Aalto-yliopisto
Perustieteiden korkeakoulu
Tietotekniikan tutkinto-ohjelma

# Mallipohjaiset käsienseurantamenetelmät ja niiden soveltuvuus viittomakielten käden konfiguraatioiden estimointiin videosta

**Kandidaatintyö**

**29. marraskuuta 2011**

**Matti Karppa**

Aalto-yliopisto
Perustieteiden korkeakoulu
Tietotekniikan tutkinto-ohjelma

KANDIDAATINTYÖN
TIIVISTELMÄ

| | |
|---|---|
| **Tekijä:** | Matti Karppa |
| **Työn nimi:** | Mallipohjaiset käsienseurantamenetelmät ja niiden soveltuvuus viittomakielten käden konfiguraatioiden estimointiin videosta |
| **Päiväys:** | 29. marraskuuta 2011 |
| **Sivumäärä:** | 31 |
| **Pääaine:** | Informaatiotekniikka |
| **Koodi:** | T3006 |
| **Vastuuopettaja:** | Ma professori Tomi Janhunen |
| **Työn ohjaaja(t):** | DI Ville Viitaniemi (Tietojenkäsittelytieteen laitos) |


Käsien seuranta on tärkeä työvaihe muun muassa viittomakielten tunnistuksessa sekä erilaisissa elepohjaisissa käyttöliittymäsovelluksissa. Tässä kandidaatintyössä tutustutaan neljään erilaiseen kirjallisuudessa esiteltyyn korkean tason malliin perustuvaan käsien seurantamenetelmään. Tutkittavat menetelmät on valittu siten, että ne soveltuvat osaksi luonnollisten videokameralla kuvattujen laajojen viittomakieliaineistojen automaattista analysointijärjestelmää.

Menetelmiä tarkastellaan eri näkökulmista. Eri menetelmien käyttämiä tapoja mallintaa käden muotoa vertaillaan toisiinsa. Tutkitaan menetelmien kykyä löytää käsi ilman aikaisempaa tietoa sen sijainnista ja kykyä palautua seurannan häiriintyessä. Menetelmiä arvioidaan niiden laskennallisen raskauden osalta. Lisäksi tutustutaan foneettiseen tapaan esittää käden muotoja viittomakielissä, ja selvitetään menetelmien soveltuvuutta tuottamaan foneettista viittomakielisignaalia.

Osoittautuu, että kaikki tarpeelliset osat kokonaisvaltaiseen käsien seurantajärjestelmään ovat jo olemassa. Yksikään menetelmä ei täytä kaikkia järjestelmän vaatimuksia, mutta menetelmien havaitaan täydentävän toistensa puutteita. Osa menetelmistä käyttää sellaisia käden muodon malleja, että niiden pohjalta päästään suoraan foneettiseen esitykseen käsiksi. Menetelmät ovat kuitenkin laskennallisesti hyvin raskaita, mikä voi osoittautua haasteeksi suurten aineistojen käsittelyssä.



| | |
|---|---|
| **Avainsanat:** | käsien seuranta, kohteen seuranta, tietokonenäkö, viittomakieli, käsimuodot, käden konfiguraatiot |
| **Kieli:** | Suomi |

Aalto University
School of Science
Degree Program of Computer Science and Engineering

ABSTRACT OF
BACHELOR'S THESIS

| | |
|---|---|
| **Author:** | Matti Karppa |
| **Title of thesis:** | Model-based hand tracking methods and their applicability to estimating sign language hand configurations from video |
| **Date:** | 29 November 2011 |
| **Pages:** | 31 |
| **Major:** | Informaatiotekniikka |
| **Code:** | T3006 |
| **Supervisor:** | Professor (pro tem) Tomi Janhunen |
| **Instructor:** | M.Sc. (Tech), Ville Viitaniemi (Department of Information and Computer Science) |

Hand tracking is an important task in sign language recognition and various gesture-based user interface applications. In this bachelor's thesis, four different hand tracking methods based on abstract high-level models are examined. Among the methods found in literature, the methods have been chosen in a way that they can be used as components in a system intended for automatic analysis of large bodies of natural sign language videos, recorded using an ordinary video camera.

The methods are discussed from different points of view. The different ways that the methods model the shape of the hand are compared with one another. The ability of different methods to discover the location of the hand without prior knowledge and the ability to recover in case the track is lost are examined. The methods are evaluated based on their computational complexity. In addition, a phonetic system described in the literature for representing the shape of the hand in different sign languages is presented, and the different methods are assessed based on their compatibility with the phonetic representation for the purpose of producing phonetic sign language signal.

It will be shown that all components required for creating a comprehensive hand tracking system already exist. No method is able to meet the requirements of the complete system on its own. However, the methods will be shown to complement the shortcomings of one another. Some methods use models for representing the shape that allow access to the phonetic information in a very straightforward way. However, the methods are computationally very expensive which may be prove to be a challenge with respect to processing large amounts of data.

| | |
|---|---|
| **Keywords:** | hand tracking, object tracking, computer vision, sign language, handshapes, hand configurations |
| **Language:** | Finnish |

# Sisältö

**1 Johdanto** **5**

**2 Käsimallin rakenne** **8**

2.1 Yleistä . . . . . 8

2.2 Kolmiulotteiset luurankomallit . . . . . 8

2.3 Käsivarsimalli . . . . . 9

**3 Käsien löytäminen** **11**

**4 Mallien sovittaminen** **12**

4.1 Yleistä . . . . . 12

4.2 Diskriminatiiviset ja generatiiviset menetelmät . . . . . 12

4.3 Athitsoksen ja Sclaroffin menetelmä . . . . . 12

4.4 Stengerin et al. menetelmä . . . . . 14

4.5 De la Gorcen et al. menetelmä . . . . . 17

4.6 Buehlerin et al. menetelmä . . . . . 18

**5 Laskennallinen vaativuus** **19**

5.1 Yleistä . . . . . 19

5.2 Approksimaatiot . . . . . 20

5.3 Kokeelliset suoritusajat . . . . . 21

**6 Soveltaminen viittomasignaalin mallintamiseen** **21**

6.1 Yleistä . . . . . 21

6.2 Käden konfiguraatioiden foneettinen esittäminen . . . . . 22

6.3 Siirtyminen laskennallisista malleista foneettisen esitykseen . . . . . 24

**7 Tarkastelu** **25**

**8 Yhteenveto** **28**

**Lähteet** **29**

# 1 Johdanto

Käsien sijainnin, muodon sekä asennon seuraaminen videosekvensseissä on tärkeä työvaihe muun muassa viittomakielten tunnistuksessa sekä erilaisissa elepohjaisissa käyttöliittymäsovelluksissa. Seuranta voidaan tehdä monella eri tavalla ja vaihtelevalla tarkkuudella. Yksinkertaiseen käyttöliittymäsovellukseen voi esimerkiksi riittää, että estimoidaan käsiosan sijaintia näkökentässä yksittäisenä pisteenä eri ajanhetkillä. Toisaalta esimerkiksi viittomakielten automaattisessa tunnistuksessa saatetaan haluta mahdollisimman tarkka käsitys käden nivelten konfiguraatiosta kullakin ajanhetkellä. Tässä kandidaatintyössä luodaan pieni vertaileva katsaus neljään erilaiseen kirjallisuudesta löytyvään käsien seurannan menetelmään, ja selvitetään, miten eri menetelmät sopivat kokonaisvaltaiseen käsien liikkeen analysointijärjestelmään viittomakielisten videoiden kontekstissa ja kuinka niitä voidaan käyttää viittomakielisignaalin irrottamiseen videosta.

Käsien konfiguraatiota ja liikettä on vanhastaan voitu mitata suoraan esimerkiksi datahansikkailla. Tällainen menettely kuitenkin edellyttää kalliita erityislaitteita ja soveltuu käytettäväksi vain laboratorio-olosuhteissa. Lisäksi datahansikkaat häiritsevät ihmisen luonnollista elehdintää. Ihminen luultavasti kokee tietokonenäkömenetelmin tapahtuvan seurannan miellyttävämmäksi ja vähemmän rajoittavaksi.

Tutkittavien menetelmien valinnan lähtökohdaksi on otettu niiden sopivuus järjestelmään, jolla pyritään foneettisesti analysoimaan suuria, tavallisella videokameralla etukäteen kuvattuja luonnollista viittomista sisältäviä viittomakielisiä aineistoja. Analysoitava videomateriaali oletetaan monokulaarisella eli yksilinssisellä kameralla kuvatuksi värilliseksi videoksi. Menetelmien tulisi toimia näkyvän valon aallonpituusalueella ($\lambda = 380$–$740$ nm) ja tavanomaisella kuvataajuudella ($f = 24$–$30$ Hz) kuvatulla videolla. Tutkittavien menetelmien ei tulisi vaatia syvyyssensorijärjestelmää, eivätkä ne saisi vaatia esimerkiksi suurinopeuskameraa, segmentoinnin avuksi infrapunakuvaa tai syvyysnäkymän päättelyyn stereopsista. Luonnollisten videoiden analysointiin käyttökelpoisten menetelmien tulisi olla siinä määrin robusteja, etteivät ne vaadi laboratorio-olosuhteita tai muita kontrolloituja apukeinoja kuten värillisiä hansikkaita.

Käsien seuranta on luonteeltaan melko haastavaa [11]. Kädet voivat tehdä hyvin monimutkaisia liikkeitä nopeudella, jonka tarkkarajainen havaitseminen ajassa on tavallisen videokameran kuvataajuuden äärirajoilla. Käsien liike on myös haastavaa nivelikkyytensä ja elastisuutensa takia; kädet eivät vain liiku jäykästi paikasta toiseen, vaan liikkeeseen liittyy usein myös käden konfiguraation muutoksia. Menetelmien tulee selviytyä okkluusioista ja itseokkluusioista eli tilanteista, joissa yksi käsi peittää toisen alleen tai käden jokin osa, kuten sormet, peittää allensa osan muuta kättä. Kädet ovat melko tasaisesti ihonvärisiä eikä niillä juurikaan ole tekstuuria, joten käsien eri osien erottaminen toisistaan on haastavaa. Myös tavanomaiset kohteen seurannan haasteet ovat

jatkuvasti läsnä, kuten syvyysinformaation katoaminen yksikameraisessa järjestelmässä, vaihtuvat valaistusolosuhteet sekä erilaiset häiriötekijät [29; 11]. Erityisesti painoarvoa on annettava rakenteellisille häiriöille, jotka johtuvat esimerkiksi kuvassa esiintyvistä ylimääräisistä esineistä (englanniksi *clutter*) ja sotkevat erityisesti reunainformaatiota. Erityisesti käden konfiguraation estimointi on laskennallisesti vaativaa, koska tila-avaruus on suuri; de la Gorce, Fleet ja Paragios arvioivat, että yhdellä käsiosalla on jopa 30 vapausastetta [11].

Tyypillinen liikkuvien kohteiden videoanalyysijärjestelmä koostuu kolmesta avainosasta: kohteen löytämisestä, sen seuraamisesta ajanhetkestä toiseen sekä liikkeen analysoimisesta ja arvioimisesta [29]. Kaikki tutkittavat menetelmät eivät vastaa jokaiseen ongelmaan, vaan osa menetelmistä olettaa, että käsien paikka on tunnettuna seurannan aloitusajankohdalla. Kaikki menetelmät eivät välttämättä pysty myös palautumaan tilanteesta, jossa kohde kadotetaan. On siis mahdollisesti tarpeen täydentää menetelmien puutteita toisilla menetelmillä. Tässä kandidaatintyössä selvitetään, miten näihin haasteisiin on vastattu.

Tässä työssä on keskitytty kehittyneisiin ja monimutkaisiin mallin sovittamiseen perustuviin menetelmiin. Yksi menetelmä perustuu käsivarren nivelten kulmien estimointiin kuvatasossa. Loput kolme menetelmää perustuvat käsiosan kolmiulotteisen luurankomallin parametrien estimointiin. Menetelmät on valittu siten, että ne kattavat jonkin tarpeellisen kokonaisjärjestelmän osa-alueen. Toisaalta tutkittavat mallit on parametrisoitu tavoilla, joilla on selkeä yhteys ihmisen fysiologiaan ja mahdollisesti viittomakielten fonetiikkaan. Menetelmiä pyritään vertailemaan ja kategorisoimaan erinäisten ominaisuuksien mukaan. Erityisesti kiinnitetään huomiota siihen, miten eri menetelmät mallintavat kättä – sen muotoa ja paikkaa – ja miten menetelmän tuottamasta datasta olisi irrotettavissa piirteitä viittomakielisen signaalin esittämiseen Liddellin ja Johnsonin [18] foneettisen järjestelmän tapaisesti. Käsiteltävät mallit ovat

- Athitsoksen ja Sclaroffin kolmiulotteinen käsiosamalli [1]
- Stengerin, Thayanathanin, Torrin ja Cipollan kolmiulotteinen käsiosamalli [24; 23]
- De la Gorcen, Fleetin ja Paragioksen kolmiulotteinen käsiosamalli [11]
- Buehlerin, Everinghamin, Huttenlocherin ja Zissermanin käsivarsimalli [5; 6].

Käden seurantaan ja kohteiden seurantaan yleisesti on esitetty myös lukuisia toisenlaisia ratkaisuja. Tässä työssä tutkitut menetelmät soveltavat monimutkaisia ja kehittyneitä, abstrakteja malleja, mutta kirjallisuudessa on kuvattu myös täysin toisenlaisia lähestymistapoja: Han, Awad ja Sutherland [15] lähestyvät ongelmaa segmentoimalla ihonvärisiä alueita ja soveltamalla Kalman-suodinta okkluusioiden havaitsemiseen. Shan, Tan ja

Wei [21] seuraavat kättä yhdistämällä *mean-shiftin* ja partikkelisuotimen. Grabner, Grabner ja Bischof [14] ja Vacchetti, Lepetit ja Fua [27] lähestyvät kohteen seurantaa avainpisteiden vastineiden löytämisongelmana. Pezzementi, Zachary ja Voros [20] seuraavat nivelikkään kohteen liikettä kolmiulotteisella, kinemaattiseksi ketjuksi jäsennetyllä mallilla, joka renderöidään kuvan päälle optimoitavan eri osien ulkoasuun perustuvan todennäköisyysfunktion evaluoimiseksi. Lu ja Hager [19] seuraavat ja segmentoivat taustasta kohteita parametrittömästi ulkoasumalleilla, joista he käyttävät nimitystä *bags of image patches*. Zhang ja Rui [30] seuraavat kohteita pikselien luokitteluun ja integrointiin perustuvalla menetelmällä. Sminchisescu, Kanaujia ja Metaxas [22] esittelevät diskriminatiivisen inferenssimallin, jota he nimittävät *Conditional Bayesian Mixture of Experts Markov Modeliksi* ($BM^3E$) ja jota he soveltavat kohteen seurantaan käyttäen kohteen silhuetista erotettuja piirteitä. Tämä ei ole kattava lista menetelmistä, vaan niitä on tuotu esille esimerkinomaisesti. Näiden menetelmien tarkempi käsittely ylittää tämän kandidaatintyön laajuuden.

Suomen kielen sana *käsi* on jokseenkin monitulkintainen, joten sen asemesta puhutaan *käsiosasta* (englanniksi *hand*), kun tarkoitetaan kättä ranteesta eteenpäin, ja toisaalta *käsivarresta* (englanniksi *arm*), kun tarkoitetaan koko kättä olkapäästä sormenpäihin. *Käden konfiguraatiolla* tarkoitetaan tässä työssä tietoa siitä, miten ihmisen käden käsiosa tuottaa jonkin tietyn muodon, esimerkiksi sormen nivelten kulmien tarkkuudella tai vastaavana informaationa laskennallisen mallin parametreista. Havainnoitsijalle tällainen tuotettu käden konfiguraatio näyttäytyy *käsimuotona*. Näin määritellyt käden konfiguraation ja käsimuodon merkitykset ja erot niiden välillä vastaavat Liddellin ja Johnsonin [18] terminologiaa, jota he käyttävät viittomakielten foneettisen mallinsa yhteydessä.

Työn rakenne jäsentyy seuraavasti: Luvussa 2 tutustutaan erilaisiin käsimalleihin ja käsien esitystapoihin. Luvussa 3 tarkastellaan menetelmiä käsien löytämiseksi yksittäisestä kuvasta, seurantajärjestelmien alustamista sekä palautumista tilanteesta, jossa kädet hukataan tilapäisesti. Luvussa 4 tutustutaan eri tapoihin, joilla mallit sovitetaan kuvaan. Luvussa 5 arvioidaan eri mallien laskennallista vaativuutta, hakuavaruuden rajoittamista sekä soveltuvuutta reaaliaikaisiin sovelluksiin. Luvussa 6 tutustutaan Liddellin ja Johnsonin foneettisen viittomakielimallin käden konfiguraatioita kuvaavaan osaan ja arvioidaan menetelmien soveltuvuutta tuottamaan järjestelmän mukaista signaalia. Luvussa 7 tarkastellaan eri malleja kokonaisuutena ja vertaillaan niitä toisiinsa. Luku 8 päättää työn. Aiemmin mainittujen kohteen seurannan kolmen vaiheen kannalta ensimmäistä eli käden löytämistä ja seurantajärjestelmän palauttamista käsitellään luvussa 3, seuraamista ruudusta toiseen erityisesti luvussa 4 ja myöhemmän prosessoinnin kannalta olennaisia käden esitystapoja luvuissa 2 ja 6. Luvut 5 ja 7 keskittyvät erityisesti esitettyjen menetelmien haasteiden ja kokeellisten tulosten tarkasteluun.

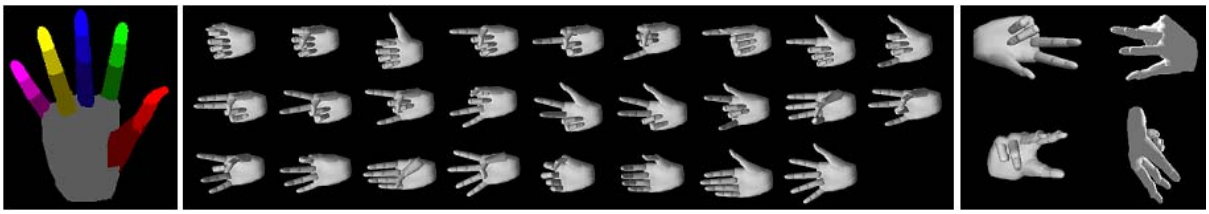
# 2 Käsimallin rakenne

## 2.1 Yleistä

Tässä luvussa tutustutaan käsiteltävien mallien käyttämiin tapoihin esittää käden konfiguraatioinformaatiota. Mallit voidaan jakaa luontevasti kahteen ryhmään: kolmiulotteisiin käsiosan luurankomalleihin sekä käsivarsimalliin. Seuraavaksi tutustutaan mallien parametrisointiin ja vapausasteiden määrään.

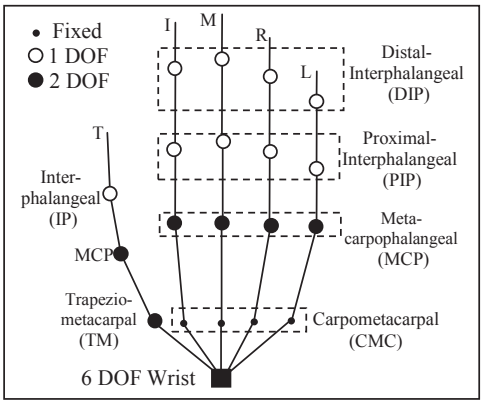
## 2.2 Kolmiulotteiset luurankomallit

Athitsos ja Sclaroff [1], Stenger et al. [24; 23] ja de la Gorce et al. [11] mallintavat käsiosaa kolmiulotteisella luurankomallilla. Athitsoksen ja Sclaroffin mallissa kukin sormi muodostuu kolmesta segmentistä. Näiden lisäksi kämmen muodostaa yhden segmentin. Yhteensä käsiosa siis koostuu 16 palasta ja 15 nivelestä. Vapausasteita itse synteettisellä käsiosalla on yhteensä 20. Tämän lisäksi he mallintavat kameran suuntaa kahdella ja kuvatason suuntaa yhdellä vapausasteella, joten kokonaisuudessaan mallissa on 23 vapausastetta. Mallin osat sekä esimerkkejä syntetisoiduista käsistä on nähtävissä kuvassa 1. [1]



Kuva 1: Vasemmalla Athitsoksen ja Sclaroffin käsimallin eri osat. Keskellä heidän käyttämänsä 26 peruskäsimuotoa, joista tietokanta koottiin. Oikealla eräs käsimuoto eri asennoissa näytteistettynä. [1]

Stenger et al. käyttävät luurankomallia, jossa on yleisessä tapauksessa on 27 vapausastetta. Näistä 21 on nivelten kulmaparametreja ja loput kuusi orientaatiolle ja spatiaaliselle sijainnille. Tämä yleinen malli on esitetty tarkemmin Erolin, Bebisin, Boylen ja Twomblyn katsauksessa [12], ja sen rakenne on nähtävissä kuvassa 2. Stenger et al. pienentävät kokeellisissa tilanteissa vapausasteiden määrää rajoittaakseen järjestelmänsä laskennallista vaativuutta. Erityisesti he viittaavat Wun, Linin ja Huangin [28] tuloksiin, joiden mukaan nivelkulmien korkean keskinäisen korrelaation vuoksi käden tila-avaruus voidaan vähentää seitsemään vapausasteeseen pääkomponenttianalyysillä. Stenger et al. osoittivat tämän pätevän kokeellisesti keräämällä koehenkilöiltä mielivaltaista




Kuva 2: Stengerin et al. käyttämän luurankomallin rakenne ja nivelten vapausasteet. [12]

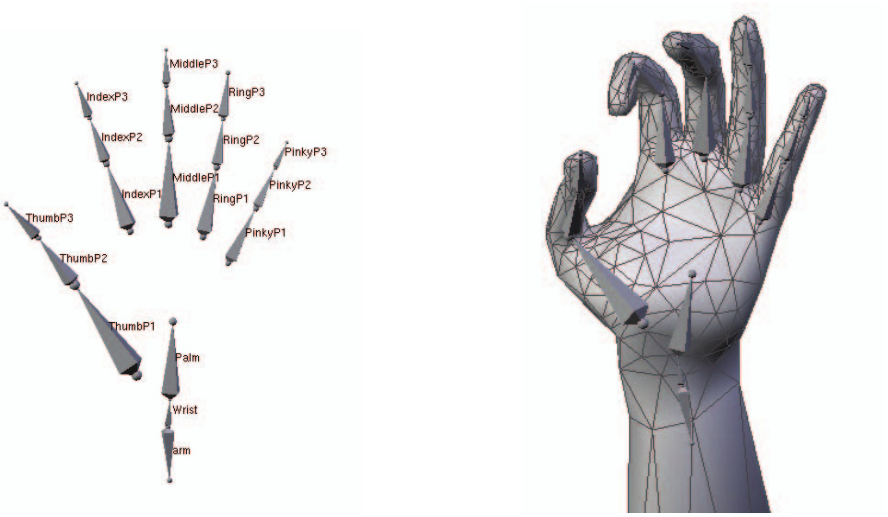
käsidataa datahansikkailla. Kaikissa 15 mittausdatajoukossa kahdeksan ensimmäistä pääkomponenttia kattoi yli 95 prosenttia varianssista, ja kymmenessä tapauksessa seitsemän pääkomponenttia ylitti saman rajan. [23]

De la Gorcen et al. luurankomalli koostuu 18 luusta. Mallilla on 22 vapausastetta, jotka vastaavat luurangon nivelten kulmia. Näiden lisäksi käsikonfiguraatiovektoriin kuuluu kolme translaatioparametriä ja yksi kvaternioparametri, joka kuvaa ranteen globaalia sijaintia ja suuntaa kameran suhteen. Kullakin luulla on lisäksi kolme morfologista parametria eli skaalauskerrointa, yhteensä siis 54 kappaletta. Morfologiset parametrit määritellään kalibrointivaiheessa ja ne määrittävät mallin eri osien pituuksien suhteita kussakin sormessa. Luurankomallin rakenne on nähtävissä kuvassa 3a. [11]

De la Gorce et al. mallintavat luurangon lisäksi käden pinnan. Pinta koostuu 1000 kolmiosta, joiden varjostusta mallinnetaan Gouraud-varjostusmallilla. Reflektanssimalli oletetaan lambertilaiseksi. Tekstuuria ja muita erikseen mallintamattomia käden ulkoasun ominaisuuksia varten he käyttävät adaptiivista albedo-funktiota. Valaistusmallin määrää neliulotteinen vektori, jonka yksi komponentti kuvaa yleistä taustavalaistusta ja kolme muuta määrittelevät pistemäisen etäisen valaistuslähteen suunnan. Verteksin valaistus määritellään summana, jonka yksi termi on taustavalaistus ja toinen termi on suunnatun valonlähteen ja pinnannormaalin skalaaritulo verteksipisteessä. Kolmiopintojen sisäinen valaistus saadaan bilineaarisella interpolaatiolla. Pisteen ulkoasu saadaan valaistuksen ja reflektanssin tulona. Käden pintamallin rakenne on nähtävissä kuvassa 3b. [11]

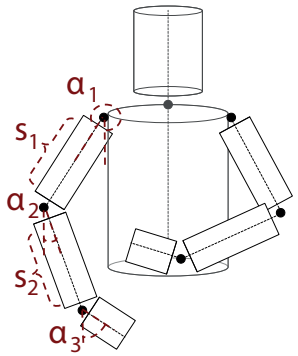
## 2.3 Käsivarsimalli

Buehler et al. lähestyvät ongelmaa mallintamalla kummankin käden käsivarret. Verrattuna suoraan käsiosan sijainnin etsintään, tämä lähestymistapa tuo lisää tarkkuutta sovitukseen erityisesti tilanteessa, jossa käsiosat osittain peittävät toisensa. Mallin




(a) (b)

Kuva 3: De la Gorcen et al. käsimallin rakenne. Kuvasta nähdään, että malliin kuuluu käsiosan lisäksi myös hieman rannetta. (a) Käden luurankomallin rakenne. (b) Käden pinnan rakenne. [11]




Kuva 4: Buehlerin et al. käsivarsimalli osineen. [5]

rakenne on esitetty kuvassa 4. Formaalisti he määrittelevät mallin konfiguraatiovektorilla $\mathbf{L} = (b, l_1, l_2, \ldots, l_6)$, missä $b$ on binäärinen termi, joka kertoo, kumpi käsi on kumman päällä, ja termit $l_i = (s_i, \alpha_i)$ määrittelevät käsivarren eri osat. Kumpikin käsi on jaettu kolmeen segmenttiin, jotka vastaavat olkavartta, kyynärvartta ja käsiosaa. Nämä segmentit muodostavat kinemaattisen ketjun. Käden osa määritellään nivelen kulmana $\alpha_i$ sekä skaalaustekijänä $s_i$, joka kertoo osan kuvassa näkyvän suhteellisen pituuden. Käsiosien kohdalla skaalaustekijä $s$ on vakio, joten kummallekin kädelle on määritelty yhteensä viisi vapausastetta. Kokonaisuudessaan vapausasteita mallilla on siis 11. [5]

Käsivarsien lisäksi Buehler, Everingham ja Zisserman tarvitsevat käsimuotoinformaatiota viittomien tunnistusoppimisongelmaansa varten. He käyttävät luvussa 4.6 esitettyä taustan ja etualan värimallia ja segmentoivat käsivarsimallin ennustamasta paikasta

käsiosan taustasta *graph cut* -menetelmällä [4]. He esittävät käsimuotoinformaation HOG-vektorina (*Histogram of Oriented Gradients*) eli suunnattujen gradienttien histogrammina [10]. Tämä esitys sisältää sekä muoto- että ulkoasuinformaatiota, kuten ulkoreunainformaatiota ja sisäistä tekstuuria, ja on samaan aikaan jossain määrin valaistusinvariantti. Tunnistuskäyttöä varten he kvantisoivat käsimuotovektorit assosioimalla ne euklidisen etäisyyden suhteen lähimpiin opetusesimerkkimuotoihin. [6]

# 3 Käsien löytäminen

Tässä luvussa tutustutaan siihen, miten eri menetelmät hallitsevat käsien löytämisen kuvasta. Käsien löytämistä tarvitaan seurantajärjestelmän alustamiseen ja toisaalta, mikäli käsien seuranta menetetään, järjestelmän palautumiseen. Kaikki menetelmät eivät ota kantaa käsien löytämiseen, kun taas osalle menetelmistä käsien löytäminen on osa mallin sovitusprosessia.

Athitsoksen ja Sclaroffin malli edellyttää manuaalista alustusaskelta jokaista videon ruutua kohti. Ihmisen tai muun toimijan on määriteltävä *bounding box* eli suorakulmainen alue, jonka sisältä käden oletetaan löytyvän annetusta ruudusta. [1]

Stengerin et al. menetelmä hallitsee kaikki kolme tärkeää vaihetta. Se osaa löytää kädet, seurata niitä ja palauttaa seurannan menettäneen seurantajärjestelmän yhtenäisellä tavalla. Mikäli järjestelmällä ei ole dynaamista tietoa edellisistä ruuduista, kuten esimerkiksi ensimmäisessä ruudussa tai heti seurannan menettämisen jälkeen, järjestelmä toimii hierarkisena detektiojärjestelmänä, jossa käsimuoto haetaan puurakenteesta. Seurannan edetessä haku saa lisäinformaatiota käden dynaamisista liikkeistä edeltävissä ruuduissa. [23]

De la Gorcen et al. esittämä menetelmä asettaa lukuisia vaatimuksia alustukselle. Menetelmä ei osaa löytää kättä lainkaan, vaan järjestelmälle tulee jotenkin toimittaa tieto käden konfiguraatiosta ensimmäisessä ruudussa – joko käsin tai jonkin toisenlaisen, esimerkiksi diskriminatiivisen, järjestelmän avulla. Diskriminatiiviset järjestelmät on määritelty luvussa 4.2. Ensimmäisessä ruudussa käden oletetaan myös olevan aseteltuna likimain kuvatason suuntaisesti, ja alussa käden pinnan albedon oletetaan olevan vakio siten, että käden ulkoasu on pelkästään muodon ja varjostuksen funktio. Käden morfologiset parametrit, jotka on esitelty luvussa 2.2, estimoidaan ensimmäisestä ruudusta ja säilytetään vakioina myöhemmissä ruuduissa. Käyttäjän tulee myös toimittaa taustakuva tai taustan histogrammi. Koska menetelmältä puuttuu kyky alustaa itsensä, se ei kykene palautumaan tilanteesta, jossa käden seuranta kadotetaan. [11]

Buehlerin et al. menetelmä hallitsee käden eri osien löytämisen. Kun malli on opetettu, se voidaan sovittaa kuvaan sellaisenaan ja käden eri osien paikat löytyvät suoraan sovituksen yhteydessä. Ajallista informaatiota hyödynnetään vain sovitustulosten parantamiseen. [5]

# 4 Mallien sovittaminen

## 4.1 Yleistä

Tässä luvussa tutustutaan siihen, miten eri eri menetelmät sovittavat mallinsa syötekuvaruutuihin. Luvussa 4.2 esitellään menetelmien jako diskriminatiivisiin ja generatiivisiin menetelmiin. Tämän jälkeen luvuissa 4.3–4.6 käydään läpi kunkin menetelmän mallin sovitus erikseen.

## 4.2 Diskriminatiiviset ja generatiiviset menetelmät

De la Gorce et al. jakavat esitetyt kolmiulotteisiin malleihin perustuvat menetelmät diskriminatiivisiin, kuten [1], ja generatiivisiin, kuten [11]. Diskriminatiiviset menetelmät nojaavat regressiomenetelmään tai luokittimeen, joka luokittelee annetussa ruudussa näkyvän käden muodon johonkin tunnettuun käsimuotoluokkaan Opetusdata kootaan joko kuvaamalla kameralla kättä tunnetuissa asennoissa tai tuottamalla mallikuvat *off-line* synteettisesti. Käsien suuri vapausasteiden määrä rajoittaa diskriminatiivisten mallien soveltamista mielivaltaisiin käsikonfiguraatioihin, koska tila-avaruus on liian suuri näytteistettäväksi. Generatiivisten mallien tapauksessa kuvia projisoidaan synteettisestä kolmiulotteisesta mallista *on-line*, ja mallin parametrejä optimoidaan suoraan ajonaikaisesti. Stengerin et al. menetelmä [23] yhdistelee mielenkiintoisesti sekä diskriminatiivisille että generatiivisille menetelmille ominaisia piirteitä, muttei käytä *on-line* synteesiä. [11]

## 4.3 Athitsoksen ja Sclaroffin menetelmä

Athitsos ja Sclaroff generoivat etukäteen mallistaan suuren määrän synteettisiä kuvia ja mallintavat sovitusongelman kuvahakuna tietokannasta. De la Gorce et al. mainitsevat erityisesti tämän menetelmän esimerkkinä diskriminatiivisestä menetelmästä [11]. Athitsos ja Sclaroff ovat pyrkineet sovitusfunktiota laatiessaan erityisesti ottamaan huomioon *clutterin* eli kuvassa esiintyvien ylimääräisten kohteiden aiheuttamat rakenteelliset häiriöt. He käyttävät kahta erilaista samankaltaisuuden mittafunktiota. Ensimmäinen mitta perustuu chamfer-etäisyyteen ja pyrkii kuvaamaan synteettisen mallin ja sovituskohteena olevan kuvan muodostamien binääristen reunakuvien välistä

samankaltaisuutta. Toinen mitta on todennäköisyyspohjainen ja pyrkii löytämään kahdesta vertailtavasta kuvasta geometrisesti samankaltaisia viivasegmenttejä, jotka olisivat mahdollisimman epätodennäköisesti syntyneet sattumalta. [1]

Suunnattu chamfer-etäisyys [2] $c(X, Y)$ joukosta $X$ joukkoon $Y$ määritellään alkion $x \in X$ keskimääräisenä etäisyytenä lähimpään joukon $Y$ alkioon

$$c(X, Y) = \frac{1}{|X|} \sum_{x \in X} \min_{y \in Y} ||x - y|| \tag{1}$$

Suuntaamaton chamfer-etäisyys määritellään suunnattujen etäisyyksien summana

$$C(X, Y) = c(X, Y) + c(Y, X) \tag{2}$$

Chamfer-etäisyyksien laskennalliset yksityiskohdat on esitetty luvussa 5.2. [1]

Athitsoksen ja Sclaroffin mukaan kuvien viivasegmenttien vastaavuuksien tutkimisessa on neljä olennaista vaihetta: pistemäisten piirteiden etsintä, suorien viivasegmenttien etsiminen niistä, segmenttien välisten vastaavuuksien määrittäminen ja vastaavuuksien laadun arviointi. Pisteet valitaan käyttämällä intensiteettigradientin lokaaleja maksimeja. Viivojen erottamiseen he näytteistävät $N_O$ suuntaa väliltä 0–180 astetta ja $N_l$ pituutta väliltä 5–100 pikseliä. Jokaisessa tutkittavassa piirrepisteessä otetaan yksi näytepari jokaisella mahdollisella suunnan ja pituuden yhdistelmällä ja arvioidaan kahta asiaa: salienssia eli silmiinpistävyyttä ja voimaa, joka määritellään gradientin suuruuden perusteella. Salienssi määritellään painottamalla viivasegmentillä sijaitsevien pistepiirteiden gradienttien ja oletetun viivan suunnan erotuksia niin, että sattumanvaraisesti vaihtelevat gradientit kumoavat toisensa ja vain oletetun suuntaiset gradientit tuovat lisää painoarvoa viivalle. Mallin viivat lasketaan *off-line*. Edellä mainittuja saliensseja ja voimakkuuksia ei tarvitse laskea malleille, koska viivojen sijainnit tiedetään niistä ilman epäselvyyksiä. [1]

Mallin ja syötekuvan viivasegmenttien keskinäisen sopivuuden laatu arvioidaan edellä mainittujen salienssi- ja voimakkuusarvojen avulla. Lisäksi huomioidaan viivasegmenttien keskipisteiden välinen etäisyys sekä suuntien ja pituuksien erotus. Tämä tuottaa kullekin mallin viivasegmentille $A$ ja syötekuvan viivasegmentin $B$ parille määritellyn viisiulotteisen laatuvektorin $Q(A, B)$. Näille laatuvektoreille on määritelty osittainen järjestys. Täydellinen järjestys saadaan arvioimalla todennäköisyyttä, jolla viivasegmenttiparin yhteensopivuus voisi tapahtua sattumalta. Hyvä mallikuvan vastine syötekuvan viivasegmentille on sellainen, joka tulisi hyvin epätodennäköisesti valittua sattumalta. Paras vastinsegmentti valitaan siten, että tämä todennäköisyys minimoituu. [1]

Lopuksi kahden kuvan viivasegmenttien keskinäistä yhteensopivuutta arvioidaan laskemalla parhaiten toisiaan vastaavien viivasegmenttien minimoitujen satunnaisen vastaavuuden todennäköisyyksien keskiarvo. Tällä lukuarvolla on taipumus suosia vähäviivaisia kuvapareja, joten arvot normeerataan jokaisen viivamäärän keskihajonnalla. [1]

Lopuksi Athitsos ja Sclaroff yhdistävät edellä mainitut viivasegmenttien vastaavuudet ja chamfer-etäisyydet kaksivaiheisella hakumenettelyllä. He yhdistävät $k$ erilaista vastaavuusmittaa syötekuvan $I$ ja tietokannan $i$:nnen mallikuvan $V_i$ välillä vastaavuusmitalla $r_{ij}$ yhdistetyksi vastaavuusmitaksi

$$M_c(V_i, I) = \sum_{j=1}^{k} \log r_{ij}$$

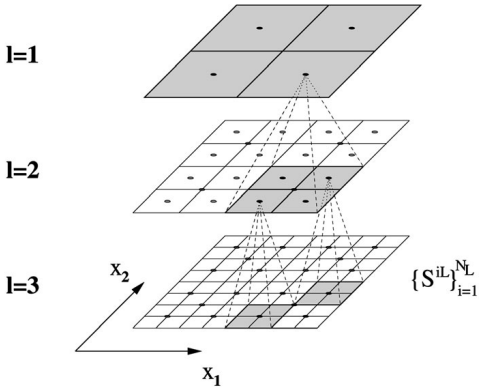
Ensimmäisessä vaiheessa karsitaan pois suurin osa mallikuvista. Vastaavuusmittoina käytetään nopeaa mallista syötteeseen chamfer-etäisyyttä ja likimääräistä syötteestä malliin chamfer-etäisyyttä, joka on esitelty luvussa 5.2. Toiseen vaiheeseen otetaan vakiomäärä parhaita mallikuvia, joiden paremmuusjärjestys lasketaan täydellistä chamfer-etäisyyttä, orientaatiohistogrammeja ja viivojen vastaavuuskustannusarvioita käyttäen. [1]

## 4.4 Stengerin et al. menetelmä

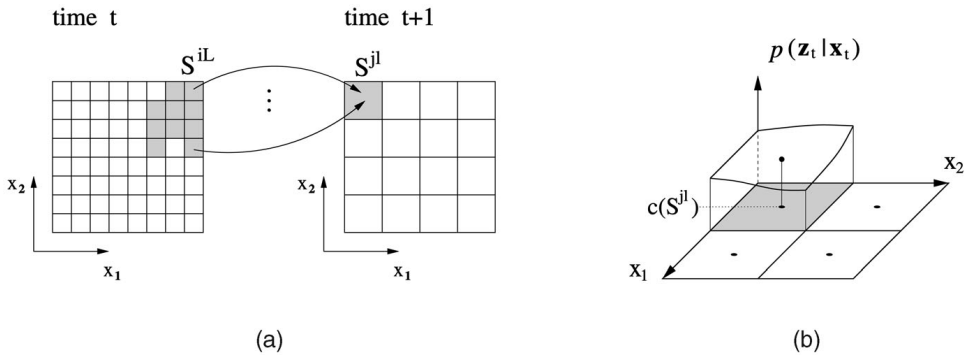
Myös Stenger et al. generoivat käsimuodoista etukäteen mallineita (englanniksi *template*). Nämä järjestetään hierarkiseksi puurakenteeksi osittamalla parametriavaruus. Kullakin ajanhetkellä eri tilaparametreille estimoidaan posteriorijakauma, joka riippuu käden dynaamisista liikkeistä edeltävinä ajanhetkinä. Mikäli tällaista tietoa ei ole saatavilla, kuten luvussa 3 on mainittu, järjestelmä toimii hierarkisena detektorina. Edeltäviä tietoja todennäköisyyksistä välitetään ajassa eteenpäin parantamaan sovituksia myöhemmillä ajanhetkillä. [23]

Stenger et al. lähtevät oletuksesta, että tilavektorit $\mathbf{x} \in \mathbb{R}^n$ sijaitsevat pienellä tila-avaruuden alueella $\mathcal{R}$, jonka rajat käsien seurannan yhteydessä määrittävät realistiset liikealueet. Tälle alueelle laaditaan moniresoluutioinen ositus $L$:ään tasoon. Kullakin tasolla $l$ on $N_l$ osaa $\{S^{i,l}\}_{i=1}^{N_l}$ siten, että $\bigcup_{i=1}^{N_l} S^{i,l} = \mathcal{R}$. Tämä on havainnollistettu kuvassa 5. Jatkuva bayesilainen todennäköisyysfunktio on diskretisoitu olettamalla se paloittain vakioksi näytteistettyjen pisteiden ympäristössä. [23]

Stenger et al. huomioivat muutokset ajassa rekursiivisella tilamuutosyhtälöllä, jota on havainnollistettu kuvassa 6. Parametrien arvot saadaan tilapuusta leveyssuuntaisella haulla. Tätä on havainnollistettu kuvassa 7. Haku aloitetaan ylimmältä tasolta, ja haarat, joiden todennäköisyys annetulla tasolla $l$ on alle kynnysarvon $\tau_t^l = \hat{p}_t^{l,min} + c_\tau(\hat{p}_t^{l,max} - \hat{p}_t^{l,min})$, jätetään tutkimatta. Kynnysarvo määrää sovituksen tarkkuuden ja vaikuttaa toisaalta myös laskennalliseen vaativuuteen. Mitä pienempi arvo on, sitä enemmän aikaa sovitukseen tarvitaan. Jos arvo on liian suuri, globaali maksimi voi jäädä haaraan, joka jätetään tutkimatta. Tällöin löydetään vain jokin muu lokaali maksimi, mikä johtaa virheelliseen tulokseen annetussa ruudussa. [23]




Kuva 5: Stengerin et al. tila-avaruuden hierarkinen osittaminen moniresoluutioiseen ruudukkoon. Todennäköisyysfunktiota approksimoidaan paloittain vakiona kullakin tasolla. [23]



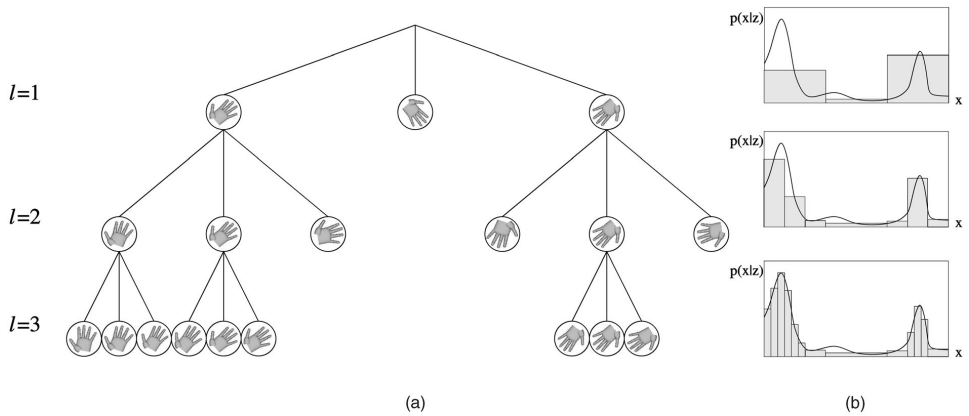

Kuva 6: (a) Siirtymät diskreeteistä tiloista toisiin eri ajanhetkien välillä. (b) Todennäköisyysfunktio evaluoidaan kunkin alueen keskipisteessä. $\mathbf{z}_t$ on havainto ja $\mathbf{x}_t$ on seurattavan kohteen tila hetkellä $t$. [23]

Stengerin et al. käyttämä havaintovektori $\mathbf{z} = (\mathbf{z}^{edge}, \mathbf{z}^{col})^T$ koostuu kahdenlaisista piirteistä: reunapiirteistä $\mathbf{z}^{edge}$ ja väripiirteistä $\mathbf{z}^{col}$. Stenger et al. olettavat nämä havainnot riippumattomiksi, joten edellä mainittu havainnon todennäköisyys annetulla tilalla $\mathbf{x}$ on muotoa $p(\mathbf{z}|\mathbf{x}) = p(\mathbf{z}^{edge}, \mathbf{z}^{col}|\mathbf{x}) \approx p(\mathbf{z}^{edge}|\mathbf{x})p(\mathbf{z}^{col}|\mathbf{x})$. Reunavektoreiden todennäköisyysmalli perustuu suunnattuun neliölliseen chamfer-etäisyyteen [2; 3]

$$d(\mathcal{A}, \mathcal{B}) = \frac{1}{N_a} \sum_{a \in \mathcal{A}} \min_{b \in \mathcal{B}} ||a - b||^2 \tag{3}$$

missä $\mathcal{A} = \{\mathbf{a}_i\}_{i=1}^{N_a}$ on mallineen pisteiden joukko ja $\mathcal{B} = \{\mathbf{b}_i\}_{i=1}^{N_b}$ syötekuvan Canny-reunapisteiden [8] joukko. Etäisyydet lasketaan erikseen $N_\gamma$ orientaatiokanavalle, jotka määräytyvät gradientin suunnan mukaan

$$d_e(\mathcal{A}, \mathcal{B}) = \frac{1}{N_a} \sum_{i=1}^{N_\gamma} \sum_{a \in \mathcal{A}^i} \min(\min_{b \in \mathcal{B}^i} ||a - b||^2, \tau) \tag{4}$$




Kuva 7: Stengerin et al. posterioritodennäköisyyden estimointi puurakenteesta. (a) Tila-avaruuden ositusten muodostama puu. Huomattavaa on, että kokonaisuudessaan epätodennäköisiä haaroja ei evaluoida kuin vain ylimmillä tasoilla. (b) Jatkuva todennäköisyysfunktio sekä sen paloittain vakio approksimaatio puun eri tasoilla. [23]

missä $\mathcal{A}^i$ ja $\mathcal{B}^i$ ovat kanavan $i$ reunapisteitä ja $\tau$ on kynnysarvo, joka määrää suurimman mahdollisen etäisyyden. Itse jakauma on metrinen eksponentiaalijakauma [26]

$$p(\mathbf{z}^{edge}|\mathbf{x}) = \frac{1}{Z}\exp(-\lambda d_e(\mathcal{A}(\mathbf{x}), \mathcal{B}(\mathbf{z}^{edge}))) \tag{5}$$

missä $\mathcal{A}(\mathbf{x})$ on tilan $\mathbf{x}$ määräämä mallineen pistejoukko ja $\mathcal{B}(\mathbf{z}^{edge})$ syötekuvan reunapistejoukko. [23]

Stenger et al. laskevat väritermin $p(\mathbf{z}^{col}|\mathbf{x})$ yksinkertaisesti siten, että jokainen kuvan pikseli jaetaan joko kuuluvaksi käsiosan silhuetin sisäpuolelle joukkoon $S(\mathbf{x})$ tai taustajoukkoon $\bar{S}(\mathbf{x})$. He olettavat havainnot pikseleittäin riippumattomiksi, jolloin termi on muotoa

$$p(\mathbf{z}^{col}|\mathbf{x}) = \prod_{k\in S(\mathbf{x})} p^s(I(k)) \prod_{k\in \bar{S}(\mathbf{x})} p^{bg}(I(k)) \tag{6}$$

missä $I(k)$ on värivektori kuvan pisteessä $k$. Ihon värimalli on gaussinen jakauma $(R, G)$-avaruudessa. Tausta oletetaan tasajakautuneeksi, mikäli parempaa mallia ei ole käytettävissä. [23]

Käden liikedynamiikkaa mallinnetaan yksinkertaisena Markovin prosessina. Koska käsiosan eri nivelten välillä on havaittavissa paljon korrelaatioita [28], oletetaan, että parametrit löytyvät pieneltä tila-avaruuden alueelta. Hakupuun eri lehtitason alueiden välille määritellään Markovin siirtymämatriisi, joka kuvaa todennäköisyyttä, että tietyltä diskreetiltä tila-alueelta liikutaan toiselle. Stenger et al. ovat keränneet nämä todennäköisyydet koehenkilöiltä datahansikkailla, kuten luvussa 2.2 mainittiin. [23]

Hakupuun lehtitason diskreettien tilajoukkojen rakentamiseen Stenger et al. esittävät kokeellisten tulosten yhteydessä kaksi erilaista tapaa: parametriavaruuden klusterointi ja ominaisavaruuden osittaminen. Ensimmäisessä tapauksessa puu rakennetaan hierarkisella k-meansilla. Jälkimmäisessä tapauksessa hyödynnetään nivelten korreloituvuutta ja projisoidaan niveldata pienemmälle määrälle pääkomponentteja. Tämä matalampidimensioinen data osioidaan, ja eri osioiden keskipisteitä käytetään puun solmuina. Osiot, joihin ei liity riittävän suurta määrää datapisteitä, jätetään kokonaan huomiotta. [23]

## 4.5 De la Gorcen et al. menetelmä

De la Gorcen et al. sovitus perustuu kohdefunktion paikalliseen optimointiin. Malli on parametrisoitu käden luurangon konfiguraatiolla $\theta$, valaistuksella $L$ ja tekstuurilla $T$. Nämä osat on esitetty tarkemmin luvussa 2.2. Parametrien estimointi tapahtuu kaksivaiheisesti. Ensimmäisessä vaiheessa estimoidaan $\theta$ ja $L$ muokatulla versiolla Broydenin-Fletcherin-Goldfarbin-Shannonin menetelmästä (BFGS) [9, s. 281–284]. Toisessa vaiheessa päivitetään käden pinnan tekstuuri $T$ iteratiivisella painotetulla pienimmän neliösumman menetelmällä (IRLS, *Iteratively Reweighted Least Squares*). [11]

De la Gorce et al. käyttävät ensimmäisessä vaiheessa kohdefunktiota

$$\begin{aligned} E(\theta, T, L) &= \int_{\Omega} R(\mathbf{x}; \theta, L, T) \mathrm{d}\mathbf{x} \\ R(\mathbf{x}; \theta, T, L) &= \begin{cases} \rho(I_{syn}(\mathbf{x}; \theta, T, L) - I_{obs}(\mathbf{x})), \forall \mathbf{x} \in S(\theta) \\ -\log p_{bg}(I_{obs}(\mathbf{x})), \forall \mathbf{x} \in \Omega \backslash S(\theta) \end{cases} \end{aligned} \tag{7}$$

missä $E$ on kohdefunktio, $\Omega$ on jatkuva kuvataso (englanniksi *image domain*), $\Omega \subset \mathbb{R}^2$, $\mathbf{x}$ on piste kuvassa, $\theta$ on käden konfiguraatiovektori, $T$ tekstuurimalli, $L$ valaistusmalli, $R$ mallin kokonaisresiduaalifunktio, $\rho$ jokin residuaalifunktio, $I_{syn}$ synteettisen käden kuva pisteessä $\mathbf{x}$, $I_{obs}$ havaittu käden kuva, $p_{bg}$ taustan todennäköisyysfunktio ja $S(\theta) \subset \Omega$ käden annetulla konfiguraatiolla $\theta$ peittämä kuvatason osa. Residuaalifunktiona $\rho$ de La Gorce et al. käyttivät tavanomaista neliövirhesummaa. Optimointia varten de la Gorce et al. johtavat jatkuvasta kohdefunktiosta diskreetin approksimaation $\bar{E}$, jonka gradientti $\nabla_\theta \bar{E}$ on laskettavissa. Optimointi suoritetaan muokatulla BFGS-menetelmällä. [11]

Kun käsikonfiguraatioparametrit $\theta$ ja valaistusparametrit $L$ on löydetty, samaa kohdefunktiota sovelletaan toisessa vaiheessa käsien tekstuurin $T$ päivittämiseen. Kohdefunktiota on muokattu lisäämällä siihen neliövirhemuotoinen sileystermi $E_{sm}$:

$$\begin{aligned} E_{texture}(T) &= \bar{E}(\theta, L, T) + \beta E_{sm}(T) \\ E_{sm} &= \sum_i \sum_{j \in \mathcal{N}_T(i)} ||T_i - T_j||^2 \end{aligned} \tag{8}$$

missä $\mathcal{N}_T(i)$ on tekselin $i$ naapurusto. De la Gorce et al. minimoivat tämän funktion iteratiivisella painotetulla pienimmän neliösumman menetelmällä. [11]

## 4.6 Buehlerin et al. menetelmä

Buehlerin et al. käsivarsimallin sovittaminen alkaa estimoimalla pään ja torson sijainnit – ja siten myös olkapäiden paikat. Tämän jälkeen luvussa 2.3 esitetty käsivarsimalli ankkuroidaan arvioituihin olkapään paikkoihin ja käsivarsien parametrit etsitään kaksivaiheisesti. Ensimmäisessä vaiheessa käsivarret sovitetaan jokaiseen ruutuun ilman ajallista jatkuvuutta. Tässä vaiheessa yritetään tunnistaa ruudut, joille konfiguraatio voidaan estimoida yksiselitteisesti. Toisessa vaiheessa video käydään ajallisesti läpi etu- ja takaperin, ja näin parannetaan sovitustuloksia niissä ruuduissa, joiden konfiguraatio jäi epävarmaksi. [5]

Pään ja torson sijainnit selvitetään sovittamalla kuvaan kaksiosaisena rakenteena pää- ja torsomallineet. Mallineet ovat binäärisiä ja niillä on neljä vapausastetta: rotaatio, skaalaus ja x- ja y-akselin suuntainen translaatio. Sovitukseen käytetään posteriorijakaumaa, jonka prioritermi pakottaa pään kaulan alueelle ja jonka uskottavuustermi on väritermi, joka on määritelty kuten myöhemmin tässä luvussa esiteltävä kustannusfunktion väritermi. Lopullinen segmentaatio tehdään valitsemalla kaikki mallineet, jotka ovat lähellä MAP-estimaattia (*Maximum A Posteriori*), painottamalla ne ulkonäkötermillä, laskemalla painotettu summa ja kynnystämällä se. Olkapäiden sijainti lasketaan vastaavasti kunkin mallineen lineaarikombinaationa. [5]

Buehler et al. käyttävät luvussa 2.3 esitellyn mallinsa sovitukseen seuraavanlaista kustannusfunktiota:

$$P(\mathbf{L}|\mathbf{I}) \propto P(\mathbf{L}) \prod_{i=1}^{N} p(x_i|\lambda_i) \prod_{j \in \{LF,RF\}} p(y_j|l_j) \tag{9}$$

$\mathbf{L}$ on mallin konfiguraatiovektori ja $\mathbf{I}$ syötekuva. $P(\mathbf{L})$ on konfiguraation prioritodennäköisyys, joka rajaa epärealistiset konfiguraatiot pois. $N$ on pikselien kokonaismäärä ja $x_i$ on havaitun pikselin värivektori. $\lambda_i = \Lambda(\mathbf{L}, i)$ on pikselin $i$ ominaisuudet selittävä mallin osa; $\lambda_i$ määritellään maalarin algoritmilla renderöimällä malli syvyysjärjestyksessä aloittaen pohjalta, okkluusiot huomioiden. $\lambda_i$ voi siis vastata kumman tahansa käden mitä tahansa kolmea osaa – käsiosaa, kyynär- tai olkavartta – tai niiden lisäksi joko päätä, torsoa tai taustaa. Käsiosat muodostavat poikkeuksen siten, että pikselin värivektorin tulee olla riittävän todennäköisesti käsiosan värinen – pelkkä sijainti käsiosan suorakulmaisella alueella ei siis riitä. $y_j$ on vasemmalle (LF) ja oikealle (RF) kyynärvarrelle laskettu HOG-deskriptori (*Histogram of Oriented Gradients*) eli suunnattu gradienttihistogrammi [10]. [5]

Buehler et al. määrittelevät väritermin $p(x_i|\lambda_i)$ eri tavalla etualan osille ja taustalle. Etualan osat määritellään useiden gaussisten jakaumien yhdistelmänä, jotka on opetettu käsin annotoidulla opetusmateriaalilla kullekin osalle erikseen. Taustaa varten he käyttävät RGB-histogrammia, jota päivitetään joka ruudun kohdalla. He käyttivät kokeissaan uutislähetyksiä, jotka kuvaruudun kulmassa esiintyvä tulkki kääntää brittiläiselle viittomakielelle (BSL, *British Sign Language*), joten histogrammin päivitys vastaamaan muuttuvaa taustaa onnistui valitsemalla pikselit viittojan alueen ulkopuolelta. [5]

HOG-termin $p(y_j|l_j)$ Buehler et al. laskevat syötteen $y_j$ ja mallineen $l_j$ välillä katkaistuna $L2$-normina, joka on normeerattu välille $[0, 1]$. Mallineet on opetettu käsin annotoidulla opetusdatalla. Kokeissaan he opettivat mallineet vain kyynärvarsiosille. [5]

Buehler et al. huomioivat käsivarsikonfiguraatioiden ajallisen jatkuvuuden kaksivaiheisesti. Ensin he tunnistavat videosekvenssistä kuvaruudut, joille estimoitu käsivarsikonfiguraatio on yksiselitteinen etukäteen valittujen sääntöjen puitteissa. Heidän suorittamassaan kokeessa 6000 ruudun videosekvenssistä tällaisia ruutuja löytyi 191 kappaletta tasaisesti jakautuneena, huonoimmillaankin muutaman sekunnin välein. Tämän jälkeen he seuraavat peräkkäisten ruutujen jatkuvuutta eteen- ja taaksepäin videosekvenssissä lisäämällä kustannusfunktioon uuden termin $P(\mathbf{L}|\mathbf{L}') = \prod_{k=1}^{n} p(l_k|l'_k)$, missä $\mathbf{L}'$ on ajallisen kulkusuunnan suhteen edeltävän ruudun käsivarren konfiguraatiovektori. Todennäköisyystermin $p(l_k|l'_k)$ arvo on lähellä yhtä, jos osat $k$ ovat lähellä toisiaan ja lähellä nollaa, mikäli muutos on fyysisesti epätodennäköinen tai mahdoton, kuten esimerkiksi valtavan suuri hyppäys kahden ruudun välissä on. Jakauman histogrammin he opettavat automaattisesti erityisellä sekvenssillä, jossa on staattinen tausta, torson ja hihojen värit poikkeavat toisistaan ja joka tarjoaa täten luotettavat sovitustulokset ilman ajallista termiä. [5]

# 5 Laskennallinen vaativuus

## 5.1 Yleistä

Tässä luvussa tutustutaan eri menetelmien laskennalliseen vaativuuteen. Menetelmien tila-avaruudet ovat tyypillisesti niin suuria, ettei niitä voida sovitusvaiheessa käydä kokonaisuudessaan läpi. Aliluvussa 5.2 tutustutaan erilaisiin approksimaatioihin, joita menetelmien yhteydessä käytetään läpikäytävän tila-avaruuden kohtuullistamiseksi. Aliluvussa 5.3 tarkastellaan menetelmien raportoituja suoritusaikoja käytännön kokeissa.

## 5.2 Approksimaatiot

Athitsoksen ja Sclaroffin mallin sovittaminen edellyttää yhtälössä 1 esitetyn suunnatun chamfer-etäisyyden laskemista syötekuvaruudusta jokaiseen mallikuvaan ja päinvastoin. Tämä on laskettavissa mallikuvista syötekuviin tehokkaasti etäisyysmuunnoksen avulla. Mikäli mallikuvia on $d$ kappaletta ja jokaisessa kuvassa $n$ reunapikseliä, aikakompleksisuus on $O(dn)$. [1]

Toisin päin tämä ei kuitenkaan toimi, koska tyypillisessä tapauksessa $d$ on liian suuri eikä etäisyysmuunnettua kopiota voida tallentaa tietokoneen keskusmuistiin jokaisesta tietokannan mallikuvasta. Suora laskennallinen vaativuus $O(dn \log n)$ olisi myös liian suuri. Athitsos ja Sclaroff ratkaisevat ongelman suorittamalla Lipschitz-upotuksen euklidiseen $\mathbb{R}^k$-avaruuteen. Tämä tapahtuu siten, että valitaan $k \ll d$ satunnaista kuvaa $r_1, r_2, \ldots, r_k$ tietokannasta ja määritellään euklidinen avaruus

$$E(g) = (c(g, r_1), c(g, r_2), \ldots, c(g, r_k)) \tag{10}$$

Kun oletetaan, että kahden lähekkäisen pisteen välinen etäisyys kolmanteen pisteeseen on lähes samansuuruinen kummassakin avaruudessa, saadaan likimääräinen chamfer-etäisyys $c'(I, B) = ||E(I) - E(B)||_1$ syötekuvan $I$ ja tietokantakuvan $B$ välille. Aikakompleksisuus pienenee tällöin arvoon $O(kn \log n + dk)$. Athitsos ja Sclaroff käyttivät kokeissaan arvoa $k = 200$. He kokeilivat myös $L_2$-normia likimääräisen etäisyyden laskemisessa, mutta tämä ei parantanut tulosta merkittävästi. [1]

Stenger et al. laskevat yhtälössä 6 esitetyn todennäköisyystermin tehokkaasti integraalikuvana eli ruudun kokoisena kumulatiivisena summataulukkona $\mathbf{B}^{sum}$ x-akselin suhteen. Tämä taulukko tarvitsee laskea vain kerran. Sen jälkeen summat voidaan laskea pinta-alojen yli kulkemalla silhuettia pitkin ja laskemalla taulukon arvojen summia tai erotuksia näissä pisteissä. [23]

Buehlerin et al. järjestelmä on myös laskennallisesti hyvin vaativa. He esittävät, että heidän mallinsa 11 vapausasteen tila-avaruus diskretisoidaan siten, että kyynär- ja olkavarsien skaalaustermit jaetaan 12 tasolle ja nivelkulmat 36 tasolle. Ranteen kulma rajataan 50° alueelle suhteessa kyynärvarteen ja diskretisoidaan 11 tasolle. Näillä parametreilla mahdollisia tiloja on noin $10^{13}$ kappaletta, joten täydellinen tila-avaruuden läpikäynti ei ole mahdollista. [5]

Tämän takia Buehler et al. käyttävät seuraavia kahta tekniikkaa: iteratiivista käsivarsisovitusta ja todennäköisten konfiguraatioiden etsimistä käyttämällä piktoriaalisia struktuureja (*pictorial structures*) [13]. Iteratiivinen käsivarsisovitus toimii kaksivaiheisesti siten, että ensiksi etsitään yhdelle kädelle paras konfiguraatio yksinään. Käsi lukitaan parhaaksi osoittautuneeseen asentoon. Sitten etsitään paras konfiguraatio toiselle kädelle, ja pidetään ensimmäistä kättä koko ajan paikallaan edellisessä vaiheessa

löydetyssä parhaassa asennossa. Lopuksi toistetaan koko operaatio vaihtamalla käsien käsittelyjärjestys. Näin saadaan mahdollisesti kaksi eri konfiguraatioehdotusta, joista valitaan parempi. Tämä vastaa laskennallisen vaativuuden tiputtamista $O(N^2)$:sta $O(N)$:ään. [5]

Piktoriaaliset struktuurit mahdollistavat sopivien konfiguraatioiden löytämisen laskennallisesti tehokkaasti, mutta painottavat todistusaineistoa välillä liikaa; tätä Buehler et al. yrittävät torjua arvioimalla konfiguraatioiden hyvyyttä varsinaisella luvussa 4.6 esitetyllä kustannusfunktiolla. Lisäksi tutkittavien konfiguraationäytteiden valinnassa käytetään hyödyksi seuraavia ominaisuuksia: Näytteistetään max-marginaalijakaumasta tavanomaisen marginaalijakauman asemesta. Värin todennäköisyystermiä muokataan huomioimalla sovellusalan eli viittomakielen ominaisuuksia okkluusioista – kädet ovat aina torson edessä ja kämmenet käsivarsien edessä. Lisäksi Buehler et al. havaitsivat kokeissaan, että tulokset paranivat, jos todennäköisyysjakaumaa terävöitetään. [5]

## 5.3 Kokeelliset suoritusajat

Mikään tarkastelluista järjestelmistä ei ole käyttökelpoinen reaaliajassa. Athitsoksen ja Sclaroffin järjestelmän suoritusaika oli 15 sekuntia ruutua kohti 2000-luvun alun työasemalla. Tämä siitä huolimatta, että heidän käyttämänsä aineiston ruutukoko oli erittäin pieni, $50 \times 80$–$120 \times 160$ pikseliä. [1] Stengerin et al. järjestelmän ajoaika oli noin kaksi sekuntia ruutua kohti 1 GHz työasemalla. Tällöin tosin vapausasteiden määrä oli rajoitettu kuuteen ja mallineita oli generoitu vain 16 055 kappaletta olettamalla käsimuoto jäykäksi. Nivelikkään liikkeen seuraamista varten generoitiin 35 000 mallinetta. Tällöin suoritusaika oli kolme sekuntia ruutua kohti 2,4 GHz työasemalla, mutta nivelten liikettä mallinnettiin kokeessa vain kahdella parametrilla. [23] De la Gorcen et al. kokeissa järjestelmän suoritusaika oli 40 sekuntia ruutua kohti nykyaikaisella työasemalla [11]. Buehlerin et al. Järjestelmän ajoaika oli kokeissa kaksi minuuttia ruutua kohti vuoden 2008 aikaisella työasemalla [5].

# 6 Soveltaminen viittomasignaalin mallintamiseen

## 6.1 Yleistä

Tässä luvussa tutustutaan ensiksi Liddellin ja Johnsonin viittomakielten foneettisen esitysmallin käden sormien konfiguraatioita käsittelevään osaan. Sen jälkeen tarkastellaan luvussa 2 esiteltyjen laskennallisten mallien parametrisointien yhteensopivuutta foneettisen mallin kanssa.

## 6.2 Käden konfiguraatioiden foneettinen esittäminen

Liddell ja Johnson ovat kehittäneet universaalin foneettisen esitysmallin viittomakielten kuvaamiseen. Järjestelmä on kielestä riippumaton, ja se on pyritty laatimaan siten, että sitä voidaan soveltaa minkä tahansa viittomakielen kuvaamiseen. Mallia voidaan pitää hyvin vaikutusvaltaisena ja kattavana esityksenä viittomakielen tutkimuksen alalla. Toistaiseksi uusin versio mallista on julkaistu vuonna 2011.

Foneettinen malli on luonteeltaan tuottamista kuvaava; Liddell ja Johnson mallintavat nimenomaan sitä, miten ihminen tuottaa annetun viittoman eivätkä niinkään sitä, miltä se havainnoitsijasta näyttää. He kutsuvat tietoa siitä, miten käsi tuottaa viittoman, *käden konfiguraatioksi* erotuksena havaitusta *käsimuodosta*. [18] Tämän työn puitteissa keskitytään nimenomaan järjestelmän sormien konfiguraatiota kuvailevaan osaan. Järjestelmän kuuluu muitakin osia, mutta niiden käsittely ei kuulu tämän kandidaatintyön laajuuteen.

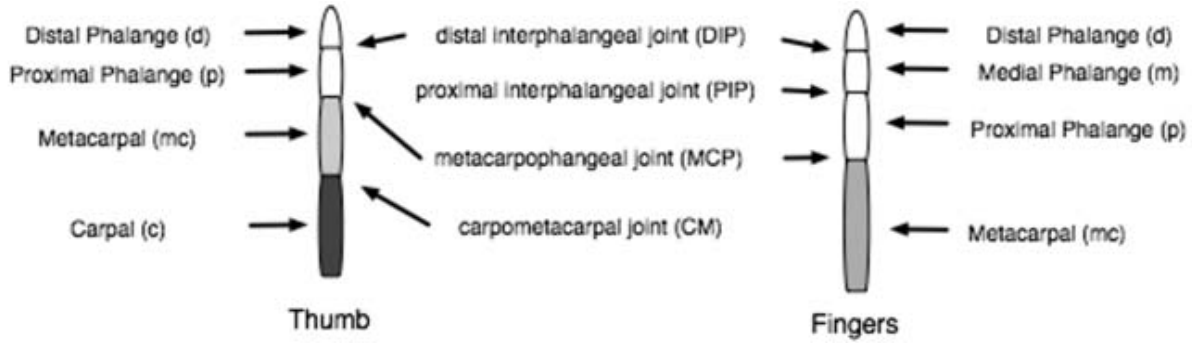
Käsikonfiguraation malli pohjautuu tietoon käsiosan luiden ja nivelten asennoissa viittoman eri hetkinä. Nämä eri osat on esitetty kuvassa 8. Sormet koostuvat neljästä luusta:

- metakarpaaliluusta eli kämmenluusta (*metacarpus*)
- proksimaalifalangiluusta eli sormen tyvijäsenestä (*phalanx proximalis*)
- mediaalifalangiluusta eli sormen keskijäsenestä (*phalanx media*)
- distaalifalangiluusta eli sormen kärkijäsenestä (*phalanx distalis*).

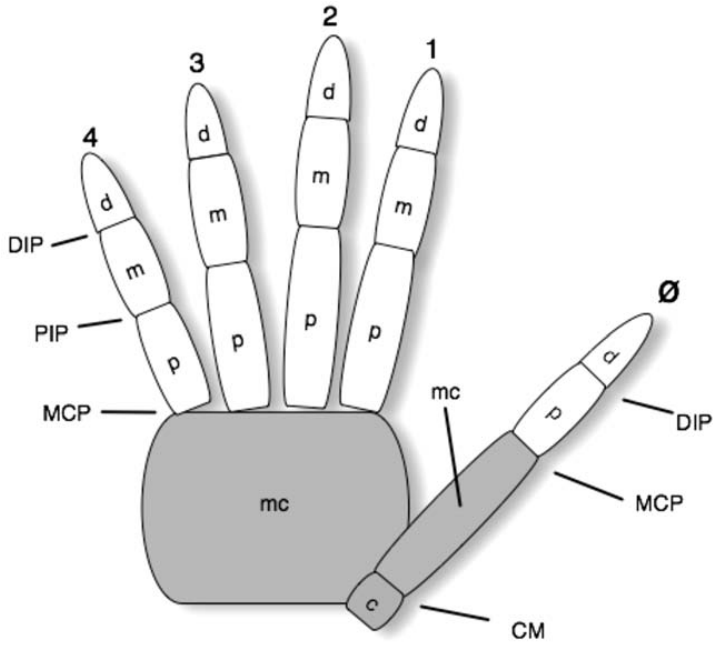
Peukalo poikkeaa muista sormista siten, että siltä puuttuu mediaalifalangiluu. Sormissa on näiden neljän luun välillä kolme niveltä:

- metakarpofalangeaalinivel (MCP, *articulatio metacarpophalangea* eli englanniksi *metacarpophalangeal joint*) eli sormen tyvinivel, joka yhdistää sormen proksimaalifalangiluun metakarpaaliluuhun
- proksimaalinen interfalangeaalinivel (PIP, *articulatio interphalangea proximalis manus* eli englanniksi *proximal interphalangeal joint*) eli sormen keskinivel, joka yhdistää mediaalifalangiluun proksimaalifalangiluuhun
- distaalinen interfalangeaalinivel (DIP, *articulatio interphalangea distales manus* eli englanniksi *distal interphalangeal joint*) eli sormen kärkinivel, joka yhdistää distaalifalangiluun mediaalifalangiluuhun.

Koska peukalossa ei ole mediaalifalangiluuta, peukalolla ei ole PIP-niveltäkään. Peukalon tapauksessa DIP-nivel yhdistää distaalifalangiluun suoraan proksimaalifalangiluuhun. [18]




(a)




(b)

Kuva 8: (a) Liddellin ja Johnsonin käyttämät merkinnät sormen eri luille ja nivelille. (b) Koko käsiosa eri osineen. [18]

Liddell ja Johnson ovat havainneet, että sormen nivelten distinktiiviset ojentuneisuuden erot – eli erot, jotka voivat muodostaa viittomien välisiä minimipareja – voidaan esittää kvantisoituna muutamalle eri tasolle. Kaikille kolmelle nivelelle yhteisiä kvantisointitasoja he nimittävät täydeksi tai osittaiseksi ojentuneisuudeksi sekä täydeksi tai osittaiseksi koukistumiseksi. Lisäksi metakarpofalangeaalinivelellä on kaksi muuta tasoa: täysi ja osittainen hyperojentuneisuus. Näistä eri tasoista he johtavat kolme nivelen tilaa kuvaavaa binääristä piirrettä, jotka ovat nimeltään taivutettuneisuus, hyperojennettuneisuus ja

lihaksen äärirajoilla olo. Jälleen hyperojennettuneisuus koskee vain metakarpofalangeaaliniveltä. Yhden sormen taivutuksella on siis seitsemän binääristä vapausastetta. [18]

Liddell ja Johnson huomioivat sormen taivutuksen lisäksi sormien vetämisen lateraalisesti irti toisistaan eli abduktion ja sormien viemisen yhteen eli adduktion. He erottavat kolme abduktion astetta. Neutraalissa tapauksessa sormet ovat hieman irti toisistaan. Tämän lisäksi sormet voivat olla adduktoituna tiukasti yhteen tai abduktoituna selvästi erilleen toisistaan. Lisäksi he huomioivat hyperadduktion eli sormien viemisen ristiin. Hyperadduktiolle he erottavat neljä eri tilaa: vähäisen ristimisen, ristimisen kosketuskontaktilla sekä ilman kosketusta ja ultraristimisen, jossa sormi viedään pitkälle toisen sormen yli koskettamaan vastakkaista puolta sormesta. Yhdelle vierekkäisten sormien parille on siis määritelty seitsemän eri abduktiotilaa. [18]

## 6.3 Siirtyminen laskennallisista malleista foneettisen esitykseen

Kolmiulotteiset luurankomallit sopivat suoraan käytettäväksi luvussa 6.2 esitetyn mallin kanssa. Liddellin ja Johnsonin ehdottamat likimääräiset kulmarajat nivelten eri taivutusasteille on esitelty taulukossa 1. De la Gorcen et al., Stengerin et al. sekä Athitsoksen ja Sclaroffin käyttämät mallit on parametrisoitu siten, että vastaavat kulmaluvut voidaan lukea niistä suoraan.

Taulukko 1: Liddellin ja Johnsonin kulmat eri nivelten taivutuksille. [18]

| Nivel | Täysin hyper-ojennettu | Osittain hyper-ojennettu | Täysin ojennettu | Osittain ojennettu | Täysin koukis-tettu | Osittain koukistettu |
|---|---|---|---|---|---|---|
| MCP | −45° | −23° | 0° | +30° | +60° | +90° |
| PIP | – | – | 0° | +30° | +60° | +90–100° |
| DIP | – | – | 0° | +10° | +20° | +45–80° |

Liddellin ja Johnsonin malli on tarkoitettu universaaliksi foneettiseksi malliksi, jota voidaan soveltaa kaikkiin viittomakieliin [18]. Mahdollisesti tästä johtuen mallin tila-avaruus on liian laaja, että sitä voitaisiin käyttää kokonaisuudessaan diskriminatiivisten mallien kanssa. Kunkin sormen taivutuksella on seitsemän binääristä vapausastetta, joista tosin kaksi on toisensa poissulkevia, eli käytännössä yksi ternäärinen vapausaste ja viisi binääristä, eli yhdellä sormella on yhteensä 96 laillista taivutusta. Neljälle sormelle saadaan kolme vierekkäistä sormiparia, joista kukin voi olla seitsemällä eri tavalla abduktoituna. Näin ollen jo pelkästään neljällä sormella ilman peukaloa on heidän mallissaan 29 132 587 008 erilaista konfiguraatiota. Tämä lienee jo liian suuri

määrä konfiguraatioita täydellisen konfiguraatiotietokannan laatimiseen. Kohdekielen tuntemuksella voitaisiin kuitenkin päästä mahdollisesti kohtuullisempaan määrään. Kirjallisuudessa on esitetty arvioita, että esimerkiksi suomalaisessa viittomakielessä esiintyisi määritelmästä riippuen noin 30–100 erilaista peruskäsimuotoa [16]. Vastaavasti amerikkalaisen viittomakielen käsimuotojen määrästä on esitetty arvioita ainakin välillä 19–150 [17; 25].

Buehlerin et al. malli ei sovellu tällaiseen käyttöön suoraan. Heidän käyttämänsä käsiosaa kuvaavat HOG-piirteet eivät ole sellaisenaan yhteensopivia tämän mallin kanssa. Tätä mallia voidaan kuitenkin käyttää diskriminatiiviseen viittomien tunnistamiseen [6; 7], kuten he ovat kokeissaan osoittaneet.

# 7 Tarkastelu

Tässä luvussa tarkastellaan tutkittuja menetelmiä kokonaisuutena. Aluksi tutustutaan eri menetelmien hyviin ja huonoihin puoliin menetelmien kehittäjien koetulosten ja heidän esittämiensä omien pohdintojen pohjalta. Sen jälkeen tehdään johtopäätöksiä käsienseurantajärjestelmän kokonaisuuden kannalta. Työssä tutkitut menetelmät on listattu lyhyesti taulukkoon 2.

Athitsoksen ja Sclaroffin menetelmä ei ole itsessään seurantamenetelmä; se ei seuraa käden liikkumista tai muodon muuttumista ajassa, vaan vain löytää käden konfiguraation annetusta yksittäisestä kuvasta. Heidän menetelmänsä ainoa asettama vaatimus on summittainen tieto käden paikasta sovitettavasta kuvasta, joskin arvioiden käsien keskipisteistä ja ko'oista olisi oltava melko tarkkoja. Olisikin kuviteltavissa, että tätä menetelmää voitaisiin siis käyttää toisena, avustavana osana järjestelmässä, jonka toinen osa seuraa kättä ruudusta toiseen. Erityisesti he mainitsevat, että järjestelmää voisi kuvitella käytettäväksi auttamaan seurantajärjestelmää palautumaan käden katoamistilanteesta tai alustamaan seuraimen. [1]

Athitsos ja Sclaroff testasivat menetelmäänsä tietokannalla, joka koostui 26 käsimuodosta generoituna 86 eri kuvakulmasta ja 48 eri asennosta kuvakulmaa kohti. Peruskäsimuotojen tietokantakuvia on nähtävissä kuvassa 1. Yhteensä tietokannassa oli 107 328 kuvaa. Heidän mukaansa menetelmän arvioiminen oli haastavaa, koska testidataa luodessa eri ihmiset arvioivat käden konfiguraation hieman eri tavoin. Sovitusta häiritsi se, että mallin antropometriset parametrit kuten sormien pituudet eivät vastanneet testiaineistossa olevia aitoja ihmisen käsiä täydellisesti. Menetelmä toimi paremmin edestä päin kuvatuille kuville. Rajoitteeksi mainitaan, ettei järjestelmä skaalaudu kuin vain joillekin kymmenille käsimuodoille. [1]

Taulukko 2: Yhteenveto käsitellyistä malleista.

| Malli | Rakenne | Alustaminen | Sovittaminen | Sovitusaika / ruutu[1] |
|---|---|---|---|---|
| Athitsos ja Sclaroff | 3D-luurankomalli käsiosasta | Manuaalinen | Kuvahaku synteettisten kuvien tietokannasta | 15 s |
| Stenger et al. | 3D-luurankomalli käsiosasta | Ei tarvita erikseen | Hierarkinen haku tila-avaruudesta | 2–3 s |
| De la Gorce et al. | 3D-luurankomalli käsiosasta | Manuaalinen | Generatiivisen mallin valaistus- ja ulkoasuparametrien optimointi | 40 s |
| Buehler et al. | Kinemaattinen ketju käsivarsista | Automaattinen | Generatiivisen mallin parametrien iteratiivinen haku renderöimällä malli eri parametrin arvoilla | 2 min |

Stenger et al. testasivat järjestelmäänsä muun muassa videoilla, joissa kuvassa on taustalla paljon häiritseviä, käteen liittymättömiä tekijöitä. He osoittivat kokeissaan, että järjestelmä kykenee selviämään tästä ja pystyy samalla seuraamaan kättä, joka peittää itseään tehdessään kuvatasosta poikkeavia rotaatioita. Lisäksi järjestelmä selviytyi käden katoamisesta ja pystyi palautumaan, kun käsi poistui ja tuli takaisin kuvaan. Kaikissa kokeissaan he rajoittivat käsien vapausasteiden määrää tekemällä oletuksia käden paikasta tai liikkeen laadusta. [23]

De la Gorcen et al. menetelmä ei edellytä opettamista *off-line*. De la Gorce et al. vertasivat omaa järjestelmäänsä Stengerin et al. järjestelmään [23] ja, toisin kuin Stengerin et al. järjestelmä, heidän järjestelmänsä pysyy laskennallisesti kohtuullisena myös ilman konfiguraatioavaruuden rajoittamista ja selviytyy käden itseokkluusioista

[1] Arvot eivät ole vertailukelpoisia, koska tulokset on saatu erilaisilla laitteistokokoonpanoilla.

paremmin. He raportoivat myös tuloksista, joiden mukaan heidän menetelmänsä yleistyy kahdelle kädelle täydellä 28 vapausasteen laajuudella kättä kohti ja selviytyy myös käsien keskinäisestä okkluusiosta. Kvantitatiivisissa kokeissa, joita he tekivät sekä luonnollisella materiaalilla että syntetisoidulla materiaalilla, menetelmän havaittiin monokulaariselle videolle ominaisesti kärsivän syvyysvirheistä; käden sijainti arvioitiin helposti väärin syvyyssuunnassa. [11]

Buehler et al. testasivat menetelmäänsä BBC:n viittomakielitulkatuilla uutislähetyksillä. Kyseinen testiaineisto oli erityisen haastavaa mielivaltaisesti vaihtelevan taustan takia. Materiaalia ei ollut tuotettu erityisesti koneellista analysointia varten, vaan se oli luonnollista. Menetelmä osoittautui heidän tekemissään kokeissa hyvin luotettavaksi. He mittasivat menetelmänsä toimivuutta laatimalla käsin pohjatotuuden 296 kuvaruudulle. Sovitetusta käsivarsikonfiguraatiosta generoitua maskia $M$ verrattiin käsin laadittuun pohjatotuusmaskiin $T$ erikseen kummankin käsivarren osalta. He myös laativat vastaavan maskin, johon he laittoivat yhtä aikaa kummankin käden käsiosat. He määrittelivät konfiguraation oikeaksi, jos pikseleiden leikkausten suhde unioniin $\frac{T \cap M}{T \cup M}$ oli yli 0,5. Kaikkia kustannusfunktion termejä – väriä, HOG-piirteitä ja ajallista seurantaa – käytettäessä heidän menetelmänsä sovitti oikein tällä kriteerillä vasemman käsivarren 91,2 %, oikean käsivarren 99,7 % ja käsiosat 95,6 % ruuduista. Osittain oikean sovituksen kriteeriksi he määrittelivät vastaavan suhteen arvon alarajaksi 0,2. Tällöin vastaavasti vasen käsivarsi, oikea käsivarsi ja käden käsiosat löytyivät 99,7 %, 100 % ja 100 % ruuduista. [5]

Kokonaisvaltaisen liikkuvien kohteiden seurantajärjestelmän kolme avainosaa Yilmazin, Javedin ja Shahin [29] mukaan ovat kohteen havaitseminen, sen seuraaminen ruudusta toiseen ja liikkeen analysointi kohteen käyttäytymisen tunnistamiseksi. Kun lähtökohdaksi otetaan viittomakielen foneettinen analysointi, kaikki tarpeelliset komponentit ovat jo olemassa: Buehlerin et al. ja Stengerin et al. järjestelmät kykenevät löytämään käsiosan kuvasta ilmeisen luotettavasti. Athitsoksen ja Sclaroffin ja toisaalta Stengerin et al. järjestelmät kykenevät alustamaan käden konfiguraation estimointijärjestelmän. De la Gorcen et al. järjestelmä kykenee estimoimaan mielivaltaisen käden konfiguraation kuvaruudusta. De la Gorcen et al. malli on parametrisoitu niin, että estimoidut parametriarvot ovat helposti muutettavissa Liddellin ja Johnsonin foneettiseen esitykseen viittomakielen tutkimusta varten.

Jokaisella järjestelmällä on puutteensa eikä yksikään tässä tutkimuksessa käsitelty järjestelmä kata kaikkia osa-alueita yksinään. Eri järjestelmät kuitenkin täydentävät toisiaan. Suora järjestelmien ketjuttaminen peräkkäin voi kuitenkin olla laskennallisesti liian raskasta. Kaikki järjestelmät ovat jo itsessään niin raskaita, että suurten aineistojen analysointi voi vaatia kohtuuttomasti laskenta-aikaa, mikäli järjestelmät ketjutetaan sellaisessa muodossa kuin ne on tässä työssä kuvattu.

Skaalautuvuus on ongelma erityisesti diskriminatiivisille kolmiulotteisiin käsimalleihin pohjautuville järjestelmille [23; 1]. Ne eivät skaalaudu mielivaltaiselle tarkkuudelle, koska tutkittavasta tila-avaruudesta tulisi liian laaja. Athitsoksen ja Sclaroffin mallin tapauksessa tämä tarkoittaisi, että tietokannasta tulisi liian kookas. Stengerin et al. järjestelmän tapauksessa hakupuusta tulisi niin suuri, ettei sitä voitaisi käydä läpi riittävällä tarkkuudella kohtuullisessa ajassa. Kuitenkin viittomakielten käsimuotojen määristä esitetyt arviot ovat niin pieniä, että kohdekielen tuntemuksen avulla voitaisiin saada aikaan toimiva yksinkertaisen peruskäsimuodon tunnistava järjestelmä. Tämä tukee ajatusta käyttää diskriminatiivista järjestelmää alustavana askeleena, jolloin seurantajärjestelmä saadaan kiinnitettyä videoon selkeän, staattisen käsimuodon näkyessä ruudulla. Käden konfiguraatio siirtymäruuduissa eri viittomien välillä sekä viittoman sisällä käsimuotojen välissä saadaan estimoitua tarkasti generatiivisella mallilla.

# 8 Yhteenveto

Työssä tutustuttiin neljään kehittyneeseen mallipohjaiseen kädenseurantamenetelmään eri näkökulmista. Käsiosan paikan ja käden konfiguraation selvittäminen osoittautui odotetusti haastavaksi tehtäväksi. Tehtävä osoittautui myös laskennallisesti erittäin vaativaksi, eikä mikään tutkituista menetelmistä sovellu reaaliaikaiseen käyttöön.

Kaikki seurantajärjestelmän olennaiset osat ovat olemassa, mutta yksikään menetelmä ei kata kaikkea täydellisesti. Olisikin mielenkiintoista selvittää, miten hyvin erilaisista lähtökohdista toimivat järjestelmät voitaisiin kytkeä toisiinsa ja testata kokeellisesti, kuinka hyvin tällainen yhdistetty järjestelmä soveltuisi esimerkiksi viittomakielisissä videoissa esiintyvien käsikonfiguraatioiden estimointiin. Buehlerin et al. menetelmä tarjoaa toimivan tavan löytää käsiosan paikka kuvaruudusta, ja de la Gorcen et al. menetelmä vaikuttaa lupaavalta varsinaisen käden konfiguraation estimointiin. De la Gorcen et al. menetelmästä puuttuva alustus voitaisiin suorittaa diskriminatiivisilla menetelmillä.

Menetelmät ovat kuitenkin niin raskaita, että olisi aiheellista tutkia laskennallisen vaativuuden keventämistä, mikäli menetelmiä aiotaan hyödyntää esimerkiksi suurten viittomakielisten aineistojen foneettiseen analysointiin.

# Lähteet


[1] V. Athitsos ja S. Sclaroff. Estimating 3d hand pose from a cluttered image. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2003)*, osa 2, sivut 432–439, 2003. doi: 10.1109/CVPR.2003.1211500.

[2] H. G. Barrow, J. M. Tenenbaum, R. C. Bolles ja H. C. Wolf. Parametric correspondence and chamfer matching: two new techniques for image matching. *Proceedings of the 5th international joint conference on Artificial intelligence*, osa 2, sivut 659–663. Morgan Kaufmann Publishers Inc., 1977.

[3] G. Borgefors. Hierarchical chamfer matching: a parametric edge matching algorithm. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 10(6):849–865, 1988. ISSN 0162-8828. doi: 10.1109/34.9107.

[4] Y. Y. Boykov ja M. P. Jolly. Interactive graph cuts for optimal boundary & region segmentation of objects in N-D images. *Proceedings of the Eighth IEEE International Conference on Computer Vision (ICCV 2001)*, osa 1, sivut 105–112, 2001. doi: 10.1109/ICCV.2001.937505.

[5] P. Buehler, M. Everingham, D. P. Huttenlocher ja A. Zisserman. Long term arm and hand tracking for continuous sign language TV broadcasts. *Proceedings of the 19th British Machine Vision Conference (BMVC 2008)*, sivut 1105–1114, 2008. ISBN 978-1-901725-36-0.

[6] P. Buehler, A. Zisserman ja M. Everingham. Learning sign language by watching TV (using weakly aligned subtitles). *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009)*, sivut 2961–2968, 2009. ISBN 978-1-4244-3992-8. doi: 10.1109/CVPR.2009.5206523.

[7] P. Buehler, M. Everingham ja A. Zisserman. Employing signed TV broadcasts for automated learning of British Sign Language. *Proceedings of 4th Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies*, sivut 33–40, 2010.

[8] John Canny. A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 8(6):679–698, 1986. ISSN 0162-8828. doi: 10.1109/TPAMI.1986.4767851.

[9] A. R. Conn, N. I. M. Gould ja P. L. Toint. *Trust-region methods*. Society for Industrial Mathematics, 2000. ISBN 0-89871-460-5.

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

[11] M. de La Gorce, D. Fleet ja N. Paragios. Model-based 3D hand pose estimation from monocular video. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(9):1793–1805, 2011. ISSN 0162-8828. doi: 10.1109/TPAMI.2011.33.

[12] A. Erol, G. Bebis, M. Nicolescu, R. D. Boyle ja X. Twombly. A review on vision-based full DOF hand motion estimation. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2005) – Workshops*, sivu 75, 2005. doi: 10.1109/CVPR.2005.395.

[13] P. F. Felzenszwalb ja D. P. Huttenlocher. Pictorial structures for object recognition. *International Journal of Computer Vision*, 61(1):55–79, 2005. doi: 10.1023/B:VISI.0000042934.15159.49.

[14] M. Grabner, H. Grabner ja H. Bischof. Learning features for tracking. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2007)*, sivut 1–8, 2007. doi: 10.1109/CVPR.2007.382995.

[15] J. Han, G. Awad ja A. Sutherland. Automatic skin segmentation and tracking in sign language recognition. *Computer Vision, IET*, 3(1):24–35, 2009. ISSN 1751-9632. doi: 10.1049/iet-cvi:20080006.

[16] T. Jantunen. Johdanto: näkökulmia viittomaan ja viittomistoon. Teoksessa *Näkökulmia viittomaan ja viittomistoon*, T. Jantunen, toimittaja, Soveltavan kielentutkimuksen teoriaa ja käytäntöä 5, sivut 11–28. Jyväskylän yliopisto, Jyväskylä, 2010. ISBN 978-951-39-3955-7.

[17] S. K. Liddell ja R. E. Johnson. American sign language: the phonological base. *Sign Language Studies*, 64:195–277, 1989.

[18] S. K. Liddell ja R. E. Johnson. Toward a phonetic representation of hand configuration: The fingers. *Sign Language Studies*, 12(1):5–45, 2011. ISSN 0302-1475.

[19] L. Lu ja G. D. Hager. A nonparametric treatment for location/segmentation based visual tracking. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2007)*, sivut 1–8, 2007. doi: 10.1109/CVPR.2007.382976.

[20] Zachary Pezzementi, Sandrine Voros ja Gregory D. Hager. Articulated object tracking by rendering consistent appearance parts. *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA 2009)*, sivut 3940–3947, 2009. doi: 10.1109/ROBOT.2009.5152374.

[21] C. Shan, T. Tan ja Y. Wei. Real-time hand tracking using a mean shift embedded particle filter. *Pattern Recognition*, 40(7):1958–1970, 2007. ISSN 0031-3203.

[22] C. Sminchisescu, A. Kanaujia ja D. N. Metaxas. $BM^3E$: Discriminative density propagation for visual tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(11):2030–2044, 2007. ISSN 0162-8828. doi: 10.1109/TPAMI.2007.1111.

[23] B. Stenger, A. Thayananthan, P. H. S. Torr ja R. Cipolla. Model-based hand tracking using a hierarchical bayesian filter. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28:1372–1384, 2006. ISSN 0162-8828. doi: 10.1109/TPAMI.2006.189.

[24] B. D. R. Stenger. *Model-Based Hand Tracking Using A Hierarchical Bayesian Filter*. Väitöskirja, University of Cambridge, UK, 2004.

[25] W. C. Stokoe. Sign language structure: An outline of the communicative systems of the american deaf. *Studies in Linguistics: Occasional Papers*, osa 8, Buffalo, NY, 1960. Department of Anthropology and Linguistics, University of Buffalo.

[26] K. Toyama ja A. Blake. Probabilistic tracking with exemplars in a metric space. *International Journal of Computer Vision*, 48(1):9–19, 2002. doi: 10.1023/A:1014899027014.

[27] L. Vacchetti, V. Lepetit ja P. Fua. Stable real-time 3d tracking using online and offline information. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(10):1385–1391, 2004. ISSN 0162-8828. doi: 10.1109/TPAMI.2004.92.

[28] Y. Wu, J. Y. Lin ja T. S. Huang. Capturing natural hand articulation. *Proceedings of the Eighth IEEE International Conference on Computer Vision (ICCV 2001)*, osa 2, sivut 426–432, 2001. doi: 10.1109/ICCV.2001.937656.

[29] A. Yilmaz, O. Javed ja M. Shah. Object tracking: A survey. *ACM Computing Surveys*, 38(4):1–45, 2006. ISSN 0360-0300. doi: 10.1145/1177352.1177355.

[30] C. Zhang ja Y. Rui. Robust visual tracking via pixel classification and integration. *Proceedings of the 18th International Conference on Pattern Recognition (ICPR 2006)*, osa 3, sivut 37–42, 2006. doi: 10.1109/ICPR.2006.1019.