelmällä 1950-luvun aikana ensi-iltaan tulleissa draamaelokuvissa esiintyviä hevosia ja autoja. Elokuvien lukumäärä tässä tutkimuksessa oli 186 ja niiden yhteiskesto yli 257 tuntia. Siten oli mahdotonta tarkistaa ihmistyönä, kuinka monessa ruudussa tunnistin oli tunnistanut objektin, vaikka sitä ei todellisuudessa

Kuva 2. Pullojen ja hevosten tunnistuksia *Nuori mylläri* -elokuvassa. Vihreät rajaukset ovat Faster R-CNN -menetelmän ja siniset YOLOn tuottamia. Tunnistimien itse tuottamat arviot tunnistuksen tarkkuudesta ovat muissa tapauksissa 0.90 tai yli, mutta ylhäällä oikealla olevalle YOLOn tulokselle se on 0.21 ja alhaalla vasemmalla olevalle YOLOn laajemmalle rajaukselle 0.58. Lähde: Xiang 2022.

sillä kohtaa elokuvaa näykään (väärä positiivinen tunnistus eli tyypin I virhe), tai jättänyt näkyvän hevosen tai auton tunnistamatta (väärä negatiivinen tunnistus eli tyypin II virhe). Tarkkaa kvantitatiivista arviota tunnistusten tarkkuudesta kummallekin objektiluokalle ei siten ollut mahdollista saada. Sen sijaan koetta tehtäessä oli tiedossa, missä elokuvissa ei lainkaan esiintynyt autoja. Näiden autottomien elokuvien lukumäärä oli 64. (Ks. Tommi Römpötin artikkeli tässä Lähikuvan numerossa.) Tehdyissä kokeissa osoittautui, että järjestelmä kykeni noin 84 % tarkkuudella tunnistamaan oikein, oliko kyseessä autoton vai autollinen elokuva. On huomattava, että tunnistimen opetuksessa käytetyt esimerkkikuvat autoista esittävät enimmäkseen automalleja, jotka poikkeavat huomattavasti 1950-luvun suomalaisten elokuvien sisältämistä autoista. Samoin osassa elokuvista saattoi olla autoista vain sisäkuvia, kun taas tunnistin oli opetettu tunnistamaan autoja vain ulkopuolelta nähtyinä.

Automaattisen tunnistuksen sovellettavuudesta

Visuaalinen objektintunnistus ei – kuten eivät mitkään kuvainformaation automaattiset tunnistusmenetelmät – voi koskaan olla täydellistä, sillä se tuottaa aina myös virheitä sekä väärinä tunnistuksina että tunnistamatta jättämisinä. Millaisiin tarkoituksiin tällaista osittain epäluotettavaa tunnistusta sitten voidaan elokuvatutkimuksessa käyttää? Ensimmäinen käyttötarkoitus voi olla etsiä tiettyä esinettä tai objektia yhdestä elokuvasta tai pienestä joukosta elokuvia. Tunnistin voi tällöin itse tuottamansa tarkkuusarvion perusteella osoittaa joukon ruutuja, joissa kyseinen objekti kaikkein *todennäköisimmin* esiintyy. Järjestelmää käyttävä ihminen voi tällöin vain pienen määrän ruutuja tarkistamalla selvittää varsin luotettavasti, esiintyykö kohteena oleva esine kyseisissä elokuvissa lainkaan. Mikäli tarkastuksessa ilmenee, että esine ylipäätään esiintyy jossain elokuvassa, on mahdollista jatkaa järjestelmän osoittamien ruutujen tarkistamista edelleen useampien tai kaikkien sellaisten kohtausten löytämiseksi, joissa esine on nähtävissä.

Toinen mahdollinen sovellus on objektien esiintymisten runsauden suurpiirteinen kvantitatiivinen arviointi. Tätä lähestymistapaa käytettiin edellä kuvatussa julkaisussa (Grósz et al. 2022), jossa aiheena oli hevosten ja autojen esiintymisten frekvenssin muutos 1950-luvun elokuvissa. Kaikkien kyseisten objektien esiintymisten tarkka tunnistaminen ei tällaisessa *kvantitatiivisessa* tutkimuksessa ole välttämätöntä, mikäli voidaan luottaa siihen, että automaattisten tunnistusten määrä on jonkinlaisessa vakiosuhteessa todellisten esiintymisten määrään. Tällainen epätarkkakin tulos tulee vielä tarkemmaksi ja käyttökelpoisemmaksi, jos lisäksi voidaan keskiarvoistaa useiden elokuvien tuloksia. Edellä mainitussa tutkimuksessa näin tehtiin kunakin vuonna ensi-iltaan tulleiden elokuvien tuloksille ennen muutostrendien laskemista.

Visuaalisen objektintunnistuksen soveltaminen hyvin suomalaiseen esineistöön

Visuaalinen objektintunnistus on mahdollista vain sellaisille objektiluokille, joista on olemassa riittävästi esimerkkikuvia käytettäväksi tunnistimen opetuksessa. Toisaalta jotkut objektityypit, kuten aiemman esimerkin autot, ovat voineet vuosien kuluessa muuttaa visuaalista muotoaan niin, että vanhojen esineiden tunnistaminen uusia esittävien kuvien perusteella on vaikeaa. Visuaalisia tunnistimia voidaan luoda täysin uusille objektiluokille keräämällä riittävä määrä kuvia ja annotoimalla kyseisen luokan esiintymiset niissä. Samoin esimerkiksi COCO-kuvatietokannassa jo olevia kuvaluokkia voidaan täydentää esimerkiksi historiallisilla tai paikallisilla näytteillä samasta aiheesta. Tällä tavoin esimerkiksi suomalaiselle kulttuuriperinnölle ominaisia esineitä, kuten heinäseipäitä ja saunavihtoja, voitaisiin alkaa tunnistaa aikaisempaa paremmin.

Kuinka paljon annotoituja esimerkkejä, siis kuvia, joihin kyseisen esineen esiintyminen on rajattu siluetilla tai suorakaiteen muotoisella alueella, sitten tarvittaisiin? COCO-tietokannassa on opetuksessa käytettäviä kuvaesimerkkejä vähimmillään 128 yhtä objektiluokkaa kohden. Tämä pienin luokka sattuu olemaan *hiustenkuivaaja*. Yleisesti onkin arvioitu, että useimmissa tapauksissa noin 100–200 esimerkkiä on riittävä määrä uuden visuaalisen tunnistimen opettamiseksi. Siten myös hyvin suomalaisten esineiden automaattinen tunnistaminen elokuvista on mahdollista, jos esimerkkien keräämiseen ja annotoimiseen halutaan ryhtyä.

Lähteet

Deng, Jia, Dong, Wei, Socher, Richard, Li, Li-Jia, Li, Kai, Fei-Fei, Li (2009) ImageNet: A large-scale hierarchical image database. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (CVPR). Saatavilla: https://image-net.org/static_files/papers/imagenet_cvpr09.pdf (linkki tarkistettu 1.11.2022).

Felzenszwalb, Pedro, Girshick, Ross, McAllester, David, ja Ramanan, Deva (2010) Object Detection with Discriminatively Trained Part Based Models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9). Saatavilla: https://cs.brown.edu/people/pfelzens/papers/lsvm-pami.pdf (linkki tarkistettu 1.11.2022).

Grósz, Tamás, Kallioniemi, Noora, Kiiskinen, Harri, Laine, Kimmo, Moisio, Anssi, Römpötti, Tommi, Virkkunen, Anja, Salmi, Hannu, Kurimo, Mikko, ja Laaksonen, Jorma (2022) Tracing Signs of Urbanity in the Finnish Fiction Film of the 1950s: Toward a Multimodal Analysis of Audiovisual Data. *Proceedings of the 6th Conference on Digital Humanities in the Nordic and Baltic Countries*. Saatavilla: http://ceur-ws.org/Vol-3232/paper05.pdf (linkki tarkistettu 1.11.2022).

Lin, Tsung-Yi, Maire, Michael, Belongie, Serge, Hays, James, Perona, Pietro, Ramanan, Deva, Dollár, Piotr, ja Zitnick, C. Lawrence (2014) Microsoft COCO: Common Objects in Context. *Proceedings of the European Conference on Computer Vision* (ECCV). Saatavilla: https://arxiv.org/abs/1405.0312 https://cocodataset.org/ (linkki tarkistettu 1.11.2022).

Redmon, Joseph, ja Farhadi, Ali (2018) YOLOv3: An Incremental Improvement. *arXiv*:1804.02767. Saatavilla: https://arxiv.org/abs/1804.02767 (linkki tarkistettu 1.11.2022).

Ren, Shaoqing, He, Kaiming, Girshick, Ross, ja Sun, Jian (2015) Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *Proceedings of the 29th Conference on Neural Information Processing Systems* (NIPS). Saatavilla: https://arxiv.org/abs/1506.01497 (linkki tarkistettu 1.11.2022).

Xiang, Wen (2022) *Object Detection in Finnish Movies*. Diplomityö, Aalto-yliopisto. Saatavilla: https://aaltodoc.aalto.fi/handle/123456789/116378 (linkki tarkistettu 1.11.2022).