HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Automatic assessment of nonverbal interaction from smartphone videos

University of Helsinki
Faculty of Medicine
Logopedics
Master's Thesis
October 2019
Laura Lehtimäki

Supervisor: Satu Saalasti

HELSINGIN YLIOPISTO

Tiedekunta - Fakultet - Faculty	Laitos - Institution – Department	
Lääketieteellinen		
Tekijä - Författare - Author		
Laura Lehtimäki		
Työn nimi - Arbetets titel		
Ei-kielellisten vuorovaikutuspiirteiden automaattinen arviointi älypuhelimella kuvatulta videolta		
Title		
Oppiaine - Läroämne - Subject		
Logopedia		
Työn laji/ Ohjaaja - Arbetets art/Handledare - Level/Instruct	Aika - Datum - Month and year	Sivumäärä - Sidoantal - Number of pages
Pro Gradu/ Satu Saalasti	Lokakuu 2019	51 s + 16 liites.

Tiivistelmä - Referat - Abstract

Ei-kielellisten vuorovaikutuspiirteiden arviointi perustuu nykyään pitkälti tarkkailuun, haastatteluihin ja kyselyihin. Määrällisiä menetelmiä ei juuri ole. Uusi teknologia tuo arviointiin uusia mahdollisuuksia, ja siihen perustuvia arviointimenetelmiä kehitetäänkin jatkuvasti. Monet teknologia-avusteisista menetelmistä perustuvat liikkeen tunnistukseen esimerkiksi sensoreiden, kameroiden tai tietokonenäön avulla. Tässä tutkimuksessa selvitetään mahdollisuutta käyttää asennontunnistusalgoritmia ei-kielellisen vuorovaikutuksen arvioimisessa. Tavoitteena on selvittää, pystytäänkö algoritmin avulla tunnistamaan videolta samat vuorovaikutuspiirteet kuin käsin annotoimalla. Tavoitteena on myös tutkia, mikä on paras tapa annotoida videot tämänkaltaisessa tutkimuksessa.

Tutkimusmateriaali koostui neljästä videosta, joissa lapsi ja vanhempi puhalsivat saippuakuplia. OpenPose-algoritmilla tunnistettiin lapsen ja vanhemman asennot jokaisesta yksittäisestä kuvasta. Näin saadut koordinaatit käsiteltiin edelleen Matlabilla siten, että niistä laskettiin lapsen ja vanhemman aktiivisuudet ja käsien läheisyys jokaisella ajanhetkellä. Videot annotoitiin kahdella eri tavalla. Perusyksiköistä annotoitiin katseiden suunnat ja saippukuplapurkin käsittely. Vuorovaikutuspiirteistä annotoitiin kommunikointialoitteet, vuorottelu ja jaetun tarkkaavuuden hetket. Algoritmin avulla laskettuja tuloksia vertailtiin annotointeihin visuaalisesti.

Kommunikaatioaloitteet ja vuorottelu näkyivät käsien läheisyytenä ja lapsen ja aikuisen aktiivisuuksien vuorotteluna. Vaihtelua käsien läheisyydessä ja aktiivisuuksissa aiheutti kuitenkin moni muukin toiminta kuin vuorovaikutus, joten pelkästään niiden avulla vuorovaikutusta ei voitu erottaa muusta toiminnasta. Kaikki vuorovaikutus ei myöskään liittynyt saippuakuplapurkin käsittelyyn, jolloin se ei näkynyt käsien läheisyytenä. Kuvausjärjestelyistä johtuen algoritmi ei pystynyt tunnistamaan videoista katseen suuntaa, joten myöskään jaetun tarkkaavuuden hetkiä ei pystytty tunnistamaan automaattisesti. Kuvausjärjestelyjä pitäisikin muuttaa niin, että kuvattavien kasvot ovat koko ajan näkyvissä. Tämän kaltaisessa tutkimuksessa kannattaa jatkossa yksittäisten vuorovaikutustekojen arvioimisen sijasta keskittyä laajempiin kokonaisuuksiin kuten synkroniaan vuorovaikutuskumppanien välillä. Paras annotointitapa riippuu tutkimuksen tavoitteesta.

Avainsanat - Nyckelord

Ei-kielellinen vuorovaikutus, automaattinen arviointi, OpenPose, video

Keywords

Säilytyspaikka - Förvaringsställe - Where deposited

Helsingin yliopiston kirjasto – Helda / E-thesis

Muita tietoja - Övriga uppgifter - Additional information

Tiedekunta - Fakultet - Faculty	Laitos - Institution – Department	
Faculty of Medicine		
Tekijä - Författare - Author		
Laura Lehtimäki		
Työn nimi - Arbetets titel		
Title		
Automatic assessment of nonverbal interaction from smartphone videos		
Oppiaine - Läroämne - Subject		
Logopedics		
Työn laji/ Ohjaaja - Arbetets art/Handledare - Level/Instruct	Aika - Datum - Month and year	Sivumäärä - Sidoantal - Number of pages
Master's Thesis / Satu Saalasti	October 2019	51 pp. + 16 appendices

Tiivistelmä - Referat - Abstract

The assessment of nonverbal interaction is currently based on observations, interviews and questionnaires. The quantitative methods for assessment of nonverbal interaction are few. Novel technology allows new ways to perform assessment, and new methods are constantly being developed. Many of them are based on movement tracking by sensors, cameras and computer vision. In this study the use of OpenPose, a pose estimation algorithm, was investigated in detection of nonverbal interactional events. The aim was to find out whether the same meaningful interactional events could be found from videos by the algorithm and by human annotators. Another purpose was to find out the best way to annotate the videos in a study like this.

The research material consisted of four videos of a child and a parent blowing soap bubbles. The videos were first run by OpenPose to track the poses of the child and the parent frame by frame. The data obtained by the algorithm was further processed by Matlab to extract the activities of the child and the parent, the coupling of the activities and the closeness of child's and parent's hands at each time point. The videos were manually annotated in two different ways: Both the basic units, such as the gaze directions and thehandling soap bubble jar, and the interactional events, such as communication initiatives, turn-taking and joint attention, were annotated. The results obtained by the algorithm were visually compared to annotations.

The communication initiatives and turn-taking could be seen as peaks in hand closeness and as alternation in activities. However, interaction events were not the only reasons that caused changes in hand closeness and in activities, so they could not be distinguished from other actions solely by these factors. There also existed interaction that was not related to jar handling, which could not be seen from the hand closeness curves. With current recording arrangements, the gaze directions could not be detected by the algorithm and therefore the moments of joint attention could not be determined either. In order to enable the detection of gaze directions in the future studies, the faces of subjects are visible all the time. Distinguishing individual interaction events may not be the best way to assess interaction, and the focus of assessment should be in global units, such as synchrony between interaction partners. The best way to annotate the videos depends on the aim of the study.

Avainsanat - Nyckelord	
Keywords	
Nonverbal interaction, automatic assessment, OpenPose, video	
Säilytyspaikka - Förvaringsställe - Where deposited	
Helsinki University Library – Helda / E-thesis	
Muita tietoja - Övriga uppgifter - Additional information	

Table of contents

TABLES

FIGURES

1 INTRODUCTION

The novel technology may allow new easier and faster ways to assess behavior and one's abilities and impairments. Also in the field of logopedics, there will be a gradual shift towards new technologies in research and in clinical work. This study examines one possible way to utilize technology when assessing nonverbal interaction.

Nonverbal interaction is an important part of interaction and communication (Burgoon & Bacue, 2003). The nonverbal interaction skills have an essential role when acquiring language and learning social interaction (De Schuymer, De Groote, Bayers, Striano & Roeyers, 2011; Tomasello & Farrar, 1986). Atypical nonverbal interaction may be a sign of language or interaction impairment or disorder, such as autism spectrum disorder (Lee & Schertz, 2019; Zwaigenbaum, Bryson & Garon, 2013). Using current methods, obtaining quantitative data on different interaction traits may be slow and laborious. Novel technology may provide faster and easier ways to perform assessments, and it may also provide methods to identify children at the risk of interaction or language disorder earlier than is currently possible.

Interaction can be assessed for example by using movement tracking. Movements can be tracked by, for example, sensors or computer vision using different algorithms (Goodwin, Intille, Albinali & Velicer, 2010; Hashemi et al. 2014; Mahdhaoui et al., 2011). The OpenPose (Cao, Hidalgo, Simon, Wei & Sheikh, 2018) is an algorithm that recognizes subjects poses from video recordings frame by frame. The aim of this study is to investigate whether the OpenPose algorithm can be used to detect nonverbal interaction events on a smartphone video.

2 NONVERBAL INTERACTION

The development of nonverbal interaction skills precedes language acquisition and have an essential role when learning communication and social interaction skills (Filipi, 2009). Nonverbal interaction remains a significant part of any social interaction also after language acquisition (Burgoon & Bacue, 2003), and for some people with learning disabilities, it may remain the only way to interact and to communicate with other people.

The nonverbal cognitive skills, including for example joint attention, promote the language acquisition, and children's language skills in early childhood correlate with the development of their non-linguistic skills (De Schuymer et al., 2011; Tomasello & Farrar, 1986). Therefore, assessment of nonverbal interaction skills provides valuable information on language development conditions or on symptoms of several disorders, for example autism spectrum disorders (Watson, Crais, Baranek, Dykstra & Wilson, 2013).

2.1 Typical development of nonverbal interaction skills

Development of interaction and communication skills begins already as newborn, when an infant learns the most essential interaction skills from his/her parents. One of the most essential skills in human interaction is turn-taking. The aim to alternate is innate as newborns and even preterm infants produce reciprocal vocalizations with parents (Caskey, Stephens, Tucker & Vohr, 2011). The vocalizations become intentional few weeks after birth (Nathani, Ertmer & Stark, 2006). The shared rhythm in talk and in bodily movement between infant and parent is also learned in early infancy (Holmlund, 1995). As the eye contact develops at about the age of two months (Launonen, 2007), the infant is ready to have proto-conversations with his/her parent (Bateson, 1979). In the proto-conversations parents strengthen infant's turn-taking skills by answering to his/her turns as if they had a meaning and adjusting their own responses to infant's turns (Yilmaz et al., 2015).

As infants get older, they learns more ways to interact. The first smile appears around the age of two months (Messinger & Fogel, 2007). At 7–9 months of age,

children start to use gestures, first in social interaction (e.g. to direct another person's attention to oneself), then to control another person's behavior (e.g. to reach out to an object or to shake head to indicate "no") and finally to joint attention (e.g. pointing and gaze, which is explained later) (Watson, Crais, Baranek, Dykstra & Wilson, 2013). Gestures and development of intentional communication are connected, as by using gestures, the child finds out s/he can have an influence on other people by his/her own actions (Bates, 1976). At the same time as the nonverbal interaction skills are learned, vocalizations change from cooing to babbling and finally to the first words usually at the age of 12–18 months (Kunnari & Savinainen-Makkonen, 2012).

Infants learn to follow parent's gaze usually at the age of 8–9 months (Corkum & Moore, 1998). The gaze following skill precedes the learning of joint attention skills that are learned about at the age of 9–12 months (Carpenter, Nagell, Tomasello, Butterworth & Moore, 1998; Corkum & Moore, 1998). Learning of joint attention means that the infant knows how to share the focus of his/her attention with an other person. The most common example of joint attention is a situation where a child alternates his/her gaze between an object and parent's eyes. When a child has learned joint attention, s/he can also check if another person is involved and use his/her gaze to direct other person's attention to what s/he wants (Carpenter & al. 1998). At that point child also understands that parent has intentions and focuses of attention that differ from his/her own (Tomasello, 1995).

Often the term "joint attention" is used to refer to the triadic joint attention, when two persons are focusing their attention together to the same object. However, also dyadic interaction can be regarded as joint attention (Reddy, 2005). Dyadic joint attention appears when two persons are focusing their attention to each other. In this study the term joint attention is used for both triadic and dyadic joint attention.

2.2 Meaning of nonverbal interaction to development of language and social interaction skills

Alternation is a basic feature of interaction, and turn-taking skills are essential to make the social interaction and communication fluent. The ability to take turns is

a necessary precursor for language acquisition as the child needs to be able to alternate his/her own attempt to produce speech with parent's utterances (Kaye, 1977). According to Lee and Schertz (2019) the turn-taking skills of children with autism spectrum disorder also correlate with the development of joint attention.

Joint attention promotes language development: Children learn new words and new gestures best when they have joint attention with their interaction partners (Morales et al., 2000; Tomasello & Farrar, 1986). The early development of joint attention promotes the development of language (Beuker, Rommelse, Donder & Buitelaard, 2013) and is also related to better social competence and lower levels of externalizing behavior in childhood (Van Hecke et al., 2007). Infants who can follow parent's gaze may have better word comprehension because they can recognize the object the parent is focused on and connect it to the words used (Brooks & Meltzoff, 2008; Baldwin, 1995). Also, the way parent follows child's focus of attention has an effect on language acquisition (Tomasello & Farrar, 1986). If parent follows child's attention and talks about the object the child's attention is focused on, new words are learned better than if parent constantly redirects child's attention.

Pointing gesture is a key feature of joint attention involved in the language development (Colonnesi, Stams, Koster & Noom, 2010). Parents give verbal responses and name objects more often when child has learned and uses the pointing gesture (Kishimoto, Shizawa, Yasuda, Hinobayashi & Minami, 2007). Besides the pointing gesture, also the use of other communicative gestures at the age of 14–15 months predicts language development later in childhood (Kuhn, Willoughby, Wilbourn, Vernon-Feagans & Blair, 2014; Rowe, Özçalıskan & Goldin-Meadow, 2008). The use of gestures supports both lexical and syntactic development, and the changes in gesture use predict also changes in language (Iverson & Goldin-Meadow, 2005).

As nonverbal interaction skills are crucial to the development of language and social interaction skills, it is important to detect atypical interaction traits as early as possible. The earlier the impairments and disorders are detected, the earlier the rehabilitation can begin, and the earlier the intervention starts, the more effective it is (Rogers & Dawson, 2010; Zwaigenbaum et. al., 2013).

2.3 Atypical interaction

The most common interaction disorders are the autism spectrum disorders (ASD). According to the DSM-5 (American psychiatric association, 2013), the international taxonomic and diagnostic tool of mental disorders, the most essential symptoms of ASD are deficits and deviances in social interaction and in both verbal and nonverbal communication. These atypical interaction traits can be detected as early as at the age of 1–2 years (Charman, 2004). One of the first signs of ASD is weakness in joint attention. A child with ASD takes less eye contact, follows less adults gaze and uses less pointing gestures than his/her typically developed peers (Watson et al., 2013). Also the amounts of spontaneous initiations to joint attention, eye contact, gestures and verbal communication are smaller (Winder, Wozniak, Parlade & Iverson, 2013).

Another rather common disorder with interaction deficits and delays in language development, is Williams syndrome (Laing et al., 2002). Children with Williams syndrome are proficient in dyadic but impaired in triadic interaction. This means they are skilled to interact with another people but their triadic joint attention skills are weak. They use fewer pointing gestures and their language development is delayed. This is at least partly due to the difficulties in joint attention. Also developmental disorders may affect in verbal and nonverbal interaction. For example, Down syndrome children make nonverbal requests for objects or assistance significantly less than their typically developed peers (Mundy, Sigman, Kasari & Yirmiya, 1988). This is associated with their deficits in expressive language.

Interaction between a parent and a child can be impaired as well. Attachment, healthy development, learning and well-being are all based on early parent-infant interaction (Davis, 2010). If the early interaction is atypical, it can affect child's skills and general well-being also later in life (Mäntymaa et al., 2003; Murray, Fiori-Cowley, Hooper & Cooper, 1996). The interaction may be atypical for example due to mother's mental health's problems (Murray et al., 1996) including maternal postnatal depression, or to mother's and child's temperament differences (Campbell, 1979).

3 ASSESSING NONVERBAL INTERACTION

The assessment of nonverbal interaction skills is currently based on observations, interviews and questionnaires (Launonen, 2007). The assessment methods used include for example the Infant-Toddler Checklist (Wetherby & Prizant, 2002), The MacArthur Communicative Development Inventories (Fenson & al., 1993), the Symbolic Play Test (Lowe & Costello, 1976) and the Early Social Communication Scales (Mundy & al., 2003). The focus varies slightly from test to test, but is mainly on gaze following, turn-taking, gestures, vocalizations and symbolic play (Laakso et al., 2011; Mundy & al., 2003; Watt, Wetherby & Shumway, 2006). The tests are checklists, that parents or therapists fill out and the results are obtained by scoring answers. Most of the test are qualitative and they rely on therapist's or parents' interpretations when answering to questions.

In interaction research, observing interaction situations and analyzing video recordings are common ways to study both verbal and nonverbal interaction, e.g. with conversation analysis (Antaki & Wilkinson, 2013; Beach, 2013; Dickerson, Rae, Stribling, Dautenhahn & Werry, 2005). Conversation analysis is used to study interaction and communication in naturally occurring conversations (Beach, 2013) instead of studying specific verbal or nonverbal skills in isolation. The research methods based on observations offer qualitative data on everyday interaction, on one's communicative skills and how they are used, and on one's communication disabilities and the ways to compensate them (Antaki & Wilkinson, 2013). They provide a lot of valuable information for example on turns, communication initiatives, responds and repairs (Beach, 2013: Dickerson, et al., 2005). However, because of their workload and the time required, it is challenging to transfer observation based methods to clinical work.

3.1 Automatic assessment methods

Automatic assessment methods are methods that utilize technology to recognize and analyze interaction traits automatically. Assessment of interaction traits using traditional methods may be slow and laborious. Novel technological methods allow new ways to evaluate interaction, and may provide an effective and rapid way to perform assessments (see e.g. Cabibihan, Javed, Aldosar, Frazier & Elbashir,

2016; Gatica-Perez, 2009; Pisharady & Saerbeck, 2015). New automatic assessment methods are constantly being studied. The aim is to develop methods that are faster, more accurate and/or easier to use than the traditional ones. As interaction is a complex entity consisting of, among others, speech, prosody, gestures and gaze, also the methods studied are diverse. Assessments can be made using for example video or audio recordings, different sensors or eye-tracking (Cabibihan et al., 2016).

There are only few existing quantitative assessment methods, especially for those who have limited communication abilities (Saulnier & Ventola, 2012). Using automatic analyzing methods, the valuable knowledge obtained by observations could be better utilized for assessing and diagnosing disorders and for evaluating the effectiveness of an intervention. Automatic assessment methods may also allow assessment of several persons at the same time, which means that the whole interaction event can be better analyzed (e.g. Avril et al., 2014). They could also facilitate and advance the detection of atypical interaction traits. For example, autism spectrum disorders are usually first suspected at the age of 1.5–2.5 years but an accurate diagnosis can usually only be made after prolonged observation and follow-up of the child (Autismikirjon häiriöt, 2019). However, using automatic detection methods, it might be possible to identify and quantify atypical interaction traits from children who are at high risk of ASD earlier than is possible nowadays (Hashemi et al., 2014; Taffoni et al, 2012).

As autism spectrum disorders are the most common interaction disorders, many studies focus on evaluating persons with ASD. For example, the attention and orienting in response to name calls of children with and without ASD was studied by using camera and computer vision (Campbell et al., 2018). Computer vision has been used also to develop a tool to assess visual attention by tracking facial features (Hashemi et al., 2014). Several screening tools using eye-tracking and gaze detection has been studied (e.g. Frazier et al., 2018; Vargas-Cuentas et al., 2017), as well as a method to detect autism spectrum disorders based on acoustic-prosodic analysis of pre-verbal vocalizations of 18 month old toddlers (Santos et al., 2013).

In fields other than ASD, the analysis of audio recordings by automatic speech sounds segmentation has been used for detecting motherese in home videos (Mahdhaoui et al., 2011). Eye-tracking methods has been used for example when studying mother's gaze direction in mother-infant interaction (De Pascalis et al., 2017) and computer vision tools in automatic recognition of facial expressions (Zhu, 2015).

3.2 Assessment methods using movement tracking

Movements can be tracked for example by using sensors like accelerometers (e.g. Goodwin et al., 2011) or 2D or 3D cameras and computer vision (e.g. Campbell & al., 2018; Hashemi & al., 2014).

Leclère and colleagues (2016) have proposed a method to measure different features of interaction and to study dyadic behaviors. They studied quality of early interaction during free play four-minute video sessions using automatic measures of individual and dyadic motion features. They had two groups of mothers and their 12–36 months old children: a group of 10 dyads with mothers showing emotional neglect and a control group of 10 dyads with normal development and without interactional difficulty. Kinetic cameras were used to obtain two-dimensional and three-dimensional data. From the 2D and 3D images obtained, several individual (e.g. quantity of movement) and dyadic parameters (e.g. head distances, time spent face to face, synchrony ratio) were extracted. As a result, Leclère and colleagues could classify 100 % of the dyads correctly in either control dyads or in dyads with mother showing neglect. The method is not fully automatized as it requires preprocessing to obtain 3D space reconstruction from the saved data.

Gesture recognition methods have been utilized for example in Praxis test, a test used when diagnosing cortical pathologies such as Alzheimer's disease (Negin et al., 2018). To evaluate dynamic and static gestures the method uses body part data obtained from RGB-D gesture videos. Also gesture recognition systems for the detection, segmentation and recognition of hand postures against natural backgrounds has been developed (Pisharady, Vadekkepat & Loh, 2013). The method is based on image features like shape, texture and color. The recognition rate of the algorithm was 94.4 %

The OpenPose algorithm (Cao et al., 2018) is a new open source software to recognize poses of multiple persons on a video. Applications taking advantage of it for different purposes are constantly being developed. It has for example been used to estimate the attention level of participants in a multi-person interaction scene (Komiya, Saitoh and Shimada, 2018). In eight simulated interaction scenes three participants were seated around a round table, and on the center of the table was placed an omnidirectional camera. The attention level of participants was calculated by estimating the head position using the horizontal direction of yaw, and estimating the gaze direction by detecting the eye center point. As the eye image size was small and often too unclear for OpenPose, the gaze direction was also estimated using convolutional neural network (CNN) based eye center point detection. The overall accuracy of the proposed method was 73 %. The biggest source of error was the small size and unclearness of the eye image.

The OpenPose algorithm has been used to find a way to infer attention automatically from videos of parents implementing Pivotal Response Treatment (PRT) for their child (Heath, Venkateswara, McDaniel and Panchanathan, 2018). The research material consisted a total of 14 videos from seven parent-child pairs, two videos from each pair, implementing PRT in free play situation. In the videos, the activities of child and parent varied from playing with toys and playing games to moving about the room and watching videos. Poses of the parent and the child were extracted using the OpenPose algorithm. Also gazes were estimated using the facial key points. However, due to several reasons, the results were not very promising. The OpenPose algorithm identified only 66 % of the body key points with an average confidence of 56 %. The confidence of facial point recognition was only 23 %. The poor body and facial key point recognition was mainly due to the poor quality of the videos, the distance of subjects from the camera, the fact that time to time subjects had his or her back to camera and/or that the clothing did the identification more difficult. The two-dimensionality of the videos made it difficult to detect the target of the gaze of the subjects.

The OpenPose algorithm has also been used to study rapport in groups of 3–4 persons from nonverbal behavior, including facial expressions, hand motion, gaze, speaker turns and speech prosody (Müller, Huang & Bulling, 2018). The

interaction of 22 groups was recorded using multi-view system consisting of 8 cameras placed behind and above each participant. The OpenPose algorithm was used to extract hand poses and motions from videos. The amount of hand movements for each participant and the synchronization of hand movements between participants were computed. When estimating the low rapport, the facial expression data extracted using OpenFace algorithm (Baltrušaitis, Robinson & Morency, 2016) performed best (average precision (AP) = 70 %), but the performance of hand motion and synchrony data extracted using OpenPose (AP = 44 %) also outperformed the baseline (AP = 25 %).

In other fields than interaction, the OpenPose algorithm has, for example, been used in assessment of Parkinson's disease (Ajay et al., 2018; Li et al., 2018), as a part of bipolar disorder classification framework (Yang et al., 2018) and to capture the key aspects of infant's general movements (Marchi et al., 2019). Li and colleagues (2018) captured ordinary two-dimensional videos of a Timed-Up-and-Go test (TUG), a test of basic functional mobility for people with Parkinson's disease. There were 24 participants with Parkinson's disease involved in the data collection. Each participant underwent 4–6 TUG tests, and the total amount of video sequences was 127. The data was analyzed using the OpenPose and the Iterative Error Feedback, another pose estimation algorithm, and the sub-tasks were classified using two machine learning models. The method using the OpenPose algorithm had an average accuracy of 93 %. The bipolar disorder classification framework of Yang and colleagues (2018) used the OpenPose to detect upper body movements and hand gestures. The amount of hand movements distinguished patients in the remission phase from those in the manic or hypomanic phase.

Marchi and colleagues (2019) assessed the general movements of infants aging from 8 to 17 months in order to recognize the infants in risk of cerebral palsy. The research material consisted of videos of 21 infants, of which 14 had typical movements and 7 had atypical movements and were later diagnosed with cerebral palsy. The poses were estimated using the OpenPose algorithm that was enhanced with a software customized to better estimate infants body-to-limb pro-

portions and movement ranges that differ from adults. Two experts rated the original videos and the skeleton videos obtained by the algorithm. Several video clips had to be removed, because the algorithm had not recognized all poses right. The main reason for this was the suboptimal quality of the videos. Finally, only seven videos were included in the assessment, four with typical and three with atypical movements. From these videos, the spatial distribution of key points was computed and infant's movements were assessed. The wrist-related movements were found to be the best classifier between the two groups.

Some researchers have combined two-dimensional pose estimations obtained by the OpenPose to three-dimensional RGB-D videos to get three-dimensional pose estimates. The 3D pose estimates have been used to develop a 3D Skinned Multi-Infant Linear body model (SMIL) (Hesse et al., 2018). The model was successfully applied in General Movements Assessment, a method used for early detection of neurodevelopmental disorders in infants. A method to estimate child's engagement during child-robot collaboration using the OpenPose and RGB-D camera has also been proposed (Hadfield et al., 2018).

3.3 Challenges with automatic assessment methods

Many of the automatic assessment methods require specific equipment like RGB-D cameras or sensors, or the assessment need to be done in very controlled laboratory settings or in other specific place (e.g. Hesse et al. 2018; Leclère et al., 2016; Negin et al., 2018). If many special devices are needed, measurements can be difficult or expensive to perform. If the examination is made in a laboratory by unfamiliar adults, a child may be shy of them, stress or feel uncomfortable, and thus not perform in the best possible way. Therefore, it would be ideal if the tests could be made at home with parents or other familiar people, at a time when the child is not tired or hungry but in a good mood.

The use of wearable sensors can disturb children, and they may refuse to use them or they may fiddle sensors and thereby cause errors in the measurements (Rodrigues, Gonçalves, Costa, & Soares, 2013). According to DSM-5 (American

psychiatric association, 2013), sensory abnormalities including tactile hypersensitivity is one of the diagnostic criteria of ASD, and thus especially children with ASD may experience wearable accessories distracting.

The OpenPose algorithm looks promising, as it does not require laboratory measurements or use of any specific equipment. However, in some situations it has been found to have challenges in detecting people correctly (Ajay et al. 2018; Heath et al., 2018; Komiya et al. 2018; Marchi et al. 2019).

In the study of Heath and colleagues (2018) the biggest problem was that some body key points were occasionally missing because of occlusion. The problem of missing body key points was solved by estimating the location from a set number of frames or using the last known location. If the facial points were missing, the gaze direction was estimated using eye, nose and neck values from the body key point set. Probably due to the occlusion, OpenPose often recognized only one person in a video frame instead of two. There were also significant amount of frames where OpenPose recognized three or more persons instead of only two. This may be due to OpenPose incorrectly recognizing objects in the background as humans.

Also Ajay and colleagues (2018) mention occlusion when reporting problems when using OpenPose. In addition, they mention the problem of moving camera, if the position of the camera is not fixed. The changes in viewing angles have to be normalized somehow for meaningful analysis. Also the distance from the camera is critical. If the person on the video is too far away from the camera, the skeleton extraction is not accurate enough. The suboptimal quality of videos is also a possible source of error (Heath et al., 2018; Komiya et al., 2018; Marchi et al., 2019), as was mentioned earlier.

3.4 Annotation

To enable the analysis of video or audio recordings, they need to be transcribed or annotated (Ochs, 1979). The choice of the schema used depends on the aim of the study, on the type of the analysis and on the material (Wagner, Malisz & Kopp, 2014). When annotating gestures, the schema may be focused either on

form or on function or take into account both of them (Jenks, 2011; Wagner, Malisz & Kopp, 2014).

Transcriptions and annotations are representations of interaction or an event (Green, Franquiz & Dixon, 1997), and therefore they are subjective and dependent on transcriber or annotator who decides what features to include or exclude from the transcription (Tilley, 2003). Therefore, annotations of different annotators may differ both in timing (the start and end points of an event) and in interpretation of events.

Annotations are currently used a lot in the interaction research. In addition to manual annotations also automatic annotations are used, and a study may also combine manual and automatic annotations to make the most of both (e.g. Beugher, Brône & Goedemé, 2018; Delaherche et al., 2013; Kumano, Otsuka, Ishii & Yamato, 2017). The complex behavior is more easily recognized by a human annotator than by machine, while the automatic annotation may well be used when annotating simple or frequent movements or actions in order to speed-up and simplify the annotation process (Bianco, Ciocco, Napoletano & Schettini, 2015; Vondrick, Patterson & Ramanan, 2013). For example, Delaherche and colleagues (2013) combined manual and automatic annotation when assessing the use of speech, hand and head gestures and gazes of children with and without autism spectrum disorder. Participants hand trajectories were tracked automatically with coupled Camshift algorithm (Bradski, 1998) and the gestural turn-taking features were calculate based on the hand velocities. The more complicated interactive traits, such as utterances, conventional gestures and pointing gestures, of child and therapist were manually annotated and divided into six different categories depending on their interactional role. Also echolalia and stereotypic movements were annotated. Based on the manual annotations, features such as the durations of speaking and pause segments, percentage of interactional time and synchrony between vocal features and gestures were calculated. As a combination of the results obtained from the manual and automatic annotations, the features characterizing therapist's gestural rhythms and the duration of gestural pauses were found to discriminate the two groups best.

The manual annotation data can also be used as input for algorithms and automated processes (Mathur, Poole, Peña-Mora, Hasegawa-Johnson & Contractor, 2012). In a study of interaction links among group members, verbal and nonverbal interaction of participants were manually categorized in four groups: 1) directions of gaze, 2) conversational distance, 3) body posture, gestures and other nonverbal cues and 4) vocalics and verbal cues. The manually annotated data was used as input in a machine learning process determining interaction links between participants. The validity of the automated process was assessed by comparing the automatically detected links with manually detected ones, and it was found to be adequate.

The annotations can also be used to control reliability of an automatic method. In a study on stereotypic movements, a wearable accelerometer tracked the movements of the participants (Gilchrist et al., 2017). The movements were recorded on video and the video recordings were annotated for repetitive behaviors. An algorithm processed the accelerometer signals, and the results were compared to annotations to check the reliability of the method. Also in a study of detecting attention of children with autism spectrum disorder in PRT videos (Heath et al., 2018), annotations were used as a reliability check. The videos were split into segments of 30 frames, which were labeled as “attentive”, “inattentive” or including “joint attention”. The reliability of the proposed assessment method was checked by comparing the results of the method to the annotations.

4 AIMS OF THE STUDY

This master's thesis is part of the Quantifying Interaction project, a study of Helsinki University and Aalto University to examine the possible applications of a pose estimation algorithm. The principal investigator is Dr. Satu Saalasti (Department of Psychology and Logopedics, Medical Faculty, University of Helsinki) and the responsible data scientist is Dr. Enrico Glerean (Aalto University).

The aim of this study is to find out whether the same interactional events can be found from a video recording by the algorithm and by human annotators. Another purpose is to examine how annotations and automatic movement tracking are related, and what kind of information is possible to get by using the algorithm and by using human annotators.

The specific research questions are as follows:

1. Is it possible to recognize the same meaningful interactional events from a video by the OpenPose algorithm and by human annotators?
2. What kind of information is obtained by using the algorithm and by using annotations?
3. What is the best way to annotate the videos in a study like this?

5 RESEARCH METHODS AND MATERIAL

5.1 Research material

The material of this study consists of four videos (duration 1–2 minutes) of dyadic interaction between a parent and a 1–3 years old typically developed child. All videos are recorded at home using a smartphone. Parents were given written instructions i.e. the play videos were semistructured (see Appendix 1).

In the videos, the parent blows soap bubbles and then closes the soap bubble jar so tightly the child can not open it by him- or herself. Then the parent places the jar on the floor between them and waits the child to take the initiative.

5.2 Data processing and analysis

5.2.1 Annotations

The videos were annotated by three researchers with speech and language pathology bachelor, masters and doctorate education. The annotations were made using ELAN (ELAN, 2018), a tool to annotate video and audio resources. Two different types of annotations were made: Basic unit annotation and annotation of interaction events. The term basic unit annotation stands for the annotation of the form, or, single movements and gaze directions. Annotation of interaction events means annotation of the function, or, interactional events such as communication initiatives, turn-taking and joint attention. Basic units were annotated in order to check the reliability of algorithm's movement tracking. Annotations of interaction events were used to identify meaningful interaction moments

Interaction events were annotated by marking the beginning and ending points of communication initiatives of the child, turn-taking and joint attention. Annotations of joint attention included both dyadic and triadic joint attention. Only the nonverbal interaction was annotated as the algorithm used in the study can not detect the vocal utterances. The annotation instructions are presented in Appendix 2. Two annotators agreed with the results.

One annotator annotated basic units: the gaze directions and the handling of the soap bubble jar of both child and parent. Consensus on the annotations was sought by one of the other annotators.

5.2.2 Pose estimation by OpenPose

In this study, the OpenPose algorithm by Cao and colleagues (2018) was used to recognize poses of all persons on an image, or, a video frame. It recognizes the poses frame by frame, so the method is very precise, although a lot of data is accumulated.

The use of the OpenPose algorithm does not require any specific devices or conditions. As input, it takes regular videos recorded, for example, on a smartphone at home by parents. The algorithm recognizes two-dimensional poses of all subjects in each frame by recognizing the determined key points (25 body key points including the major limb joints, eyes, nose, neck and feet, 70 facial and 2x21 hand key points). In this study, the 25 body key points were used. They are presented in Figure 1.

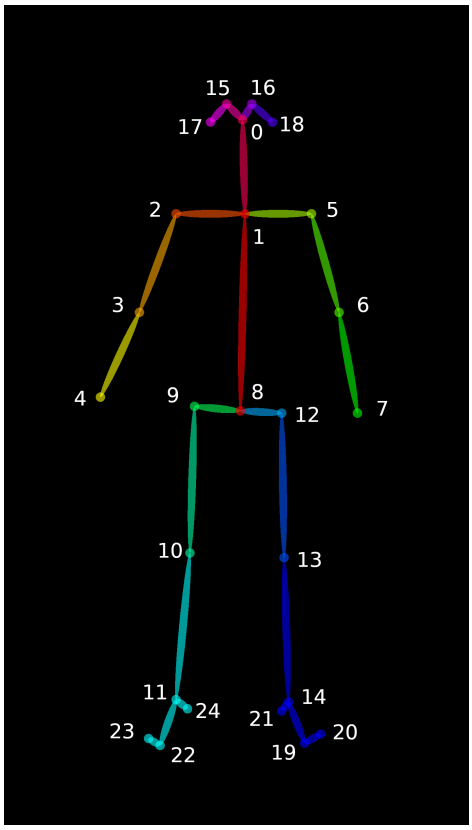


Figure 1. The OpenPose body key points

The key point recognition is done by predicting a set of 2D confidence maps of body part locations and a set of 2D vector fields (part affinity fields, PAFs), that encode the degree of association between parts. To define the 2D key points of all subjects on a frame, the confidence maps and the PAFs are parsed by greedy inference. By tracking the key points, it is possible to monitor subjects' poses and movements on a video.

Preprocessing

The videos used in this study were run using the OpenPose algorithm in Department of Neuroscience and medical engineering due to computational resources. As result a json file for each frame was obtained. A json file contains the x- and y-coordinates and the detection confidence of each key point in the frame in question. The json files were further processed with Matlab (MathWorks Inc.). The code related to analysis is available at GitHub "https://github.com/eglerean/quid". First, all json files of one video were loaded and converted into four matrixes, two for each subject. One matrix contained the x-coordinates and the other y-coordinates of key point locations over time.

The key points that were present in less than 90 % of frames were excluded. Also the first and last 2 seconds of each video were excluded from the analysis because of missing key point values and possible distractions in the beginning and at the end of the video. To get rid of noise, Savitzky-Golay filtering (Savitzky & Golay, 1964) was performed with filtering window of half second.

Challenges and solutions

Vertical videos are problematic for the algorithm, but three of the four videos were shot vertically. In order to fix the issue, x- and y-coordinate values were swapped. Furthermore, the algorithm occasionally mixed the key points of the child and the parent. Due to the semistructured play situation, parent was always on the other side of the frame and child on the other side, and the problem with mixing key points was overcome by determining that the key points on one side of the frame belong to one person and the key point on the other side belong to the other person. Furthermore, all the key points were not visible in every frame. The locations of missing key points were estimated from their location in previous frames

by linear interpolation. The non-numeric values were removed in order to make the interpolation and later filtering and plotting work properly.

The gaze directions could not be determined by algorithm with these recording arrangements. As the gaze direction is essential when recognizing joint attention, the joint attention could not be recognized by the algorithm either. Therefore, the gaze direction and joint attention are excluded from the results. The limitations are discussed more in detail in chapter 7.

Total activities, coupling and hand closeness

When all the preprocessing was done, the key point distances in pixels from the previous location were calculated. From the sum of the distances the **total activities** of child and parent were calculated at every time instance. The total activity tells how much the person's body parts have moved altogether compared to the previous frame. The **coupling** of the activities was calculated by multiplying the activities. The value is high when both child and parent are active at the same time, and at its lowest when neither of them is moving. The values are relative and therefore not comparable between videos.

Distances between the four hand pairs (child's left hand–parent's left hand, child's left–parent's right, child's right–parent's left and child's right–parent's right) were determined. Then the overall closeness of all the hands, or **hand closeness,** was calculated from the hand pair distances. The hand closeness is a relative value, and not comparable between different video recordings.

5.2.3 Analysis

The analysis was mainly qualitative, as the aim was to find out, what could be found out from videos using automatic methods and to see if the meaningful moments of interaction could be identified.

A quality check of the algorithm results was made by the author by comparing the peaks of hand closeness curves visually to corresponding video frames. The reliability was further checked by comparing the hand closeness curves to the annotated moments of interactional jar handling, such as child giving the jar toward parent and parent reaching toward child or taking the jar from the child.

The graphs of the activities of parent and child, the coupling, and the hand closeness graphs were visually compared to the annotations of interaction events. For this reason, figures containing both annotated interaction events and results obtained by the algorithm were made. The aim was to find out if, based on the changes in the curves, the interactional events can be differentiated from other activities.

5.3 Ethical aspects

The Quantifying Interaction project has been subject to a prior ethical evaluation by Helsinki University Ethical Review Board in the Humanities and Social and Behavioural Sciences. The photographs in this thesis have been edited to conceal the identity of the subjects.

6 RESULTS

6.1 Annotations

Interaction events were annotated by two annotators. The two annotators agreed with the results. Each video contained 2–3 communication initiatives by child, 2–3 turn-taking situations and 2–5 moments of joint attention. The annotated interaction events are presented in Appendices 3–4.

Basic unit annotation contained parent's and child's gaze directions and jar handling. Parent's jar handling was divided into reaching toward and taking the jar, holding the soap bubble stick, blowing bubbles, opening and closing the jar and placing it on the floor (Table 1). Child's jar handling annotations contained taking the jar, holding it, trying to open it, showing it to the parent and giving it to the parent. The gazes were directed to the jar, to the stick, to the bubbles, to other person, to the camera, straight ahead or on the side. The annotations of basic units (Table 1) are presented in Appendices 5–8.

Table 1. Basic unit annotations

Gaze direction	Jar handling, parent	Jar handling, child
to other person	reaching toward the jar	taking the jar
to the bubbles	taking the jar	trying to open the jar
to the jar	opening the jar	holding the jar
to the stick	closing the jar	showing the jar
to the camera	placing the jar on the floor	giving the jar
straight ahead	holding the stick	
on the side	blowing bubbles	

The annotations of gaze and joint attention were not reliable due to the challenges in the recognition of the gaze direction.

Comparing the basic unit annotations and the annotations of interaction events

In order to find out the overlap of the basic unit annotations and the annotations of interaction events they were compared to each other. The moments of *dyadic*

joint attention could be seen in basic unit annotations as moments, when the child and the parent looked at each other at the same time. Most of the *communication initiatives* and *turn-takings* were seen either as child giving the jar to parent or as parent reaching the jar. However, all the communication initiatives and turn-taking were not related to jar handling (e.g. pointing, kissing), so they were not visible in the basic unit annotations.

The *turn-taking* was described more in detail in basic unit annotations. In basic unit annotations all jar handling related turns were visible, while in the interaction event annotations, it was only annotated that turn-taking is happening but not who is active and how many turns the turn-taking situation included. Besides the interaction related movements, the basic unit annotations included a lot of additional information, like holding the soap bubble stick and blowing the bubbles.

6.2 Algorithm

The OpenPose algorithm (Cao & al., 2018) recognizes the key point locations for all persons on a video frame. One video frame with detected key points is presented in Figure 2. The body key points that were present in less than 90 % of the frames were excluded from the following calculations. In video 1 four key points, in videos 2 and 4 three key points and in video 3 eleven key points were excluded. The excluded key points were typically the key points of feet or the ear that was facing away from the camera.

Figure 2. A video frame with body key points detected with the OpenPose (Cao et al., 2018). Subjects have given permission to use the picture.

6.2.1 Reliability of the results obtained by the algorithm

The reliability of the results obtained by algorithm was confirmed by comparing the hand closeness curves with the videos. In Figures 3–4 are presented the hand closeness curves of the videos 2 and 4 and screenshots from the corresponding videos at the moments when the hand closeness was high. As can be seen from these figures (3–4), the algorithm tracks the movements quite reliably. Both the jar handling and the random hand closeness is seen from the screenshots. As the hand closeness is a result of the closeness of all four hand pairs, it can be high even though the hands are not touching but also if all four hands are close to each other. Even though the hand closeness tracking is quite reliable, in video 4 there can be seen some false hand closeness peaks. At 6.9 s and at 50.5 s the hands of the parent and the child are not close to each other, contrary to what the curve implies. As can be seen from Figure 5, the algorithm has not detected the key points right at these time instances. Same applies to videos 1 and 3. In video 1 there are false peaks at 77–80 s and from 87 s on and in video 3 at 12.5 s, 13.2 s, 47 s and 52.2 s. At these moments, the key points are either detected wrong or they are missing because of occlusion or because they are outside the image.

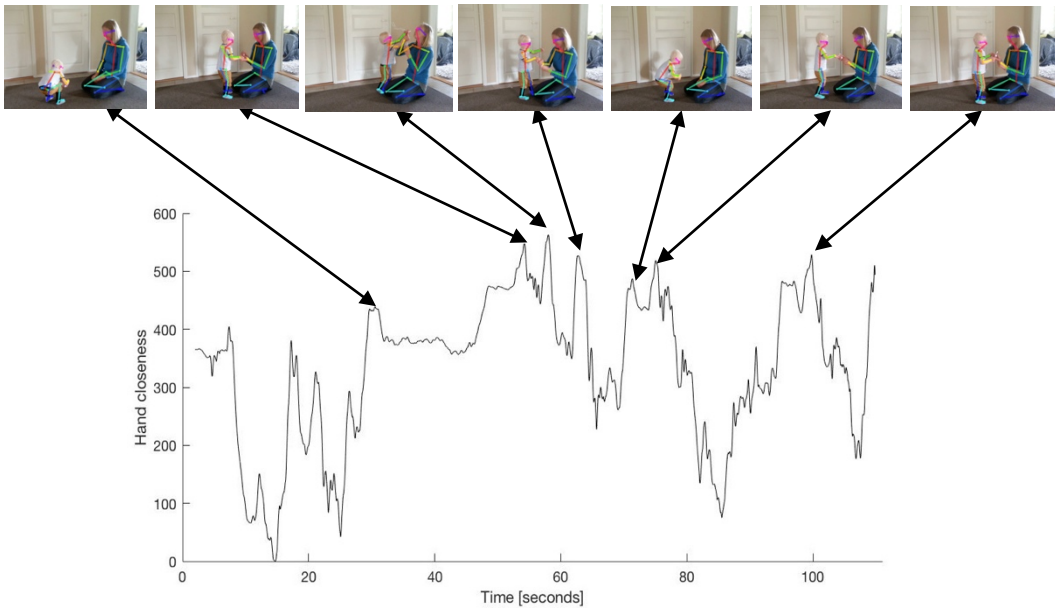


Figure 3. Video 2: Hand closeness curve and screenshots at time instances when the hand closeness is high. Faces of the subjects have been blurred to conceal their identity.

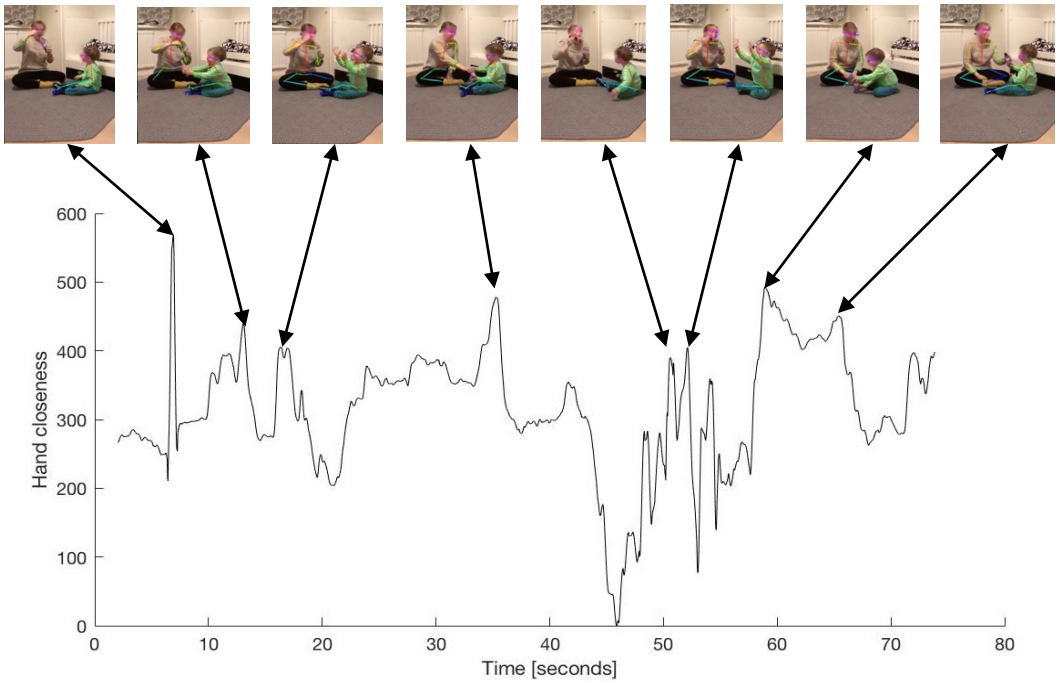


Figure 4. Video 4: Hand closeness curve and screenshots at time instances when the hand closeness is high. Faces of the subjects have been blurred to conceal their identity.

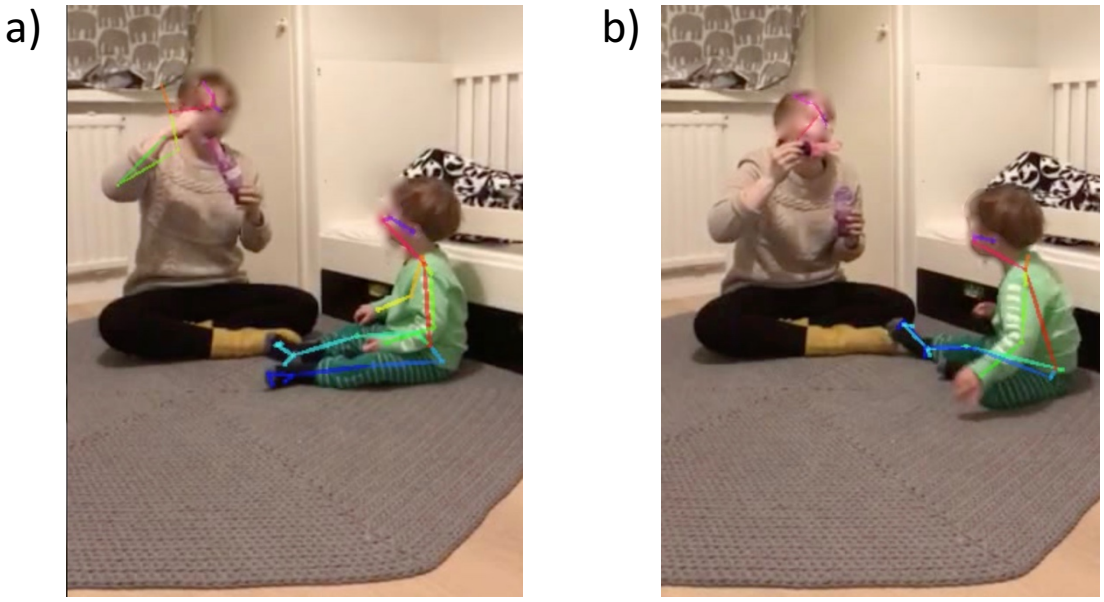


Figure 5. Video 4: Key points detected by OpenPose (Cao et al., 2018) at 6.9 s (a) and 50.5 s (b). Faces of the subjects have been blurred to conceal their identity.

The hand closeness curves were also compared to the basic unit annotations. In Figure 6 are presented the hand closeness curves and the annotated moments when the child hands the soap bubble jar to the parent. The hand closeness is high at these moments. The variation in peak heights is due to the fact that the child and the parent may use either one or two hands to give and to take the jar, and the overall closeness is a result of the closeness of all four hand pairs. As there is happening much more than just handing over the jar, the curves track also events that are not related to interaction or to jar handling, for example when the child catches bubbles near the parent.

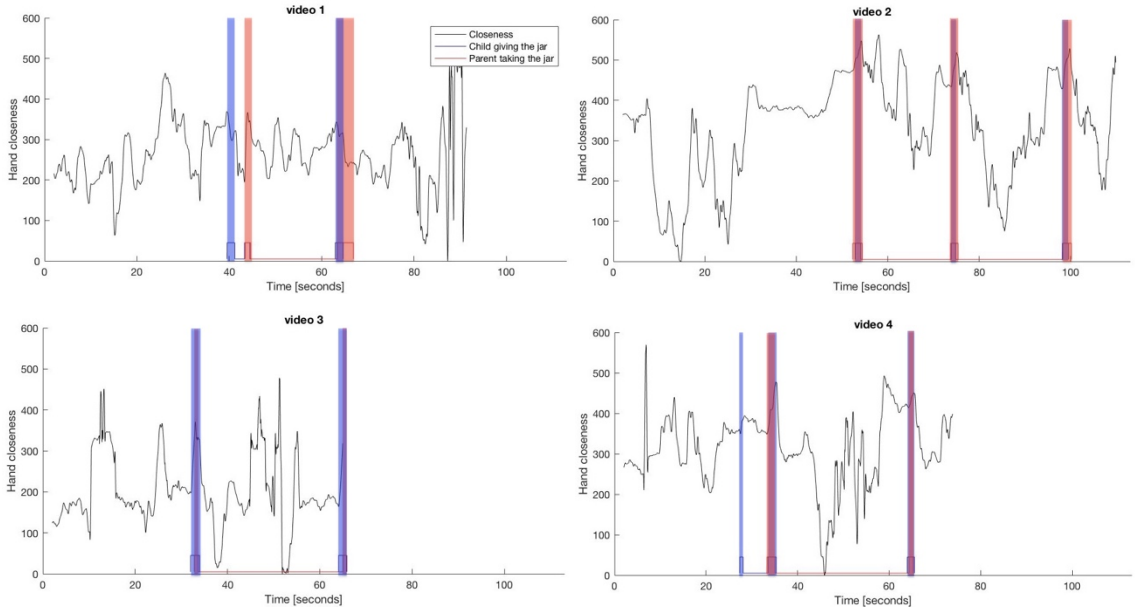


Figure 6. Overall hand closeness by algorithm (black) and jar handling annotations (red and blue). Annotations overlapped to indicate meaningful events of interaction. The higher the hand closeness curve, the closer the hands are.

6.2.2 Activities, coupling, hand pair distances and hand closeness

The **activities** of child and parent and the **coupling** of the activities are presented in Figure 7 and in Appendices 9–12. In videos 1 and 3 the child is sitting still and reaches bubbles only by stretching out his/her hands. The **activity** of the child in these videos is low and with no high peaks. In video 4 the child first sits still but starts moving at 45–55 s which can be seen from the curves as higher activity level. The child in video 2 is standing and runs after bubbles, and therefore the activity curve of the child is almost all the time at higher level than in the other

videos. The parent in video 1 is more active, and as the child is calm, the activity curve of the parent is on average at higher level than the activity of the child. In the three other videos, parents are monitoring children's activities more sedately.

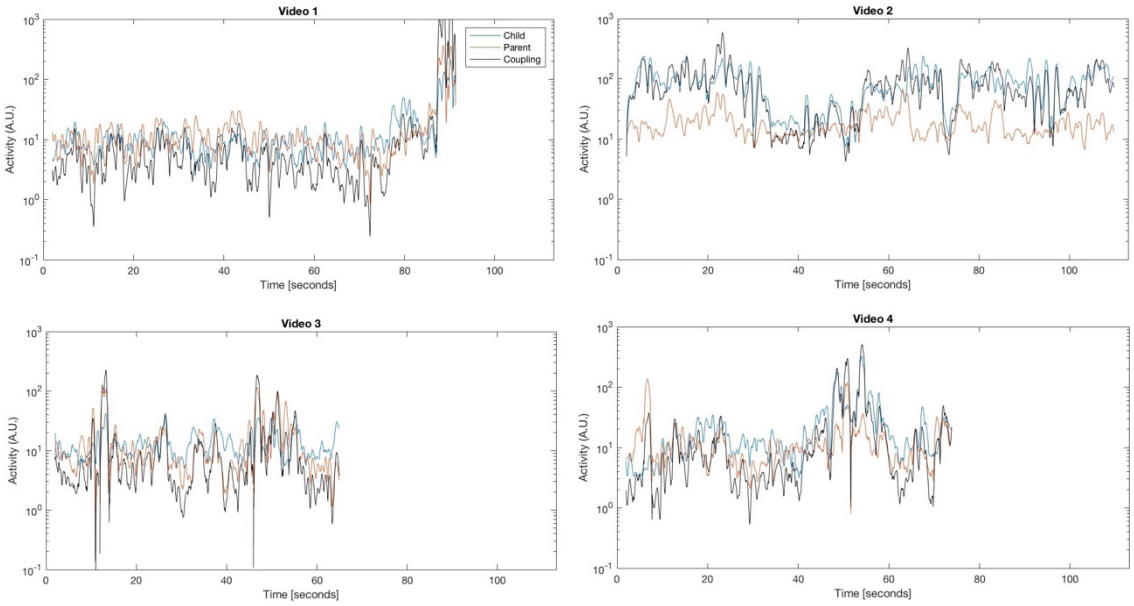


Figure 7. Activities of the child (blue) and parent (red) and the coupling of activities (black).

The value of **coupling** (Figure 7) is high when both the child and the parent are active and lower when only one of them is active. The value is at its lowest, when both child and parent are stay still. In video 2 the coupling follows closely child's activity. In that video, the activity level of the parent is quite constant whereas the activity level of the child varies a lot. In the three other videos, there is no such clear relation between the coupling and the activity of either the child or the parent.

The **distances between hand pairs** and the overall **hand closeness** are presented in Figure 8 and in Appendices 13–16. In video 2, the child runs after the bubbles and is occasionally much further from the parent than in the three other videos. Video 2 is also the only one shot vertically, so the shooting setup was wider. Therefore, the hand pair distances vary much more and the maximum hand pair distance is bigger than in the three other videos. In video 1, the parent sits with hands on knees, and the hands are close to the child almost all the time. This is why the peaks of the hand closeness curve are not as clear as in the other

videos. Instead, in video 3 the hand closeness peaks are particularly clear, because the child is very calm, and moves his hands only to reach the bubbles or to handle the soap bubble jar.

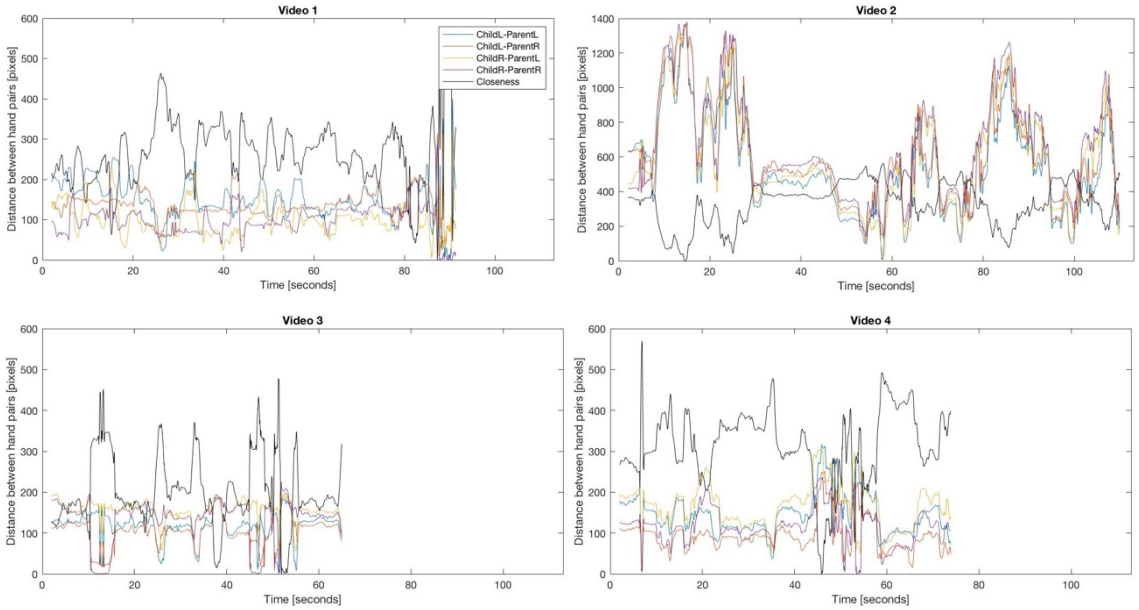


Figure 8. Hand pair distances (blue, red, yellow and purple) and overall hand closeness (black). The lower the hand distance curve, the smaller the distance between the hand pairs. The higher the hand closeness curve (black), the closer all the hands are to each other. Notice the different y-axis scale in video 2.

6.3 Comparing the annotations and the OpenPose results

The **activities** of child and parent and the annotated *communication initiatives* and *turn-takings* are presented in Figure 9. The **hand closeness** and the *communication initiatives* and *turn-takings* are presented in Figure 10. Based on this data, the **coupling** does not seem like a good indicator of interaction events.

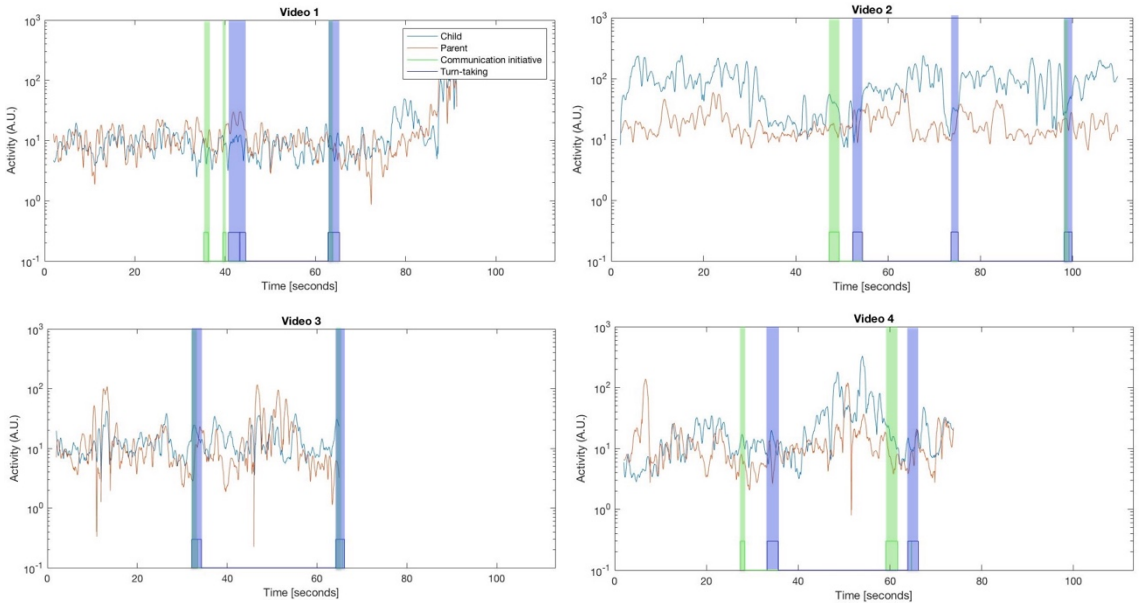


Figure 9. Activity curves of parent (red) and child (blue) show peaks and alternation in activities. Annotated events of turn-taking (blue) and communication initiatives (green) over-lapped to indicate meaningful events of interaction.

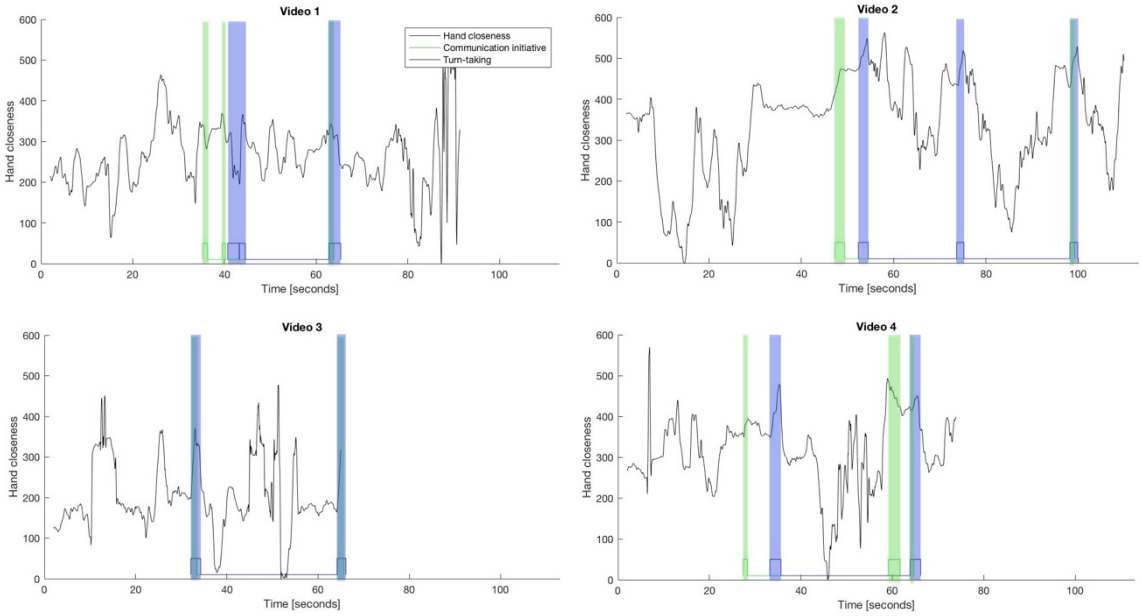


Figure 10. Overall hand closeness (black) show the closeness of parent's and child's hands. Annotated communication initiatives (green) and turn-takings (blue) overlapped to indicate meaningful events of interaction.

For all communication pairs, *communication initiatives* were seen as **hand closeness** and as **child's activity**. *Turn-taking* was seen as **hand closeness** and as **alternation in child's and parent's activities**. The **hand closeness** (Figure 10) was typically high, or, the distance between the hand pairs was typically low when there was interaction between the child and the parent. This was because in most of the cases the interaction was related to giving and taking the soap bubble jar. During communication initiatives child's **activity** was usually higher (Figure 9) and during turn-taking, the activities of child and parent alternated depending on who's turn it was.

However, as the interaction is not the only reason why hands are close to each other, **hand closeness** peaks (Figure 10) occurs also in other time instances. Therefore, interaction can not be identified simply by the hand closeness. Hands can be close to each other for example because the child reaches the bubbles near the parent or because the parent places the jar on the floor in front of the child. Also, the algorithm handles images as two-dimensional although the reality is three-dimensional. This is why the hands may seem to be close to each other even if in reality the hands of one person are further away from the camera than the hands of the other.

Identification of interaction can not be made solely on the basis of **activities** either, as communication initiatives and turn-taking are not the only reasons to causes changes in them. **Combination of the hand closeness and the activity** data predicts interaction a little better, as can be seen from Figure 11. Still it does not distinguish interaction from other activities well enough.

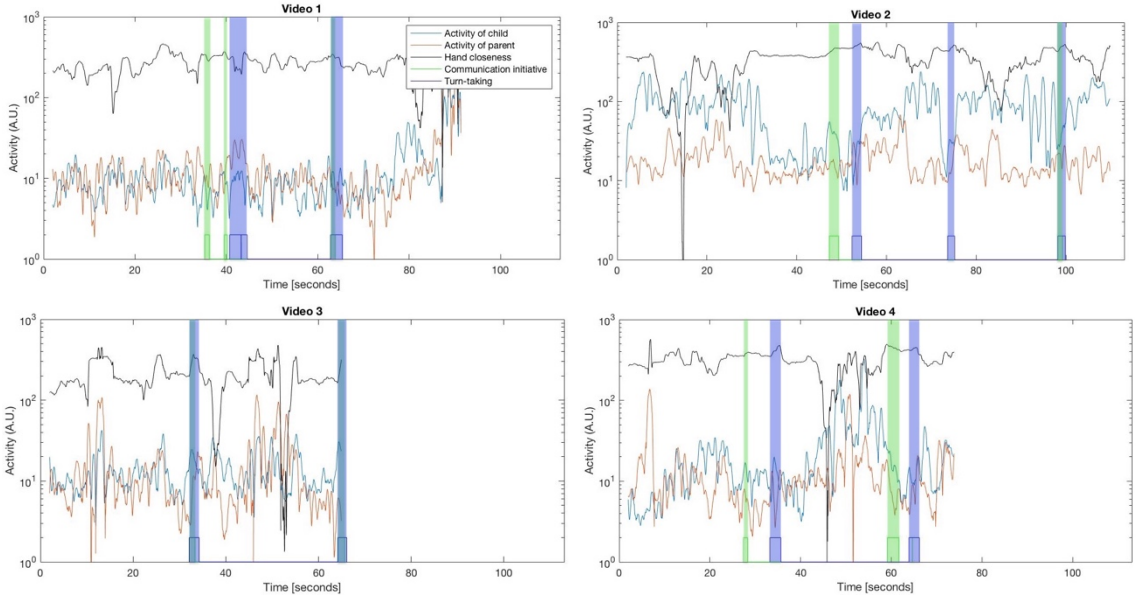


Figure 11. Hand closeness (black) and activities of parent (red) and child (blue). Annotated communication initiatives (green) and turn-takings (blue) overlapped to indicate meaningful events of interaction. The scale of y-axis is logarithmic, so the hand closeness curves look a little different from the previous figures.

Some communication initiatives and turn-takings are missed if only the activity and hand closeness curves are looked (Figures 9–11). For example, in video 1 the child points at the bubbles and later responds to the parent's kiss. These interaction situations are not based on hand closeness and can not be seen from the selected variables here. Also, the results vary a lot depending on how active the child and the parent are, as can be seen for example from Figure 12. The hand closeness peaks are more distinct in video 3 than in videos 1 or 2.

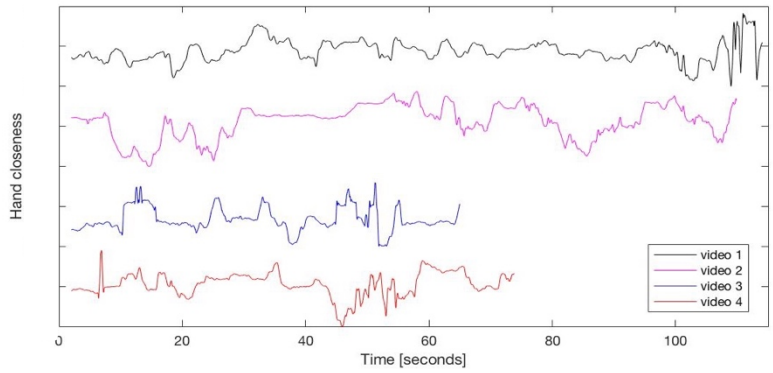


Figure 12. The hand closeness curves in videos 1–4. The y-axis has no unit, and the values of the graphs are not comparable with each other

6.4 Adding the annotated gaze direction

The algorithm could not determine the gaze direction from the videos. However, knowing when child's gaze is directed to parent might help to distinguish interaction from other events. Therefore, the moments when the child was looking at the parent were picked from the annotation data and combined to hand closeness curves and the child's and parent's activity curves. The results are presented in Figures 13–14.

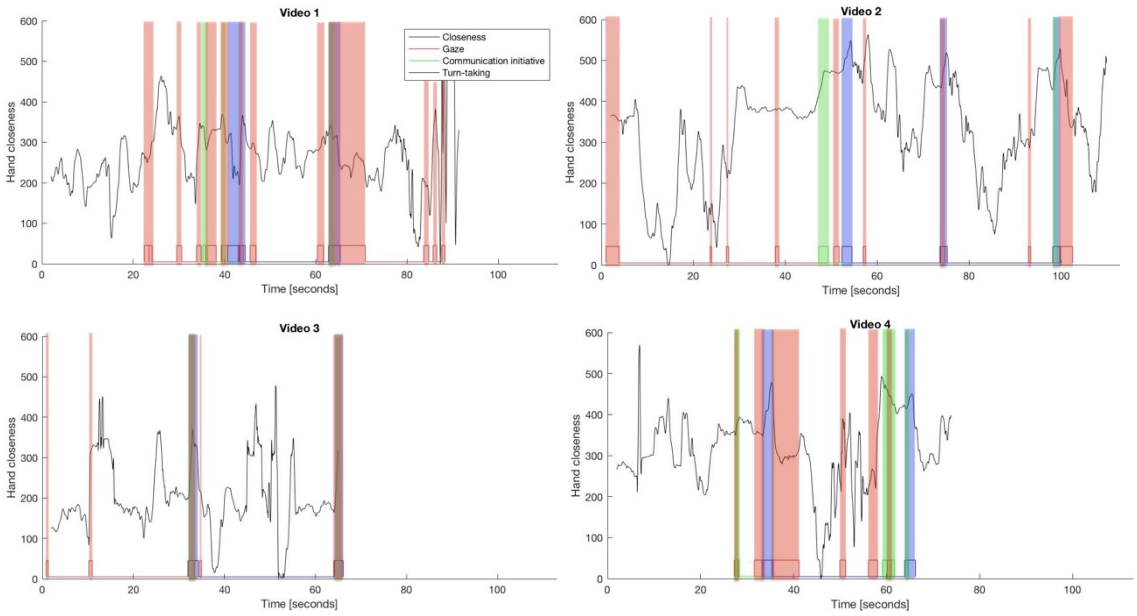


Figure 13. Hand closeness (black) and annotated gaze (red curve and area). Communication initiatives (green) and turn-taking (blue) overlapped to indicate meaningful events of interaction.

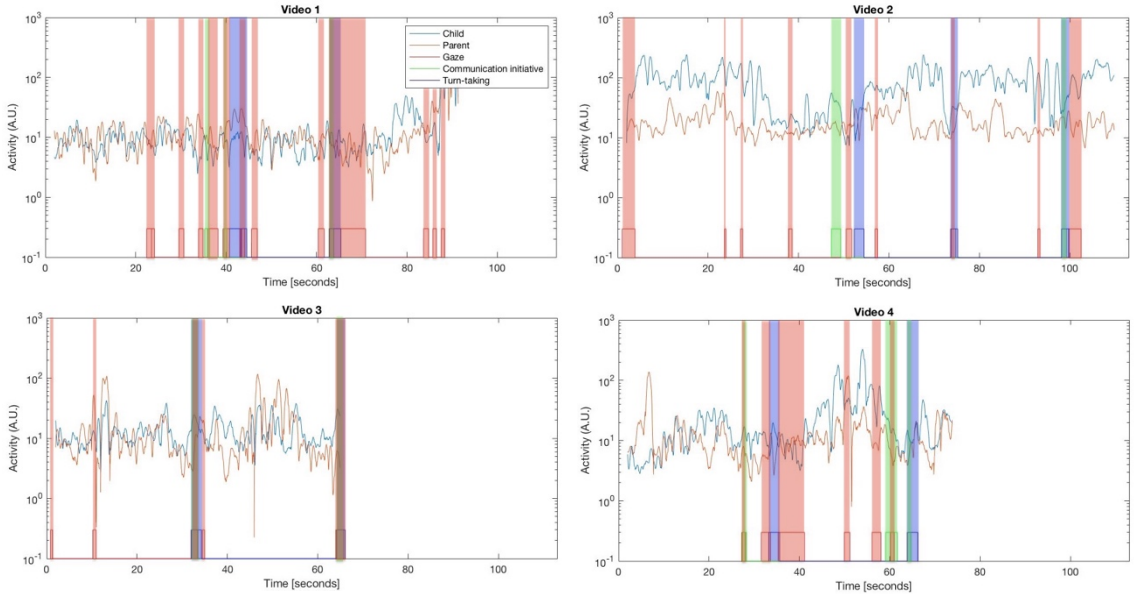


Figure 14. Child's (blue) and parent's (upper red curve) activities and annotated gaze (lower red curve and area). Communication initiatives (green) and turn-taking (blue) overlapped to indicate meaningful events of interaction.

As can be seen from Figure 13, the combination of child's gaze information and hand closeness predicts communication initiatives and turn-takings better than the hand closeness data solely. 65 % of the moments, when there was both peak in hand closeness and the child looked at the parent, involved interaction. All interaction events were not detected by this method either, as they were either not related to jar handling or did not contain child's gaze toward the parent. As was the case with the hand closeness curve alone (Figure 10), the accuracy varied considerably depending on child's and parent's activities. In video 3 the combination of the gaze and the hand closeness distinguished nearly all the communication initiatives and turn-takings from other events, while in video 2 it missed most of the interaction events. Adding the gaze data to the activity curves did not provide corresponding benefit (Figure 14).

7 DISCUSSION

7.1 The relation of annotated interaction events and automatic movement tracking

The algorithm tracked hand movements reliably. Most of the communication initiatives and turn-takings annotated by human annotator could be identified by the algorithm as peaks in hand closeness and as changes in activities. However, the monitored interaction traits were not the only things causing peaks and changes in the curves, and therefore the interaction events could not be distinguished from other events on the basis of these graphs alone. Also, the general activity of the child and the parent influenced on how reliable the results were. With a calm child and a passive parent, the hand closeness and activity curves predicted the interaction events better than with active ones.

Some interaction events, such as pointing and kissing, were not related to hand closeness, and thus they were not visible on the hand closeness curves. On the other hand, the algorithm detected hand closeness also when the hands were adjacent in the image but in fact at different distances from the camera, for example when the child was reaching bubbles behind or in front of the parent. The algorithm can not detect three-dimensional elements from a two-dimensional image. This could be, in theory, solved by using a RGB-D camera as for example Hesse and colleagues (2018) did, but it is against the aim of avoiding the use of special devices or laboratory measurements.

Although the automatic TUG test (Li et al., 2018) and the bipolar disorder classification framework (Yang et al., 2018) using the OpenPose algorithm were successful, the automatic recognition of interaction events in the current study was not reliable enough. Unlike recognition of predefined movements or tracking the movements of predefined body key points as was the case in the studies of Li and colleagues (2018) and Yang and colleagues (2018), distinguishing individual interaction events reliably enough is much more complicated. The interaction events are diverse and communication initiatives and turn-takings can take many different forms. Human annotator can identify them easily, but an algorithm would require more than just one or two factors to recognize them. In this study, the

combination of hand closeness and annotated gaze direction led to better results than the hand closeness data solely, but was still not accurate enough. To distinguish individual interaction events from other events, several features should be tracked simultaneously.

In the current study, only hand closeness and activity data were extracted from the algorithm data. However, also other parameters could be calculated from the key point location data obtained by OpenPose. For example, activities of some determined body key points, recognition of gestures or the synchrony of child's and parent's movements could be useful. Had the videos been shot closer and had the faces been more visible, the facial key points could have been used to identify gaze directions and maybe also facial expressions.

7.2 Limitations and challenges

The limitations and sources of error in this study were partly due to the video recordings and the recording arrangements and partly due to the operation of the OpenPose algorithm and the Matlab code.

The quality of the videos was not always optimal. First, the OpenPose algorithm is designed for horizontal videos, but three of the four videos were shot vertically. Although the problem was fixed by swapping the x- and y-coordinates while processing the data with Matlab, the results from vertical videos caused challenges for analysis and therefore are not ideal. The request of horizontal videos should be clearly mentioned in the recording instructions. In some videos the hand closeness peaks were not very clear, as the hands of the child and the parent were close to each other almost all the time. This was partly due to the fact that the child and the parent were situated so close to each other and partly dependent on how the parent hold his or her hands when not operating the soap bubble jar. The horizontal videos might help, at least partly, solve this problem, as in horizontal videos there is more space on the side and the child and the parent are not forced to sit so close to each other.

For some reason parent's key points were occasionally recognized as child's key points and vice versa. The problem was easily fixed, as it was known that the child was always situated on the other side of the frame and the parent on the

other side. However, mixing of the key points may cause more problems, if in the future also unstructured playing videos are used.

False hand closeness peaks

Although the hand movement recognition was reliable, the algorithm recognized some false hand closeness peaks in three of the four videos, even the hands were not close to each other. About 40 % of the false peaks were at random moments, and all of them were in videos 3 and 4 where the parent was wearing background-colored clothing and the lighting was not bright enough. Therefore, the random false peaks were probably caused by the algorithm not detecting all key points right due to the low contrast between clothing and background. Similar problems came up in Marchi's and colleagues' (2019) and Komiya's and colleagues' (2018) studies. They mentioned that the too low quality of the videos caused problems when recognizing key points. Marchi and colleagues (2019) had to reject several video clips due to bad quality. In this study the problem was not as severe since all the videos were usable, only some key points had to be excluded from the calculations.

There were false hand closeness peaks also when hand key points were missing due to occlusion or because they were outside the image. Probably the estimation of missing key points was not accurate enough. The code should be rechecked in order to dispose these sources of error. In this study the problem of missing key points was not as bad as in Heath's and colleague's (2018) study, where they had recorded children and parents playing freely. In their study, the OpenPose algorithm sometimes found only one person instead of two, and sometimes it, incorrectly, found three or more persons. In our videos, the bubble blowing situation was more structured and thus not as much occlusion occurred. There were neither toys or other objects on the background, that could have been recognized as people and so most of the key points were recognized right.

Gaze

The gaze direction is essential when evaluating joint attention, but it helps also to identify other interaction events, as people often look at each other when communicating. In this study gaze direction could not be detected by the algorithm. The use of OpenPose's facial key points requires videos, that are shot close to

face. Similarly to Komiya's and colleagues' study (2018), the distance from camera was too long, and the size of the eyes was too small and quality of the image too low to use them for gaze tracking in the current study. Instead of facial key points, gaze direction could be evaluated from the body key points of eyes, nose and ears. However, if the subjevts are sitting sideways to the camera, as was the case in our videos, some key points are hidden most of the time and the latter method does not work either. In their study, Heath and colleagues (2018) had similar problem with gaze recognition. The recording arrangements in Müller and colleagues (2018) study were much better for facial recognition as they had placed a camera in front of each subject. Therefore, the faces were visible all the time. Even though they used a different algorithm to recognize facial key points, similar placement of cameras would be ideal also when using OpenPose for facial key point recognition.

In the future, it is possible to utilize gaze directions by modifying recording arrangements. The videos should be shot closer, and better shooting angle should be chosen. The subjects should be more angled and the faces should be more towards the camera. Still, following the gaze direction of a moving child can remain challenging.

7.3 Annotation

The two different annotation schemas provided different kind of information about the videos. The basic unit annotations were used to check the reliability of the movement tracking. They included lot of additional information besides the useful information on interactive events: The jar handling annotations contained the information of, for example, holding the stick, blowing bubbles and child's attempts to open the jar, which had no use in the study. The annotation of additional events meant lot of irrelevant work when annotating, and the vast amount of information also complicated processing of the results. Moreover, as all the communication initiatives and turn-takings were not related to jar handling, they could not be seen in the jar handling annotations. So, even if the annotation of basic units was accurate, it did not catch all interaction. The annotation of interaction events was

used to identify meaningful interaction moments. However, the turn-taking annotations were not detailed enough, as they did not include the information on how many turns the event included and who was active at which time instance.

There are always differences between annotators on how they interpret data (Tilley, 2003), and the annotation of small details and their timing is not as reliable as the annotation of larger entities. In the videos used in this study, the interaction events were rather simple and easy to recognize. Therefore, instead of basic units, which hopefully in the future could be recognized by the algorithm, it might be better to concentrate on interaction events. To be able to correlate annotations with the OpenPose movement time series the annotations need to be continuous.

7.4 Recommendations for the future research

7.4.1 Aims of the future studies

Based on these data there is a lot of variation between child-parent pairs on how, for example, communication initiatives and turn-taking appear, and thus it may be difficult to find a general way to recognize them. Also, interaction is always dyadic or multilateral, and the interaction partner has an influence on the behavior of the other party (Kenny & Malloy, 1988). Therefore, distinguishing individual interaction events of one person may not be the most appropriate way to assess interaction. Parent's involvement and actions affect child's behavior, and with different interaction partner the child could act in a different way.

This study focused on predefined interaction events, but there is interaction also between the monitored interaction situations. Visualization of the hand closeness and the activity data showed that they differ between dyads also outside the annotated interaction moments. Therefore, it is probably more beneficial to study the whole interaction situation or global units such as synchrony between child and parent instead of individual interaction events and interaction traits of one subject.

Another reason for focusing on larger entities is that in some cases, the best way to detect atypical interaction is not by assessing the interaction skills of an individual, but by monitoring the behavior of his/her interaction partner. This was the

case in the study of Delaherche and colleagues (2013). The factors discriminating best a group of children with ASD from a control group, were the gestural rhythm and the duration of gestural pauses of the therapists, not the parameters related to child's behavior. Also, by concentrating on only child's behavior, the root cause of the problem is not necessarily reached, as the impaired interaction can also originate for example from parent's mental health problems (Murray et al., 1996) or temperament differences between child and parent (Campbell, 1979).

In their study, besides individual parameters, Leclère and colleagues (2016) used dyadic parameters like overlap ratio, pause ratio and percentage of time used face-to-face. By computational methods, similar parameters might be possible to use also with OpenPose algorithm. For example, although the coupling was not useful in this study that used visual analyzing, it may offer interesting results if computational analysis is used instead.

Depending on the aim of the future studies, the annotation schemas need to be reconsidered, as distinguishing predefined interaction events from other movements requires different annotation than the aim to study larger interaction entities. Also, if in future the researchers end up to study large entities, it might be worth to reconsider the interaction situation used in videos. There should be larger variety of play situations, and also unstructured free play. However, it should be kept in mind, that the free play situations cause more occlusion and probably make gaze detection harder. Therefore, the use of unstructured videos need to be planned carefully.

7.4.2 Gaze

The main disadvantage in the current study was that the faces of subjects were not fully visible on the videos, and therefore the gaze directions could not be estimated. Therefore, as described earlier in limitations, the recording instructions should be modified to enable the gaze direction estimation. If the recognition of gaze directions were to work, the moments of dyadic joint attention could probably be recognized by the algorithm as moments when child and parent are looking at each other. The triadic joint attention may not be as easy to recognize only by

gaze directions as it involves also gazes to a third party. Therefore, it probably would require annotation.

7.4.3 Voice

As the method in this study was based solely on images, verbal communication was ignored and could not be seen on the results. However, especially the older children used also language in their communication initiatives and turn-taking. The vocalizations had an influence on how the child or the parent acted, and further, also on how the interaction was decoded. For example, in video 2 the child makes verbal communication initiative saying “Äiti avaa” (“Mummy opens”) before the parent reaches toward the jar. Only after that the child gives the jar to the parent. The communication initiative was made by the child, but it can not be seen on the annotations or detected by the algorithm as it was verbal. Instead, the child’s behavior was seen as turn-taking. Some interaction events may be completely missed if the voice is ignored. As voice is an essential part of communication it should be taken into account when assessing interaction. However, this requires use of additional, voice or speech based analyzing method besides the OpenPose, as was done in the studies of Kim and colleagues (2017) and Müller and colleagues (2018).

7.4.4 Recommendations for recording instructions

The current study revealed that data quality can be improved with more accurate recording instructions. Therefore, they need to be refined. First, it should be clearly mentioned that the videos should be recorded horizontally. The videos need to be shot close enough and the subjects should sit at an angle to the camera so that their faces are visible. The clothing must preferably be of a color that is not too dark and stands out from the background. If the aim is to study child’s behavior, the parent should act restrained while waiting the child to do the initiatives and keep the hands close to the body. This may, however, be an unnatural way for a parent to act. Therefore, if in the future the aim is to study global units or the overall interaction between the child and the parent, the parent should act in a way that is natural to her/him.

7.5 Conclusion

The current study suggests that data obtained by the algorithm is reliable. The recording conditions were not always optimal, which occasionally lead to incorrect results. The accuracy of the hand closeness recognition was found to be good, but mainly due to poor lighting conditions and background-colored clothing of the subjects the key point recognition was sometimes difficult. Also occlusion and occasional missing of some key points caused problems. Therefore, the estimation of missing key points should be improved.

This study concentrated on separate interactional events and child's nonverbal communication abilities, but based on the results, it would be more fruitful to focus on global units such as synchrony. The recording arrangements should be redesigned and the annotation schema should be chosen to support this aim.

REFERENCES

Ajay, J., Song, C., Wang, A., Langan, J., Li, Z & Xu, W. (2018). A pervasive and sensor-free Deep Learning system for Parkinsonian gait analysis. *2018 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)*, 108–111.

American Psychiatric Association (2013). *Diagnostic and statistical manual of mental disorders* (5th edition). Washington, DC: American Psychiatric Association.

Antaki, C. & Wilkinson, R. (2013). Conversation analysis and the study of atypical populations. In J. Sidnell & T. Stivers (eds), *The handbook of conversation analysis*. Chichester: Wiley-Blackwell.

Autismikirjon häiriöt (2019, August 10). Retrieved from http://www.hus.fi/sairaanhoito/lasten-sairaanhoito/lastenneurologia/Neurokognitiiviset_hairiot/Autismi/Sivut/default.aspx

Avril, M., Leclère, C., Viaux, S., Michelet, S., Achard, C., Missonnier, S., … & Chetouani, M. (2014). Social signal processing for studying parent–infant interaction. *Frontiers in Psychology, 5*.

Bates, E. (1976). *Language and context: the acquisition of pragmatics*. New York: Academic Press.

Bateson, M.C. (1979). “The epigenesist of conversational interaction”: a personal account of research development. In M. Bullowa (ed), *Before Speech: The beginning of interpersonal communication*. Cambridge: Cambridge University Press.

Baldwin, D.A. (1995). Understanding the link between joint attention and language. In C. Moore & P.J. Dunham (Eds.) *Joint attention: Its origins and role in development* (pp.131-158). Hillsdale, NJ: Erlbaum.

Baltrušaitis, T., Robinson P. & Morency, L.-P. (2016). OpenFace: an open source facial behavior analysis toolkit. *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*, 1–10.

Beach, W.A. (2013). Conversation analysis and communication. In J. Sidnell & T. Stivers (eds), *The handbook of conversation analysis* (pp. 674–687). Chichester: Wiley-Blackwell

Beugher, S., Brône, G. & Goedemé, T. (2018). A semi-automatic annotation tool for unobtrusive gesture analysis. *Language Resources and Evaluation, 52*(2), 433–460.

Beuker, K.T., Rommelse, N.N.J., Donders, R. & Buitelaar, J.K. (2013). Development of early communication skills in the first two years of life. *Infant Behavior & Development, 36.* 71–83.

Bianco, S., Ciocca, G., Napoletano, P. Schettini, R. (2015). An interactive tool for manual, semi-automatic and automatic video annotation. *Computer Vision and Image Understanding, 131*, 88.

Bradski, G.R. (1998). Computer vision face tracking for use in a perceptual user interface. *Intel Technology Journal, 2*, 12–21.

Brooks, R. & Meltzoff, A.N. (2008). Infant gaze following and pointing predict accelerated vocabulary growth through two years of age: a longitudinal, growth curve modeling study. *Journal of Child Language, 35*(1), 207–220.

Burgoon, J.K. & Bacue, A.E. (2003). Nonverbal communication skills. In J.O. Green & B.R. Burleson (2003). *Handbook of communication and social interaction skills (*pp. 179–220). Mahwah, N.J.: L. Erlbaum Associates.

Cabibihan, J., Javed, H., Aldosari, M., Frazier, T. & Elbashir, H. (2016). Sensing technologies for autism spectrum disorder screening and intervention. *Sensors, 17*(1).

Campbell, K., Carpenter, K.L.H., Hashemi, J. Espinosa, S., Marsan, S. Schaich Borg, J.S., … & Dawson, G. (2018). Computer vision analysis captures atypical

attention in toddlers with autism. *Autism.*
https://doi.org/10.1177/1362361318766247

Campbell, S. (1979). Mother-infant interaction as a function of maternal ratings of temperament. *Child Psychiatry and Human Development, 10*(2), 67–76.

Carpenter, M., Nagell, K., Tomasello, M., Butterworth, G. & Moore, C. (1998). Social cognition, joint attention, and communicative competence from 9 to 15 months of age. *Monographs of the Society for Research in Child Development, 63*(4), 174.

Cao, Z., Hidalgo, G, Simon, T., Wei, S.-E. & Sheikh, Y. (2018). OpenPose: Realtime multi-person 2D pose estimation using part affinity fields. arXiv:1812.08008 [cs.CV]

Caskey, M., Stephens, B., Tucker, R. & Vohr, B. (2011). Importance of parent talk on the development of preterm infant vocalizations. *Pediatrics, 128*(5), 910–916.

Charman, T. (2004). Why is joint attention a pivotal skill in autism? In U. Frith & E. L. Hill (2004), *Autism: Mind and brain* (pp. 67–87). London: Oxford University Press.

Colonnesi, C., Stams, G., J.J.M., Koster, I. & Noom, M.J. (2010). The relation between pointing and language development: A meta-analysis. *Developmental Review, 30*(4), 352–366.

Corkum, V. L., & Moore, C. (1998). The origins of joint visual attention in Infants. *Developmental Psychology, 34*, 28–38.

Davis, D.W. (2010). *Maternal sensitivity: a scientific foundation for practice*. Hauppauge, NY: Nova Science Publishers.

Delaherche, E., Chetouani, M., Bigouret, F., Xavier, J., Plaza, M. & Cohen, D. (2013). Assessment of the communicative and coordination skills of children with autism spectrum disorders and typically developing children using social signal processing. *Research in Autism Spectrum Disorders, 7*(6), 741–756.

De Schuymer, L., De Groote, I., Beyers, W., Striano, T. & Roeyers, H. (2011). Preverbal skills as mediators for language outcome in preterm and full term children. *Early Human Development, 87*(4), 265–272.

Dickerson P., Rae J., Stribling P., Dautenhahn K., Werry I. (2005) Autistic children's co-ordination of gaze and talk: Re-examining the 'asocial' autist. In K. Richards & P. Seedhouse (eds), *Applying Conversation Analysis*. Palgrave Macmillan, London.

ELAN (Version 5.4) [Computer software]. (2018). Nijmegen: Max Planck Institute for Psycholinguistics. Retrieved from https://tla.mpi.nl/tools/tla-tools/elan/

Filipi, A. (2009*). Toddler and parent interaction: the organisation of gaze, pointing and vocalisation.* Amsterdam; Philadelphia: John Benjamins Pub. Co.

Fenson, L., Dale, P. S., Reznick, J. S., Thai, D., Bates, E., Hartung, J. P., Pethick, S., & Reilly, J. S. (1993). *The MacArthur Communicative Development Inventories: User's guide and technical manual*. San Diego: Singular Publishing Group.

Frazier, T.W., Klingemier, E.W., Parikh, S., Speer, L., Strauss, M.S., Eng, C., … & Youngstrom, E.A. (2018). Development and validation of objective and quantitative eye tracking−based measures of autism risk and symptom levels. *Journal of the American Academy of Child & Adolescent Psychiatry, 57*(11), 858–866.

Gatica-Perez, D. (2009). Automatic nonverbal analysis of social interaction in small groups: A review. *Image and Vision Computing, 27*(12), 1775–1787.

Gilchrist, K., Hegarty-Craver, M., Christian, R., Grego, S., Kies, A. and Wheeler, A. (2018). Automated detection of repetitive motor behaviors as an outcome measurement in intellectual and developmental disabilities. *Journal of Autism and Developmental Disorders, 48*(5), 1458–1466.

Goodwin, M. S., Intille, S. S., Albinali, F. & Veliced, W. F. (2011). Automated detection of stereotypical motor movements. *Journal of Autism and Developmental Disorders, 41*(6), 770-782.

Green, J., Franquiz, M. & Dixon, C. (1997). The myth of the objective transcript: Transcribing as a situated act. *TESOL Quarterly, 31*(1), 172–176.

Hadfield, J., Chalvatzaki, G., Koutras, P., Khamassi, M., Tzafestas, C.S. & Maragos, P. (2018). *A deep learning approach for multi-view engagement estimation of children in a child-robot joint attention task*. arXiv:1812.00253 [cs.RO].

Hashemi, J., Tepper, M., Vallin Spina, T., Esler, A., Morellas, V., Papanikolopoulos, N. & Egger, H. (2014). Computer vision tools for low-cost and noninvasive measurement of autism-related behaviors in infants. *Autism Research and Treatment, 2014.*

Heath, C. D., Venkateswara, H., McDaniel, T., & Panchanathan, S. (2018). Detecting attention in Pivotal Response Treatment video probes. In A. Basu & S. Berretti (eds), *Smart Multimedia. ICSM 2018. Lecture Notes in Computer Science, 11010*. Springer, Cham. http://doi.org/10.1007/978-3-030-04375-9_21

Van Hecke, A.V., Mundy, P.C., Acra, C.F., Block, J.J., Delgado, C.E.F., Parlade, M.V., … & Pomares, Y.B. (2007). "Infant joint attention, temperament, and social competence in preschool children". *Child Development, 78*(1), 53–69.

Hesse, N., Pujades, S., Black, M.J., Aresns, M., Hofmann, U.G. & Schroeder, S. (2018) Learning and tracking the 3D body shape of freely moving infants from RGB-D sequences. *CoRR*, *abs/1810.07538.*

Hilbrink, E. E., Gattis, M., & Levinson, S. C. (2015). Early developmental changes in the timing of turn-taking: a longitudinal study of mother–infant interaction. *Frontiers in Psychology, 6* (1492), 246–257.

Holmlund, C. (1995). Development of turntaking as a sensorimotor process in the first 3 months: A sequential analysis. In K. E. Nelson and Z. Réger (eds), *Children's Language Vol.8*, 41–64. New Jersey: Lawrence Erlbaum Associates.

Iverson, J.M. & Goldin-Meadow, S. (2005). Gesture paves the way for language development. *Psychological Science, 16*(5), 367–371.

Jenks, C. J. (2011). *Transcribing talk and interaction: Issues in the representation of communication data*. Amsterdam: John Benjamins Pub. Co.

Kaye, K. (1977). “Toward the origin of dialogue.” In H. R. Schaffer (ed), *Studies in Mother-Infant Interaction* (pp. 89–118). New York: Academic Press.

Kenny, D. & Malloy, T. (1988). Partner effects in social interaction. *Journal of Nonverbal Behavior, 12*(1), 34–57.

Kishimoto, T., Shizawa, Y., Yasuda, J., Hinobayashi, T. & Minami, T. (2007). Do pointing gestures by infants provoke comments from adults? *Infant Behavior and Development, 30*(4), 562–567.

Komiya, R., Saitoh, T & Shimada, K. (2018) Image-based attention level estimation of interaction scene by head pose and gaze information. *IEEE/ACIS 17th International Conference on Computer and Information Science (ICIS)*, 497–501.

Kuhn, L.J., Willoughby, M.T., Wilbourn, M.P., Vernon-Feagans, L. & Blair, C.B. (2014). Early communicative gestures prospectively predict language development and executive function in early childhood. *Child Development, 85*(5), 1898–1914.

Kumano, S., Otsuka, K., Ishii, R. & Yamato, J. (2017). Collective first-person vision for automatic gaze analysis in multiparty conversations. *IEEE Transactions on Multimedia, 19*(1), 107–122.

Kunnari, S. & Savinainen-Makkonen, T. (2012). *Pienten sanat: Lasten äänteellinen kehitys*. Jyväskylä: PS-Kustannus.

Laakso, M.-L., Eklund, K., & Poikkeus, A.-M. (2011). *Esikko – Lapsen esikielellisen kommunikaation ja kielen ensikartoitus*. Jyväskylä: Niilo Mäki Instituutti.

Laing, E., Butterworth, G., Ansari, D., Gsödl, M., Longhi, E., Panagiotaki, G., … & Karmiloff-Smith, A. (2002) Atypical development of language and social communication in toddlers with Williams syndrome. *Developmental Science, 5*(2), 233–246.

Launonen, K. (2007). *Vuorovaikutus –kehitys, riskit ja tukeminen kuntoutuksen keinoin*. Helsinki: Kehitysvammaliitto.

Leclère, C., Avril, M., Viaux-Savelon, S., Bodeau, N., Achard, C., Missonnier, S., … & Cohen, D. (2016). Interaction and behaviour imaging: a novel method to measure mother–infant interaction using video 3D reconstruction. *Translational Psychiatry, 6*(5).

Lee, K. & Schertz, H.H. (2019). Brief Report: Analysis of the Relationship Between Turn Taking and Joint Attention for Toddlers with Autism. *Journal of autism and developmental disorders*,
https://doi-org.libproxy.helsinki.fi/10.1007/s10803-019-03979-1

Li, T., Chen, J., Hu, C., Ma, Y., Wu, Z, Wan, W., … & Li, L. (2018). Automatic timed up-and-go sub-task segmentation for Parkinson's disease patients using video-based activity classification. *IEEE Transactions on Neural Systems and Rehabilitation Engineering, 26(*11), 2189–2199.

Lowe, M. & Costello. A.J. (1976). *Manual for the symbolic play test.* Windsor, UK: NFER.

Mahdhaoui, A., Chetouani, M., Cassel, R.S., Saint-Georges, C., Parlato, E., Laznik, M.C., … & Cohen, D. (2011). Computerized home video detection for motherese may help to study impaired interaction between infants who become autistic and their parents. *International journal of methods in psychiatric research, 20*(1), 6–18.

Marchi, V., Hakala, A., Knight, A., D'Acunto, F., Scattoni, M.L., Guzzetta, A. & Vanhatalo, Sampsa (2019). Automated pose estimation captures key aspects of General Movements at eight to 17 weeks from conventional videos. *Acta paediatrica. 18.*

Mathur, S., Poole, M.S., Peña-Mora, F., Hasegawa-Johnson, M. & Contractor, N. (2012). Detecting interaction links in a collaborating group using manually annotated data. *Social Networks, 34*(4), 515–526.

Messinger, D. & Fogel, A. (2007). The interactive development of social smiling. *Advances in Child Development and Behavior, 35*, 328–366.

Morales, M., Mundy, P., Delgado, C.E.F., Yale, M., Messinger, D., Neal, R. & Schwartz, H.K. (2000). Responding to joint attention across the 6- through 24-month age period and early language acquisition. *Journal of Applied Developmental Psychology, 21*(3), 283-298.

Mundy, P., Delgado, C., Block, J., Venezia, M., Hogan, A & Seiber. J. (2003). *A manual for the abridged early social communication scales*. Coral Gable. FL: University of Miami.

Mundy, P., Sigman, M., Kasari, C. & Yirmiya, N. (1988). Nonverbal communication skills in Down syndrome children. *Child Development. 59*(1), 235–249.

Murray, L., Fiori-Cowley, A., Hooper, R. & Cooper, P. (1996). The impact of postnatal depression and associated adversity on early mother-infant interactions and later infant outcome. *Child Development, 67*(5), 2512–2526.

Müller, P., Huang, M.X. & Bulling, A. (2018). Detecting low rapport during natural interactions in small groups from non-verbal behaviour. *IUI.*

Mäntymaa, M., Puura, K., Luoma, I., Salmelin, R., Davis, H., Tsiantis, J., … & Tamminen, T. (2003). Infant–mother interaction as a predictor of child's chronic health problems. *Child: Care, Health and Development, 29*(3), 181–191.

Nathani, S., Ertmer, D.J. & Stark, R.E. (2006). Assessing vocal development in infants and toddlers. *Clinical Linguistics & Phonetics, 20*(5), 351–369.

Negin, F., Rodriguez, P., Koperski, M., Kerboua, A., Gonzàlez, J., Bourgeois, J., … & Bremond, F. (2018). PRAXIS: Towards automatic cognitive assessment using gesture recognition. *Expert Systems With Applications, 106*, 21–35.

Ochs, E. (1979). Transcription as theory. In E. Ochs & B. Schieffelin (eds), *Developmental Pragmatics*. New York: Academic Press.

De Pascalis, L., Kkeli, N., Chakrabarti, B., Dalton, L., Vaillancourt, K., Rayson, H., … & Murray, L. (2017). Maternal gaze to the infant face: Effects of infant age

and facial configuration during mother-infant engagement in the first nine weeks. *Infant Behavior and Development, 46*, 91–99.

Pisharady, P.K. & Saerbeck, M. (2015). Recent methods and databases in vision-based hand gesture recognition: A review. *Computer Vision and Image Understanding, 141*, 152.

Pisharady, P.K., Vadakkepat, P. & Loh, A. (2013). Attention based detection and recognition of hand postures against complex backgrounds. *International Journal of Computer Vision, 101*(3), 403–419.

Reddy, V. (2005). Before the third element: Understanding attention to self. In N. Eilan, C. Hoerl, T. McCormack & J. Roessler (Eds.), *Joint attention: Communication and other minds* (pp. 85–109). Oxford, UK: Oxford University Press.

Rodrigues, J. L., Gonçalves, N., Costa, S. & Soares, F. (2013). Stereotyped movement recognition in children with ASD. *Sensors & Actuators: A.Physical, 202*, 162-169.

Rogers, S. J. & Dawson, G. (2010). *Early start Denver model for young children with autism*. New York: The Guilford Press.

Rowe, M.L., Özçalıskan, S. & Goldin-Meadow, S. (2008). Learning words by hand: Gesture's role in predicting vocabulary development. *First language, 28*(2), 182–199.

Santos, J.F., Brosh, N., Falk, T.H., Zwaigenbaum, L., Bryson, S.E., Roberts, W., … & Brian, J.A. (2013). Very early detection of autism spectrum disorders based on acoustic analysis of pre-verbal vocalizations of 18-month old toddlers. *IEEE International Conference on Acoustics, Speech and Signal Processing*, 7567–7571.

Saulnier, C. A. & Ventola, P. E. (2012). *Essentials of autism spectrum disorders evaluation and assessment.* Hoboken: John Wiley and Sons.

Savitzky, A. & Golay, M.J.E. (1964). Smoothing and differentiation of data by simplified least squares procedures. *Analytical Chemistry, 36*(8), 1627–39.

Taffoni, F., Focaroli, V., Formica, D., Gugliemelli, E., Keller, F. & Iverson, J. M. (2012). Sensor-based technology in the study of motor skills in infants at risk for ASD. *Proceedings of the IEEE/RAS-EMBS International Conference on Biomedical Robotics and Biomechatronics August 2014*, 244-251.

Tilley, S.A. (2003). “Challenging” research practices: Turning a critical lens on the work of transcription. *Qualitative Inquiry, 9*(5), 750–773.

Tomasello, M. (1995) Joint attention as social cognition. In C. Moore & P. Dunham (1995) (Eds.). *Joint attention: Its origins and role in development* (pp.103–130). Hillsdale (N.J.): Erlbaum.

Tomasello, M. & Farrar, M.J. (1986). Joint attention and early language. *Child Development, 57*(6), 1454–1463.

Yang, L., Li, Y., Chen, H., Jiang, D., Oveneke, M.C. & Sahli, H. (2018). Bipolar disorder recognition with histogram Features of arousal and body gestures. *AVEC’18*.

Vargas-Cuentas, N.I., Roman-Gonzalez, A., Gilman, R.H., Barrientos, F. Ting, J., Hidalgo, D., … & Zimic, M. (2017). Developing an eye-tracking algorithm as a potential tool for early diagnosis of autism spectrum disorder in children. *PLoS ONE Vol.12*(11).

Vondrick, C., Patterson, D. & Ramanan, D. (2013). Efficiently scaling up crowdsourced video annotation. *International Journal of Computer Vision, 101*(1), 184–204.

Wagner, P., Malisz, Z. & Kopp, S. (2014). Gesture and speech in interaction: An overview. *Speech Communication, 57*, 209–232.

Watson, L.R., Crais, E.R., Baranek, G.T., Dykstra, J.R. & Wilson, K.P. (2013). Communicative Gesture Use in Infants with and without Autism: A Retrospective Home Video Study *American Journal of Speech-Language Pathology,.22*(1), 25–39.

Watt, N., Wetherby, A. & Shumway, S. (2006). Prelinguistic predictors of language outcome at 3 years of age. *Journal of Speech, Language, and Hearing Research, 49*(6), 1224–1237.

Wetherby A, Prizant B (2002). *The Infant Toddler Checklist from the Communication and Symbolic Behavior Scale*s. Baltimore: Brookes Publishing.

Winder, B.M., Wozniak, R.H., Parlade, M.V. & Iverson, J.M. (2013). Spontaneous initiation of communication in infants at low and heightened risk for autism spectrum disorders. *Developmental Psychology, 49*(10), 1931–1942.

Yilmaz, E., Parlato-Oliveira, E., Gratier, M., Devouche, E., Guellaï, B. & Infanti, R. (2015). Early development of turn-taking in vocal interaction between mothers and infants. *Frontiers in Psychology, 6.*

Zhu, S. (2015). Automatic recognition of facial expression based on computer vision. *International Journal on Smart Sensing and Intelligent Systems, 8*(3), 1464.

Zwaigenbaum, L., Bryson, S. & Garon, N. (2013). Early identification of autism spectrum disorders. *Behavioural Brain Research, 251*, 133–146.

APPENDICES

APPENDIX 1. Recording instructions

In the Bubble

The purpose is to videotape a natural, short play situation between your child and you, or someone else close to your child. You need a smartphone and an assistant who can video record the play situation, and a bubble jar. Choose a moment that is pleasant for both you and your child, and a place (preferably on the floor) where you can easily play together face-to-face. It is good to clear other toys away so that they will not distract your child's attention.

Sit on the floor face-to-face with your child. The video should clearly show both the adult and the child.

To engage your child's attention and interest, start blowing bubbles.

When the child's attention is on you and the bubbles, stop, and close the jar tight enough that the child will not be able to open it alone and place the jar between you on the floor. Wait for the child's reaction. It is important that you do not prompt your child either verbally (e.g. "Give me the bubbles so I will blow some more!") or with gestures (e.g., by pointing the soap bubbles or stretching out your hand). Wait for your child's reaction.

After the child has in some way indicated that s/he wants more you should continue blowing the bubbles. You may continue this few times.

Thank you!

APPENDIX 2. The instructions for annotating the interaction events.

Instructions for annotating interaction events

Annotation is made by using ELAN, an annotator tool. Annotate the child's communication initiatives, turn-taking and joint attention in each video. When annotating, take into account gestures and other nonverbal interaction. Speech and vocalizations are ignored.

- Create a new file for every video.
- Create three tiers, one for each interaction event (communication initiative, turn-taking, joint attention).
- It is recommended to watch the whole video first before doing the annotation, so you will know what to expect.
- Then watch the video and examine one feature at a time
- Annotate the following information:
    - Communication initiatives
        - Mark the beginning and ending time of the initiative
        - Tell what the child is doing
        - Other comments if necessary
    - Turn-taking
        - Mark the begining and ending time of turn-taking
        - Tell what the child and the parent are doing
        - Other comments (e.g. uncertainty or if the child is runnign after bubbles at the same time).
    - Joint attention
        - The beginning and ending times of the activity that implies joint attention.
        - Tell what the child and the parent are doing
        - Other comments (e.g. uncertainty)
- Save the files as NnNn_nameOfTheVideo.eaf (NnNn = Two first letters of first and family names). Save the file also in txt-format (Export as… -> Tab-limited text).

APPENDIX 3. Annotations of interaction events, videos 1–2

Video 1

Interaction event	Starting point	Ending point	Duration	Comments
Communication initiative	00:35.3	00:36.3	00:01.0	Child points the bubbles
Communication initiative	00:39.5	00:40.2	00:00.7	C gives the jar
Communication initiative	01:02.8	01:03.8	00:01.0	C gives the jar
Turn-taking	00:40.7	00:43.2	00:02.5	Dad kisses, child leans towards dad
Turn-taking	00:43.2	00:44.5	00:01.3	C gives the jar, D takes it
Turn-taking	01:02.8	01:05.4	00:02.5	C gives the jar, D takes it
Joint attention	00:24.2	00:24.7	00:00.5	Looking at each other
Joint attention	00:29.6	00:30.2	00:00.6	C looks at D after popping bubbles
Joint attention	00:34.7	00:38.0	00:03.3	C looks at D, points bubbles and looks back
Joint attention	01:02.8	01:06.8	00:04.0	Looking each other
Joint attention	01:23.1	01:25.5	00:02.4	Dad points, C looks at the same direction

Video 2

Interaction event	Starting point	Ending point	Duration	Comments
Communication initiative	00:47.2	00:49.4	00:02.1	Child walks to mother and shoes the jar
Communication initiative	01:38.2	01:39.2	00:01.1	C gives the jar to M
Turn-taking	00:52.3	00:54.5	00:02.1	M hands out a hand, C gives the jar
Turn-taking	01:13.6	01:15.2	00:01.6	M hands out a hand, C gives the jar
Turn-taking	01:38.2	01:39.9	00:01.7	C hands out the jar, M takes it
Joint attention	00:31.4	00:32.3	00:00.9	Unsure: C glances at M
Joint attention	00:37.9	00:38.7	00:00.8	C tries to open, glances at M
Joint attention	00:50.5	00:52.0	00:01.5	C tries to open, glances at M
Joint attention	01:13.8	01:14.9	00:01.1	Looking at each other
Joint attention	01:39.5	01:41.5	00:02.0	Looking at each other

APPENDIX 4. Annotations of interaction events, videos 3–4

Video 3

Interaction event	Starting point	Ending point	Duration	Comments
Communication initiative	00:32.1	00:33.4	00:01.3	Child hands out the jar to mother
Communication initiative	01:04.1	01:05.7	00:01.6	C hands out the jar to M
Turn-taking	00:32.1	00:34.3	00:02.1	C hands out the jar, M takes it
Joint attention	00:32.8	00:33.5	00:00.7	Looking at each other
Joint attention	01:04.2	01:06.1	00:01.9	Looking at each other

Video 4

Interaction event	Starting point	Ending point	Duration	Comments
Communication initiative	00:27.4	00:28.3	00:01.0	Child hands out the jar to mother
Communication initiative	00:59.1	01:01.7	00:02.6	C takes the jar, shows it to M and says: "Äiti aukee tää"
Communication initiative	01:03.9	01:04.7	00:00.8	L hands out the jar saying: "Äiti aukee"
Turn-taking	00:33.2	00:35.6	00:02.4	M hands out a hand, C gives the jar
Turn-taking	01:03.9	01:06.2	00:02.3	C hands out the jar, M takes it
Joint attention	00:27.6	00:28.2	00:00.7	C looks at M while showing the jar to her
Joint attention	00:32.1	00:33.7	00:01.6	Looking at each other
Joint attention	01:00.8	01:01.2	00:00.4	C glances at M while showing the jar

APPENDIX 5. Annotation of basic units, video 1

Basic unit	Starting point	Ending point	Comments	Basic unit	Starting point	Ending point	Comments
C gaze	00:00,0	00:00.5	jar	P gaze	00:34.2	00:36.2	child
C gaze	00:00.5	00:01.4	stick	P gaze	00:36.2	00:37.2	up
C gaze	00:01.4	00:06.7	bubbles	P gaze	00:37.2	00:38.5	child
C gaze	00:07.1	00:08.2	dad	P gaze	00:38.5	00:40.4	jar?
C gaze	00:08.2	00:09.1	jar	P gaze	00:40.4	00:44.2	child
C gaze	00:09.1	00:09.6	stick	P gaze	00:44.2	00:47.0	jar
C gaze	00:09.6	00:22.4	bubbles	P gaze	00:47.0	00:49.1	stick
C gaze	00:22.4	00:23.5	dad	P gaze	00:49.1	00:51.3	jar
C gaze	00:23.5	00:24.1	dad's shirt?	P gaze	00:51.3	00:53.1	stick
C gaze	00:24.1	00:29.5	bubbles	P gaze	00:53.1	00:55.1	jar
C gaze	00:29.5	00:30.6	dad	P gaze	00:55.1	00:56.8	stick
C gaze	00:30.6	00:33.9	dad's hand + bubbles	P gaze	00:56.8	01:02.1	jar?
C gaze	00:33.9	00:34.9	dad	P gaze	01:02.1	01:08.3	child
C gaze	00:34.9	00:35.9	bubbles	P gaze	01:08.3	01:10.1	jar?
C gaze	00:35.9	00:38.2	dad	P gaze	01:10.1	01:13.4	stick
C gaze	00:38.2	00:39.3	jar	P gaze	01:13.4	01:24.1	child
C gaze	00:39.3	00:43.0	dad	P gaze	01:24.1	01:24.8	bubble
C gaze	00:43.4	00:44.1	dad	P gaze	01:24.8	01:26.1	child
C gaze	00:44.1	00:45.6	camera	P gaze	01:26.1	01:26.9	jar
C gaze	00:45.6	00:46.9	dad	C jar handling	00:01.0	00:01.9	
C gaze	00:46.9	00:47.5	camera	C jar handling	00:06.0	00:09.2	
C gaze	00:47.5	00:47.9	?	C jar handling	00:38.9	00:39.5	takes
C gaze	00:48.0	00:57.5	bubbles	C jar handling	00:39.5	00:41.2	hands out
C gaze	01:00.4	01:01.7	dad	C jar handling	00:41.2	00:43.3	holds
C gaze	01:01.7	01:02.9	jar	C jar handling	00:43.3	00:44.4	hands out
C gaze	01:02.9	01:10.8	dad	C jar handling	00:49.3	00:51.8	-
C gaze	01:10.8	01:11.2	stick	C jar handling	01:02.2	01:02.9	takes
C gaze	01:11.2	01:12.7	bubbles	C jar handling	01:02.9	01:04.7	hands out
C gaze	01:12.7	01:13.7	?	C jar handling	01:27.1	01:27.9	takes
C gaze	01:13.7	01:17.0	bubbles	C jar handling	01:27.9	01:29.3	pulls along the floor
C gaze	01:22.9	01:23.7	bubble?	P jar handling	00:00.0	00:01.0	stick handling
C gaze	01:23.7	01:24.8	dad	P jar handling	00:01.0	00:01.9	blows
C gaze	01:24.8	01:25.7	bubble	P jar handling	00:01.9	00:04.6	holds stick
C gaze	01:25.7	01:26.5	dad	P jar handling	00:04.6	00:06.0	blows
C gaze	01:26.5	01:27.6	jar	P jar handling	00:06.0	00:09.2	stick handling
C gaze	01:27.6	01:28.3	dad	P jar handling	00:09.2	00:11.7	blows
P gaze	00:00.0	00:00.3	jar	P jar handling	00:11.7	00:15.4	stick handling
P gaze	00:00.3	00:02.0	stick	P jar handling	00:15.4	00:16.7	blows
P gaze	00:02.0	00:04.0	bubbles	P jar handling	00:16.7	00:19.1	stick handling
P gaze	00:04.0	00:06.7	stick	P jar handling	00:19.1	00:20.5	blows
P gaze	00:06.7	00:08.4	child	P jar handling	00:20.5	00:24.4	puts away
P gaze	00:08.4	00:11.7	stick	P jar handling	00:43.5	00:44.6	takes
P gaze	00:11.7	00:12.3	child?	P jar handling	00:44.6	00:46.3	opens
P gaze	00:12.3	00:14.5	jar	P jar handling	00:46.3	00:48.1	stick handling
P gaze	00:14.5	00:16.6	stick	P jar handling	00:48.1	00:49.3	blows
P gaze	00:16.6	00:17.6	jar	P jar handling	00:49.3	00:51.8	stick handling
P gaze	00:17.6	00:18.0	child?	P jar handling	00:51.8	00:53.1	blows
P gaze	00:18.0	00:18.4	jar	P jar handling	00:53.1	00:55.6	stick handling
P gaze	00:18.4	00:21.7	stick	P jar handling	00:55.6	00:56.7	blows
P gaze	00:21.7	00:22.5	jar	P jar handling	00:56.7	01:02.0	puts away
P gaze	00:22.5	00:23.1	child	P jar handling	01:03.4	01:06.9	takes
P gaze	00:23.1	00:23.8	bubble	P jar handling	01:06.9	01:09.2	opens
P gaze	00:23.8	00:24.2	jar	P jar handling	01:09.2	01:11.0	stick handling
P gaze	00:24.2	00:25.2	child	P jar handling	01:11.0	01:12.7	blows
P gaze	00:25.2	00:27.4	jar?	P jar handling	01:12.7	01:17.7	closes
P gaze	00:27.4	00:30.4	child?	P jar handling	01:17.7	01:26.1	holds jar
P gaze	00:30.4	00:34.2	own hand + "bubble"	P jar handling	01:26.1	01:26.8	puts away

APPENDIX 6. Annotation of basic units, video 2

Basic unit	Starting point	Ending point	Comment	Basic unit	Starting point	Ending point	Comment
P gaze	00:00.0	00:01.5	child	P gaze	01:48.4	01:51.0	child
P gaze	00:01.5	00:05.3	stick	C gaze	00:00.0	00:01.0	camera
P gaze	00:05.3	00:05.9	child	C gaze	00:01.0	00:03.9	mother
P gaze	00:05.9	00:06.9	jar	C gaze	00:07.7	00:23.7	bubbles
P gaze	00:06.9	00:07.6	child	C gaze	00:23.7	00:24.0	mother
P gaze	00:07.6	00:10.5	jar	C gaze	00:24.0	00:27.2	bubbles
P gaze	00:10.5	00:12.0	child	C gaze	00:27.2	00:27.7	mother
P gaze	00:12.0	00:12.4	stick	C gaze	00:27.7	00:37.8	jar
P gaze	00:12.4	00:12.8	child	C gaze	00:37.8	00:38.6	mother
P gaze	00:12.8	00:19.0	stick	C gaze	00:38.6	00:50.5	jar
P gaze	00:19.0	00:20.6	bubbles	C gaze	00:50.5	00:51.8	mother
P gaze	00:20.6	00:21.5	jar	C gaze	00:51.8	00:56.9	jar
P gaze	00:21.5	00:22.0	bubbles	C gaze	00:56.9	00:57.5	mother
P gaze	00:22.0	00:22.9	jar	C gaze	00:57.5	00:57.7	stick
P gaze	00:22.9	00:23.4	bubbles	C gaze	00:57.7	01:00.9	bubbles
P gaze	00:23.4	00:24.4	child	C gaze	01:00.9	01:03.8	jar
P gaze	00:24.4	00:25.2	bubbles	C gaze	01:03.8	01:08.3	bubbles
P gaze	00:25.2	00:28.7	child	C gaze	01:08.3	01:13.6	jar
P gaze	00:28.8	00:31.0	jar	C gaze	01:13.6	01:13.8	to the side
P gaze	00:31.0	00:34.5	child	C gaze	01:13.8	01:14.8	mother
P gaze	00:34.5	00:38.4	jar	C gaze	01:14.8	01:17.4	jar
P gaze	00:38.4	00:38.8	child	C gaze	01:17.4	01:32.9	bubbles
P gaze	00:38.8	00:40.5	jar	C gaze	01:32.9	01:33.4	mother
P gaze	00:40.5	00:42.0	child?	C gaze	01:33.4	01:39.7	jar
P gaze	00:42.0	00:48.4	jar	C gaze	01:39.7	01:42.5	mother
P gaze	00:48.4	00:48.8	child	C gaze	01:42.5	01:49.4	bubbles
P gaze	00:48.8	00:50.7	jar	C gaze	01:49.4	01:50.2	jar
P gaze	00:50.7	00:51.2	child	P jar handling	00:00.0	00:01.6	opens
P gaze	00:51.2	00:57.6	jar	P jar handling	00:01.6	00:04.4	stick handling
P gaze	00:57.6	00:59.2	bubbles	P jar handling	00:04.4	00:05.0	blows
P gaze	00:59.2	01:00.3	jar	P jar handling	00:05.0	00:16.3	stick handling
P gaze	01:00.3	01:01.9	child	P jar handling	00:16.3	00:17.4	blows
P gaze	01:01.9	01:04.0	jar	P jar handling	00:17.4	00:18.6	stick handling
P gaze	01:04.0	01:08.8	bubbles	P jar handling	00:18.6	00:22.7	puts away
P gaze	01:08.8	01:09.4	jar	P jar handling	00:52.4	00:54.4	reaching towards c
P gaze	01:09.4	01:09.8	child	P jar handling	00:54.4	00:56.3	opens
P gaze	01:09.8	01:11.7	jar	P jar handling	00:56.3	00:57.4	stick handling
P gaze	01:11.7	01:12.7	child	P jar handling	00:57.4	00:58.1	blows
P gaze	01:12.7	01:18.7	jar	P jar handling	00:58.1	01:00.8	stick handling
P gaze	01:18.7	01:20.0	bubbles	P jar handling	01:00.8	01:03.5	puts away
P gaze	01:20.0	01:20.8	jar	P jar handling	01:13.7	01:15.4	reaching towards c
P gaze	01:20.8	01:21.7	bubbles	P jar handling	01:15.4	01:16.9	opens
P gaze	01:21.7	01:22.5	child	P jar handling	01:16.9	01:18.3	stick handling
P gaze	01:22.5	01:26.0	bubbles	P jar handling	01:18.3	01:19.2	blows
P gaze	01:26.0	01:29.9	child	P jar handling	01:19.2	01:21.2	stick handling
P gaze	01:29.9	01:31.4	bubbles	P jar handling	01:21.2	01:24.0	puts away
P gaze	01:31.4	01:32.4	child	P jar handling	01:38.2	01:40.2	reaching towards c
P gaze	01:32.4	01:32.8	bubbles	P jar handling	01:40.2	01:41.8	opens
P gaze	01:32.8	01:34.6	child	P jar handling	01:41.8	01:42.9	stick handling
P gaze	01:34.6	01:37.0	jar	P jar handling	01:42.9	01:43.6	blows
P gaze	01:37.0	01:38.5	child	P jar handling	01:43.6	01:45.9	stick handling
P gaze	01:38.5	01:39.6	jar	P jar handling	01:45.9	01:51.1	puts away
P gaze	01:39.6	01:41.2	child	C jar handling	00:29.9	00:32.2	raising the jar
P gaze	01:41.2	01:43.1	stick	C jar handling	00:32.3	00:52.9	opening the jar
P gaze	01:43.1	01:43.9	bubbles	C jar handling	00:52.9	00:54.1	giving the jar
P gaze	01:43.9	01:44.2	child	C jar handling	01:10.6	01:11.8	raising the jar
P gaze	01:44.2	01:45.0	jar	C jar handling	01:11.8	01:13.9	opening the jar
P gaze	01:45.0	01:45.8	child	C jar handling	01:13.9	01:14.9	giving the jar
P gaze	01:45.8	01:46.7	bubbles	C jar handling	01:35.2	01:38.3	raising the jar
P gaze	01:46.7	01:47.8	child	C jar handling	01:38.3	01:39.6	opening & giving
P gaze	01:47.8	01:48.4	bubbles				

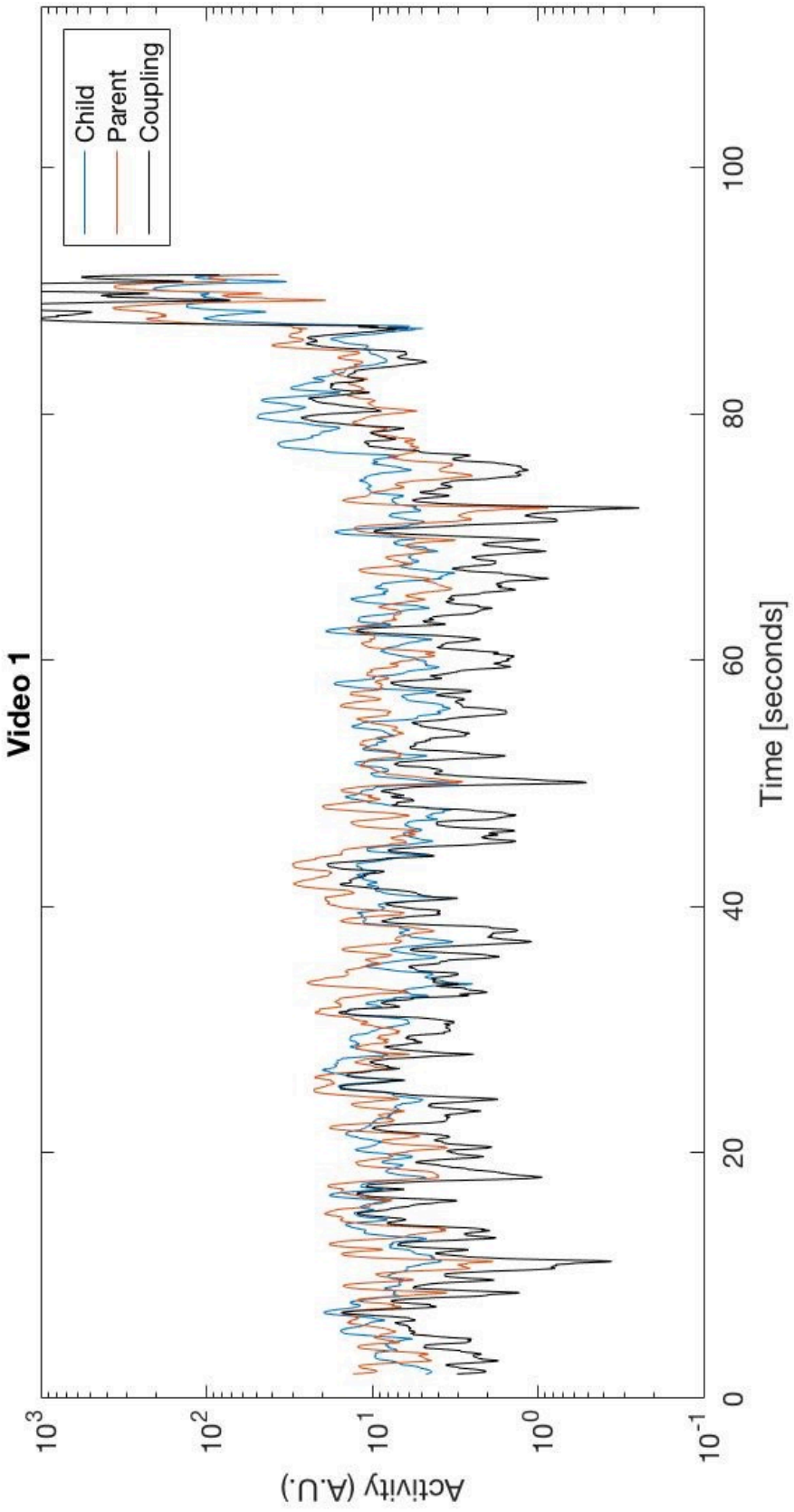
APPENDIX 7. Annotation of basic units, video 3

Basic unit	Starting point	Ending point	Comment	Basic unit	Starting point	Ending point	Comment
C gaze	00:00,0	00:00,3	jar	P gaze	00:36,8	00:37,3	child
C gaze	00:00,3	00:00,9	camera	P gaze	00:37,3	00:38,3	jar
C gaze	00:00,9	00:01,3	mother	P gaze	00:38,3	00:42,9	stick
C gaze	00:01,3	00:02,6	jar	P gaze	00:42,9	00:46,2	jar
C gaze	00:02,6	00:04,0	at own lap	P gaze	00:46,2	00:49,4	stick
C gaze	00:04,0	00:05,4	stick	P gaze	00:49,4	00:50,3	jar
C gaze	00:05,4	00:09,0	bubbles?	P gaze	00:50,3	00:54,1	child
C gaze	00:09,0	00:10,2	bubbles?	P gaze	00:54,1	00:56,0	jar
C gaze	00:10,2	00:11,0	mother	P gaze	00:56,0	01:02,7	child
C gaze	00:11,0	00:12,3	jar	P gaze	01:02,7	01:04,0	jar
C gaze	00:12,3	00:12,9	stick	P gaze	01:04,0	01:05,3	child
C gaze	00:12,9	00:17,2	bubbles	P gaze	01:05,3	01:06,5	jar
C gaze	00:17,2	00:18,2	jar	P gaze	01:06,5	01:07,5	ahead
C gaze	00:18,2	00:24,1	bubbles	C jar handling	00:25,7	00:28,4	lifting the jar
C gaze	00:24,1	00:32,0	jar	C jar handling	00:28,4	00:32,0	holding jar
C gaze	00:32,0	00:33,5	mother	C jar handling	00:32,0	00:34,0	giving
C gaze	00:33,5	00:34,5	ahead	C jar handling	00:54,8	00:56,5	lifting the jar
C gaze	00:34,5	00:35,0	mother	C jar handling	00:57,0	01:04,1	holding
C gaze	00:35,0	00:36,2	jar	C jar handling	01:04,1	01:05,9	giving
C gaze	00:36,2	00:37,3	to the side	P jar handling	00:00,0	00:02,8	opening
C gaze	00:37,3	00:38,7	stick	P jar handling	00:02,8	00:05,2	stick handling
C gaze	00:38,7	00:44,4	bubbles	P jar handling	00:05,2	00:07,7	blowing
C gaze	00:44,4	00:47,1	jar	P jar handling	00:07,7	00:12,7	stick handling
C gaze	00:47,1	00:52,6	bubbles	P jar handling	00:12,7	00:14,3	blowing
C gaze	00:52,6	00:54,1	ahead	P jar handling	00:14,3	00:18,7	stick handling
C gaze	00:54,1	01:04,0	jar	P jar handling	00:18,7	00:20,3	blowing
C gaze	01:04,0	01:06,1	mother	P jar handling	00:20,3	00:21,4	stick handling
C gaze	01:06,1	01:07,5	ahead	P jar handling	00:21,4	00:23,3	closing
P gaze	00:00,0	00:00,3	child	P jar handling	00:23,3	00:24,2	putting away
P gaze	00:00,3	00:02,5	jar	P jar handling	00:24,2	00:25,4	closing
P gaze	00:02,5	00:02,8	child	P jar handling	00:25,4	00:26,1	putting away
P gaze	00:02,8	00:03,5	jar	P jar handling	00:32,8	00:33,6	reaching towards c
P gaze	00:03,5	00:07,7	stick	P jar handling	00:33,6	00:37,2	opening
P gaze	00:07,7	00:12,1	jar	P jar handling	00:37,2	00:38,7	stick handling
P gaze	00:12,1	00:14,3	stick	P jar handling	00:38,7	00:42,8	blowing
P gaze	00:14,3	00:18,3	jar	P jar handling	00:42,8	00:46,9	stick handling
P gaze	00:18,3	00:20,2	stick	P jar handling	00:46,9	00:48,2	blowing
P gaze	00:20,2	00:27,0	jar	P jar handling	00:48,2	00:49,9	stick handling
P gaze	00:27,0	00:33,5	child	P jar handling	00:49,9	00:54,2	closing
P gaze	00:33,5	00:35,0	jar	P jar handling	00:54,2	00:55,0	putting away
P gaze	00:35,0	00:36,3	child	P jar handling	01:05,0	01:05,8	reaching towards c
P gaze	00:36,3	00:36,8	jar	P jar handling	01:05,8	01:07,4	opening

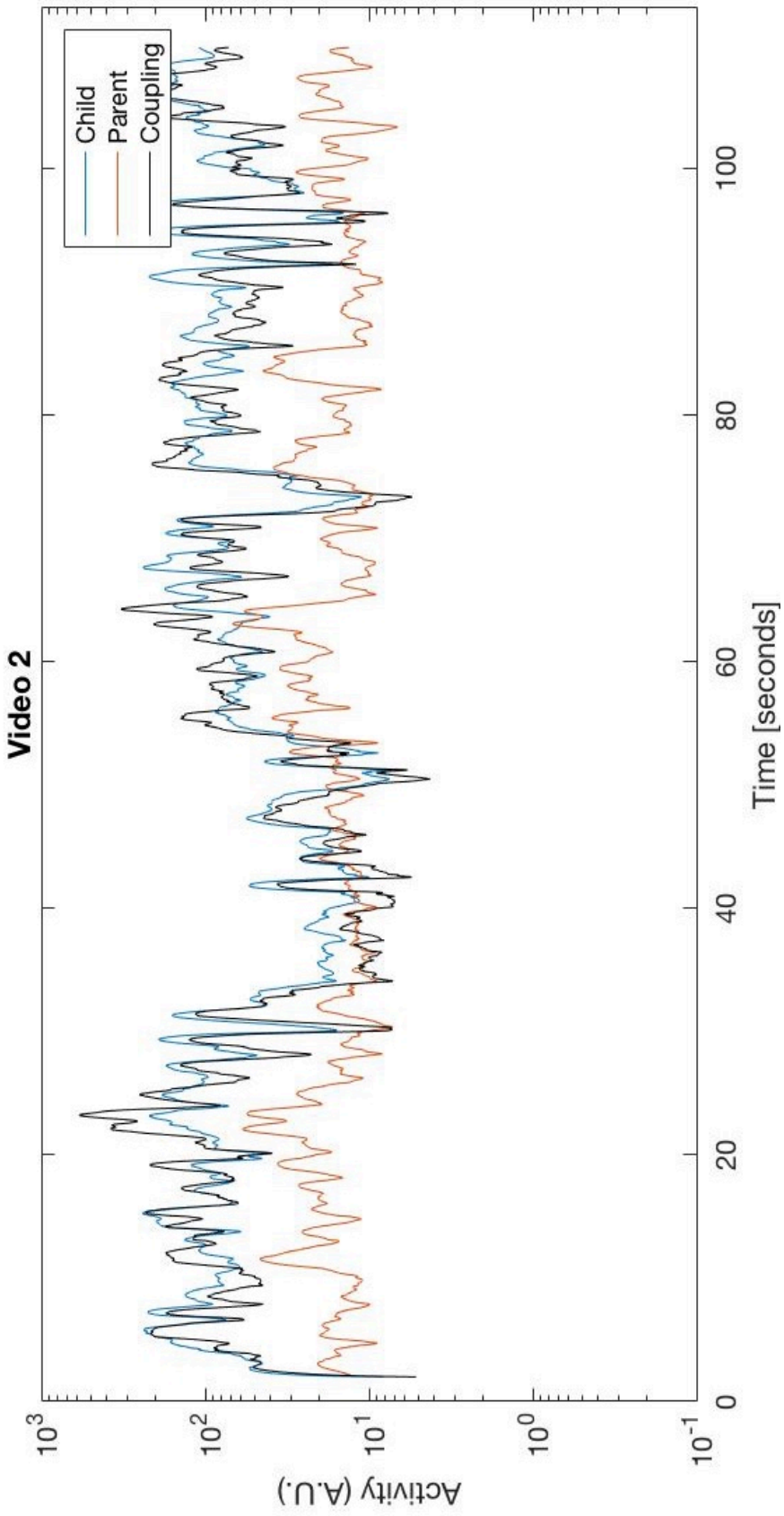
APPENDIX 8. Annotation of basic units, video 4

Basic unit	Starting point	Ending point	Comment	Basic unit	Starting point	Ending point	Comment
P gaze	00:00,2	00:03,4	child	C gaze	00:41,2	00:45,6	bubble
P gaze	00:03,4	00:06,9	jar	C gaze	00:50,0	00:51,2	parent?
P gaze	00:06,9	00:11,8	stick	C gaze	00:51,2	00:53,1	bubble
P gaze	00:11,8	00:14,6	jar	C gaze	00:54,5	00:54,8	jar?
P gaze	00:14,6	00:16,3	stick	C gaze	00:54,8	00:56,1	down, bubble?
P gaze	00:16,3	00:16,8	child	C gaze	00:56,1	00:58,1	parent?
P gaze	00:16,8	00:18,3	jar	C gaze	00:58,1	01:00,1	jar
P gaze	00:18,3	00:21,6	child	C gaze	01:00,1	01:01,0	parent
P gaze	00:21,6	00:22,9	jar	C gaze	01:01,0	01:03,9	jar
P gaze	00:22,9	00:23,1	child	C gaze	01:03,9	01:06,2	parent
P gaze	00:23,1	00:27,4	jar	C gaze	01:06,2	01:10,0	jar
P gaze	00:27,4	00:34,1	child	C gaze	01:10,0	01:10,9	stick?
P gaze	00:34,1	00:34,8	jar	C gaze	01:10,9	01:16,4	bubble
P gaze	00:34,8	00:37,6	child	P jar handling	00:00,0	00:00,5	holding
P gaze	00:37,6	00:39,7	jar	P jar handling	00:00,5	00:03,6	opening
P gaze	00:39,7	00:41,7	stick	P jar handling	00:03,6	00:08,9	stick handling
P gaze	00:41,7	00:44,5	jar	P jar handling	00:08,9	00:11,9	blowing
P gaze	00:44,5	00:45,2	stick	P jar handling	00:11,9	00:15,2	stick handling
P gaze	00:45,2	00:47,3	child	P jar handling	00:15,2	00:15,8	blowing
P gaze	00:47,3	00:52,9	stick	P jar handling	00:15,8	00:17,1	stick handling
P gaze	00:52,9	00:53,1	bubble	P jar handling	00:17,1	00:21,4	closing
P gaze	00:53,1	00:54,1	stick	P jar handling	00:21,4	00:22,5	putting away
P gaze	00:54,1	00:57,8	child	P jar handling	00:33,3	00:35,1	taking
P gaze	00:57,8	00:59,7	jar	P jar handling	00:35,1	00:36,2	holding
P gaze	00:59,7	01:05,4	child	P jar handling	00:36,2	00:38,8	opening
P gaze	01:05,4	01:06,0	jar	P jar handling	00:38,8	00:40,6	stick handling
P gaze	01:06,0	01:06,9	child	P jar handling	00:40,6	00:41,7	blowing
P gaze	01:06,9	01:08,6	jar?	P jar handling	00:41,7	00:45,8	stick handling
P gaze	01:08,6	01:09,2	jar	P jar handling	00:45,8	00:50,8	holding stick&jar
P gaze	01:09,2	01:09,6	child	P jar handling	00:50,8	00:51,8	blowing
P gaze	01:09,6	01:10,2	jar	P jar handling	00:51,8	00:54,4	stick handling
P gaze	01:10,2	01:11,6	stick	P jar handling	00:54,4	00:57,8	closing
P gaze	01:11,6	01:11,8	bubble	P jar handling	00:57,8	00:58,5	putting away
P gaze	01:11,8	01:12,4	stick	P jar handling	01:04,3	01:05,3	taking
P gaze	01:12,4	01:13,0	jar	P jar handling	01:05,3	01:06,1	holding
P gaze	01:13,0	01:13,5	child	P jar handling	01:06,1	01:09,4	opening
P gaze	01:13,5	01:16,0	bubble	P jar handling	01:09,4	01:10,8	stick handling
P gaze	01:16,0	01:17,0	child	P jar handling	01:10,8	01:11,8	blowing
C gaze	00:00,0	00:01,1	parent?	P jar handling	01:11,8	01:12,9	stick handling
C gaze	00:01,1	00:06,4	jar	P jar handling	01:12,9	01:16,6	closing
C gaze	00:06,4	00:09,6	stick?	C jar handling	00:22,2	00:23,3	taking
C gaze	00:09,6	00:13,3	bubble	C jar handling	00:23,3	00:24,2	holding
C gaze	00:13,3	00:15,7	stick	C jar handling	00:24,2	00:27,4	trying to open
C gaze	00:15,7	00:20,8	bubble	C jar handling	00:27,4	00:28,1	handing out
C gaze	00:20,8	00:27,3	jar	C jar handling	00:28,1	00:33,6	holding
C gaze	00:27,3	00:28,1	parent	C jar handling	00:33,6	00:35,3	handing out
C gaze	00:28,1	00:31,6	jar	C jar handling	00:58,3	00:59,0	taking
C gaze	00:31,6	00:33,6	parent	C jar handling	00:59,0	01:04,0	holding
C gaze	00:33,6	00:35,3	jar	C jar handling	01:04,0	01:05,5	handing out

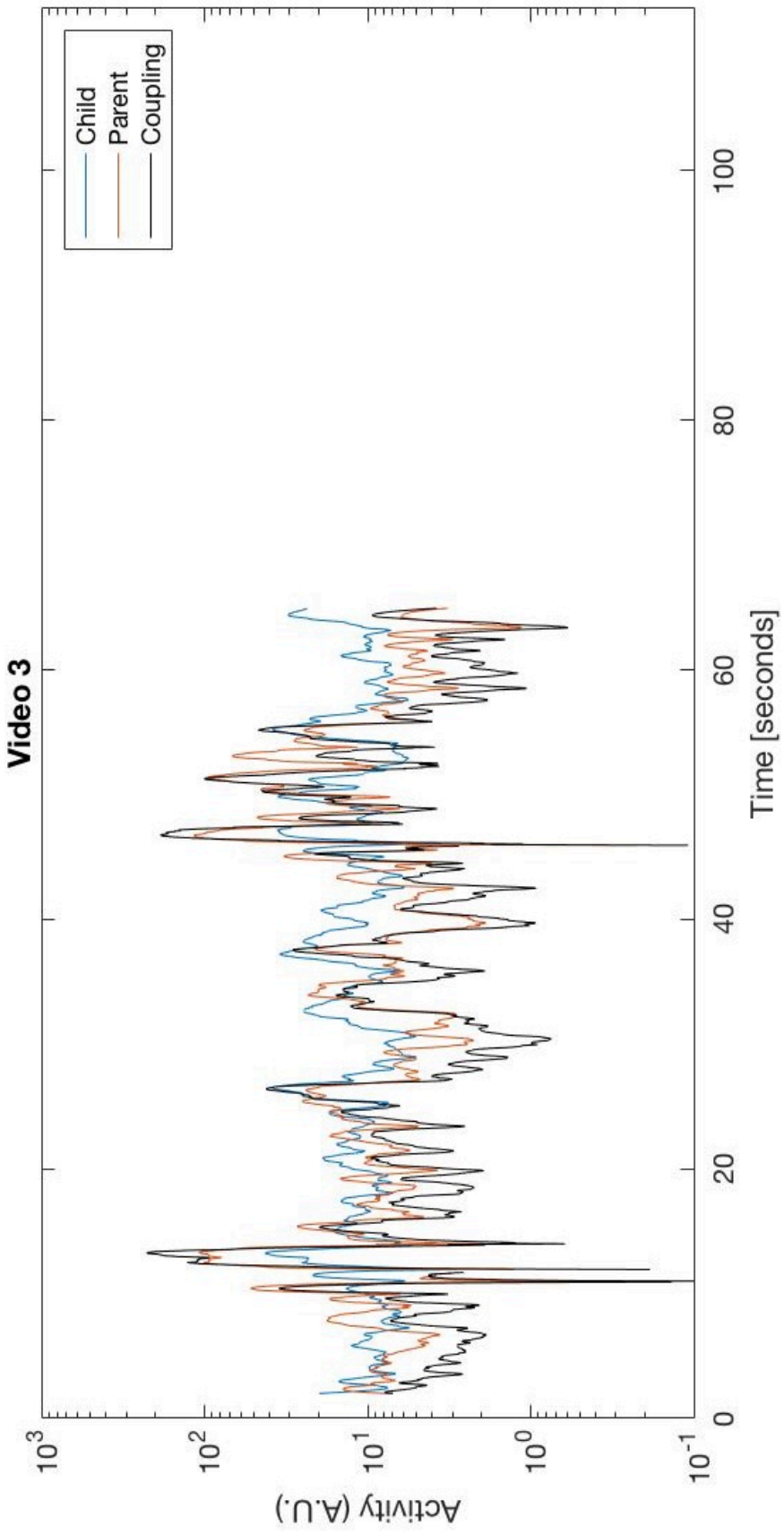
APPENDIX 9. Child's and parent's activities and coupling of the activities, video 1



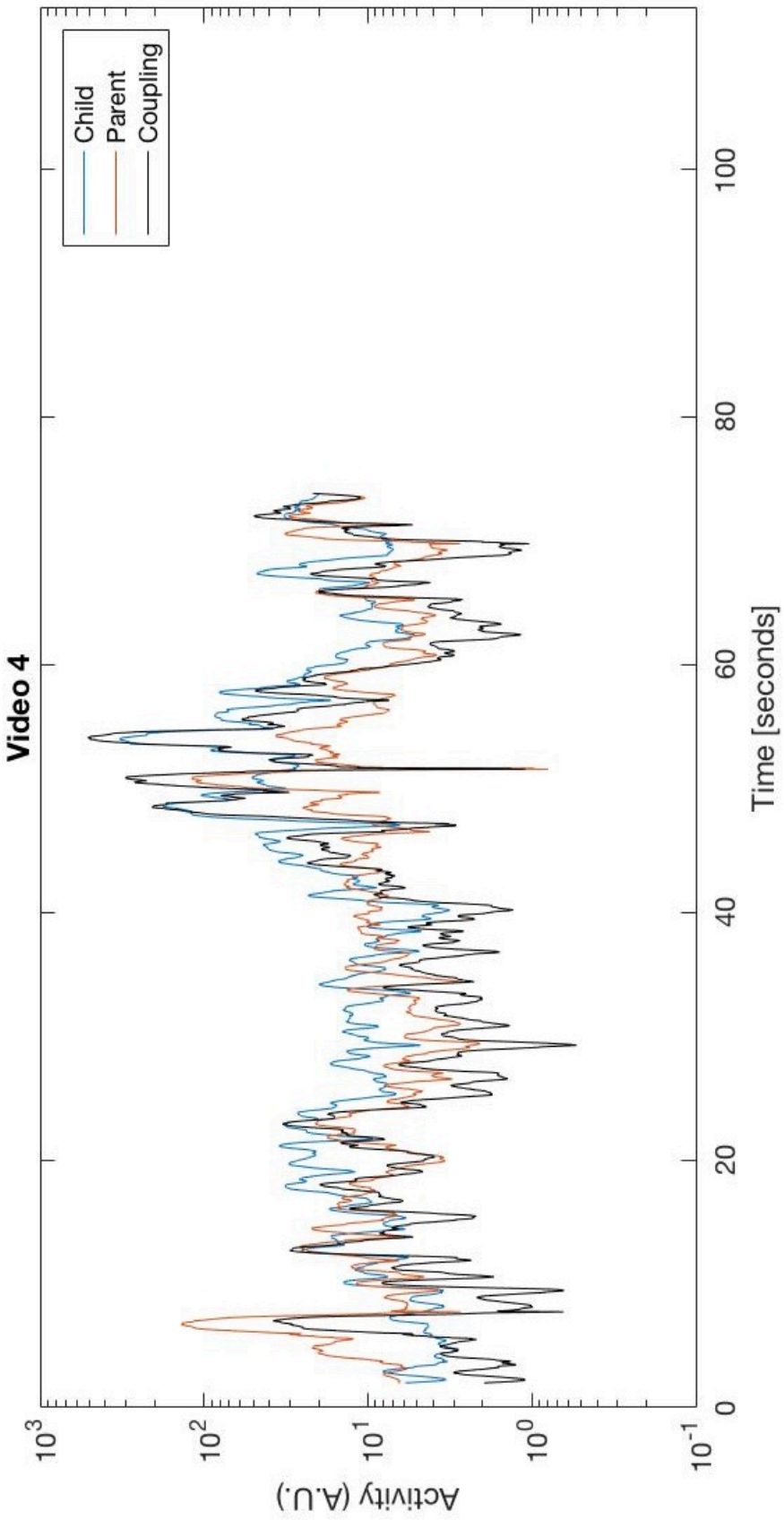
APPENDIX 10. Child's and parent's activities and coupling of the activities, video 2



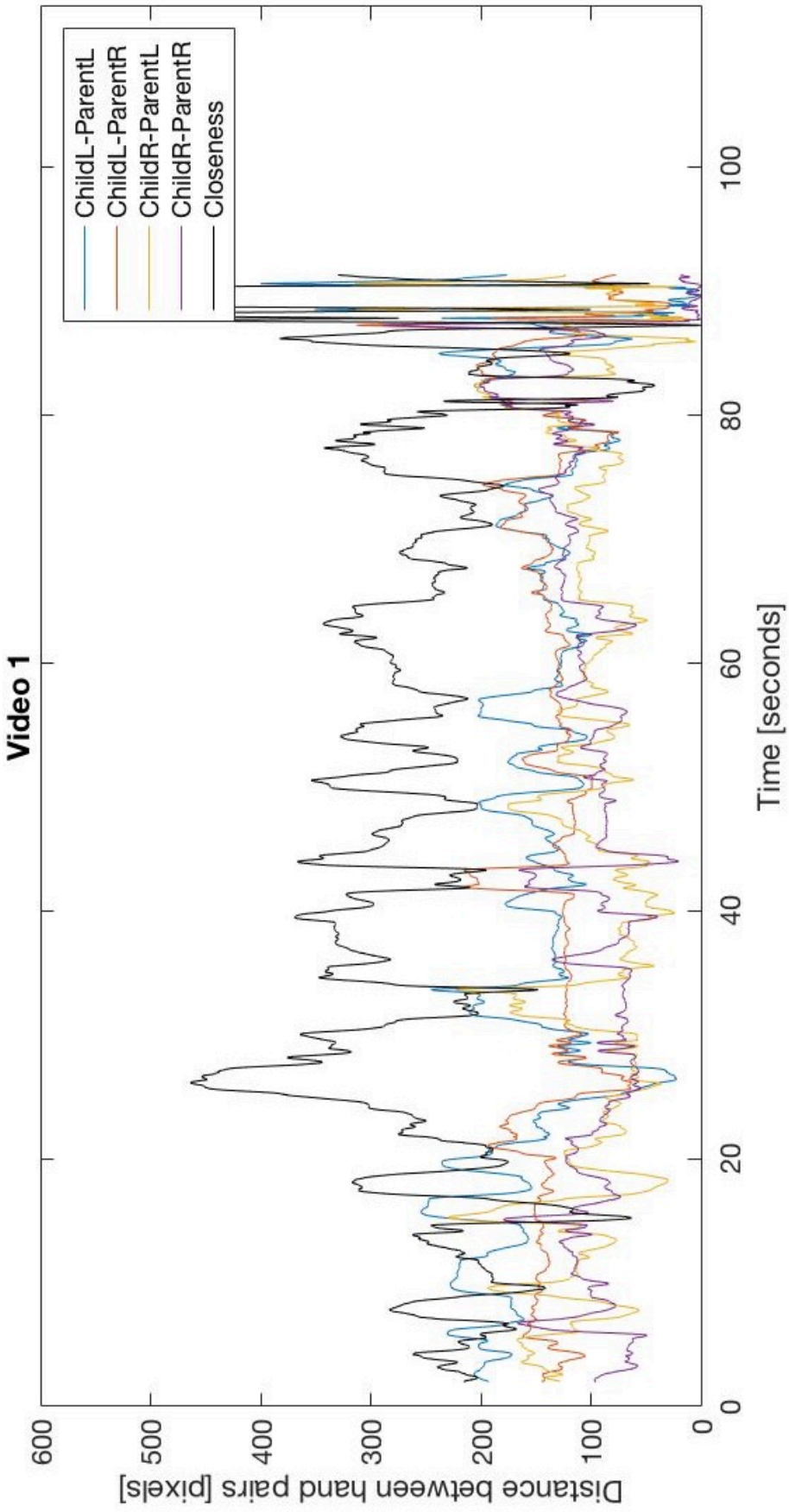
APPENDIX 11. Child's and parent's activities and coupling of the activities, video 3



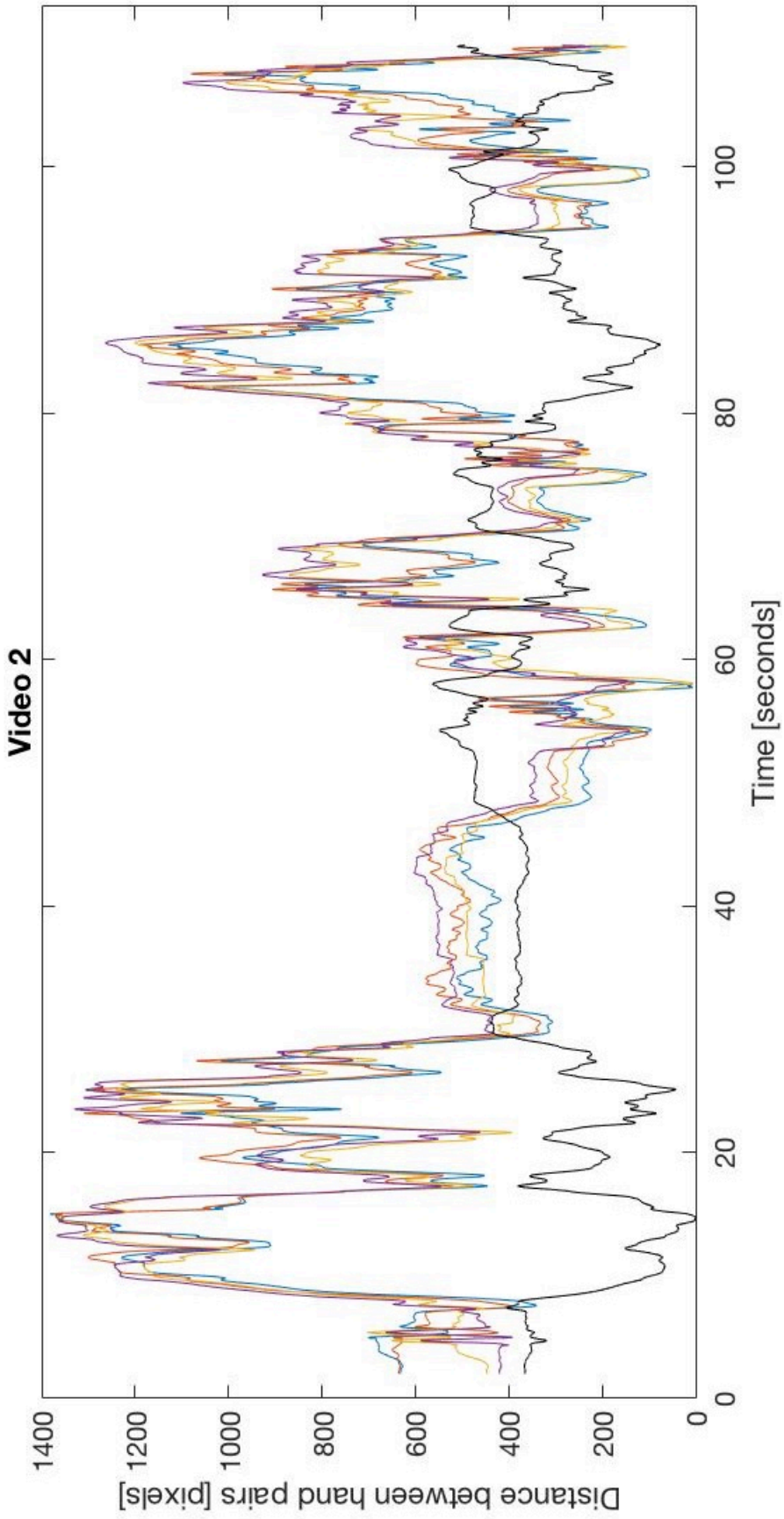
APPENDIX 12. Child's and parent's activities and coupling of the activities, video 4



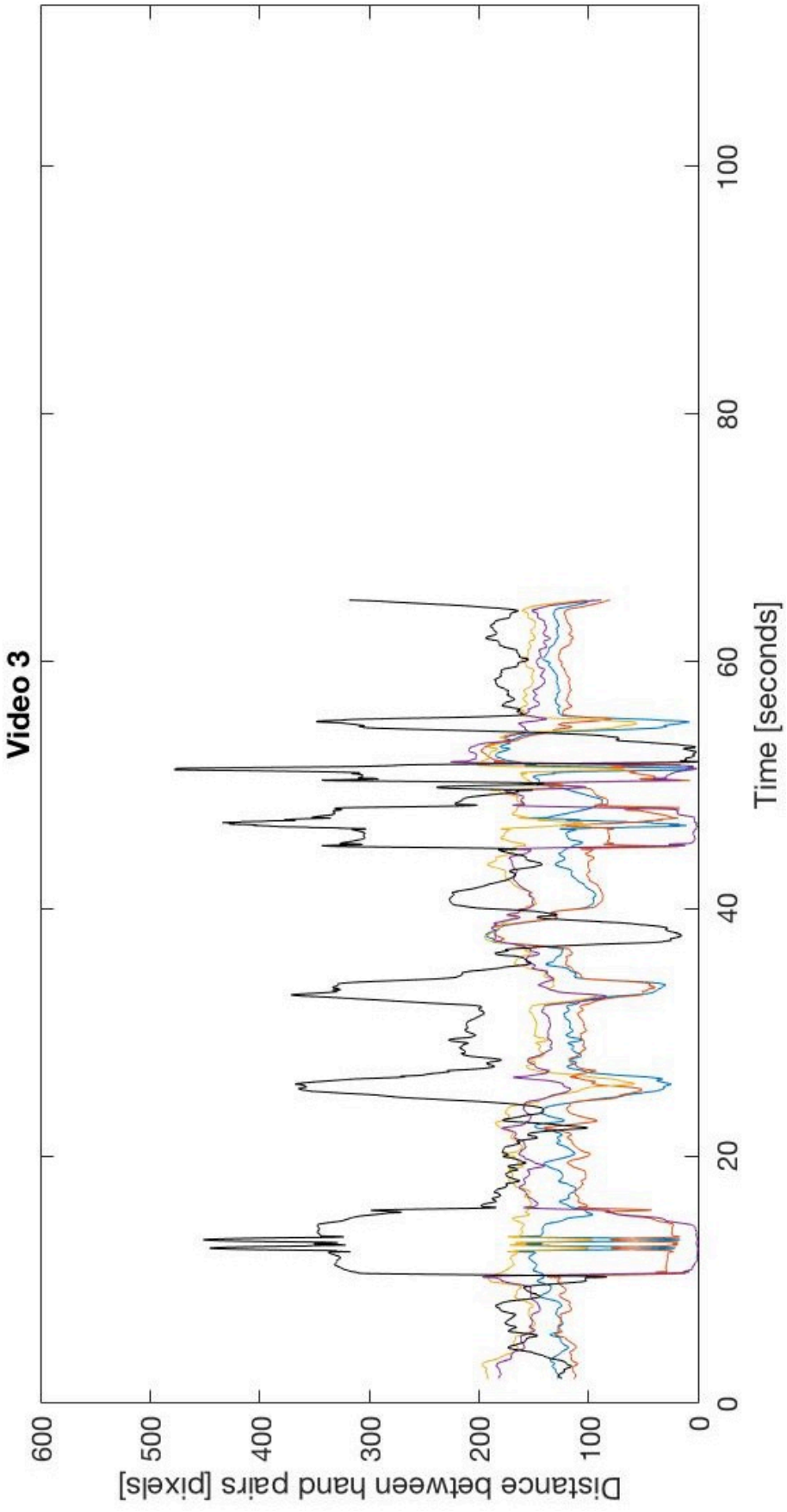
APPENDIX 13. Hand pair distances and the overall hand closeness, video 1



APPENDIX 14. Hand pair distances and the overall hand closeness, video 2



APPENDIX 15. Hand pair distances and the overall hand closeness, video 3



APPENDIX 16. Hand pair distances and the overall hand closeness, video 4