Aalto-yliopisto
Perustieteiden korkeakoulu
Tietotekniikan koulutusohjelma

Eleentunnistus Kinect-sensorilla

Kandidaatintyö

25. huhtikuuta 2013

Liisa Sailaranta

Aalto-yliopisto
Perustieteiden korkeakoulu
Tietotekniikan koulutusohjelma

KANDIDAATINTYÖN
TIIVISTELMÄ

Tekijä:	Liisa Sailaranta
Työn nimi:	LaTeX-pohja kandidaatintyötä varten ohjeiden kera ja varuilta kokeillaan vähän ylipitkää otsikkoa
Päiväys:	25. huhtikuuta 2013
Sivumäärä:	Kirjoita tähän oikea määrä, tässä esimerkissä 23
Pääaine:	Tietojenkäsittelytiede
Koodi:	IL3010
Vastuuopettaja:	Ma professori Tomi Janhunen
Työn ohjaaja(t):	Ohjaajantitteli Sinun Ohjaajasi (Poimi tähän ohjaajasi laitos, DEPT, main.tex)
Avainsanat:	avain, sanoja, niitäkin, tähän, vielä, useampi, vaikkei, niitä, niin, montaa, oikeasti, tarvitse
Kieli:	Suomi

Sisältö

1 Johdanto

Tämä kandinaatintyö käsittelee eleentunnistusta Kinect-syvyyskameran avulla. Työn tarkoitus on tutustua eleentunnistusongelmaan ja 3D-videokuvan käyttöön eleentunnistustuksessa.

Kuvaa ja videokuvaa on tutkittu paljon, mutta eleentunnistus on edelleen suuri haaste. Ihmisen eleet ovat monimutkaisia ja niiden esitystapa vaihtelee esiintyjästä ja tilanteesta riippuen. Eleentunnistuksella on kuitenkin monia käyttötarkoituksia esimerkiksi elekäyttöliittymissä. (Guyon et al., 2012a) Toistaiseksi eleentunnistusmenetelmät eivät ole olleet riittävän luotettavia ja nopeita, jotta niitä olisi voitu laajasti hyödyntää kuluttajasovelluksissa.

Kehittynyt tekniikka kuten Microsoftin Kinect-sensori ja kehittynyt laskentateho tuovat alalle uusia mahdollisuuksia. Kinect-sensori on Microsoftin kehittämä 3D-kamera, joka tarjoaa 3D-videokuvaa eli tavallisen värikuvan lisäksi Kinect-kamera antaa infrapunakameralla mitattua syvyyskuvaa. (Latotzky, 2011) 3D-videokuva tuo monia etuja eleentunnistukseen verrattuna 2D-videokuvaan. 3D-kuva ei ole kuitenkaan vielä toistaiseksi radikaalisti kehittänyt tai muuttanut olemassa olevia eleentunnistusmenetelmiä.

Lisätäkseen kiinnostusta 3D-kuvaan järjestettiin ChaLearn Gesture Challenge -eleentunnistuskilpailu, jossa kilpailijat kehittivät Kinect-sensorin datalle suunnattuja eleentunnistusmenetelmiä. Tämä kandinaatintyö tutustuu kilpailutöihin ja kartoittaa niiden avulla alan uusinta tutkimusta. Kilpailu on hyvä tutkimuskohde, sillä sen avulla voidaan puolueettomasti vertailla erilaisia menetelmiä. Eleentunnistukselle on tyypillistä, että menetelmät oppivat liiankin hyvin opetusdatajoukkoon, eivätkä ole enää yleistettävissä muille datajoukoille. Jotta saadaan vertailtavissa olevia tuloksia, on eri menetelmiä testattava samalla ja mielellään kokonaan uudella datajoukolla.

ChaLearn Gesture Challenge -kilpailussa jaettiin Kinect-sensorilla kuvattuja näytteitä erilaisista eleistä. Jokaisesta eleestä annettiin yksi näyte, koska kilpailijoita haluttiin kannustaa kehittämään yhdestä eleestä oppimiseen sopivia menetelmiä (One Shot Learning). Kilpailutöiden menetelmät olivat pääasiallisesti sellaisia, että ne soveltuisivat myös tavalliselle 2D-videokuvalle. Menetelmät yhdistelivät tunnettuja menetelmiä kuvan-, videon- ja äänentunnistuksesta.

Luvussa kaksi esitellään tarkemmin eleentunnistusongelmaa ja 3D-videokuvaa. Luvussa

kolme kerrotaan ChaLearn Gesture Challenge -kilpailusta ja luodaan yleiskatsaus kilpailutöihin. Luvussa neljä esitellään tarkemmin kolme kilpailutöistä. Kilpailutyöt on valittu menestyneiden kilpailutöiden joukosta edustamaan erilaisia näkökulmia ongelmaan.

Työ pyrkii kartoittamaan minkälaisia keinoja eleentunnistuksessa 3D-videokuvalla voidaan käyttää ja miten ne vastaavat eleentunnistuksen haasteisiin. Työssä esitellään jonkin verran hahmontunnistuksen käsitteitä ja käytäntöjä, mutta pääasiallisesti lukijan oletetaan tuntevan hahmontunnistuksen peruskäsitteet. Työ ei ota kantaa menetelmien tekniseen toteutukseen, vaan keskittyy teorian kuvaamiseen. Työssä ei myöskään keskitytä Kinect-kameran teknisiin ominaisuukseen, vaan kameraa esitellään ainoastaan ongelman ymmärtämisen kannalta välttämätön määrä.

2 Eleentunnistus videokuvalta

2.1 Eleentunnistus ongelmana

Eleentunnistuksella tarkoitetaan tässä työssä ihmisen suorittaman eleen tunnistamista videokuvalta. Eleitä voisivat olla esimerkiksi viittomakielen eleet tai yksinkertaiset toiminnot kuten istuminen. Eleentunnistus on haastava ongelma. Ihmisen eleet ovat monimutkaisia ja videokuvalla on paljon muuttujia kuten valaistus, tila tai kohteen etäisyys kamerasta (Wang et al., 2003). Luokan sisällä on paljon vaihtelua eli sama ele näyttää erilaiselta videonäytteestä riippuen. Yhtenä haasteena on ollutkin riittävän monipuolisen opetus- ja testitietokannan kerääminen. (Laptev et al., 2008) Microsoftin Kinect-sensorin tapaisten syvyyskameroiden avulla haasteisiin voidaan kuitenkin vastata entistä tehokkaammin (Guyon et al., 2012b).

Eleentunnistus on hyvä erottaa asennontunnistus-ongelmasta (Pose Estimation). Asennontunnistuksessa pyritään tunnistamaan ihmisen asento videolla yhdessä pysäytyskuvassa käyttämättä lainkaan ajallista tietoa. Tunnistuksessa pyritään usein tunnistamaan ihmishahmon ruumiinosat esimerkiksi nivelet, joiden pohjalta arvioidaan asento. (Shotton et al., 2011) Ongelmana asennontunnistus on tietyssä mielessä helpompi kuin eleentunnistus, sillä hahmontunnistus kuvalta on yksinkertaisempaa kuin videokuvalta ja sitä on tutkittu enemmän. Yksittäisiä asentoja voidaan käyttää tunnistamaan kokonainen ele, joskin se on laskennallisesti raskasta.

Kiinnostus eleentunnistusta kohtaan on lisääntynyt viime vuosina sen monien käyttötarkoitusten vuoksi. Eleentunnistusta voidaan käyttää monenlaisissa elekäyttöliittymis-

sä. Yksittäisiä eleitä voidaan käyttää esimerkiksi kodinkoneiden ohjailuun. (Wang et al., 2003) Toisaalta eleentunnistusta voidaan käyttää hyödyksi tunnistamaan erilaisia vaaratilanteita. Esimerkiksi potilaan tilaa voidaan seurata eleentunnistuskameralla mahdollisten poikkeavien eleiden varalta (Guyon et al., 2012a).

Videokuvan tunnistuksessa voidaan hyödyntää perinteisiä hahmontunnistusmenetelmiä. Monet menetelmistä ovat kuitenkin laskennallisesti liian raskaita reaaliaikaiseen videokuvan tunnistukseen, jota vaaditaan elekäyttöliittymissä (Wang et al., 2003). Monet hahmontunnistusmenetelmät vaativat myös paljon opetusdataa. Kuluttajille suunnatuissa sovelluksissa olisi toivottavaa, että uuden eleen voi opettaa muutaman testinäytteen perusteella (Wang et al., 2003). Eleentunnistusmenetelmien on kyettävä vastaavaan näihin haasteisiin.

Eleentunnistuksessa, kuten hahmontunnistuksessa yleensä, korkeimpana tavoitteena on jäljitellä ihmisen toimintamalleja. Ihmisen kyky tunnistaa ja oppia hahmoja on erinomainen. Ihminen kykenee oppimaan eleet yhden opetusnäytteen perusteella ja tunnistaa eleet tehokkaasti ulkoisista muuttujista riippumatta. Käytännössä huimaa vauhtia kehittynyt tekniikka on kuitenkin viime aikoina ajanut tutkimuksen ohi ja monet menetelmät on kehitetty nopeasti lähinnä vastaamaan käytännön tarpeisiin. Samalla ihmisen jäljittely -näkökulma on unohdettu. (Guyon et al., 2012a)

2.2 3D-videokuva

Kinect-sensori on Microsoftin kehittämä kaupallinen 3D-kamera. Se on tarkoitettu Microsoftin Xbox-pelikonsolin lisäosaksi. Kamera kehitettiin ensisijaisesti viihdekäyttöön parantamaan Xbox-pelien käyttökokemusta. Microsoft on kuitenkin avannut Kinectille ohjelmointirajapintoja, joiden avulla Kinect-kameralle on kehitetty lukuisia alkuperäisestä käyttötarkoituksesta irrallisia sovelluksia. (Microsoft, 2013)

Tavallisen värikuvan lisäksi Kinect-sensori tarjoaa syvyyskuvaa kohteesta. Syvyyskuva kertoo kohteen etäisyyden kamerasta ja luo näin kolmiulotteista videokuvaa. Microsoft tarjoaa Kinectille myös ohjelmistokehitystyökaluja, jotka sisältävät erilaisia hahmontunnistustyökaluja. Niiden avulla kehittäjä saa käyttöönsä esimerkiksi ranganseurauksen eli tiedon ihmishahmon asennoista videokuvan eri hetkillä. (Microsoft, 2013)
Myöhemmin esiteltävässä ChaLearn Gesture Challenge -kilpailussa kilpailijat eivät kuitenkaan hyödyntäneet Microsoftin tarjoamia valmiita hahmontunnistustyökaluja, vaan kilpailun tarkoitus oli kehittää omia menetelmiä.

Eleentunnistuksen näkökulmasta syvyyskuvalla on monia etuja verrattuna värikuvaan. Syvyyskuva on yksiväristä eli siitä on riisuttu erilaiset värit ja tekstuurit, jotka usein aiheuttavat ongelmia värikuvan tunnistamisessa. Syvyyskuvan värisävyt on pakotettu tietylle asteikolle, mikä helpottaa kuvien vertailtavuutta. (Shotton et al., 2011) Hahmo on helposti erotettavissa taustastaan ja eleet, jotka eroavat toisistaan ainoastaan syvyyssuuntaisen liikkeen perusteella, on mahdollista erottaa huomattavasti luotettavammin kuin pelkän värikuvan perusteella.

Kuvassa 1 on esimerkki Kinectin värikuvasta ja syvyyskuvasta. Kuvat ovat samasta tilanteesta. Syvyyskuvasta näkee hyvin syvyyskameran hyödyt. Hahmo on helposti erotettavissa taustasta ja hahmon käsi, joka on vartalon edessä voidaan helposti erottaa omaksi raajakseen.

Kuva 1: Esimerkki Kinectin syvyyskuvasta (vasemmalla) ja värikuvasta (oikealla). (Latotzky, 2011)

2.3 Eleentunnistus video- ja 3D-videokuvalta

Eleentunnistus videokuvalta noudattaa tavallisia hahmontunnistuksen vaiheita: esikäsittely, piirrevalinta ja luokittelu. Eleentunnistuksen erityishaasteet on huomioitava eri tavoin eri työvaiheissa. 3D-videokuva tuo oman lisänsä, mutta se ei merkittävästi muuta työvaiheita. Eleentunnistus on hahmontunnistuksen termein luokitusongelma eli mahdolliset luokat tunnetaan ennalta. (Guyon et al., 2012b)

Esikäsittelyvaiheessa videokuvalta poistetaan häiriötä, jotka voisivat haitata eleentunnistusta. Näitä voivat olla esimerkiksi kuvassa esiintyvät ylimääräiset objektit tai videokuvan virheet kuten kohina. Kuvaa voidaan myös pienentää tai pakata laskennan nopeuttami-

seksi. Esikäsittelyssä pyritään usein myös erottamaan ihmishahmo taustasta. Tämä on haastavaa, sillä ihmishahmo ei välttämättä erotu esimerkiksi väritykseltään taustasta. Kinectin syvyyskameran avulla taustan irrotus onnistuu kuitenkin luotettavammin kuin pelkän värikuvan avulla. Ihmishahmon erottaminen taustasta helpottaa tunnistusta, sillä tällöin ihminen ei sekoitu taustaansa tai mitään taustassa olevaa esinettä ei erehdytä pitämään ihmisen osana. Esikäsittelyssä voidaan suorittaa myös jonkinlaista ajallista jakoa videolle. Videokuvaa voidaan esimerkiksi jakaa ajallisiin jaksoihin perustuen videokuvan samankaltaisuuteen. Ajallisen jaon tarkoitus on auttaa hahmottamaan kuvalla tapahtuvaa liikesarjaa ja näin helpottaa tunnistusta. (Guyon et al., 2012b)

Hahmontunnistuksessa ratkaiseva vaihe on usein oikeiden piirteiden valinta eli piirreirrotus. Videokuvasta voidaan valita piirteeksi esimerkiksi tietyn suuntainen liike ajan funktiona. Liike näkyy peräkkäisten pysäytyskuvien välisenä erona. Tutkimalla liikettä videokuvat voidaan tiivistää liikekuviin, joita voidaan luokitella yksinkertaisilla luokittelualgoritmeillä. Videokuvaa voidaan tarkastella myös yksittäisten pysäytyskuvien kautta. Tällöin voidaan hyödyntää valokuvien tunnistuksessa käytettyjä menetelmiä. Pysäytyskuvista voidaan arvioida kontrastivaihteluita ja sitä kautta hahmottaa viivoja tai muotoja kuvassa. (Guyon et al., 2012b)

Tunnistusvaiheessa näytteitä verrataan opetusdatan kuvaamiin luokkiin. Tunnistusmenetelmä riippuu valituista piirteistä. Jos videokuvaa käsitellään kokonaisuutena esimerkiksi liikekuvan avulla, kuvia voidaan luokitella yksinkertaisilla luokittelualgoritmeillä. Näitä voisivat olla esimerkiksi k-lähimmän naapurin luokitin. Jos videokuva esitetään yksittäisillä pysäytyskuvilla, on tunnistuksessa huomioitava videon aikaulottuvuus. On käytettävä rakenteellista mallia, jonka avulla voidaan tarkastella piirteen arvoa tietyllä ajanhetkellä. Tähän tarkoitukseen on erilaisia graafisia malleja, joita esitellään tarkemmin luvussa kolme. (Guyon et al., 2012b)

Luokittelun jälkeen menetelmälle lasketaan virheprosentti. Virheprosentti lasketaan testidatan avulla. Testidatassa on opetusdatan tavoin tiedossa näytteiden oikeat luokat, jolloin on mahdollista laskea, kuinka suuri prosentti näytteistä on luokiteltu oikeisiin luokkiin. Menetelmää voidaan kehittää edelleen kokeilemalla erilaisia opetusdatajoukkoja ja valitsemalla joukko, joka tuottaa pienimmän virheprosentin testidatalle. (Guyon et al., 2012b)

3 ChaLearn Gesture Challenge -kilpailu

3.1 Kilpailun esittely

ChaLearn Gesture Challenge -kilpailun tarkoituksena oli lisätä kiinnostusta eletunnistukseen syvyyskameralla. Kilpailun järjesti ChaLearn-yhteisö. ChaLearn-yhteisö on useiden eri yliopistojen asiantuntijoista koostuva yhdistys, jonka tavoitteena on herättää kiinnostusta koneoppimiseen ja hahmontunnistukseen. (Guyon et al., 2011) Kilpailu alkoi vuoden 2011 lopuilla ja se päättyi loppuvuonna 2012. Kilpailuun osallistui ensimmäisellä kierroksella 50 ryhmää, pääasiallisesti yliopistojen tutkimusryhmiä. Kilpailijoille annettiin tietokanta, joka sisälsi 50 000 Kinect-sensorilla kuvattua videonäytettä. Videonäytteet sisälsivät yksittäisiä eleitä, esimerkiksi viittomia tai poliisin käsimerkkejä. Kilpailijoiden tarkoitus oli kehittää eleentunnistusmenetelmä, jonka avulla eleet opitaan yhdestä opetusnäytteestä. Eleitä oli jaettu kategorioihin käyttötilanteen mukaan. Esimerkiksi poliisin käsimerkit olivat yksi kategoria. (Guyon et al., 2012b)

Annetuilla videonäytteillä esiintyi aina yksi ihminen kerrallaan suorittamassa tiettyä elettä. Kuva rajattiin yläruumiiseen ja eleet tehtiin pääasiallisesti käsillä. Liikkeet lopetettiin ja aloitettiin aina samasta lepoasennosta. Videonäytteet sisälsivät syvyyskamerakuvan sekä värikuvan, mutta eivät ranganseurausta tai muuta valmista hahmontunnistustietoa. Haasteita toivat vaihtelevat taustat ja valaistukset videokuvalla. (Guyon et al., 2012b)

Kilpailijoille jaettiin kolme datajoukkoa: opetusdata, validointidata ja lopullinen arviointidata. Opetusdatan näytteille tarjottiin oikeat luokat, joiden avulla järjestelmän opetus onnistui. Sekä validointidatassa, että lopullisessa arviointidatassa jokaisesta eleestä annettiin ainoastaan yksi opetusnäyte eli näyte, jolle oli paljastettu oikea luokka. Kilpailun erityishaasteena olikin yhdestä otoksesta oppiminen (One Shot Learning). Tarkoituksena oli kehittää järjestelmä, joka oppii tunnistamaan eleet mahdollisimman pienestä määrästä opetusdataa. Kilpailijoiden odotettiin soveltavan tässä siirtovaikutusoppimista (Transfer learning). (Guyon et al., 2012b)

Siirtovaikutuksen ajatus on, että aiemmin opittuja tietoja hyödynnetään seuraavassa oppimistehtävässä. Tässä jäljitellään ihmisen oppimiskykyä. Ihminen oppii nopeasti tunnistamaan uusia hahmoja, jos hänellä on kokemusta vastaavista tehtävistä. Siirtovaikutuksessa tietoja siirretään edellisestä oppimisprosessista uuteen oppimistehtävään. Siirretyt tiedot saattavat olla esimerkiksi aiemmassa opetustehtävässä valitut piirteet tai jopa yksittäisiä datanäytteitä. (Pan ja Yang, 2010)

Järjestelmä, jonka avulla kilpailijat pystyivät testaamaan menetelmäänsä validointidataa vastaan, oli auki koko kehitysjakson ajan. Varsinainen testidata, joilla kilpailutöitä arvosteltiin paljastettiin vasta kilpailun lopuksi. Kilpailijoilla oli muutama päivä aika testata menetelmäänsä lopullista testidataa vastaan. Testidata sisälsi eri eleitä kuin opetusdata, mutta samoista kategorioista. Kilpailijoiden oli siis opetettava menetelmänsä uudelleen lopullisen testidatan avulla. Kilpailijoita pyydettiin lopuksi palauttamaan lista, joka sisälsi oikeat luokat testidatalle esitettynä merkkijonona. Lopullinen virheprosentti saatiin laskemalla Levensteinin etäisyys oikeita luokkia kuvaavan merkkijonon ja kilpailijoiden antaman vastausmerkkijonon välillä. (Guyon et al., 2012b)

3.2 Katsaus kilpailutöihin

3.2.1 Ensimmäinen kierros

Kilpailijoiden metodeja selvitettiin ensimmäisen kierroksen jälkeen lyhyellä kyselyllä, johon vastasi 20 ryhmää 22 parhaan ryhmän joukosta. Ryhmiltä kysyttiin muun muassa minkälaista esikäsittelyä he olivat tehneet videokuvalle, minkälaista tunnistusmenetelmää tai mitä piirteitä oli käytetty ja mikä oli heidän menetelmänsä suoritusaika. Kyselyn tarkoituksena oli saada yleiskatsaus kilpailutöihin, sillä monet kilpailijat eivät halunneet julkaista yksityiskohtaista kuvausta menetelmästään kilpailun ollessa vielä kesken. (Guyon et al., 2012b).

Vastauksista kävi ilmi, että lähes kaikki ryhmät tekivät jonkinlaista kuvan esikäsittelyä. Videokuvasta poistettiin häiriötä, asiaankuulumattomia kohteita tai ihmishahmon taustа. Huomioitavaa on kuitenkin, että jotkin hyvin menestyneistä ryhmistä eivät tehneet minkäänlaista esikäsittelyä kuvalle. (Guyon et al., 2012b)

Suurin osa osallistujista käytti HOG/HOF-piirteitä (Histogram of oriented Gradients/ Histogram of Flow), SIFT/STIP-piirteitä (Scale Invariant Feature Transformation/Space-time interest points), kulmien tai nurkkien tunnistusta tai kehitti omia, tälle datalle soveltuvia piirteitä.(Guyon et al., 2012b)

Käytetyt piirteet perustuvat pääosin kuvan värityksen intensiteettivaihteluun. Esimerkiksi HOG-piirteet kuvaavat kuvan intensiteettivaihtelun gradienttien suuntaa. Kuva jaetaan pieniin alueisiin, soluihin, joissa tarkastellaan alueen värityksen intensiteettivaihtelua. Soluille lasketaan intensiteettivaihtelun gradienttien suuntien histogrammi. Ajatuksena on päätellä, minkä suuntaisia viivoja tai nurkkia alueelta voidaan erottaa. (Dalal ja Triggs, 2005) Histogrammit kertovat kuvassa esiintyvistä muodoista, eivätkä ne ole riippuvaisia

hahmon sijainnista kuvassa tai kuvan yleisestä värimaailmasta. HOG-piirteet soveltuvatkin hyvin tämänkaltaiseen hahmontunnistusongelmaan, jossa kohteen sijainti ja väritys voivat vaihdella kuvalla. HOF-piirteet toimivat samoin kuin HOG-piirteet, paitsi yksittäisten kuvien sijaan ne tutkivat peräkkäisten kuvien vaihtelua eli liikettä videolla (Pers et al., 2010).

SIFT-piirteet toimivat HOG-piirteitä hienostuneemmin valiten kuvista tärkeät pisteet. Tärkeät pisteet valitaan niin, että ne ovat riippumattomia kuvan muutoksista kuten kiertämisestä tai skaalauksesta. Esimerkiksi kuvassa, jossa näkyy ovi, tärkeitä pisteitä voisivat olla oven kulmat. Vaikka kuvaa kierrettäisiin tai sen kokoa muutettaisiin, tärkeät pisteet eli kulmat voidaan löytää kuvasta. Pisteiden valinnassa hyödynnetään värityksen intensiteettivaihtelua ja tilastollisia menetelmiä. (Lowe, 1999) STIP-piirteet perustuvat samankaltaiseen menetelmään, mutta huomioivat myös videon aikaulottuvuuden (Laptev ja Lindeberg, 2003).

Piirteitä voidaan tutkia syvyys- ja värikuvasta. Syvyyskuvan etu verrattuna värikuvaan on, että siinä ei esiinny värejä tai tekstuureja, jotka häiritsisivät hahmon erottumista tai videoiden vertailua. Suurin osa kilpailijoista käyttikin töissään pelkkää syvyyskuvaa. Osa käytti sekä väri- että syvyyskuvaa. Mielenkiintoista on, että ensimmäisen kierroksen toisen sijan voittaja käytti työssään pelkkää värikuvaa. Kaikki kilpailijat käyttivät jonkinlaista piirteiden tiivistystä tai kuvausta toiseen lineaariavaruuteen. (Guyon et al., 2012b)

Ajallisen rakenteen mallintamiseen käytettiin erilaisia graafisia malleja kuten Markovin piilomuuttujaa ja Conditional Random Fields -menetelmää. Kaikki tunnistustusmenetelmät eivät kuitenkaan huomioineet videon ajallista rakennetta. (Guyon et al., 2012b)

Markovin muuttuja kuvaa havainnon sarjana tiloja eli tässä tapauksessa videon sarjana pysäytyskuvia. Tilat esitetään sopivien piirteiden avulla eli esimerkiksi kuvat voidaan esittää HOG/HOF-piirteiden avulla. Menetelmä kertoo kuinka todennäköisesti annettu havainto kuuluu tiettyyn luokkaan. Ajatuksena on, että annetun havainnon luokka on tuntematon, mutta se voidaan löytää sen etsimällä todennäköisin luokka. Luokat saadaan opetusdatasta.

Luokkien tiheysfunktiot lasketaan suurimman todennäköisyyden (Most Likelihood) -menetelmällä. Kyseessä on Bayesilainen menetelmä eli jakauma tunnetaan, mutta ei parametreja. Parametrien arvot optimoidaan niin, että todennäköisyys opetusliikkeelle kuulua

kuvaamaansa luokkaan on mahdollisimman suuri. Luokan tiheysfunktion avulla voidaan laskea kuinka todennäköisesti tietty tilasarja esiintyy tässä luokassa. Luokassa on kuitenkin useita mahdollisia tilasarjoja. Todennäköisyys annetulle havainnolle $O = O_1, O_2...O_n$ kuulua luokkaan λ saadaan siis seuraavasti:

$$P(O, Q|\lambda) = \sum_{kaikkiQ:t} P(O|Q, \lambda)P(Q|\lambda) \quad (1)$$

jossa $Q = Q_1, Q_2...Q_n$ eli Q on määrätyn mittainen tilasarja. Todennäköisyys havainnolle O esittää tiettyä tilasarjaa Q kerrotaan todennäköisyydellä $P(Q|\lambda)$ eli todennäköisyydellä tilasarjalle Q esiintyyä luokassa λ. Lopuksi summataan yhteen todennäköisyydet sarjalle O kuulua luokkaan λ kaikilla tilasarjoilla Q. Näin saadaan lopullinen todennäköisyys havainnolle O kuulua luokkaan λ. On huomioitava, että menetelmä vaatii kaikkien näytteiden olevan samanpituisia eli koostuvan tietystä ennalta määrätystä määrästä tiloja. Tämä saattaa asettaa rajoituksia menetelmän sovelluksissa.(Rabiner, 1989)

Conditional Random Fields -menetelmä perustuu samankaltaiseen intuitioon. Lähtökohta molemmissa on, että yksittäisen datapisteen sijaan luokitellaan datajoukkoja, joilla on sisäinen, tässä ajallinen, rakenne. (He et al., 2004)

Kilpailutöissä, joissa ei käytetty edellämainittuja ajallisen rakenteen mallintamista menetelmiä, luokittelussa käytettiin k-lähimmän naapurin luokitinta tai muita yksinkertaisia luokittelumenetelmiä.

Kilpailun järjestäjien odottamaa metodia, siirtovaikutusta käytettiin vähäisesti, eikä kukaan menestyneistä kilpailijoista käyttänyt sitä. Kilpailun varsinainen haaste, yhdestä eleestä oppiminen, jäi siis vähemmälle huomiolle. (Guyon et al., 2012b)

Kahdeksan menestyneintä työtä esitellään taulukossa 1. Taulukosta huomataan, että parhaiten menestyneiden töiden joukossa suurin osa käytti tunnistuksesa menetelmää, joka huomioi videon ajallisen rakenteen. Tällöin videokuvasta valitaan piirteet, joiden muutosta seurataan ajan funktiona. Poikkeuksen tekevät ryhmät Zonga ja Xiaozhuwudi, joiden menetelmä perustuu videon käsittelyyn erilaisten liikekuvien avulla. (Guyon et al., 2012b)

Taulukko 1: ChaLearn Gesture Challenge -kilpailun kahdeksan parhaiten sijoittunutta ryhmää

Ryhmän nimi	Menetelmä
Alfnie	Motion Signature analyses
Pennect	Markovin piilomallin tapainen menetelmä ja HOG/HOF-piirteet.
One Million Monkeys	Markovin piilomalli ja kulmien tunnistus
Immortals	Markovin piilomalli ja HOG/HOF-piirteet
Zonga	Pienimmänneliösumman menetelmä ja HOSVD -menetelmä
Balazs Godeny	Thumbnail Dynamic Time Warping" (DTW) ja HOG/HOF-piirteet sekä kulmien tunnistus.
SkyNet	Dynamic Time Warping(DTW) ja kulmien tunnistus
Xiaozhuwudi	MHI-kuva johon on lisätty GEI- ja INV-kuvat

3.2.2 Toinen kierros

Kilpailun toinen kierros toteutettiin samoilla järjestelyillä kuin ensimmäinen. Koska kierros loppui vasta tämän kandinaatintyön kirjoittamisen aikoihin, on se jätetty työssä vähemmälle tarkastelulle.

Toisen kierroksen jälkeen kilpailijoiden metodeja selvitettiin kyselyllä samoin kuin ensimmäisen kierroksen jälkeen. Kyselyyn vastasi 28 ryhmää. Vastausten perusteella toisella kierroksella menestyneet menetelmät olivat hyvin samantapaisia kuin ensimmäisen kierroksen menestyneet menetelmät. HOG/HOG-piirteet sekä muut kuvan intensiteettivaihteluihin perustuvat piirteet yhdistettynä Markovin piilomuuttujaan tai muuhun vastaavaan malliin olivat suosituin menetelmä. (Guyon et al., 2012a)

Molemmilla kierroksilla oli sama voittaja, ryhmä Alfnie. Voittajaryhmä väittää työnsä matkivan ihmisen hahmontunnistuskykyä. Työtä ei ole kuitenkaan vielä tämän kandinaatintyön kirjoittamisen aikaan julkaistu, joten siihen ei päästä tutustumaan tarkemmin.

Työ perustuu ryhmän ilmoituksen mukaan jonkinlaiseen mukautettuun Markovin piilomuuttujaan. Mielenkiintoista on, että huolimatta hyvästä luokittelukyvystään menetelmä oli myös yksi kilpailun nopeimmista. Vaikka menetelmä oli siis luokittelumielesssä tehokas, se ei kuitenkaan ollut laskennallisesti erityisen raskas. (Guyon et al., 2012a)

Toisen kierroksen toiseksi sijoittunut, ryhmä Turtle Tamers, käytti samankaltaista menetelmää kuin ensimmäisen kierroksen toisen sijan voittaja, ryhmä Pennect. Molemmat ryhmät käyttivät HOG/HOF-piirteitä sekä Markovin piilomuuttujaa. Kolmannen sijan saavuttanut ryhmä Joewan sen sijaan käytti hyvin erilaista menetelmää. Ryhmä käytti Bag of MOSIFT -piirteitä yhdistettynä lähimmän naapurin luokittimeen. (Guyon et al., 2012a) Bag of MOSIFT -piirteet on muokattu versio yleisestä Bag of features -menetelmästä. Bag of features -menetelmät kuvaavat tietyn piirteen avulla kuinka usein jokainen arvo esiintyy näytteessä. Ne mittaavat siis ainoastaan kuinka usein tietty arvo esiintyy näytteessä eivätkä välitä näytteen sisäisestä rakenteesta. (Nowak et al., 2006)

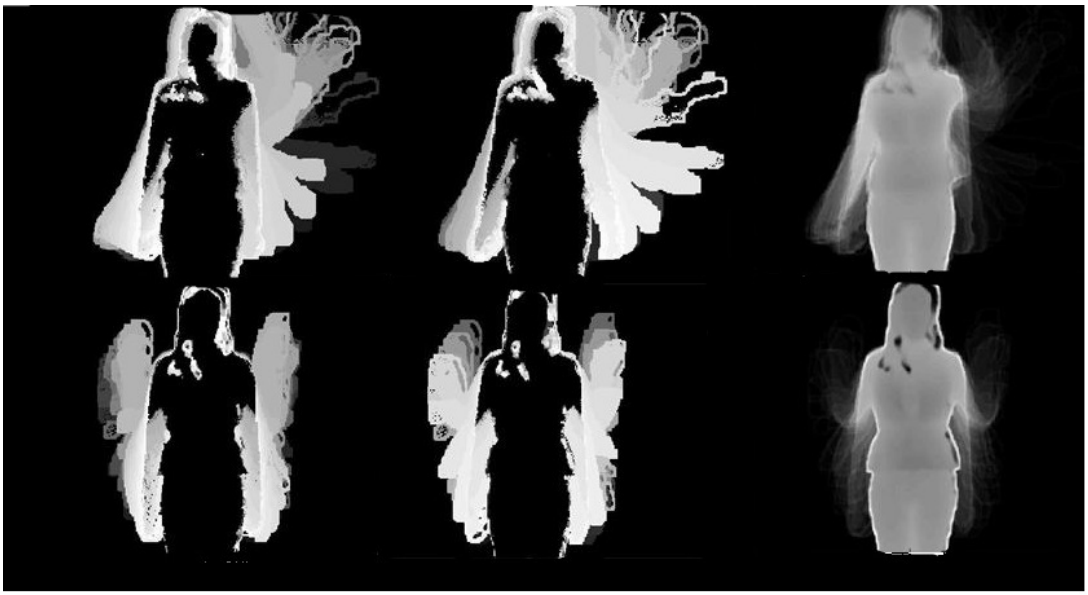
Varsinaisen validoinnin lisäksi kilpailutöille suoritettiin palautuksen jälkeen vielä yksi testaus. Tässä testauksessa tutkittiin kuinka hyvin kilpailijoiden menetelmät tunnistivat eleen, jos videokuvaa oli tietoisesti käännetty hieman. Oikeissa sovelluksissa on tärkeää, että ele pystytään tunnistamaan, vaikka se eroaisi hieman alkuperäisestä näytteestä esimerkiksi kuvakulmaltaan. Tässäkin testissä voittajaryhmä menestyi hyvin, kun taas esimerkiksi kaksi seuraavaa ryhmää menestyivät manipuloidulla datalla huomattavasti huonommin kuin varsinaisella kilpailudatalla. Tämä lisää ennestään mielenkiintoa voittajatyötä kohtaan. (Guyon et al., 2012a)

4 Katsaus eräisiin menestyneisiin kilpailutöihin

Tässä luvussa esitellään kolme menestyneistä töistä tarkemmin. Työt ovat eräitä esimerkkejä toimivista ratkaisuista. Ne on valittu tähän, koska ne edustavat erilaisia näkökulmia ongelmaan. Valintamahdollisuuksia rajoitti se, että kaikki kilpailijat eivät olleet vielä julkaisseet menetelmäänsä tämä työn kirjoittamisen aikana. Kolmesta valitusta työstä yksi, Immortals edustaa kilpailun yleislinjaa ja kaksi muuta valittua työtä, Zonga ja Xiaozhuwudi ovat esimerkkeinä omaperäisemmistä menetelmistä.

4.1 Ryhmä Xiaozhuwudi ja laajennettu MHI-menetelmä

Ryhmä Xiaozhuwudi lähti liikkeelle MHI eli Motion History Image -menetelmästä (Wu et al., 2012). MHI-kuva esitää liikkeen määrän videokuvalla. Videopätkä tiivistetään yhteen liikekuvaan, joka kuvaa liikkeen viimeaikaisuutta. Kohdat, joissa videokuvalla on ollut liikettä esitetään harmaasävyillä. Mitä viimeaikaisempaa liike on ollut sitä valkoisempana se näkyy kuvassa. Liikkumattomat alueet näkyvät täysin mustana. Videokuvalta tutkitaan siis vain liikettä, eikä pyritä esimerkiksi tunnistamaan kuvalla olevia kohteita tai ihmiskehon osia. Tämä menetelmä matkii ihmisen tapaa tunnistaa eleitä. Ihminen tunnistaa erittäin hyvin inhimilliset eleet sumealtakin videokuvalta vaikka ei yksittäisestä pysäytyskuvasta tunnistaisi edes ihmishahmoa. (Bobick ja Davis, 2001)



Kuva 2: Kuvassa vasemmalta oikealle MHI-, INV- ja MEI-kuva. (Wu et al., 2012)

Xiaozhuwudi-ryhmä tunnistaa MHI-kuvassa kuitenkin ongelmia. MHI-kuvan avulla on vaikeaa tunnistaa eleitä, jotka sisältävät toistuvaa liikkeitä, esimerkiksi vilkutusta. Liikkeen toistuessa MHI-kuva muuttuu helposti sekavaksi ja on vaikea erottaa tarkkaa liikerataa. Ryhmä ehdottaakin MHI-kuvan laajentamista INV(Inversed Recording)- ja GEI(Gait Energy Image)-kuvilla. INV-kuva on MHI-kuvalle käänteinen kuva. INV-kuvassa katsotaan videokuvaa alusta loppuun päin. Mitä aikaisemmin liike esiintyy videolla sitä vaaleampana se näytetään kuvassa. INV-kuvan avulla saadaan kuvaus videon alkutilanteesta, mikä täydentää MHI-kuvaa. GEI-kuva esittää paikallaan pysyviä kohteita. Mitä enemmän kohde on paikoillaan videolla sitä vaaleampana tämä alue näkyy kuvassa. Kohdissa joissa on liikettä, näkyy harmaa sävyä ja täysin paikallaan olevat kohteet ovat valkoisia. Tausta on erotettu kohteesta ja jätetty mustaksi. GEI-kuvan avulla

liikkeestä saadaan hyvä kokonaiskuva ja se on hyödyllinen etenkin toistuvan liikkeen tunnistuksessa. (Wu et al., 2012)

Kuvassa 2 on esitetty kahdelle liikkeelle MHI-, INV- ja GEI-kuvat. Kuvat havainnollistavat hyvin miten INV- ja GEI-kuvat täydentävät MHI-kuvaa. Pelkän MHI-kuvan perusteella on vaikea erottaa liikkeet toisistaan. INV- ja GEI-kuvien avulla liikkeet erottuvat kuitenkin selkeämmin.

Datan esikäsittelyssä Xiaozhuwudi hyödynsi Kinectin syvyyskuvaa poistamalla taustan ihmishahmolta. Lisäksi esikäsittelyssä poistettiin häiriöitä. MHI-, GEI- ja INV-kuville suoritettiin dimensioiden vähennys ja piirreirroitus. Eleiden tunnistukseen käytettiin Maximum Correlation Coeffient -luokittelijaa, joka perustuu kuvien väliseen korrelaatioon. (Wu et al., 2012)

Ryhmä väittää työnsä suoriutuvan eleentunnistusongelmasta huomattavasti nopeammin kuin muut paikalliseen tietoon perustuvat eleentunnistusmenetelmät (Guyon et al., 2012c). Lisäksi ryhmän esittämä luokittelualgoritmi vastaa ryhmän mukaan paremmin yhdestä eleestä oppimisen -haasteeseen kuin esimerkiksi Markovin piilomuuttujaan perustuvat menetelmät (Wu et al., 2012).

4.2 Ryhmä Immortals ja Markovin piilomuuttuja

Ryhmä Immortals esittää kilpailutyössään oletuksen, että ele koostuu ennen kaikkia useista yksittäisistä liikkeistä. Sen mukaan eleet tunnistetaan parhaiten käsittelemällä elettä sarjana liikkeitä. Tämä eroaa ryhmän mukaan tyypillisestä tavasta lähestyä ongelmaa. (Malgireddy et al., 2012)

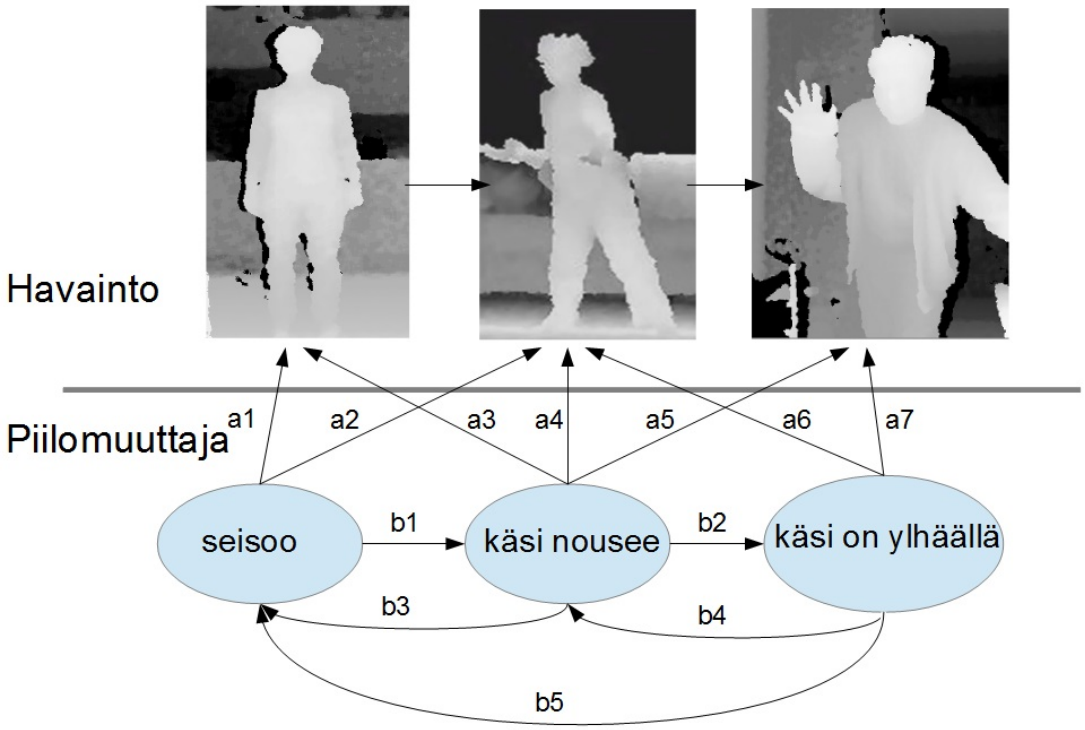
Ryhmä lähti liikkeelle opetusvaiheessa yksittäisistä liikkeistä. Yksittäisille liikkeille luodaan allekirjoitus eli malli, jonka avulla ne voidaan tunnistaa. Allekirjoituksen luominen on monivaiheinen operaatio. Ensin kuvista poimitaan niin sanotut tärkeät pisteet eli pisteet joilla on merkitystä liikkeen tunnistamisen kannalta. Tässä Immortals hyödynsi Kinectin syvyyskuvaa. Immortals arvioi, että ne kohdat kuvasta, joissa on tapahtunut syvyyssuuntaista muutosta syvyyskameran kuvassa ovat kyseisen videon pysäytyskuvan tärkeitä pisteitä. Tärkeille pisteille lasketaan HOG(Histogram of Oriented Gradients)- ja HOF(Histogram of Flow)-histogrammit. Tämän jälkeen kaikkien kuvien kaikki histogrammit ryhmitellään tavallisen ryhmittelyalgoritmin avulla. Histogrammeja kutsutaan ryhmittelyn jälkeen ”visuaalisiksi sanoiksi”. Yhdessä ryhmässä ovat kaikki tietyn sanan esiintymät. Tarkoituksena on tutkia visuaalisten sanojen esiintymistä yhdessä ja muo-

dostaa niistä aihepiirejä. Yksittäistä pysäytyskuvaa voidaan kuvata sillä, mitä sanoja ja mistä aihepiireistä siinä esiintyy. (Malgireddy et al., 2012)

Liikkeelle luodaan sanojen perusteella perusteella malli, jota käytetään tunnistusvaiheessa. Mallin perustana on Markovin piilomuuttuja eli HMM (Hidden Markov Model). Koska tutkitaan kahta piirrettä, HOG- ja HOF-piirrettä, käytetään monikanavaista Markovin piilomuuttujaa eli McHMM (Multi Channel Hidden Markov Model). HOG- ja HOF-piirteet paljastavat erilaista tietoa havainnosta ja tukevat tässä hyvin toisiaan. McHMM-muuttujan parametreja ovat alkutila, todennäköisyys tilojen väliselle siirtymälle sekä tilan todennäköisyys ja tilan kuvaus visuaalisten sanojen eli HOG- ja HOF-piirteiden avulla. Tilalla tarkoitetaan tässä yksittäistä pysäytyskuvaa. Malli opetetaan parametrien avulla niin, että se tunnistaa tietyn liikkeen eli tietyn sarjan pysäytyskuvia.(Malgireddy et al., 2012) Kuvassa 3 on tarkennettu vielä Markovin piilomuuttujan toimintaa. Kuvassa on esitetty havainto ja Markovin piilomuuttujan mahdolliset tilat, sekä niiden väliset siirtymätodennäköisyydet. Tarkoitus on löytää tilasarja, joka todennäköisimmin on muodostanut tämän havainnon. Parametrit eli tilat ja todennäköisyydet on opetettu mallien perusteella.

Tunnistusongelma pelkistyy lopulta kysymykseen: Mikä liikesarja kaikken todennäköisimmin on muodostanut tämän videonäytteen? Tämän tyylinen ongelma voidaan ratkaista Viterbin algoritmillä. Tämä vaatii kuitenkin, että ele rajataan koostumaan tietystä määrästä liikkeitä. Tässä tapauksessa on määritelty, että jokainen elenäyte sisältää viisi liikettä. Viterbin algoritmi pyrkii löytämään todennököismmän polun eri liikkeiden välillä. Algoritmi käy videota läpi liike kerrallaan ja laskee mikä on mallien perusteella todennäköisin liike. Lopuksi saadaan liikesarja, josta videonäyte todennäköisimmän koostuu. Liikesarja liitetään tunnistusvaiheessa tiettyyn eleeseen. On huomiotava, että Viterbin algoritmiä käytetään ryhmän menetelmässä kahdella tasolla. Menetelmän alimmalla tasolla algoritmiä käytetään Markovin piilomuuttujan kanssa tunnistamaan yksittäinen liike pysäytyskuvien perusteella. Toisaalta Viterbin algoritmiä käytetään myös ylemmällä tasolla tunnistamaan todennäköisin liikesarja eli ele liikkeiden perusteella. (Malgireddy et al., 2012)

Ryhmä Immortals kertoo lähestymistapansa olevan peräisin puheentunnistusmenetelmistä. Ryhmä on kokeillut menetelmää menestyksekkäästi myös muille videotietokannoilla, jotka sisälsivät pelkää värivideokuvaa. (Guyon et al., 2012c)



Kuva 3: Markovin piilomuuttujan käyttö eleentunnistuksessa. Viivan yläpuolella havainto on esitetty pysäytyskuvien avulla. Viivan alapuolella ovat mahdolliset tilat, sekä niiden väliset siirtymätodennäköisyydet b1-bn. Todennäköisyydet a1-an kuvaavat kuinka todennäköisesti tila esittää havaittua tilaa. Tarkoituksena on löytää tilasarja, joka kaikkein todennäköisimmin muodostaa havainnon. Kuvan esittämässä tilanteessa todennäköisin tilasarja olisi varmaankin: seisoo, käsi nousee, käsi on ylhäällä. On huomioitava, että oikessa tilanteessa tutkitaan täysin perättäisiä pysäytyskuvia, kun tässä esitetyt kuvat ovat selkeyden vuoksi liikkeen eri vaiheista.

4.3 Ryhmä Zonga ja pienimmän neliösumman menetelmä sovelluttuna monistoon

Ryhmä Zonga käyttää kehittämäänsä menetelmää, joka soveltuu yleisesti videokuvan luokitteluun. Menetelmää on hieman mukautettu eleentunnistusta varten, mutta lähtökohdiltaan se on hyvin matemaattinen eikä juuri hyödynnä perinteisiä kuvankäsittelymenetelmiä. (Lui, 2012a)

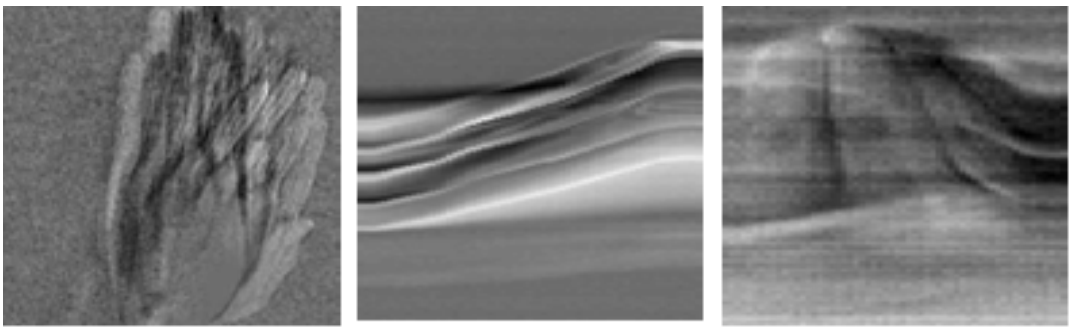
Videokuva on helppo mieltää kolmiulotteiseksi datajoukoksi. Videokuvan ulottuvuuksia ovat korkeus, leveys ja aika. Ryhmä Zonga käsitteleekin videokuvaa kolmiulotteisena tensorina. Tensorin ulottuvuudet vastaavat videon ulottuvuuksia. Voidaan ajatella, että yksi matriisi kuvaa yhtä pysäytyskuvaa, jolloin tensorin tuoma kolmas ulottuvuus on aikau-

lottuvuus. (Lui, 2012a)

Tensorin käsittely sellaisenaan on hankalaa sen suuren datamäärän vuoksi. Helpottaakseen videon käsittelyä ryhmä laskee tensorille HOSVD(Higher-order singular value decomposition)-hajotelman. Hajotelma on muokattu versio singulaarihajotelmasta. Hajotelma avaa tensorin kolmeksi matriisiksi. Matriisit kuvaavat videon vaakasuoraa liikettä, pystysuoraista liikettä ja summakuvan videon yli. (Chu et al., 2003) Tensori hajotetaan tekijöihinsä aputensorin (Core Tensor) S avulla:

$$A = S *_1 V^{(1)}_{appearance} *_2 V^{(2)}_{h-motion} *_3 V_{v-motion}*^{(3)} \tag{2}$$

jossa A on havaintomatriisi, S on aputensori ja V-matriisit ovat tensorin hajotelma. $V_{appereance}$-matriisi on videon summakuva ajan yli. $V_{h-motion}$-matriisi kuvaa vaakasuuntaista liikettä videolla ja $V_{v-motion}$-matriisi kuvaa pystysuoraa liikettä videolla. Kuvassa 4 esitetty hajotelman matriisit pikselikuvina. Kuvasta nähdään selkeästi, että matriisihajotelman tekijät esittävät liikkeen kolmena kuvana. (Lui, 2012a)



Kuva 4: Videon tensoriesitys on hajotettu HOSVD-hajotelman avulla kolmeen tekijään. Vasemmalta oikealle: summakuva koko videolle, videon vaakasuuntainen liike ja videon pystysuuntainen liike. (Lui, 2012a)

Matriisihajotelman avulla video voidaan kuvata pisteenä kolmiulotteisessa monistossa (manifold). Moniston ulottuvuudet vastaavat tensorihajotelman ulottuvuuksia. Monisto säilyttää videon alkuperäisen geometrisen rakenteen Euklidista avaruutta paremmin. (Lui, 2012b) Monistokuvauksen avulla videoita voidaan käsitellä yksittäisinä pisteinä, jolloin niiden luokittelu helpottuu. Monistojen käyttö videonkuvan kanssa ei ole uusi asia eleentunnistuksessa. Ryhmä Zonga kuitenkin yhdistää monistokuvaukseen pienimmän neliösumman menetelmän, joka tekee ryhmän mukaan heidän lähestymistavastaan ainutlaatuisen.(Lui, 2012a)

Pienimmän neliösumman menetelmä on regressio-ongelma, eli siinä etsitään jonkinlaista suhdetta havainnon ja luokan välille. Opetusvaiheessa tunnetaan havainto ja sen luokka, joiden välille pyritään löytämään funktio. Funktiota kutsutaan regressiofunktioksi. Tunnistusvaiheessa havaintojen luokat lasketaan regressiofunktion avaulla.(Rao ja Toutenburg, 1999)

Regressio-ongelma on muotoa $y = A * \beta$, jossa y-vektori esittää pisteet tulosavaruudessa, A-matriisi on havaintomatriisi ja β-vektori on painovektori, joka kuvaa havaintomatriisin pisteet tulospisteiksi. Opetusvaiheessa pyritään löytämään painovektori, joka kuvaa havainnon mahdollisimman lähelle oikeaa luokkaa tulosavaruudessa. (Lui, 2012a) Pienimmän neliösumman menetelmässä pyritään minimoimaan luokitteluvirheen neliötä eli oikean tuloksen ja arvioidun tuloksen erotuksen neliötä (Rao ja Toutenburg, 1999). Minimoidaan siis funktiota:

$$R(\beta) = ||y - A\beta||^2 \tag{3}$$

Painovektorin avulla muodostetaan regressiofunktio, jonka avulla havainnot kuvataan samaan tulosavaruuteen kuin opetuspisteet. Havainnot luokitellaan tämän jälkeen etäisyyden perusteella. Luokittelufunktio on muotoa:

$$j* = argmin_j D(Y, \psi_j(Y)) \tag{4}$$

jossa Y on annettu havainto, j on luokka ja ψ_j on luokan j regressiofunktio. D-funktio laskee annetun havainnon ja regression välisen erotuksen. Tarkoitus on löytää luokka, joka antaa pienimmän etäisyyden.

Ryhmä Zonga kertoo menetelmänsä olevan yksinkertainen ja helppo toteuttaa. Verrattuna muihin töihin se huomioi eyhmän mukaan erityisen hyvin eleen luonnollisen tila-avaruuden. (Guyon et al., 2012c)

4.4 Yhteenveto ryhmien kilpailutöistä

Ryhmä Xiaozhuwudi sijoittui kilpailussa kahdeksanneksi, ryhmä Immortals viidenneksi ja ryhmä Zonga kuudenneksi ensimmäisellä kierroksella viidenkymmenen kilpailijan joukosta. Vaikka valitut kilpailutyöt hyödynsivät varsin erilaisia menetelmiä ja panostivat eri vaiheisiin eleentunnistuksessa, niiden tuloksissa ei ollut huomattavaa eroa (Guyon et al., 2012b). Ryhmä Immortals edusti menetelmällään kilpailun yleistä suuntausta, kun taas ryhmät Zonga ja Xiaozhuwudi poikkesivat kilpailun valtavirrasta. Vaikka ryhmien Zonga ja Xiaozhuwudi menetelmät olivat keskenään kokonaisuudessaan melko erilaiset, ne muistuttivat kuitenkin toisiaan lähestymistavaltaan ja piirrevalinnan osalta.

Ryhmä Immortals kiinnitti menetelmässään erityistä huomioita piirreirroitukseen. Ryhmän esittämä piirreirroitus on monivaiheinen operaatio, jossa hyödynnettiin paitsi paikkatietoa HOG-piirteiden avulla, myös liiketietoa HOF-piirteiden avulla. Ajallisen rakenteen mallintamiseen ryhmä käyttää Markovin piilomuuttujaa. Kaksi muuta ryhmää, Zonga ja Xiaozhuwudi mallinsivat eleitä yksittäisten summakuvien avulla ohittaen videon ajallisen rakenteen. Molemmat ryhmät hyödynsivät tunnistuksessa ensisijaisesti liiketietoa. Summakuvat esittivät videolla tapahtunutta liikettä. Ryhmä Xiaozhuwudi ilmoittaa tehneensä summakuville vielä HOG-piirreirroituksen ja dimensioiden vähennyksen LDA-menetelmällä (Linear Diskriminant Analysis) (Wu et al., 2012). Ryhmä Zonga ei ilmoittanut muuta piirreirroitusta kuin HOSVD-hajotelman. Ryhmän Zonga menetelmä poikkeaakin piirreirroituksen osalta muista kilpailutöistä ja tyypillisestä lähestymistavasta kuvan- tai videokuvankäsittelyssä. Tämä näkyy myös videon esikäsittelyssä. Ryhmät Immortals ja Xiaozhuwudi esikäsittelivät videokuvaa eli muun muuassa irrottivat ihmishahmon taustastaan videokuvalla, mutta ryhmä Xiaozhuwudi ei tehnyt videokuvalle minkäänlaista esikäsittelyä (Guyon et al., 2012c).

Ryhmien käyttämät tunnistusmenetelmät heijastelivat piirrevalintoja. Ryhmät Xiaozhuwudi ja Zonga, jotka tiivistivät näytteet yksittäisiin kuviin käyttivät luokittelumenetelmiä, jotka perustuvat etäisyyden laskemiseen Euklidissa tai sen kaltaisessa tilassa. Ryhmä Immortals, joka lähti oletuksesta, että videokuva koostuu ennen kaikkea joukosta perättäisiä yksittäisiä liikkeitä käytti tunnistuksessa Markovin piilomuuttujaa ja Viterbin algoritmiä kahdella tasolla. Ryhmä Immortals jakoi videokuvaa ajallisesti Viterbin algoritmin käyttöä varten. Myös ryhmät Xiaozhuwudi ja Zonga ilmoittivat tehneensä videolle jonkinlaista ajallista jakoa luokittelua varten (Guyon et al., 2012c).

Esitellyistä töistä ryhmän Immortals työ vaikuttaa laskennallisesti raskammailta, mutta ryhmän ilmoituksen perustella laskenta-aika on vain lineaarinen suhteessa näytteiden määrään. Ryhmä Xiaozhuwudi ilmoitti saman lukeman. Ryhmä Zonga ilmoittaa laskenta-ajan olevan paitsi lineaarinen suhteessa näytteiden määrään myös neliöllinen suhteessa kuvan kokoon. (Guyon et al., 2012c) Tästä näkökulmasta ryhmän Zonga työ on suorituksista heikoin, sillä monet eleentunnistussovellukset vaativat nopeaa, lähes reaaliaikaista suoritusta.

Ryhmät Xiaozhuwudi ja Zonga ilmoittivat käyttäneensä työssään kilpailussa toivottua siirtovaikutusoppimista. Molemmat ryhmät ilmoittivat, että opetusdataa oli käytetty eleiden mallintamiseen. Ryhmät eivät kuitenkaan tarkemmin esitelleet tätä näkökulmaa menetelmiensä kuvauksissa. Ryhmä Immortals ei ilmoittanut käyttänäänsä siirtovaikutusta. (Guyon et al., 2012c)

Kaikki kolme ryhmää hyödynsivät tunnistuksessa sekä väri- että syvyyskuvaa. Kaikki kolme kuvattua menetelmää olivat kuitenkin sellasia, että ne soveltuisivat ainakin pienin muokkauksin myös pelkän värikuvan luokitteluun. Mikään menetelmistä ei perustunutt ihmishahmon tai sen osien tunnistukseen, vaan videokuvaa käsiteltiin pikselidatana, vailla informaatiota sen esittämästä kohteesta. Kaikki menetelmät sopisivat siis pienin muokkauksin myös muihin videokuvan luokitteluongelmiin kuin eleentunnistukseen. Ryhmien Xiaozhuwudi ja Zonga menetelmät perustuivat kuitenkin lähes yksinomaan liikkeen määrään videokuvalla, joten ne eivät välttämättä soveltuisi kaikkiin luokitteluongelmiin.

Taulukossa 2 on vielä esitetty tiivistetysti ryhmien Xiaozhuwudi, Immortals ja Zonga menetelmät. Taulukosta näkee, miltä osin menetelmät ovat samankaltaisia ja miltä osin ne eroavat toisistaan. Siinä missä ryhmät Zonga ja Xiaozhuwudi käyttivät piirrevalinnassa samaa lähtökohtaa (summakuvat ja liike), muistuttivat Immortals ja Xiaozhuwudi toisiaan enemmän piirreirroituksen (HOG/HOF-piirteet) ja kuvien käsittelyn osalta.

Taulukko 2: Vertailussa ryhmien Xiaozhuwudi, Immortals ja Zonga kilpailutyöt. (Guyon et al., 2012c)

	Xiaozhuwudi	Immortals	Zonga
Kuvan esikäsittely	Taustan poisto, melun poisto, värimaailman tasoittaminen	Taustan poisto	Ei esikäsittelyä
Piirreirroitus	HOG/HOF-piirteet	HOG/HOF-piirteet	HOSVD-hajotelma
Dimensioiden pienennys	Tekijöihin jako (LDA)	Datan ryhmittely	Tekijöihin jako (HOSV-menetelmällä)
Ajallinen jako	Perustuu kuvan eroon lepotilan välillä	Viterbin jako	Perustuu kuvan eroon lepotilan välillä
Eleen esitys	Summakuva	Joukko piirteitä	Kolmiuloitteinen tensori (josta hajotelma liikekuviin)
Luokittelu	Maksimikorrelaatioon perustuva luokittelija	Markovin piilomuuttuja	Lähimmän naapurin luokittelija
Siirto-oppiminen	Kehitysdataa käytetty eleiden mallintamiseen	Ei huomioitu	Kehitysdataa käytetty eleiden mallintamiseen
Suoritusaika	Lineaarinen näytteiden määrään	Lineaarinen näytteiden määrään	Neliöllinen kuvan kokooon, lineaarinen näytteiden määrään

5 Johtopäätökset

Kinect-kameran ja muiden vastaavien kameroiden 3D-videokuva on tuonut ratkaisuja eleentunnistuksen ongelmiin. Yksivärinen syvyyskuva pienentää erilaisista tekstuureista ja väreistä johtuvaa vaihtelua, jolloin luokkien sisäinen vaihtelu pienenee ja luokittelutehtävä helpottuu. Syvyyskuvan avulla saadaa lisätietoa eleestä, jolloin toisiaan muistuttavat eleet voidaan erottaa entistä varmemmin. Syvyyskamera on jo itsessään helpottanut eleentunnistusta ja saattaa antaa parempia tuloksia vanhoilla, jo tunnetuilla eleentunnistusmenetelmillä. Uusia menetelmiä pyritään kuitenkin kehittämään, jotta syvyyskameran saadut edut saataisiin entisestä paremmin hyödynnettyä.

ChaLearn Gesture Challenge -kilpailun kilpailutöiden perusteella 3D-videokuvan tunnistuksessa käytetäänkin pitkälti samoja menetelmiä kuin 2D-videokuvan tunnistuksessa. Kilpailutyöt eivät huomioineet 3D-kuvaa erityisesti piirrevalinnassa tai tunnistusmenetelmissä tai ainakaan tätä näkökulmaa ei tuotu erityisesti esille kilpailutöiden kuvauksissa. 3D-videokuvaa hyödynnettiin kilpailutöissä lähinnä esikäsittelyvaiheessa taustan irrottamiseen (lähes kaikki kilpailutyöt) ja ainakin yhdessä työssä (Immortals) tunnistuksen kannalta tärkeiden pisteiden valinnassa. Lähes kaikki kilpailijat toki käyttivät syvyyskuvaa, osa jopa pelkästään sitä, mutta eivät juurikaan kuvailleet miten heidän menetelmänsä olisi eronnut, jos käytössä olisi ollut pelkästään värikuva. Kuvaavaa on, että kilpailussa toiseksi tullut työ hyödynsi pelkkää värikuvaa.

Kilpailutöissä esitettyjä menetelmiä olivat ajallisen rakenteen mallintamiseen käytetyt graafiset mallit ja erilaiset liikekuvat. Piirteinä käytettiin kuvankäsittelystä tuttuja, kuvan intensiteettivaihteluihin perustuvia piirteitä. Piirteet valitiin niin, että ne ovat mahdollisimman vähän riippuvaisia värikuvan yleisestä värimaailmasta, hahmon sijainnista tai koosta kuvalla tai muista videokuvasta riippuvista seikoista. Esimerkiksi liikekuvat huomioivat vain videokuvalla tapahtuneen liikkeen, välittämättä kuvan värityksestä tai hahmon muodosta. Liikekuvalla sama liike näyttää melko samanlaiselta esittäjästä riippumatta. Liikekuvalle tai videokuvalle voidaan tehdä piirreirroitus esimerkiksi HOG-piirteiden avulla. Kun HOG-histogrammit luokitellaan Bag of features -esityksen avulla, katoaa tieto esimerkiksi hahmon sijainnista kuvalla. Bag of features -esityksessä säilyy vain tieto siitä minkä suuntaisia viivoja kuvassa esiintyy ja kuinka paljon. Tällöin Bag of features -esitys on sama riippumatta esimerkiksi siitä onko elettä tekevä hahmo kuvan oikeassa vai vasemmassa reunassa.

Suurin osa menestyneistä kilpailutöistä perustui ajallisen rakenteen mallintamiseen Markovin piilomuuttujan avulla yhdistettynä HOG/HOF-piirteisiin. Erityisiä perusteluita tälle ei kuitenkaan esitetty. Kilpailussa menestyi hyvin myös muutama työ, jotka käyttivät täysin tästä poikkeavia menetelmiä.

Olisi mielenkiintoista selvittää, kuinka hyvin kilpailutyöt pärjäisivät muulle kuin tässä kilpailussa annetulle datalle. Toisen kierroksen kilpailutöiden tuloksissa oli viitteitä siitä, että menetelmät olivat ”ylioppineet”tälle datalle (Guyon et al., 2012a). Silloin ne tuottavat hyviä tuloksia juuri tälle datalle, mutta eivät menestyisi yhtä hyvin muille datajoukoille. Olisi mielenkiintoista tutustua myös voittajatyöhön, joskaan sen käyttämät menetelmät eivät pinnallisen kuvauksen perusteella juuri eronneet kilpailun yleisestä suuntauksesta.

Yhteenvetona voisi todeta, että syvyyskamera on tuonut eleentunnistukseen uusia mah-

dollisuuksia, joita ei välttämättä osata vielä edes täysin hyödyntää. Tulevaisuuden haasteita on kehittää entistä varmempia ja nopeampia eleentunnistusmenetelmiä, jotka soveltuvat kuluttajasovelluksiin. Myös ChaLearn Gesture Challenge -kilpailun järjestäjien esittämä yhdestä eleestä oppimisen -ongelma on yksi haasteista, johon eleentunnistusmenetelmien on vastattava tulevaisuudessa. Alan kasvupotentiaali on huikea, sillä eleentunnistuksella on paljon käyttökohteita. Ehkä tulevaisuudessa elekäyttöliittymät ovat kosketuskäyttöliittymien tavoin arkipäivää. Lisätutkimusta on kuitenkin vielä tehtävä.

Lähteet

A.F. Bobick ja J.W. Davis. The recognition of human movement using temporal templates. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 23(3):257–267, 2001. ISSN 0162-8828. doi: 10.1109/34.910878.

Delin Chu, Lieven De Lathauwer ja Bart De Moor. A qr-type reduction for computing the svd of a general matrix product/quotient. *Numerische Mathematik*, 95:101–121, 2003. ISSN 0029-599X. doi: 10.1007/s00211-002-0431-z. URL `http://dx.doi.org/10.1007/s00211-002-0431-z(viitattu3/2013)`.

N. Dalal ja B. Triggs. Histograms of oriented gradients for human detection. *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, osa 1, sivut 886–893 vol. 1, 2005. doi: 10.1109/CVPR.2005.177.

I. Guyon, V. Athitsos, P. Jangyodsuk, B. Hamner ja H.J. Escalante. Kilpailun nettisivut. *Kilpailun nettisivut*, 2011. URL `http://gesture.chalearn.org/2011-one-shot-learning(viitattu04/2013)`.

I. Guyon, V. Athitsos, H. Jangyodsuk, P. Escalante ja B Hamner. Results and analysis of the chalearn gesture challenge 2012. *Results and Analysis of the ChaLearn Gesture Challenge 2012*, sivut 1–17, 2012a. URL `http://eprints.pascal-network.org/archive/00009716/(viitattu3/2013)`.

I. Guyon, V. Athitsos, P. Jangyodsuk, B. Hamner ja H.J. Escalante. Chalearn gesture challenge: Design and first results. *Computer Vision and Pattern Recognition Workshops (CVPRW), 2012 IEEE Computer Society Conference on*, sivut 1–6, 2012b. doi: 10.1109/CVPRW.2012.6239178.

I. Guyon, V. Athitsos, P. Jangyodsuk, B. Hamner ja H.J. Escalante. Kysely ensimmäisen kierroksen töistä. *Kysely ensimmäisen kierroksen töistä*, 2012c. URL `https://docs.google.com/a/chalearn.org/viewer?a=v&pid=sites&srcid=Y2hhbGVhcm4ub3JnfGdlc3R1cmVjaGFsbGVuZ2V8Z3g6MmYyY2YwMmUwODY2NmE0YQ(viitattu04/2013)`.

Xuming He, R.S. Zemel ja M.A. Carreira-Perpindn. Multiscale conditional random fields for image labeling. *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*, osa 2, sivut II–695–II–702 Vol.2, 2004. doi: 10.1109/CVPR.2004.1315232.

I. Laptev ja T. Lindeberg. Space-time interest points. *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*, sivut 432–439 vol.1, 2003. doi: 10.1109/ICCV.2003.1238378.

I. Laptev, M. Marszalek, C. Schmid ja B. Rozenfeld. Learning realistic human actions from movies. *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, sivut 1–8, 2008. doi: 10.1109/CVPR.2008.4587756.

David Latotzky. Intelligent wheelchair research group, freie universität berlin. *Intelligent Wheelchair Research Group, Freie Universität Berlin*, 2011. URL `http://userpage.fu-berlin.de/~latotzky/wheelchair/(viitattu3/2013)`.

D.G. Lowe. Object recognition from local scale-invariant features. *Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference on*, osa 2, sivut 1150–1157 vol.2, 1999. doi: 10.1109/ICCV.1999.790410.

Yui Man Lui. A least squares regression framework on manifolds and its application to gesture recognition. *Computer Vision and Pattern Recognition Workshops (CVPRW), 2012 IEEE Computer Society Conference on*, sivut 13–18, 2012a. doi: 10.1109/CVPRW.2012.6239180.

Yui Man Lui. Advances in matrix manifolds for computer vision. *Image and Vision Computing*, 30(6–7):380 – 388, 2012b. ISSN 0262-8856. doi: 10.1016/j.imavis.2011.08.002. URL `http://www.sciencedirect.com/science/article/pii/S0262885611000692(viitattu3/2013)`.

M.R. Malgireddy, I. Inwogu ja V. Govindaraju. A temporal bayesian model for classifying, detecting and localizing activities in video sequences. *Computer Vision and Pattern Recognition Workshops (CVPRW), 2012 IEEE Computer Society Conference on*, sivut 43–48, 2012. doi: 10.1109/CVPRW.2012.6239185.

Microsoft. Microsoft:n viralliset sivut kinect-sensorillle. *Microsoft:n viralliset sivut Kinect-sensorillle*, 2013. URL `http://www.microsoft.com/en-us/kinectforwindows/(viitattu3/2013)`.

Eric Nowak, Frédéric Jurie ja Bill Triggs. Sampling strategies for bag-of-features image classification. Teoksessa *Computer Vision – ECCV 2006*, Aleš Leonardis, Horst Bischof ja Axel Pinz, toimittajat, osa 3954 sarjasta *Lecture Notes in Computer Science*, sivut 490–503. Springer Berlin Heidelberg, 2006. ISBN 978-3-540-33838-3. doi: 10.1007/11744085_38. URL `http://dx.doi.org/10.1007/11744085_38`.

Sinno Jialin Pan ja Qiang Yang. A survey on transfer learning. *Knowledge and Data Engineering, IEEE Transactions on*, 22(10):1345–1359, 2010. ISSN 1041-4347. doi: 10.1109/TKDE.2009.191.

Janez Pers, Vildana Sulic, Matej Kristan, Matej Pers, Klemen Polanec ja Stanislav Kovacic. Histograms of optical flow for efficient representation of body motion. *Pattern Recognition Letters*, 31(11):1369 – 1376, 2010. ISSN 0167-8655. doi: 10.1016/

j.patrec.2010.03.024. URL http://www.sciencedirect.com/science/article/pii/S0167865510001121.

L. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989. ISSN 0018-9219. doi: 10.1109/5.18626.

C. Radhakrishna Rao ja Helge Toutenburg. Linear models: Least squares and alternatives, second edition. *Linear Models: Least Squares and Alternatives, Second Edition*. Springer-Verlag New York, Inc, 1999. ISBN 0-387-98848-3 (painettu).

J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman ja A. Blake. Real-time human pose recognition in parts from single depth images. *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, sivut 1297–1304, 2011. doi: 10.1109/CVPR.2011.5995316.

Liang Wang, Tieniu Tan, Huazhong Ning ja Weiming Hu. Silhouette analysis-based gait recognition for human identification. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 25(12):1505–1518, 2003. ISSN 0162-8828. doi: 10.1109/TPAMI.2003.1251144.

Di Wu, Fan Zhu ja Ling Shao. One shot learning gesture recognition from rgbd images. *Computer Vision and Pattern Recognition Workshops (CVPRW), 2012 IEEE Computer Society Conference on*, sivut 7–12, 2012. doi: 10.1109/CVPRW.2012.6239179.