Aalto-yliopisto
Perustieteiden korkeakoulu
Tietotekniikan koulutusohjelma

Eleentunnistus Kinect-sensorilla

Kandidaatintyö

28. marraskuuta 2011

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:	28. marraskuuta 2011
Sivumäärä:	Kirjoita tähän oikea määrä, tässä esimerkissä 23
Pääaine:	Tähän sinun pääaineesi nimi, kts main.tex
Koodi:	Txxxx tai ILyyyy
Vastuuopettaja:	Ma professori Tomi Janhunen
Työn ohjaaja(t):	Ohjaajantitteli Sinun Ohjaajasi (Poimi tähän ohjaajasi laitos, DEPT, main.tex)

Tiivistelmä on muusta työstä täysin irrallinen teksti, joka kirjoitetaan tiivistelmälomakkeelle vasta, kun koko työ on valmis. Se on suppea ja itsenäinen teksti, joka kuvaa olennaisen opinnäytteen sisällöstä. Tavoitteena selvittää työn merkitys lukijalle ja antaa yleiskuva työstä. Tiivistelmä markkinoi työtäsi potentiaalisille lukijoille, siksi tutkimusongelman ja tärkeimmät tulokset kannattaa kertoa selkeästi ja napakasti. Tiivistelmä kirjoitetaan hieman yleistajuisemmin kuin itse työ, koska teksti palvelee tiedonvälitystarkoituksessa laajaa yleisöä.

Tiivistelmän rakenne: teksti jäsennetään kappaleisiin (3–5 kappaletta); ei väliotsikkoja; ei mitään työn ulkopuolelta; ei tekstiviitteitä tai lainauksia; vähän tai ei ollenkaan viittauksia työhön (ei ollenkaan: “luvussa 3” tms., mutta koko työhön voi viitata esim. sanalla “kandidaatintyössä”; ei kuvia ja taulukoita.

Tiivistelmässä otetaan “löysät pois”: ei työn rakenteen esittelyä; ei itsestäänselvyyksiä; ei turhaa toistoa; älä jätä lukijaa nälkäiseksi, eli kerro asiasisältö, älä vihjaa, että työssä kerrotaan se.

Tiivistelmän tyypillinen rakenne: (1) aihe, tavoite ja rajaus (heti alkuun, selkeästi ja napakasti, ei johdattelua); (2) aineisto ja menetelmät (erittäin lyhyesti); (3) tulokset (tälle enemmän painoarvoa); (4) johtopäätökset (tälle enemmän painoarvoa).

Avainsanat:	avain, sanoja, niitäkin, tähän, vielä, useampi, vaikkei, niitä, niin, montaa, oikeasti, tarvitse
Kieli:	Suomi

Sisältö

1 Johdanto **4**

2 Eleentunnistus videokuvalta **5**

2.1 Eleentunnistus ongelmana 5

2.2 3D-videokuva 6

2.3 Eleentunnistus video- ja 3D-videokuvalta 7

3 ChaLearn Gesture Challenge -kilpailu **9**

3.1 Kilpailun esittely 9

3.2 Katsaus kilpailutöihin 10

3.2.1 Ensimmäinen kierros 10

3.2.2 Toinen kierros 13

4 Katsaus menestyneisiin kilpailutöihin **14**

4.1 Xiaozhuwudi ja laajennettu MHI-menetelmä 14

4.2 Immortals ja Markovin piilomalli 15

4.3 Zonga ja pieninimmän neliösumman menetelmä sovelluttuna monistoon . 16

4.4 Yhteenveto menestyneistä kilpailutöistä 18

5 Johtopäätökset **18**

6 Loppuluku **20**

Lähteet **21**

1 Johdanto

Tämä kandinaatintyö käsittelee eleentunnistusta Kinect-syvyyskameran avulla. Työn tarkoitus on tutustua erilaisiin eleentunnistusmenetelmiin ja erityisesti Chalearn Gesture Challenge -kilpailun kilpailutöihin.

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 erilaisissa elekäyttöliittymissä. Toistaiseksi eleentunnistusmenetelmät eivät ole olleet riittävän luotettavia ja nopeita, jotta niitä olisi voitu laajasti hyödyntää kuluttujasovelluksissa.

Kehittynyt tekniikka kuten Microsoftin Kinect-sensori ja kehittynyt laskentateho tuovat alalle uusia mahdollisuuksia. Kinect-sensori on Microsoftin kehittämä 3D-kamera, joka tarjoaa ns. 3D-videokuvaa eli tavallisen värikuvan lisäksi Kinect-kamera antaa infrapunakameralla mitattua syvyyskuvaa. 3D-videokuvan avulla eleitä voidaan tunnistaa 2D-videokuvaa luotettavammin. 3D-kuva ei ole kuitenkaan vielä toistaiseksi radikaalisti kehittänyt tai muuttanut olemassa olevia eleentunnistusmenetelmiä.

Lisätäkseen kiinnostusta 3D-kuvaan järjestettiin ChaLearn Gesture Challenge -eleentunnistuskilpailu, joissa kilpailijat kehittivät Kinect-sensorin datalle suunnattuja eleentunnistusmenetelmiä. Tämä kandinaatin työ 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 datajoukkoille. 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). Tämä näkökulma näkyi kuitenkin kilpailutöissä vähän. Kilpailutöiden menetelmät olivat pääasiallisesti sellaisia, että ne soveltuisivat myös tavalliselle 2D-videokuvalle. Menetelmät yhdistelivät tunnettuja menetelmiä kuvan, videon ja äänentunnistuksesta.

Luvussa kaksi esitellään vielä tarkemmin eleentunnistusongelmaa ja 3D-videokuvaa. Luvussa kolme kerrotaan ChaLearn Gesture Challenge -kilpailusta ja luodaan yleiskatsaus

kilpailutöihin. Luvussa neljä esitellään tarkemmin kolme kilpailutöistä.

Työ pyrkii kartoittamaan minkälaisia keinoja 3D-videokuvan tunnistuksessa voidaan käyttää ja mitä uutta 3D-kuva tuo verrattuna 2D-kuvaan. Työssä esitellään jonkinverran hahmontunnistuken 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., Dec.). 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 tehokkaammiin (Guyon et al., June).

Eleentunnistus on hyvä erottaa asennontunnistus(Pose Estimation)-ongelmasta. Asennontunnistuksessa pyritään tunnistamaan ihmisen asento videolla yhdessä pysäytyskuvassa käyttämättä lainkaan ajallista tietoa. Tunnistuksessa pyritään usein tunnistamaan ihmishahmon ruumiinosat esimerkiksi nivelet, joiden pohjalta arvioidaan asento. 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. (Shotton et al., June)

Kiinnostus eleentunnistusta kohtaan on lisääntynyt viime vuosina sen monien käyttötarkoitusten vuoksi. Eleentunnistusta voidaan käyttää monenlaisissa elekäyttöliittymissä. Yksittäisiä eleitä voidaan käyttää esimerkiksi kodinkoneiden ohjailuun. (Wang et al., Dec.) Toisaalta eleentunnistusta voidaan käyttää hyödyksi tunnistamaan erilaisia vaarati-

lanteita. Esimerkiksi potilaan tilaa voidaan seurata eleentunnistuskameralla mahdollisten poikkeavien eleiden varalta.(Guyon et al., 2012)

Videokuvan tunnistuksessa voidaan hyödyntää perinteisiä hahmontunnistusmenetelmiä. Monet menetelmistä ovat kuitenkin laskennallisti liian raskaita reaaliaikaseen videokuvan tunnistukseen, jota vaaditaan elekäyttöliittymissä. (Wang et al., Dec.) Monet hahmontunnistusmenetelmät vaativat myös paljon opetusdataa. Kuluttajille suunnatuissa sovelluksissa olisi toivottavaa, että uuden eleen voi opettaa muutaman testinäytteen perusteella (Wang et al., Dec.). 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., 2012)

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 kolmiuloista videokuvaa. Microsoft tarjoaa Kinectille ohjelmistokehitystyökaluja, jotka sisältävät erilaisa 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 ai-

heuttavat ongelmia värikuvan tunnistamisessa. Syvyyskuvan värisävyt on pakotettu tietylle asteikolle, mikä helpottaa kuvien vertailtavuutta. (Shotton et al., June) Hahmo on helposti erotettavissa taustastaan ja eleet, jotka eroavat toisistaan ainostaan syvyysuuntaisen 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 luokittely. Eleentunnistuksen erityishaasteet on huomioitava erityisesti piirrevalinnassa. 3D-videokuva tuo oman lisänsä, mutta se ei merkittävästi muuta työvaiheita. Eleentunnistus on hahmontunnistuksen termein luokitusongelma eli mahdolliset luokat tunnetaan ennalta. (Guyon et al., June)

Esikäsittely aiheessa 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 nopeuttamiseksi. Esikäsittelyssä pyritään usein myös erottamaan ihmishahmo taustasta. Tämä on haastavaa sillä ihmishahmo ei välttämättä erotu esimerkiksi väritykseltään taustas-

ta. Kinectin syvyyskameran avulla tämä 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 tai tiivistystä. Videokuvaa voidaan esimerkiksi jakaa ajallisiin jaksoihin perustuen videokuvan samankaltaisuuteen. Ajallisen jaon tarkoitus on auttaa hahmottamaan kuvalla tapahtuvaa liikesarjaa ja helpottaa tunnistusta.(Guyon et al., June)

Hahmontunnistuksessa ratkaiseva vaihe on usein oikeiden piirteiden valinta eli piirreirroitus. 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 luokittelualkoritmeillä. Videokuvaa voidaan tarkastella myös yksittäisten pysäytyskuvien kautta. Tällöin voidaan hyödyntää valokuvien tunnistustuksessa käytettyjä menetelmiä. Pysäytyskuvista voidaan arvioida kontrastivaihteluita ja sitä kautta hahmottaa viivoja tai muotoja kuvassa. (Guyon et al., June)

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. Tässä voitaisiin käyttää esimerkiksi Markovin piilomuuttujaa. (Guyon et al., June)

Luokittelun jälkeen menetelmälle lasketaan virheprosentti. Virheprosentti lasketaaan testidatan avulla. Testidatassa on opetusdatan tavoin annettu 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., June)

3 ChaLearn Gesture Challenge -kilpailu

3.1 Kilpailun esittely

ChaLearn Gesture Challenge -kilpailun tarkoituksena oli lisätä kiinnostusta eletunnistukseen syvyyskameralla. Kilpailu alkoi vuoden 2011 aikana ja se päättyi loppuvuonna 2012. Kilpailussa tarjottiin 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. Esimerksi poliisin käsimerkit olivat yksi kategoria. (Guyon et al., June)

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 muita valmista hahmontunnistietoa. Haasteita toivat vaihtelevat taustat ja valaistukset videokuvalla. (Guyon et al., June)

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 siirtovaikutus-oppimista (Transfer learning). (Guyon et al., June) Siirtovaikutuksen ajatus on, että aiemmin opittuja tietoja hyödynnetään seuraavassa oppimistehtävässä. Tässä jäljitellään ihmisen oppimiskykyä. Ihminen oppii nopeammin tunnistamaan uusia hahmoja, jos hänellä on kokemusta vastaavista tehtävistä. 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 tehtiin laskemal-

la Levensteinin etäisyys oikeita luokkia kuvaavan merkkijonon ja kilpailijoiden antaman vastausmerkkijonon välillä. (Guyon et al., June)

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 kilpalun ollessa vielä kesken. (Guyon et al., June).

Vastauksista kävi ilmi, että lähes kaikki ryhmät tekivät jonkinlaista kuvan esikäsittelyä. Videokuvasta poistetiin häiriötä, asiaanliittymättömiä kohteita tai ihmishahmon tausta. Huomioitavaa on kuitenkin, että jotkin hyvin menestyneistä ryhmistä eivät tehneet minkäänlaista esikäsittelyä kuvalle. (Guyon et al., June)

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

Käytetyt piirteet perustuvat pääosin kuvan värityksen intensiteettivaihteluun. HOG-piirteet kuvaavat kuvan intensitettivaihtelun 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, June) 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.

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 voi-

sivat olla oven kulmat. Vaikka kuvaa kierrettäisiin tai sen kokoa muutettaisiin, tärkeät pisteet eli kulmat voidaan löytää kuvasta. Pisteiden valinnassa hyödynnetään värityksen intensiteettivaihtelua ja tilastolllisia menetelmiä. (Lowe)

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 jonkilaista piirteiden tiivistystä tai kuvausta toiseen lineaariavaruuteen. (Guyon et al., June)

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

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 Bayesialainen 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:

$$P(O, Q|\lambda) = \sum_{allQ} P(O|Q, \lambda)P(Q|\lambda) \tag{1}$$

jossa $Q = Q_1, Q_2...Q_n$ eli Q on määrätyn mittainen tilasarja. Todennäköisyys havainnolle O esittää tiettyä tilasarjaa Q kerrotaan todenäköisyydellä $P(Q|\lambda)$ eli todenäköisyydellä 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. (Rabiner, 1989)

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

Itse luokittelu tapahtui k-lähimmän naapurin luokittimella tai muilla yksinkertaisilla luokittelumenetelmillä. Kilpailun järjestäjien odottamaa metodia Transfer learning -metodia käytettiin vähäisesti, eikä kukaan menestyneistä kilpailijoista käyttänyt sitä. Kilpailun varsinainen haaste, yhdestä eleestä oppiminen, jäi siis vähemmälle huomiolle. (Guyon et al., June)

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

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 tunnisus
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

Luvussa kaksi 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.

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.

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 kustomoituun Markovin piilomuuttujaan. Mielenkiintoista on, että huolimatta hyvästä sijoituksestaan meneltemä oli myös yksi kilpailun nopeimmista.

Toisen sijan voittaja toisella kierroksella, ryhmä Turtle Tamers, käytti samankaltaista menetelmää kuin ensimmäisen kierroksen toisen sijan voittaja, ryhmä Pennect. Molemmat ryhmät käyttivät HOG/HOF-piirteitä sekä Markovin piilomallia. Kolmannen sijan saavuttanut ryhmä Joewan sen sijaan käytti hyvin erilaista menetelmää. Ryhmä käytti Bag of MOSIFT - piirteitä yhdistettynä lähimmän naaprin luokittimeen. Bag of MOSIFT -piirteet on kustoimoitu versio yleisestä Bag of features -menetelmästä. Bag of features -menetelmät kuvaavat kuinka usein tietty arvo esiintyy näytteessä. Ne hajottavat näytteen sisäisen rakenteen ja eroavat siis Markovin piilomallista, joka säilyttää näytteen sisäisen rakenteen ().

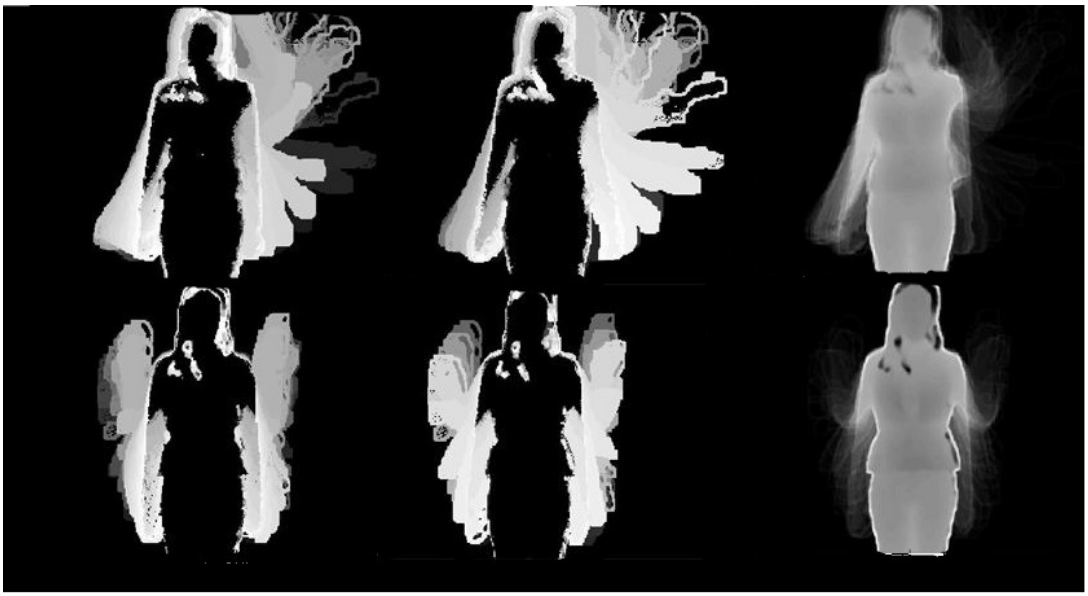
Varsinaisen validoinnin lisäksi kilpailutöille suoritettiin vielä yksi testaus. Tässä testauksessa tutkittiin kuinka hyvin kilpailijoiden menetelmät tunnistivat eleen, jos videokuvaa oli tietoisesti käännetty hieman. Oikeissa sovelluksissa on tärkeää, että ele pystytään tunnistamaan, vaikka se eroaisi hieman alkuperäisestä näytteestä esimerkiksi kuvakulmal-

taan. Tässäkin testissä voittajaryhmä pärjäsi hyvin, kun taas esimerkiksi kaksi seuraavaa ryhmää pärjäsit manipuloidulla datalla huomattavasti huonommin kuin varsinaisella kilpailudatalla. Tämä lisää ennestään mielenkiintoa voittajatyötä kohtaan.

4 Katsaus menestyneisiin kilpailutöihin

4.1 Xiaozhuwudi ja laajennettu MHI-menetelmä

Ryhmä xiaozhuwudi lähti liikkeelle MHI eli Motion History Image -tekniikasta. (Wu et al., June) MHI tutkii liikkeen määrää videokuvalla. Videopätkä tiivistetään yhteen liikekuvaan, joka kuvaa liikkeen viimeaikaisuutta. Kohdat, joissa videokuvalla on ollut liikettä esitetään harmaasävyillä. Mitä viimeaikasempaa 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 framesta tunnistaisi edes ihmishahmoa. (Bobick ja Davis, Mar)



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

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 liikeratataa. Xiaozhuwudi ehdottaakin MHI-kuvan laajentamista INV- ja GEI-kuvilla. INV-kuva

on MHI-kuvalle käänteinen kuva. INV-kuvassa katsotaan videokuvaa alusta loppuun päin. Mitä aikasemmin liike esiintyy videolla sitä vaalempana se näytetään kuvassa. INV:n avulla saadaan kuvaus videon alkutilanteesta, mikä täydentää MHI kuvaa. GEI-kuva esittäää liikkeen määrää keskimäärin koko videon aikana. Siinä summataan koko videon liike yhdelle kuvalle ja jaetaan lopuksi koko aikavälille. GEI muistuttaa MEI:tä (Motion Energy Image), jossa myös lasketaan liikkeelle summa. Voidaan ajatella, että siinä missä MHI- ja INV-kuvat kuvaavat liikettä , MEI- ja GEI-kuvat mittaavat energiaa, joka on kulunut liikkeeseen. GEI-kuvan avulla liikkeestä saadaan hyvä kokonaiskuva ja se on hyödyllinen etenkin toistuvan liikkeen tunnistuksessa. (Wu et al., June)

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. Eleiden tunnistukseen käytettiin Maximum Correlation Coeffient -luokittelijaa. (Wu et al., June)

4.2 Immortals ja Markovin piilomalli

Ryhmä Immortals esittää kilpailutyössään oletuksen, että ele koostuu ennen kaikkia useista yksittäisistä liikkeistä. Sen mukaan eleet tunnistetaan parhaiten käsittelmällä elettä sarjana liikkeitä. Tämä eroaa ryhmän mukaan tyypillisestä tavasta lähestyä ongelmaa.

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). 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 muodostaa niistä aihepiirejä.

Yksittäistä pysäytyskuvaa voidaan kuvata sillä, mitä sanoja ja mistä aihepiireistä siinä esiintyy.

Liikkeelle luodaan sanojen perusteella perusteella malli, jota käytetään tunnistusvaiheessa. Mallin perustana on Markovin muuttuja eli HMM (Hidden Markov Model). Koska tutkitaan kahta piirrettä, HOG- ja HOF-piirrettä, käytetään monikanavaista Markovin piilomallia eli McHMM. HOG- ja HOF-piirteet paljastavat erilaista tietoa havainnosta ja tukevat tässä hyvin toisiaan. McHMM parametreja ovat alkutila, todennäköisyys tilojen väliselle muutokselle sekä tilan todennäköisyys ja tilan kuvaus HOG ja HOF-piirteiden avulla. Tilalla tarkoitetaan tässä yksittäistä pysäytyskuvaa. Malli opetetaan parametrien avulla niin, että se tunnistaa tietyn liikkeen.

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. Kilpailussa on määritelty, että jokainen elenäyte sisältää viisi liikettä. Viterbin alkorytmi pyrkii löytämään todennököismmän polun eri liikkeiden välillä. Alkorytmi käy videota läpi liike kerrallaan ja laskee mikä on mallien perusteella todennäköisin liike. Lopuksi saadaan liikesarja, josta videonäyte todennäköisimmin koostuu. Liikesarja liitetään tunnistusvaiheessa tiettyyn eleeseen. (Malgireddy et al., June)

4.3 Zonga ja pieninimmän neliösumman menetelmä sovelluttuna monistoon

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

Videokuva on helppo mieltää kolmiuloitteiseksi datajoukoksi. Videokuvan ulottuvuuksia ovat korkeus, leveys ja aika. Ryhmä Zonga käsitteleekin videokuvaa kolmiulotteisena tensorina. Tensorin ulottovuudet vastaavat videon ulottuvuuksia. Voidaan ajatella, että yksi matriisi kuvaa yhtä pysäytyskuvaa, jolloin tensorin tuoma kolmas ulottuvuus on aikaulottovuus. (Lui, June)

Tensorin käsittely sellaisenaan on hankalaa sen suuren datamäärän vuoksi. Helpot-

taaksen videon käsittelyä ryhmä laskee tensorille HOSVD(Higher-order singular value decomposition)-hajotelman. Hajotelma on kustomoitu 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. (Lui, June)

Matriisihajotelman avulla video voidaan kuvata pisteenä kolmiulotteisessa monistossa (manifold). Moniston ulottuvuudet vastaavat tensorihajotelman ulottuvuuksia. Monisto säilyttää videon alkuperäisen geometrisen rakenteen Euklidista avaruutta paremmin. (Lui, 2012) 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, June)

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.(**?**)

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, June) Pienimmän neliösumman menetelmässä pyritään minimoimaan luokitteluvirheen neliötä eli oikean tuloksen ja arvioidun tuloksen erotuksen neliötä (**?**). Minimoidaan siis funktiota:

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

Paintovektorin avulla muodostetaan regressiofunktio, jonka avulla havainnot kuvataan samaan tulosavaruuteen. Havaintonnot kuvataan regressiöfunktion avulla samaan tulosavaruuteen, jossa ne luokitellaan 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.

4.4 Yhteenveto menestyneistä kilpailutöistä

Menestyneet kilpailutyöt lähestyivät ongelmaa varsin erilaisin menetelmin ja panostivat eri vaiheesiin eleentunnistuksessa.

Ryhmä Xiaozhuwudi käytti työssään laajennettua MHI-kuvaa eli tutki videolla tapahtuvaa liikettä ohittaen yksittäiset pysäytyskuvat ja videon ajallisen rakenteen. Ryhmä Zonga valitsi piirteiksi horisontaalisen ja vertikaalisen liikkeen videolla sekä kuvan ”summan”videon yli. Ryhmä immortals sen sijaan lähti tutkimaan yksittäisiä pysäytyskuvia ja tutki kuvista muotoja HOG ja HOF -piirteiden avulla.

Tunnistusmenetelmät heijastelivat ryhmien piirrevalintoja. Xiaozhuwudi ja Zonga, jotka tiivistivät näytteet yksittäisiin kuviin käyttivät luokittelumenetelmiä, jotka perustuvat etäisyyden laskemiseen euklidissa tai sen kaltaisessa tilassa.

Immortals lähti oletuksesta, että videokuva koostuu ennen kaikkea joukosta perättäisiä yksittäisiä liikkeitä. Markovin piilomallin avulla kuvattiin liikkeiden ajallinen rakenne. Luokittelussa pyrittiin löytämään liikesarja, joka kaikken todenäköisimmin esiintyi videolla. Ryhmän Immortals menetelmä edusti kilpailutöiden yleistä suuntausta.

Immortals sijoittui kilpailussa viidenneksi, Zonga kuudenneksi ja Xiaozhuwudi kahdeksanneksi. Parhaiten menestyneen Immortalssin menetelmä vaikuttaa raskaimmalta, mutta ilmoituksen perusteella se on vain lineaarinen suhteessa opetusnäytteiden määrään. Muut ryhmät ilmoittivat samankaltaisia lukemia.

Kaikki kolme ryhmää käyttivät sekä väri että syvyyskuvaa.

5 Johtopäätökset

Kilpailutöiden perusteella 3D-videokuvan tunnistuksessa käytetään pitkälti samoja keinoja kuin 2D-videokuvan tunnistuksessa. Kilpailutyöt eivät huomioineet 3D-kuvaa erityisesti piirrevalinnassa tai tunnistusmentelmissä tai ainakaan tätä näkökulmaa ei tuotu erityisesti esille kilpailutöiden kuvauksissa. 3D-videokuvaa hyödynnettiin esikäsittelyvaiheessa taustan irroittamiseen (lähes kaikki kilpailutyöt) ja ainakin yhdessä työssä (Immor-

Taulukko 2: Vertailussa ryhmien Xiaozhuwudi, Immortals ja Zonga kilpailutyöt

	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	Datan ryhmittely	Tekijöihin jako
Ajallinen jako	Perustuu kuvan eroon lepotilan välillä	Viterbin jako	Perustuu kuvan eroon lepotilan välillä
Eleen esitys	Summakuva	Joukko piirteitä	Kolmiuloitteinen tensori
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

tals) 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.

Kilpailijoiden käyttämistä menetelmistä lähes kaikki perustuivat ajallisen rakenteen mallintamiseen Markovin piilomuuttujan tai muun vastaavan menetelmän avulla. Piirteinä käytettiin erilaisia kuvan intensiteettivaihteeluun perustuvia piirteitä. Erityisiä perusteluita tälle ei kuitenkaan esitetty. Kilpailussa menestyi hyvin myös muutama työ, jotka käyttivät täysin tästä poikkeavia menetelmiä.

Jatkotutkimuksen kannalta 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.(Chalearn2)

Silloin ne tuottavat hyviä tuloksia juuri tälle datalle, mutta eivät menestyisi yhtä hyvin muille datajoukoille. Esimerkiksi näytteille, jotka on esimerkiksi kuvattu kauempaa kohteesta tai jollain muulla tapaa eroavat tästä datasta.

Olisi mielenkiintoista tutustua myös voittajatyöhön, joskaan sen käyttämät menetelmät eivät pinnallisen kuvauksen perusteella juuri eronneet kilpailun yleisestä suuntauksesta.

6 Loppuluku

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, Mar. 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`.

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, June. doi: 10.1109/CVPR.2005.177.

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, 2012. URL `http://eprints.pascal-network.org/archive/00009716/`.

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, June. doi: 10.1109/CVPRW.2012.6239178.

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, 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/`.

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. 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, June. 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, 2012. ISSN 0262-8856. doi: 10.1016/j.imavis.2011.08.002. URL `http://www.sciencedirect.com/science/article/pii/S0262885611000692`.

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, June. doi: 10.1109/CVPRW.2012.6239185.

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

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.

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.

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, June. 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, Dec. 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, June. doi: 10.1109/CVPRW.2012.6239179.