AUTOMATED GRADING OF PERIPHERAL FACIAL PALSY

TIMEN TEN HARKEL

AUTOMATED GRADING OF PERIPHERAL FACIAL PALSY

TIMEN TEN HARKEL

Scan the QR-code for the digital version or visit
https://www.radboudumc.nl/3dlab/phd/timentenharkel

ISBN
978-94-6473-724-0

Design
Timen ten Harkel

Print
Ipskamp Printing

Cover illustration
Timen ten Harkel

AUTOMATED GRADING OF PERIPHERAL FACIAL PALSY

Proefschrift ter verkrijging van de graad van doctor
aan de Radboud Universiteit Nijmegen
op gezag van de rector magnificus prof. dr. J.M. Sanders,
volgens besluit van het college voor promoties
in het openbaar te verdedigen op

maandag 7 april 2025
om 9.30 uur precies

door

Timen Christian ten Harkel
geboren op 31 mei 1990
te Hefshuizen

Promotoren
Prof. dr. T.J.J. Maal
Prof. dr. H.A.M. Marres

Copromotoren
Dr. K.J.A.O. Ingels
Dr. C.M. Speksnijder (UMC Utrecht)

Manuscriptcommissie
Prof. dr. D.J.O. Ulrich
Prof. dr. P.J. van der Wees
Prof. dr. P.P.G. van Benthem (Universiteit Leiden)

AUTOMATED GRADING OF PERIPHERAL FACIAL PALSY

Dissertation to obtain the degree of doctor
from Radboud University Nijmegen
on the authority of the Rector Magnificus prof. dr. J.M. Sanders,
according to the decision of the Doctorate Board
to be defended in public on

Monday, April 7, 2025
at 9:30 am

by

Timen Christian ten Harkel
born on May 31, 1990
in Hefshuizen, the Netherlands

PhD supervisors
Prof. dr. T.J.J. Maal
Prof. dr. H.A.M. Marres

PhD co-supervisors
Dr. K.J.A.O. Ingels
Dr. C.M. Speksnijder (UMC Utrecht)

Manuscript Committee
Prof. dr. D.J.O. Ulrich
Prof. dr. P.J. van der Wees
Prof. dr. P.P.G. van Benthem (Leiden University)

CONTENTS

CHAPTER 1

GENERAL INTRODUCTION

INTRODUCTION

The human face is a fascinating and complex topic, as it serves as the primary means of verbal and nonverbal communication for the majority of individuals. A key component in normal facial functioning is the seventh cranial nerve, appropriately called the facial nerve. The facial nerve generally terminates into five motor branches; the temporal, zygomatic, buccal, marginal mandibular, and cervical branch, which control the muscles most important for facial expression [1–3]. Although certain muscles might be innervated by multiple terminal branches, each branch has a general region of innervation. The temporal branch innervates the area of the forehead including the eyebrows, the zygomaticus innervates the eyes, the buccal branch innervates the middle part of the face including nose, cheeks, and upper lip, the marginal mandibular branch innervates the lower lip and chin, and the cervical branch innervates the neck [1–3]. As the facial nerve plays a major role in a wide variety of facial movements an impairment of the facial nerve can result in a significant impact on the physical, social, and emotional quality of life of a patient [4–8]. A unilateral peripheral facial palsy (PFP) is a symptom caused by a lesion to the facial nerve on the ipsilateral side of the face. The PFP may result in either a partial loss of nerve function which limits the range of voluntary movement (paresis), or a complete loss of nerve function resulting in the inability to move (paralysis). Apart from the palsy itself, other symptoms can accompany a PFP, such as an altered facial sensation, vestibular dysfunction, excessive tearing, hyperacusis, phonophobia, dry eyes, and taste disorders [9,10].

To diagnose a PFP it is first important to determine whether the symptoms relate to a central or peripheral facial palsy [9,11,12]. In case of a central lesion, a patient can still move their forehead and close their eyes, as the upper half of the face is bilaterally innervated [3]. When a central facial palsy is excluded, the cause for the PFP needs to be determined. As there are over 50 etiologies that can affect the facial nerve function, the patient history and physical examination will play an important role to determine the cause of the PFP [10,11,13–15]. Some of the most common causes of a PFP can be categorized as infectious (Lyme, otitis media), viral (varicella zoster virus, herpes simplex virus), neoplastic (acoustic neuroma, parotid malignancy), and traumatic (fracture, birth trauma) [10,11,13,14,16]. In case of a lacking explanation of the PFP it is called an idiopathic PFP. The idiopathic PFP is the major contributor of the PFP cases with 40% to 70% of the cases and an annual incidence rate of 10 to 40 cases per 100,000 individuals [10,13,14,16,17].

The treatment plan will depend on the exact cause of the PFP, but eye care is generally the most important factor during any acute episode of a PFP due to the potential inability for the patient to close their eye [11,13,18]. To prevent the cornea from drying out and become susceptible to injury several interventions can be recommended such as the usage of eye drops, wearing an eye patch, or adding weights to the top eyelid [11–14]. Additionally, a corticosteroid therapy such as prednisolone is regularly administered in case of an acute idiopathic PFP with the most common dosage of 60 mg prednisone per day for approximately 5 to7 days followed by a 5-day taper [14,16,18,19]. In case of idiopathic PFP, antiviral drugs can be used alongside corticosteroids, such as acyclovir and valacyclovir, although there is no consensus regarding the optimal dosage [13,14,18]. If Lyme disease is diagnosed, antibiotics such as ceftriaxone or doxycycline can be administered [13,14,18]. Apart from pharmacological intervention, there are several methods to possibly improve rehabilitation, such as surgical intervention, mime therapy, physical therapy, biofeedback, and electrotherapy, although the latter is thought to be obsolete nowadays [4,18,20].

During rehabilitation, an estimated 40% to 70% of patients will fully recover from a PFP [10,16]. In case of idiopathic PFP the recovery rate can be as high as 85% to 94% [10,16]. The rate of recovery will also depend on other factors such as the cause of the PFP, the initial severity of the PFP, voluntary activity in needle electromyography, and patient age [10,16,17]. The onset of recovery can be relatively fast where 85% of patients with a PFP see an improvement in facial function within three weeks [10]. In the remaining 15%, some facial functioning generally begins to return after 3 to 5 months. Four percent of the patients remain with severe sequelae, including hemifacial spasms, contractures, or synkinesis. Synkinesis is the involuntary and undesirable activation of facial muscles during the execution of a desired voluntary facial movement of a different facial muscle group [21]. An example is the closing or narrowing of the eye during a smile. Synkinesis is thought to be caused by regrowth of the facial nerve fibres in the incorrect region, resulting in the innervation of an unwanted facial muscle [21].

Sunnybrook Facial Grading System

Resting Symmetry

Compared to normal side

Eye (choose one only)	
Normal	0
Narrow	1
Wide	1
Eyelid surgery	1
Cheek (naso-labial fold)	
Normal	0
Abscent	2
Less pronounced	1
More pronounced	1
Mouth	
Normal	0
Corner dropped	1
Corner pulled up / out	1

Total ☐

Resting symmetry score — Total x 5 ☐

Symmetry of Voluntary Movement

Degree of muscle EXCURSION compared to normal side

Standard expressions	Unable to initiate movement/no movement	Initiates slight movement	Initiated movement with mild excursion	Movement almost complete	Movement complete	
Forehead wrinkle (FRO)	1	2	3	4	5	☐
Gentle eye closure (OCS)	1	2	3	4	5	☐
Open mouth smile (ZYG/ RIS)	1	2	3	4	5	☐
Snarl (LLA / LLS)	1	2	3	4	5	☐
Lip Pucker (OOS / OOI)	1	2	3	4	5	☐

Gross asymmetry	Severe asymmetry	Moderate asymmetry	Mild asymmetry	Normal Symmetry	
				Total	☐

Voluntary movement score — Total x 4 ☐

Synkinesis

Rate the degree of INVOLUNTARY MUSCLE CONTRACTION associated with each expression

Standard expressions	**NONE:** No synkinesis or mass movement	**MILD:** Slight synkinesis	**MODERATE:** Obvious but not disfiguring synkinesis	**SEVERE:** Disfiguring synkinesis / Gross mass movement of several muscles	
Forehead wrinkle (FRO)	0	1	2	3	☐
Gentle eye closure (OCS)	0	1	2	3	☐
Open mouth smile (ZYG/ RIS)	0	1	2	3	☐
Snarl (LLA / LLS)	0	1	2	3	☐
Lip Pucker (OOS / OOI)	0	1	2	3	☐

Synkinesis score — Total ☐

Voluntary movement score ☐ **−** **Resting symmetry score** ☐ **−** **Synkinesis score** ☐ **=** **Composite score** ☐

Figure 1. Overview of the Sunnybrook Facial Grading System (SFGS) which analyses 13 elements of a unilateral peripheral facial palsy (adapted from Ross et al. [27]). The 13 elements are divided into three weighted subscores: the resting symmetry, the symmetry of volunty movement, and the synkinesis. The composite SFGS score is calculated by subtracting the resting symmetry subscore and synkinesis subscore from the symmetry of voluntary movement subscore. The SFGS includes brief instructions on how to assess each subscore and its elements, along with their corresponding scores. Additionally, the muscles most important during each of the standard expressions are indicated on the SFGS, consisting of the frontalis (FRO), orbicularis oculi (OCS), zygomaticus (ZYG), risorius (RIS), levator labii alaeque nasi (LLA), levator labii superioris (LLS), orbicularis oris superior (OOS), and orbicularis oris inferior (OOI).

GRADING OF A PERIPHERAL FACIAL PALSY

During the initial onset and progression of the PFP it is crucial to be able to assess the severity of the PFP. This requires a grading system that is sensitive to clinically relevant changes in facial nerve functioning, whilst being reliable and easy to implement in clinical practice. There are over 19 grading systems available to determine the severity of a PFP where the conventional subjective grading system was the House-Brackmann scale which was introduced in 1985 [12,19,22–25]. The House-Brackmann scale categorizes a PFP on a scale from normal functioning (Grade 1) to a complete palsy (Grade 6) [24]. By representing the PFP with a single global score, the overall grade is often assigned to the poorest functioning muscle group. This may not be representative of all muscular function and might be insensitive to changes during the rehabilitation of the PFP. Therefore, alternative grading systems have become more popular in the last decade.

The Sunnybrook Facial Grading System (SFGS) is emerging as one of the most popular PFP grading systems, which was introduced in 1996 [12,25–27]. During this time, the SFGS has been recommended multiple times as the standard grading system for PFP due to its clinical relevance, sensitivity, and reliable measuring method [12,19,22,23,25]. Therefore, the SFGS is routinely used at the Radboudumc in the Department of Otorhinolaryngology and the Department of Physical Therapy for the grading of patients with a PFP. The SFGS achieves the clinical relevance by assessing the facial nerve function of the muscles most important for facial expression and compares the function between the healthy and palsy side of the face. A total of 13 individual elements are assessed which are grouped into three subcomponents; the resting symmetry (3 elements), symmetry of voluntary movement (5 elements), and synkinesis (5 elements) [27]. The resting asymmetry assesses the eye, the cheek (naso-labial fold), and the corners of the mouth at rest. The symmetry of voluntary movement assesses five voluntary movements consisting of the forehead wrinkle, gentle eye closure, open mouth smile, snarl, and lip pucker. During the grading of the voluntary movements both the degree of muscle movement and the degree of asymmetry are compared to the healthy side of the face. The same five voluntary movements are used to determine the degree of synkinesis. The patient might be requested to perform the voluntary movements multiple times to complete the entire SFGS. After the grading of these 13 elements, each of the three subcomponents will results in its own weighted subscore. The resting symmetry subscore ranges from 0 to 20, the symmetry of voluntary movement subscore ranges from 20 to 100, and the synkinesis subscore ranges from 0 to 15. The composite score is then calculated by subtracting the resting symmetry subscore and synkinesis subscore from the symmetry of voluntary movement subscore. This results in a composite score ranging from 0 to 100, where a score of 0 indicates a complete flaccid PFP (without synkinesis) and a score of 100 indicates normal functioning of the facial muscles. A complete breakdown of the SFGS is shown in Figure 1.

AUTOMATION OF THE SUNNYBROOK FACIAL GRADING SYSTEM

As the SFGS is a subjective grading system, the grading is influenced by the individual input of an observer which could bias the assessment of the PFP. Although the SFGS is found to be a reliable grading system, there is a learning curve to achieve the optimal reliability, which makes the SFGS inaccessible to untrained observers [23,28–33]. This might make it unfeasible to increase the frequency in grading when monitoring the rehabilitation of a patient. These limitations of the SFGS could be alleviated by automating the grading of the SFGS. This automated system could be used independently by the patient, during online consultations in an eHealth environment, or by untrained co-workers. To increase the likelihood of adoption by clinicians and researchers, the barrier of entry for an automated SFGS should be as low as possible. Therefore, the automated SFGS should be relatively inexpensive, portable, non-invasive, reliable, and fast. Finally, the automated system should generate the same output as the manual SFGS, in order to keep the clinical relevance, validation, and experience gained over the years with the SFGS.

Imaging of the facial surface

A form of input data is required for the automation of the SFGS. As the SFGS relies on visual examination, one of the initial considerations would be to use non-ionizing and non-invasive imaging techniques, such as two-dimensional (2D) photos, 2D video, three-dimensional (3D) photos, or 3D video (4D). Due to the complex geometric structure of the human face and the significant amount of motion in the anterior-posterior plane during facial expression, the inclusion of depth data with 3D and 4D imaging could be beneficial in the assessment of facial asymmetry [34–37]. The most common methods for high accuracy 3D facial surface imaging are based on stereophotogrammetry, structured light imaging, or laser scanning [38]. Laser scanners are generally not suitable for dynamic measurements due to the relatively long acquisition times and are therefore not suitable for the automation of the SFGS. In contrast, stereophotogrammetry and structured light systems can achieve capture times in the millisecond range both suitable for 3D and 4D imaging [34,38–44].

Stereophotogrammetry

Stereophotogrammetry creates a depth image based on the information from two or more 2D images of a scene. These images can be captured by a single moving camera or multiple (synchronized) cameras. In a clinical setup, the most common practice is to use multiple cameras placed laterally apart, pointed towards a subject near the centre of the viewpoints of the cameras. Using this setup, it is possible to create depth images with an accuracy around 0.2 mm or better [34,36,40,43,45]. The calculation of

the depth image relies on the automatic detection of matching key points between two or more images. Without proper lighting conditions or insufficient overlap between the multiple images, the detection of matching key points will be affected. This could result a decreased accuracy of the depth image or even result in missing depth data [34]. To overcome the reliance on lighting conditions or lack of key points it is possible to use active stereophotogrammetry, in contrast to passive stereophotogrammetry. With active stereophotogrammetry an emitter is used to project a random pattern on the objects in the scene which will generate key points on the object which can be used for the reconstruction of the depth image. This projected pattern is usually in the infra-red (IR) range, as not to disturb the colours on the image visible to the human eye. However, both systems based on passive or active stereophotogrammetry can achieve a high accuracy with proper lighting conditions and camera setup [45].

Structured light imaging

Instead of relying on key points from multiple images to create a depth image, structured light imaging uses the distortion of a structured light pattern to determine the depth image [46]. An IR emitter is used to project a pattern on objects in the scene, similar to active stereophotogrammetry. However, in this case a structured light pattern is emitted, compared to a random pattern in active stereophotogrammetry. A second sensor will capture the reflected light pattern, where the reflected pattern will be distorted due to the geometry and position of objects in the scene. By comparing the original projected pattern to the distorted pattern, it is possible to reconstruct the depth image [46]. The usage of a structured light pattern generally results in a high depth accuracy, 0.2 mm or better, even in low light conditions or lack of natural key points in the scene, due to the use of an IR emitter [41–43,46]. However, the accuracy is correlated to the strength of the emitter and the absence of interfering signals in the IR range of the projected pattern, such as the sun or other emitters used in the same environment [38,41–44,47].

RealSense 4D camera

Traditional 3D and 4D imaging systems tend to be expensive (tens of thousands of US dollars), bulky, or overly complicated, which limits their usage to dedicated healthcare centres [34,37,41]. However, technical developments have made it possible to create inexpensive, portable 4D cameras [47]. One of the more recent developments in this space is the Intel RealSense™ (Intel®, Santa Clara, USA) camera range [48–51]. The RealSense F200 is a structured light 4D camera specifically developed for close range imaging and was released in 2015 for $100 USD [49]. The camera is the size of a webcam and consists of five core elements: the image processor, RGB colour sensor, IR sensor, IR laser projector, and a stereo microphone. The RealSense F200 allows for simultaneous capture of colour (1920 x 1080 pixels) and depth images (640 x 480 pixels),

with a frame rate around 30 frames per second (FPS). The output of the depth image is a set of individual points with a X, Y, and Z-coordinate, resulting in a point cloud. The 2D colour image can be projected onto the 3D point cloud, generating a 3D colour image. After a first generation of 4D cameras the RealSense D415 was launched in 2018 for $150 USD [49]. The RealSense D415 includes multiple hardware and software improvements compared to the RealSense F200. The biggest change is the switch to an active stereophotogrammetry based camera in the RealSense D415 with a higher depth resolution (1280 x 720 pixels) compared to the RealSense F200. Due to the switch to active stereophotogrammetry, the RealSense D415 is also better suited to be used in conditions with direct sunlight.

Data analysis

The next step in the automation of the SFGS is to convert the facial surface image data into the SFGS scores. This conversion from input data to a desired output is related to the field of machine learning. Broadly speaking, machine learning refers to the development and usage of algorithms which can be trained to perform a certain task without explicit programming. During the training stage of these algorithms, input data is supplied to identify patterns in the dataset and generalize this knowledge. When implemented successfully, the trained algorithm can properly execute the task even when unseen data from different situations is introduced. In case of the automated SFGS the task would be to accurately determine the SFGS score when a new patient with PFP is presented to the model.

Facial landmarks and anthropometric measurements

Historically, objective measurements of the face were performed with a calliper or measuring tape. With these tools specific features of the face were measured, such as the distance and angles between facial landmarks. This method of objectively quantifying and evaluating surface morphology on the human body is called anthropometry [40,52,53]. The direct anthropometric measurements are a reliable and affordable method to quantify the human face [40]. However, the execution of direct anthropometric measurements can be a very time-consuming task. Additionally, all necessary measurements must be taken during the initial assessment, as there is no opportunity to add or repeat measurements based on the patient's original condition at a later time. In case of the SFGS it even is infeasible for the patient to stay in maximum exertion during the voluntary movements to perform the anthropometric measurements. Due to these limitations, there has been a shift from direct anthropometric measurements towards digital 2D and 3D anthropometric measurements, where 3D measurements have been found to be a reliable alternative to direct anthropometric measurements [40,54]. The digital 3D measurements consist of placing the landmarks on the 3D image from which

the anthropometric measurements can be calculated, similar to direct anthropometric measurements. The 3D image also allows for more extended analysis of the face where the entire surface or volume of the face can be considered. Traditional machine learning algorithms are regularly based on features such as facial landmarks and anthropometric measurements, for the automation of a multitude of facial analysis tasks, which make these facial features an interesting option for the automation of the SFGS [55–57].

Deep learning

Despite the successful applications of traditional machine learning algorithms, the selection of the features for the input of the algorithms is crucial in this process to prevent the loss of valuable information [58,59]. To alleviate the issue of manual feature selection, a subset of machine learning, deep learning can be applied. Deep learning is a type of neural network, where the neural network consists of interconnected nodes which are organized into layers [60]. In deep learning, multiple layers are used in the neural network, hence the term deep. The nodes connect to each other in between the layers, with associated weights and thresholds, which can be adjusted during the training phase of the network. If a node's output exceeds the threshold, it activates and passes data to the next layer. Each layer processes and transforms the input data to automatically detect relevant features from the raw input data, removing the need for manual feature selection. A type of deep learning network, the convolutional neural network (CNN) is especially suited for the feature selection from images [60,61]. There have been many successful implementations of CNNs in (medical) image processing tasks where implementations can even exceed human performance [62–67]. Therefore, CNNs could prove to be a valuable tool for the automation of the SFGS.

RESEARCH QUESTIONS

This thesis investigates the automation of the SFGS using the RealSense F200 and D415 for the recording of patients with a PFP. The first part of the thesis validates the depth data of the RealSense cameras including their derived landmarks and anthropometric measurements, answering the following research questions:

1. What is the depth accuracy of the RealSense F200 (Chapter 2) and the RealSense D415 (Chapter 3) during the SFGS poses?
2. What is the reliability of 3D landmark placement on RealSense D415 images during the SFGS poses? (Chapter 3)
3. What is the reliability and agreement of 3D anthropometric measurements on RealSense D415 images during the SFGS poses? (Chapter 3)

The second part of the thesis implements an automated SFGS addressing the following research questions:

4. What is the reliability of an automated SFGS grading system based on a CNN compared to human observers? (Chapter 4)
5. What is the impact on the reliability of the automated SFGS by adding a facial landmark layer to the CNN? (Chapter 5)

THESIS OUTLINE

The SFGS is one of the major grading systems used to determine the severity and progression of a PFP but requires a trained observer for optimal reliability. Therefore, this thesis investigates the automation of the SFGS with the long-term aim to develop a user-friendly system that could be used by the patient at home without any assistance, whilst ideally exceeding the inter-rater reliability of human observers.

The changes in the facial surface during the SFGS poses can be captured in real-time when using a 4D imaging system, such as the portable and inexpensive RealSense F200 camera. Due to the difference in price and complexity compared to a professional setup, a lower image quality of the RealSense F200 is expected. Therefore, **Chapter 2** determines the depth accuracy of the RealSense F200 during the maximum exertion of the SFGS poses in a cohort of 34 patients with a PFP. The depth accuracy is validated by using the clinically validated 3dMD system (3dMDface, 3dMD, Atlanta, USA) as the gold standard. The results from **Chapter 2** can be used to determine if the RealSense F200 is a viable 4D camera for the implementation of the automated SFGS.

The RealSense F200 was superseded by the RealSense D415 with improvements on both the hardware and software level. Therefore, the clinical validation of the depth accuracy was repeated for the RealSense D415 in **Chapter 3**. A major group of facial analysis is based on facial landmarks or their derived anthropometric measurements, which can be useful for the automation of the SFGS. Therefore, the reliability of landmark placement and the reliability and agreement of the anthropometric measurements are determined for the RealSense D415 using the 3dMD system as the gold standard. **Chapter 3** is split into two separate parts, where both parts are based on the same dataset of 30 patients with a PFP. Part 1 of **Chapter 3** will discuss the measurements for the patients at rest and Part 2 will discuss the measurements during the voluntary movements of the SFGS. The validation of the RealSense D415 in **Chapter 3** enables the camera to be used in a clinical setting with a known impact on the depth accuracy, reliability of landmark placement, and reliability and agreement of anthropometric measurements.

A first version of the automated SFGS is implemented in **Chapter 4** based on a dataset of 116 patients with a PFP and 9 healthy subjects recorded with the RealSense D415. The automated SFGS uses CNNs to automatically score the 13 elements of the SFGS. From these elements the three subscores and composite score of the SFGS are determined, which are then compared to the manual grading based on three human observers all experienced in the grading with the SFGS. These results give a first indication of the expected reliability of the automated SFGS in a clinical setting.

Chapter 5 continues the development of the automated SFGS by adding a facial landmark layer to the CNN implemented in **Chapter 4**. In order to compare the results between the two chapters, as many potentially confounding variables are kept consistent. These results determine whether the facial landmarks can improve the reliability of the automated SFGS without increasing the size of the underlying dataset.

REFERENCES

1. Chhabda, S., Leger, D. S. & Lingam, R. K. Imaging the facial nerve: A contemporary review of anatomy and pathology. Eur. J. Radiol. 126, 108920 (2020).
2. Kochhar, A., Larian, B. & Azizzadeh, B. Facial Nerve and Parotid Gland Anatomy. Otolaryngol. Clin. North Am. 49, 273–284 (2016).
3. Ho, M.-L., Juliano, A., Eisenberg, R. L. & Moonis, G. Anatomy and Pathology of the Facial Nerve. Am. J. Roentgenol. 204, W612–W619 (2015).
4. Luijmes, R. E. et al. Quality of life before and after different treatment modalities in peripheral facial palsy: A systematic review. Laryngoscope 127, 1044–1051 (2017).
5. Bruins, T. E. et al. Interpreting Quality-of-Life Questionnaires in Patients with Long-Standing Facial Palsy. Facial Plast. Surg. aesthetic Med. 24, 75–80 (2022).
6. Tavares-Brito, J., van Veen, M. M., Dusseldorp, J. R., Bahmad, F. & Hadlock, T. A. Facial Palsy-Specific Quality of Life in 920 Patients: Correlation With Clinician-Graded Severity and Predicting Factors. Laryngoscope 129, 100–104 (2019).
7. Kleiss, I. J., Beurskens, C. H. G., Stalmeier, P. F. M., Ingels, K. J. A. O. & Marres, H. A. M. Quality of life assessment in facial palsy: validation of the Dutch Facial Clinimetric Evaluation Scale. Eur. Arch. Otorhinolaryngol. 272, 2055–2061 (2015).
8. Kleiss, I. J., Hohman, M. H., Susarla, S. M., Marres, H. A. M. & Hadlock, T. A. Health-related quality of life in 794 patients with a peripheral facial palsy using the FaCE Scale: A retrospective cohort study. Clin. Otolaryngol. 40, 651–656 (2015).
9. Ferreira-Penêda, J. et al. Peripheral facial palsy in emergency department. Iran. J. Otorhinolaryngol. 30, 145–152 (2018).
10. Peitersen, E. Bell's Palsy: The Spontaneous Course of 2,500 Peripheral Facial Nerve Palsies of Different Etiologies. Acta Otolaryngol. Suppl. 4–30 (2002).
11. Garro, A. & Nigrovic, L. E. Managing Peripheral Facial Palsy. Ann. Emerg. Med. 71, 618–624 (2018).
12. Kim, S. J. & Lee, H. Y. Acute Peripheral Facial Palsy: Recent Guidelines and a Systematic Review of the Literature. J. Korean Med. Sci. 35, e245 (2020).
13. Hohman, M. H. & Hadlock, T. A. Etiology, Diagnosis, and Management of Facial Palsy: 2000 Patients at a Facial Nerve Center. Laryngoscope 124, E283-93 (2014).
14. Steinhäuser, J. et al. Multidisciplinary Care of Patients with Facial Palsy: Treatment of 1220 Patients in a German Facial Nerve Center. J. Clin. Med. 11, (2022).
15. Butler, D. P., Morales, D. R., Johnson, K. & Nduka, C. Facial palsy: when and why to refer for specialist care. Br. J. Gen. Pract. 69, 579–580 (2019).
16. Geißler, K. et al. Non-idiopathic peripheral facial palsy: prognostic factors for outcome. Eur. Arch. Oto-Rhino-Laryngology 278, 3227–3235 (2021).
17. Sullivan, F. M. et al. Early Treatment with Prednisolone or Acyclovir in Bell's Palsy. N. Engl. J. Med. 357, 1598–1607 (2007).

18. Shokri, T., Azizzadeh, B. & Ducic, Y. Modern Management of Facial Nerve Disorders. Semin. Plast. Surg. 34, 277–285 (2020).
19. Fattah, A. Y. et al. Facial nerve grading instruments: Systematic Review of the Literature and Suggestion for Uniformity. Plast. Reconstr. Surg. 135, 569–579 (2015).
20. Nakano, H. et al. Physical therapy for peripheral facial palsy: A systematic review and meta-analysis. Auris. Nasus. Larynx 51, 154–160 (2024).
21. Lannadère, E. et al. Contribution of the Synkinesis Assessment Questionnaire and the Sunnybrook Facial Grading System to the evaluation of synkinesis after peripheral facial palsy: A STROBE observational study. Eur. Ann. Otorhinolaryngol. Head Neck Dis. 140, 8–12 (2022).
22. Samsudin, W. S. W. & Sundaraj, K. Evaluation and Grading Systems of Facial Paralysis for Facial Rehabilitation. J. Phys. Ther. Sci. 25, 515–519 (2013).
23. Niziol, R., Henry, F. P., Leckenby, J. I. & Grobbelaar, A. O. Is there an ideal outcome scoring system for facial reanimation surgery? A review of current methods and suggestions for future publications. J. Plast. Reconstr. Aesthetic Surg. 68, 447–456 (2015).
24. House, J. W. & Brackmann, D. E. Facial nerve grading system. Otolaryngol. – Head Neck Surg. 93, 146 (1985).
25. Berner, J. E., Kamalathevan, P., Kyriazidis, I. & Nduka, C. Facial synkinesis outcome measures: A systematic review of the available grading systems and a Delphi study to identify the steps towards a consensus. J. Plast. Reconstr. Aesthetic Surg. 72, 946–963 (2019).
26. Lapidus, J. B. et al. Too much or too little? A systematic review of postparetic synkinesis treatment. J. Plast. Reconstr. aesthetic Surg. 73, 443–452 (2020).
27. Ross, B. G., Fradet, G. & Nedzelski, J. M. Development of a sensitive clinical facial grading system. Otolaryngol. neck Surg. 114, 380–386 (1996).
28. Waubant, A., Franco-Vidal, V. & Ribadeau Dumas, A. Validation of a French version of the Sunnybrook facial grading system. Eur. Ann. Otorhinolaryngol. Head Neck Dis. 139, 119–124 (2022).
29. Mengi, E. et al. Comparison of the Reliability of the House- Brackmann, Facial Nerve Grading System 2.0, and Sunnybrook Facial Grading System for the Evaluation of Patients with Peripheral Facial Paralysis. J. Int. Adv. Otol. 20, 14–18 (2024).
30. Coulson, S. E., Croxson, G. R., Adams, R. D. & O'Dwyer, N. J. Reliability of the 'Sydney,' 'Sunnybrook,' and 'House Brackmann' facial grading systems to assess voluntary movement and synkinesis after facial nerve paralysis. Otolaryngol. - Head Neck Surg. 132, 543–549 (2005).
31. Neely, J. G., Cherian, N. G., Dickerson, C. B. & Nedzelski, J. M. Sunnybrook facial grading system: reliability and criteria for grading. Laryngoscope 120, 1038–1045 (2010).

32. Kayhan, F. T., Zurakowski, D. & Rauch, S. D. Toronto facial grading system: Interobserver reliability. Otolaryngol. - Head Neck Surg. 122, 212–215 (2000).
33. van Veen, M. M., Bruins, T. E., Artan, M., Werker, P. M. N. & Dijkstra, P. U. Learning curve using the Sunnybrook Facial Grading System in assessing facial palsy: An observational study in 100 patients. Clin. Otolaryngol. 45, 823–826 (2020).
34. Heike, C. L., Upson, K., Stuhaug, E. & Weinberg, S. M. 3D digital stereophotogrammetry: a practical guide to facial image acquisition. Head Face Med. 6, 18 (2010).
35. Maal, T. J. J. et al. Registration of 3-Dimensional Facial Photographs for Clinical Use. J. Oral Maxillofac. Surg. 68, 2391–2401 (2010).
36. Maal, T. J. J. et al. Variation of the face in rest using 3D stereophotogrammetry. Int. J. Oral Maxillofac. Surg. 40, 1252–1257 (2011).
37. Knoops, P. G. M. et al. Comparison of three-dimensional scanner systems for craniomaxillofacial imaging. J. Plast. Reconstr. Aesthetic Surg. 70, 441–449 (2017).
38. Mai, H.-N., Kim, J., Choi, Y.-H. & Lee, D.-H. Accuracy of Portable Face-Scanning Devices for Obtaining Three-Dimensional Face Models: A Systematic Review and Meta-Analysis. Int. J. Environ. Res. Public Health 18, (2021).
39. Nguyen, C., Nicolai, E. S. J., He, J. J., Roshchupkin, G. V & Corten, E. M. L. 3D surface imaging technology for objective automated assessment of facial interventions: A systematic review. J. Plast. Reconstr. Aesthet. Surg. 75, 4264–4272 (2022).
40. Dindaroğlu, F., Kutlu, P., Duran, G. S., Görgülü, S. & Aslan, E. Accuracy and reliability of 3D stereophotogrammetry: A comparison to direct anthropometry and 2D photogrammetry. Angle Orthod. 86, 487–494 (2016).
41. Schipper, J. A. M. et al. Reliability and validity of handheld structured light scanners and a static stereophotogrammetry system in facial three-dimensional surface imaging. Sci. Rep. 14, 8172 (2024).
42. Quinzi, V. et al. Facial Scanning Accuracy with Stereophotogrammetry and Smartphone Technology in Children: A Systematic Review. Children 9, (2022).
43. Pellitteri, F., Scisciola, F., Cremonini, F., Baciliero, M. & Lombardo, L. Accuracy of 3D facial scans: a comparison of three different scanning system in an in vivo study. Prog. Orthod. 24, 44 (2023).
44. Bohner, L. et al. Accuracy of digital technologies for the scanning of facial, skeletal, and intraoral tissues: A systematic review. J. Prosthet. Dent. 121, 246–251 (2019).
45. Wesselius, T. S., Verhulst, A. C., Xi, T., Ulrich, D. J. O. & Maal, T. J. J. Effect of skin tone on the accuracy of hybrid and passive stereophotogrammetry. J. Plast. Reconstr. Aesthetic Surg. 72, 1564–1569 (2019).
46. Geng, J. Structured-light 3D surface imaging: a tutorial. Adv. Opt. Photonics 3, 128–160 (2011).
47. Halmetschlager-Funek, G., Suchi, M., Kampel, M. & Vincze, M. An empirical evaluation of ten depth cameras: Bias, precision, lateral noise, different lighting conditions and

materials, and multiple sensor setups in indoor environments. IEEE Robot. Autom. Mag. 26, 67–77 (2019).
48. Carfagni, M. et al. Metrological and Critical Characterization of the Intel D415 Stereo Depth Camera. Sensors 19, (2019).
49. Siena, F. L., Byrom, B., Watts, P. & Breedon, P. Utilising the Intel RealSense Camera for Measuring Health Outcomes in Clinical Research. J. Med. Syst. 42, 53 (2018).
50. Servi, M. et al. Metrological Characterization and Comparison of D415, D455, L515 RealSense Devices in the Close Range. Sensors 21, (2021).
51. da Silva Neto, J. G., da Lima Silva, P. J., Figueredo, F., Teixeira, J. M. X. N. & Teichrieb, V. Comparison of RGB-D sensors for 3D reconstruction. in 2020 22nd Symposium on Virtual and Augmented Reality (SVR) 252–261 (2020). doi:10.1109/SVR51698.2020.00046.
52. Mocini, E. et al. Digital Anthropometry: A Systematic Review on Precision, Reliability and Accuracy of Most Popular Existing Technologies. Nutrients 15, (2023).
53. Caple, J. & Stephan, C. N. A standardized nomenclature for craniofacial and facial anthropometry. Int. J. Legal Med. 130, 863–879 (2016).
54. Liu, J. et al. Accuracy of 3-dimensional stereophotogrammetry: Comparison of the 3dMD and Bellus3D facial scanning systems with one another and with direct anthropometry. Am. J. Orthod. Dentofac. Orthop. 160, 862–871 (2021).
55. Kaur, P., Krishan, K., Sharma, S. K. & Kanchan, T. Facial-recognition algorithms: A literature review. Med. Sci. Law 60, 131–139 (2020).
56. Dang, K. & Sharma, S. Review and comparison of face detection algorithms. in 2017 7th International Conference on Cloud Computing, Data Science & Engineering - Confluence 629–633 (2017). doi:10.1109/CONFLUENCE.2017.7943228.
57. Kumar, M. & Hussaini, T. Face Recognition Algorithm based on Traditional and Artificial Intelligence: A Systematic Review. in 2021 International Conference on Intelligent Technologies (CONIT) 1–5 (2021). doi:10.1109/CONIT51480.2021.9498476.
58. Wang, S. et al. Machine/Deep Learning for Software Engineering: A Systematic Literature Review. IEEE Trans. Softw. Eng. 49, 1188–1231 (2023).
59. Verdonck, T., Baesens, B., Óskarsdóttir, M. & vanden Broucke, S. Special issue on feature engineering editorial. Mach. Learn. (2021) doi:10.1007/s10994-021-06042-2.
60. Li, F. & Du, Y. Introduction: A Brief History of Deep Learning and Its Applications in Power Systems. in Deep Learning for Power System Applications: Case Studies Linking Artificial Intelligence and Power Systems 1–13 (Springer International Publishing, Cham, 2024). doi:10.1007/978-3-031-45357-1_1.
61. Krizhevsky, A., Sutskever, I. & Hinton, G. E. ImageNet Classification with Deep Convolutional Neural Networks. in Advances in Neural Information Processing Systems (eds. Pereira, F., Burges, C. J., Bottou, L. & Weinberger, K. Q.) vol. 25 (Curran Associates, Inc., 2012).

62. Taghizadeh, M. & Chalechale, A. A comprehensive and systematic review on classical and deep learning based region proposal algorithms. Expert Syst. Appl. 189, 116105 (2022).
63. Wang, Y. et al. A systematic review on affective computing: emotion models, databases, and recent advances. Inf. Fusion 83–84, 19–52 (2022).
64. Alzubaidi, L. et al. Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. J. Big Data 8, 53 (2021).
65. Suganyadevi, S., Seethalakshmi, V. & Balasamy, K. A review on deep learning in medical image analysis. Int. J. Multimed. Inf. Retr. 11, 19–38 (2022).
66. Sharifani, K. & Amini, M. Machine Learning and Deep Learning: A Review of Methods and Applications. World Inf. Technol. Eng. J. 10, 3897–3904 (2023).
67. Singh, S. P. et al. 3D Deep Learning on Medical Images: A Review. Sensors 20, (2020).

CHAPTER 2

DEPTH ACCURACY OF THE REALSENSE F200: LOW-COST 4D FACIAL IMAGING

Published as: Timen C. ten Harkel, Caroline M. Speksnijder, Ferdinand van der Heijden, Carien H.G. Beurskens, Koen J.A.O. Ingels, & Thomas J.J. Maal. Depth accuracy of the RealSense F200: Low-cost 4D facial imaging. Scientific Reports 7, 16263 (2017).

DOI:10.1038/s41598-017-16608-7

ABSTRACT

The RealSense F200 represents a new generation of economically viable 4-dimensional imaging (4D) systems for home use. However, its 3D geometric (depth) accuracy has not been clinically tested. Therefore, this study determined the depth accuracy of the RealSense, in a cohort of patients with a unilateral peripheral facial palsy (PFP, n = 34), by using the clinically validated 3dMD system as a gold standard. The patients were simultaneously recorded with both systems, capturing six poses of the Sunnybrook Facial Grading System (SFGS). This study has shown that the RealSense depth accuracy was not affected by the PFP (1.48 ± 0.28 mm), compared to the healthy side of the face (1.46 ± 0.26 mm). Furthermore, the SFGS poses did not influence the RealSense depth accuracy ($p = 0.76$). However, the distance of the patients to the RealSense was shown to affect the accuracy of the system, where the highest depth accuracy of 1.07 mm was measured at a distance of 35 cm. Overall, this study has shown that the RealSense can provide reliable and accurate depth data when recording a range of facial movements. Therefore, when the portability, low-costs, and availability of the RealSense are taken into consideration, the camera is a viable option for 4D close range imaging in telehealth.

INTRODUCTION

Three-dimensional (3D) and 4-dimensional (4D) imaging is extensively used in routine clinical practice, ranging from surgical planning and evaluation to patient monitoring, and rehabilitation [1–4]. A significant advantage of 4D imaging over 3D imaging, is that it can create multiple 3D images over time, which is especially suited for dynamic measurements, such as the movement of limbs or facial expressions [5]. Despite this, traditional 4D imaging systems tend to be bulky, expensive, or overly complicated for self-patient use. Thus, their use has been limited to dedicated healthcare centres [1,6]. Technical developments have made it possible to create inexpensive, portable 4D cameras such as the RealSense F200 (which will be referred to as the RealSense). This may allow the shift of current 3D and 4D imaging tasks into telehealth applications [7]. However, before such an imaging device can be implemented in a clinical setting, it is crucial to evaluate the accuracy of the system.

The RealSense is a portable 4D imaging device composed of five core elements: the image processor, colour sensor, infrared (IR) sensor, IR laser projector, and a stereo microphone. This device was developed for close range imaging, with a recommended user range of 20 to 120 cm, which allows the user to capture detailed areas such as the face or hand [8]. Typically, the RealSense will simultaneously capture colour and depth images, with a frame rate around 30 frames per second (FPS). One single frame consists of a 2-dimensional (2D) colour image, captured by the light sensor (Figure 1a) and a depth image, containing geometrical 3D information (Figure 1b – d). The depth image is generated with the IR laser projector and the IR sensor. First, the IR laser projector emits a structured light pattern. Subsequently, the IR sensor captures the reflected light pattern from the object or person. The reflected pattern will be used to reconstruct the 3D surface, by a technique called triangulation [9]. The generated depth data consists of individual points with X, Y, Z coordinates resulting in a point cloud (Figure 1d).

A possible telehealth application for the RealSense is the development of an automated grading system for patients with a unilateral peripheral facial palsy (PFP), for monitoring rehabilitation progress at home. Currently, there exist over 19 subjective and objective grading systems to grade the severity of a PFP [10–13]. One of the recommended subjective grading systems is the Sunnybrook Facial Grading System (SFGS). This grading system is a well-established sensitive method for evaluating facial movement outcomes, both at rest and through five key voluntary movements (forehead wrinkle, gentle eye closure, open mouth smile, snarl, and lip pucker) [14]. Therefore, the SFGS is one of the most robust manual measuring methods currently in clinical use [10,11]. Thus, to

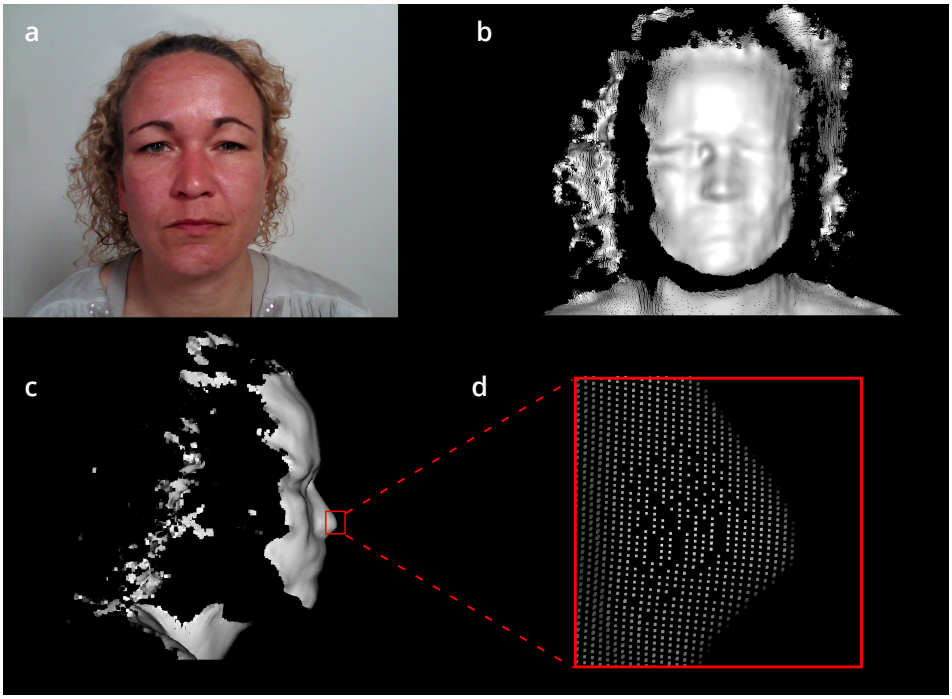

Figure 1. A single frame from a RealSense F200 recording is shown, which simultaneously captures both the colour image (a) and the depth image (b), by the colour sensor and the IR sensor, respectively. During this study, the RealSense captured 27 of these frames per second. Although the recording was performed from a frontal position, it is possible to show the depth data from multiple angles, such as a lateral perspective (c), visualizing the additional available information. The individual points of the point cloud become visible when zooming in on the image (d). The colour frame was cropped and shading was added to the depth data for visualization purposes.

incorporate the positive aspects of the SFGS, and to make its clinical implementation easier, it would be valuable to create an automated scoring system based on the SFGS. Since facial expressions consist of a significant amount of anterior-posterior movement [15], a 4D system such as the RealSense could capture the information in this direction. However, currently there is no data available on the depth accuracy of the RealSense point cloud.

Therefore, the goal of this study was to determine the depth accuracy of the RealSense in a cohort of patients with a PFP capturing the face at rest with five additional voluntary movements based on the SFGS. In addition, as this study was conducted in patients with a PFP, the healthy side of the face of the patient was used to determine the depth accuracy of the RealSense in a healthy situation.

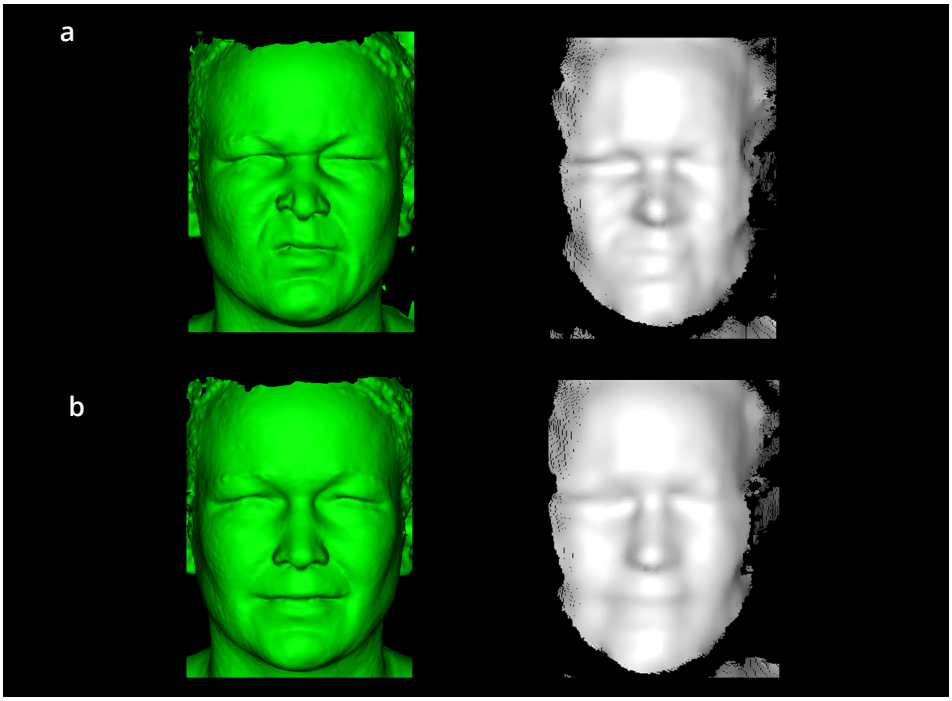

Figure 2. Comparison of the depth data between the simultaneously captured 3D reference image (3dMD system; green) and the RealSense F200 depth image (white). A total of 6 poses based on the Sunnybrook Facial Grading System were captured for each patient with a unilateral peripheral facial palsy (n = 34), where the snarl (a) and smile (b) are shown as an example for a single patient. The 3D reference image acted as the gold standard, to determine the depth accuracy obtained by the RealSense F200.

MATERIALS & METHODS

Population

In this study, patients presenting to the Radboud University Medical Centre (Radboudumc, Nijmegen, the Netherlands) with a unilateral PFP were included, irrespective of etiology, severity, and the time since onset of the PFP. The exclusion criteria were the presence of a bilateral PFP and an age younger than 18 years. Approval of this study was authorized by the Ethics Committee of the Radboudumc (2015-1829). This study was conducted in compliance with the World Medical Association Declaration of Helsinki on medical research ethics. All subjects provided written informed consent before data acquisition. Additionally, a written informed consent was obtained from the patient shown in this paper, to publish the images in an online open-access publication.

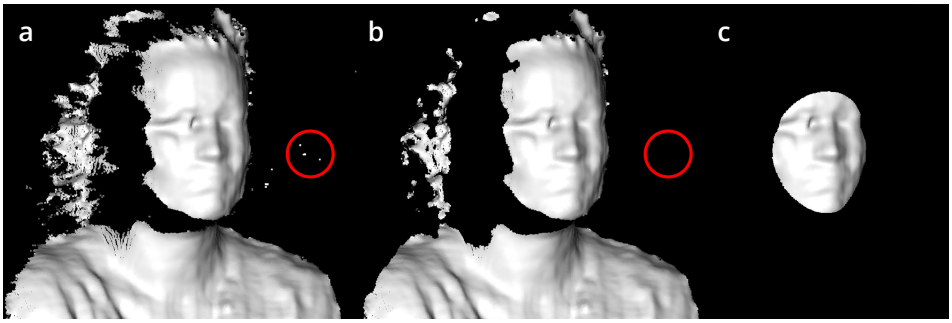

Figure 3. Overview of the pre-processing of RealSense F200 depth images. The original image (a) showing the raw depth data. As exemplified in the red circle, this image still contains spurious background noise. Using a statistical outlier filter, the background noise was removed from the original RealSense depth data (b). After correcting for statistical outliers, a region of interest (ROI) was selected based on a sphere centred at the pronasale (c). The radius of the sphere was determined by the distance between the pronasale and the exocanthion, with an additional margin of 10%, to include the complete eye region.

Data acquisition

Continuous RealSense recordings were acquired with the RealSense F200 (depth camera manager version 1.4.27 and RealSense Software Development Kit, RSSDK, version 7.0.23.8048, Intel, Santa Clara, USA) using a colour resolution of 1920 × 1080 pixels and depth resolution of 640 × 480 pixels. RealSense recordings were captured with an average frame rate of 27 FPS. Simultaneously, a two-pod 3dMD system (3dMDface, 3dMD, Atlanta, USA) was used to capture single static 3D images, acting as the reference clinical standard (Figure 2). Patients were first positioned in front of the 3dMD camera. Subsequently, the RealSense camera was positioned in front of the patient at eye level on a tripod. The minimum distance from camera to patient was determined by the RealSense facial tracking algorithm from the RSSDK. The distance between patient and camera was increased if this was required due to physical limitations, such as body size. All recordings were acquired in a windowless room used for clinical 3D imaging at the Department of Oral and Maxillofacial Surgery. A diffuse lighting environment was created with two Diva Light 400 lights (Kinoflo Lighting Systems, Los Angeles, USA), which was the only light source in the room. Finally, a single RealSense recording was made for each patient, capturing six different poses based on the SFGS, which includes the face at rest and five facial expressions based on voluntary movements (forehead wrinkle, gentle eye closure, open mouth smile, snarl, and lip pucker) [14]. The patient was asked to hold each pose at maximum exertion of the voluntary movement, until the static 3D image was taken. A total of six static 3D images were captured with the 3dMD system during a single RealSense recording. The static 3D images made by the 3dMD system will be referred to as the 3D reference images.

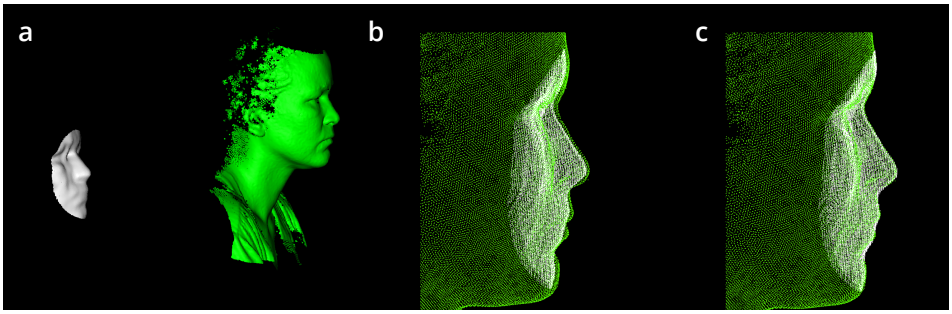

Figure 4. A registration pipeline was used to align the cropped RealSense F200 image (white) with the 3D reference image (green). A rough alignment was performed with a Procrustes analysis (b). Subsequently a refined alignment was performed with the Iterative Closest Point Algorithm (c). The 3D reference image was cropped in subfigure b & c for clarity.

Data processing

To determine the accuracy of the RealSense, the RealSense depth data was compared to the 3D reference images for each SFGS pose. Since the RealSense recording consisted of a continuous data stream, six frames from the RealSense recording were selected at the capture time of the 3D reference image. The frame selection was based on the flash from the 3dMD system that was visible on the RealSense recording. To prevent RealSense depth data distortion due to the 3dMD flash, the RealSense frame immediately prior to the 3dMD flash was used in the final analysis. From the six selected RealSense frames, which captured the facial movements at maximum exertion, the depth data was exported with the RSSDK as individual point clouds in X, Y, Z coordinates (Figure 2). After exporting the point clouds, the pre-processing was performed using the Point Cloud Library (PCL, version 1.8.0) [16]. Due to a limited field of view of the RealSense depth image compared to the 3D reference image (Figure 2), a region of interest (ROI) was selected from the RealSense image. To remove possible noise within the ROI, a statistical outlier filter was applied (Figure 3a & b) [16]. Next, the ROI was selected with a sphere centred at the pronasale (Figure 3c). The radius of the sphere was determined by the maximum Euclidean distance between the pronasale and the left or right exocanthion, based on manual landmarks placed on the 3D reference image. The sphere radius was increased by 10% to include the eye region completely. No pre-processing was applied to the 3D reference image. After the pre-processing stage, initial registration was performed between the RealSense point cloud and 3D reference image by the Procrustes algorithm implemented by libigl (Figure 4a & b) [17]. The Procrustes algorithm was performed with manually placed landmarks, at the exocanthion and pronasale, at the RealSense and 3D reference image. During this registration, no scaling or reflection was applied. The initial registration was followed by a refined registration with the Iterative Closest Point (ICP) algorithm implemented by PCL (Figure 4c) [18], set to a rigid registration without scaling, as not to deform the RealSense point cloud.

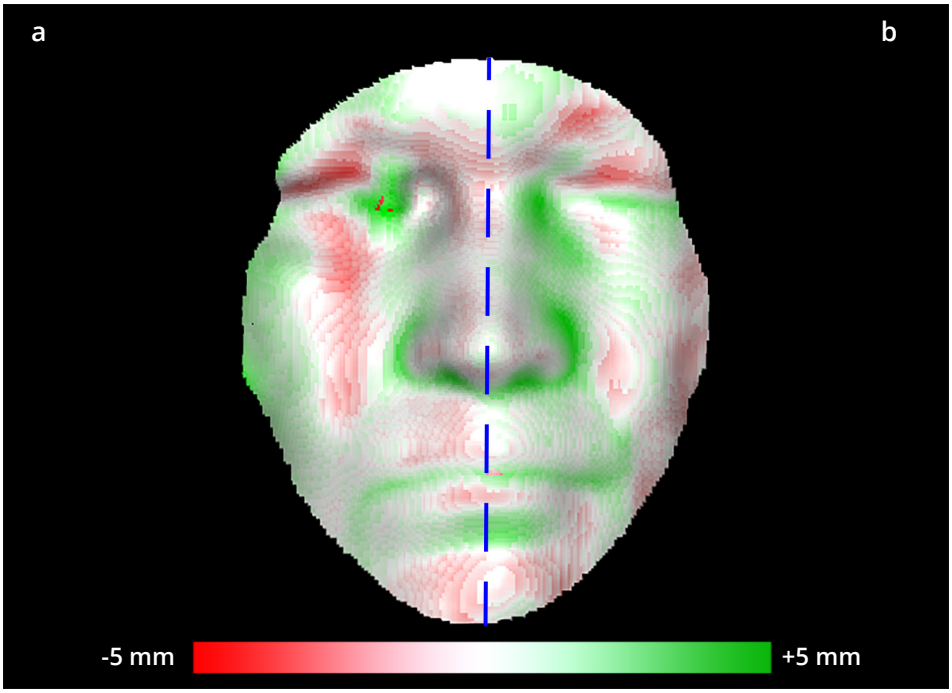

Figure 5. A distance map generated from a patient at rest. The distance map was created by calculating the closest distance between the RealSense F200 image and the 3D reference image, for each RealSense point. The white areas represent a perfect match with the 3D reference image (0 mm), with areas in red and green showing distances between ± 5 mm. Using the midsagittal plane (blue dashed line), the distance map was calculated separately for the healthy (a) and palsy side (b) of the face.

Data analysis

To determine the accuracy of the RealSense point cloud, a distance map was calculated between the RealSense and the 3D reference image with PCL (Figure 5). The distance map was created by calculating the Euclidean distance between each point of the RealSense point cloud to the closest point on the 3D reference image. The final depth accuracy was defined as the root mean square (RMS) of the distance map, where the 3D reference image was considered as the gold standard. This analysis was performed separately for the healthy and palsy side of the patient, determined by the midsagittal plane (Figure 5), and each of the six SFGS poses (rest, forehead wrinkle, gentle eye closure, open mouth smile, snarl, and lip pucker) [14]. A paired Student's t-test was performed for each pose comparing the depth accuracy of the healthy and palsy side. Additionally, a one-way analysis of covariance (ANCOVA) was performed to determine if there were significant differences in depth accuracy between the SFGS poses. In this analysis, the SFGS poses were categorized as six different groups, with the depth accuracy acting as

Table 1. Depth accuracy of the RealSense F200 depth data in patients with a unilateral peripheral facial palsy grouped by the 6 poses of the Sunnybrook Facial Grading System (SFGS) with the healthy and palsy side combined (n = 34 for each pose). Depth accuracy is expressed as the root mean square (RMS).

SFGS pose	RMS ± SD (mm)	p95 (mm)
Rest	1.48 ± 0.22	1.95
Forehead wrinkle	1.49 ± 0.20	1.90
Gentle eye closure	1.46 ± 0.24	1.89
Open mouth smile	1.53 ± 0.22	2.04
Snarl	1.49 ± 0.22	1.93
Lip pucker	1.48 ± 0.22	1.83

the dependent variable, and the RealSense camera distance as the covariate. The data from the paired Student's t-test and the ANCOVA analysis was tested for normality using the Kolmogorov-Smirnov test with Lilliefors significance correction [19]. Additionally, the homogeneity of variances was tested with Levene's test for the ANCOVA analysis [20]. A p-value of <0.05 was considered as statistically significant. Statistical analysis was performed using IBM SPSS Statistics, Version 22 (IBM Corp., Armonk, NY, USA).

RESULTS

A total of 34 patients were included in this study (age: 53 ± 13 years, gender: 71% female, left sided PFP: 53%). Each patient was captured with the face at rest and with five additional voluntary movements based on the SFGS (forehead wrinkle, gentle eye closure, open mouth smile, snarl, and lip pucker), where the patient was simultaneously recorded with the RealSense and the 3dMD system. This resulted in the comparison of 204 RealSense point clouds with their associated 3D reference image (Figure 3).

Firstly, the depth accuracy for the healthy side of the face was calculated for all SFGS poses combined, which lead to an average RMS of 1.48 mm (standard deviation (SD) = 0.28 mm; 95th percentile (p95) = 2.08 mm). The palsy side of the face resulted in an RMS of 1.46 mm (SD = 0.26, and p95 = 1.93). The depth accuracies of the healthy and palsy side of the separate poses are shown in Figure 6. When the Kolmogorov-Smirnov test was applied to any of the SFGS poses, the results were not significant. Therefore, the data was assumed to be normally distributed. The paired Student's t-test showed no statistically significant differences between the accuracy of the healthy and palsy side for any of the six poses (rest, p = 0.25; forehead wrinkle, p = 0.96; gentle eye closure, p = 0.63; smile, p = 0.22; snarl, p = 0.41; lip pucker, p = 0.63).

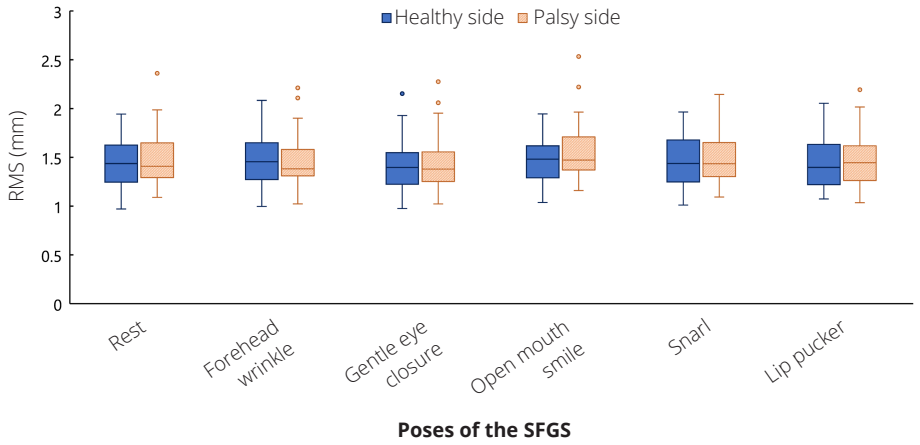

Figure 6. Average accuracy RealSense F200 depth data (n = 34) comparing the healthy and palsy side of the patient for each pose of the Sunnybrook Facial Grading System (SFGS). Depth accuracy is expressed as the root mean square (RMS). Any values greater than 1.5 times the interquartile range were considered as outliers for the boxplot.

As we have shown that there is no significant difference in accuracy between the two sides of the face, the data from the healthy and palsy side of the face were combined in order to determine the depth accuracy in between the SFGS poses (see Table 1). No significant results were found in either the Kolmogorov-Smirnov test or Levine's test in the data of the ANCOVA analysis. The ANCOVA analysis showed no significant differences in depth accuracy between the poses ($F = 0.53$, $p = 0.76$). When combining the data from the six poses with the ANCOVA analysis, an average linear regression of $y = 0.003 x + 0.1715$ ($r = 0.78$, $p = 0.00$) was found as shown in Figure 7, where x is the distance to the camera in mm and y the depth accuracy in mm.

DISCUSSION

In this study, the depth accuracy of the RealSense was determined in a cohort of 34 patients with a PFP. The patients were recorded with the face at rest with five additional voluntary movements (forehead wrinkle, gentle eye closure, open mouth smile, snarl, and lip pucker), which were based on the SFGS [14]. No significant differences were found in depth accuracy when comparing the healthy and palsy side of the patients, with an average RMS of 1.48 ± 0.28 mm and 1.46 ± 0.26 mm, respectively. Additionally, no significant differences were found in depth accuracy between the SFGS poses ($p = 0.76$).

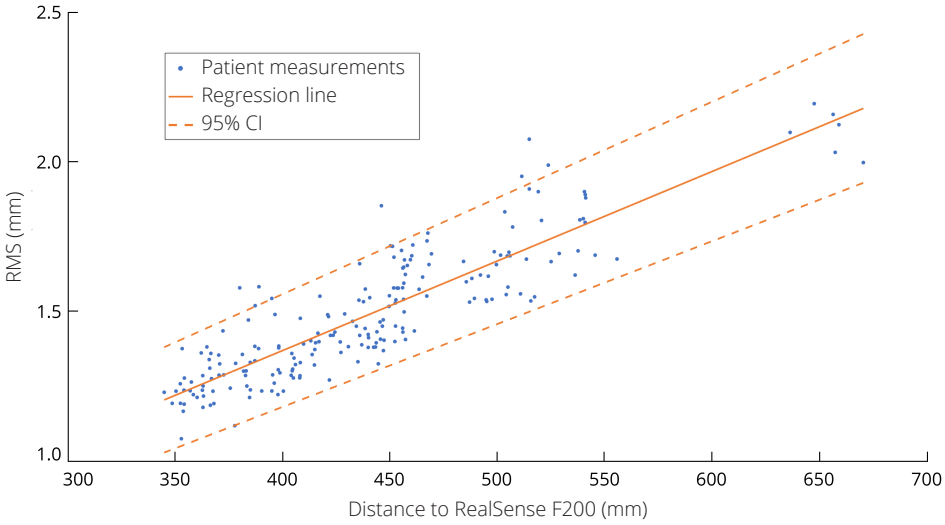

Figure 7. Correlation between the distance from patient to the RealSense F200 camera and the depth accuracy of the RealSense F200. The patient measurements include all the 6 poses from the Sunnybrook Facial Grading System for the 34 patients (n = 204). Depth accuracy is expressed as the root mean square (RMS) in mm. The regression line from the ANCOVA analysis is shown (y = 0.003x + 0.1715 with r = 0.78 and p = 0.00), including the 95% confidence interval (CI).

To the best of our knowledge, this is the first study investigating the depth accuracy of the RealSense. Therefore, no direct comparison can be made with other studies investigating the depth accuracy of the RealSense. Although other "off-the-shelf" 4D imaging systems such as the Kinect (Version 1 and 2, Microsoft, USA), have been used to reconstruct a face using multiple frames (RMS accuracy of 0.84 to 2.00 mm [21–25]). No comparable studies have used a single frame of the face and compared this to a clinical reference standard, such as the 3dMD system. Therefore, our study has shown that the RealSense depth accuracy lies within the range of the Kinect accuracy when imaging the face, but it is expected to be higher due to use of only a single frame for the analysis of the RealSense.

The depth accuracy of the RealSense camera was determined by comparing the RealSense data to the 3D reference image from the two-pod 3dMD system. The reference image has a known accuracy of 0.20 to 0.25 mm when imaging the face at rest, which is considered as a sufficient accuracy for a range of clinical implementations [26–29]. Therefore, the RealSense would have a similar accuracy as the 3dMD system if the depth accuracy of the RealSense had been in the range of 0.25 mm. However, the RealSense is an order of magnitude more inaccurate, with an average accuracy of 1.48 mm for the healthy face at

rest. This resulted in a smoother RealSense image compared to the 3D reference image (Figure 2). Therefore, a decrease in accuracy at regions with a higher curvature, such as the mouth and nose, is expected and can be seen in Figure 5. The inaccuracies around the nose region can partially be explained by the blocked view around the alar groove due to the frontal positioning of the RealSense, whereas the 3D reference image was captured by two pods from the side.

The accuracy of the RealSense was shown to decrease when imaging the eye region. However, it is known that the accuracy of the 3D reference image decreases when capturing specular surfaces, such as the eye and teeth [30]. This inaccuracy in the gold standard was not corrected for during this study, since the average accuracy of the 3D reference image is 0.38 ± 0.34 mm around the eye [31], which is still an order of magnitude more accurate than the RealSense. The impact on the depth accuracy can for example be seen between the neutral pose, and the gentle eye closure, where the specular area of the eye is covered, which resulted in a non-significant difference in the average accuracy in this study (Table 1). Overall, the apparent difference in depth accuracy between the RealSense image and the 3D reference image was an expected result, considering the difference in cost, size, and complexity of the two systems.

Further analysis compared the accuracy between the healthy and palsy side for the RealSense. A possible difference in accuracy, between the healthy and palsy side, could have been found due to the asymmetrical nature of the face in patients with a PFP. For example, patients with a PFP can experience a dropped corner of the mouth, or a pronounced labial fold, in the palsy side of the face at rest [14]. This can lead to an increased complexity of the facial surface. However, no significant differences in depth accuracy were found in this study for any of the SFGS poses when comparing the healthy side to the palsy side. This indicates that the RealSense is able to capture the depth information of the asymmetrical features of the patients for all the SFGS poses. Although this study found that the RealSense has an average accuracy ranging between 1.46 mm and 1.53 mm, the average facial movement is expected to be 6.49 mm in the vertical direction and 5.49 mm in anterior-posterior direction in a healthy situation [32]. Therefore, the surface differences between the healthy and palsy side of the face seem to be large enough to be detected by the RealSense.

An important consideration when analysing the depth accuracy of the RealSense is the influence of patient to camera distance. Cameras that acquire depth data with structured light patterns, such as the RealSense, are expected to increase their depth accuracy at closer distances [33]. During this study, the minimal camera distance to the patient was determined by a facial tracking algorithm built in with the software development kit of

the camera, with an operating range of 30 to 100 cm [8]. However, since the patients needed to be captured simultaneously with the 3dMD system, the available imaging space was limited. Due to physical limitations, such as body size, the camera distance needed adjustments for each patient to make the recording possible. This resulted in the majority of patients being recorded in a range of 35 to 55 cm, with one patient being measured at a distance of 65 cm. However, when the healthy side of the face was compared to the palsy side, the distance to the camera was approximately the same since the patients were positioned perpendicular to the RealSense. Therefore, the intra-patient accuracy was minimally influenced. In contrast, the inter-patient accuracy is heavily influenced by the distance to the camera, as can be seen in Figure 7 ($r = 0.78$, $p = 0.00$). Therefore, the average RMS, SD, and p95 reported in this study highly depend on the distance to the camera and the selected distance range. This will represent a realistic scenario for certain real-world clinical implementations where the distance to the camera will vary between measurements. For example, in this study patients moved in between the captured SFGS poses, as can be seen in Figure 7 at the patient measured at 65 cm. Therefore, an ANCOVA analysis was applied to correct for the distance to the camera, when comparing depth accuracy between the six poses. Since no significant differences were found ($p = 0.76$), the RealSense was tested in a wide range of facial motion, without showing significant differences in depth accuracy.

The current study design has several limitations that should be taken into account. First of all, the cohort consisted of patients with a PFP with an age older than 18 years, making the depth accuracy unknown for children, healthy adults, and other diseases. Additionally, the accuracy of structured light cameras is known to be influenced by different light sources [33], which was not investigated in this study. When applying the RealSense in telehealth applications more various lighting conditions can be expected. Therefore, future research should determine the influence of the lighting in the room on the accuracy of the RealSense. The current measurement setup used a single RealSense camera, compared to the two pods of the 3dMD system. This resulted in a more limited field of view for the RealSense (Figure 2), possibly losing valuable information of the face. To overcome this limitation, it is possible to use multiple synchronized RealSense cameras, positioned at different angles. However, this will increase the complexity of the measurement setup that needs to be used at home. Therefore, this study used a single RealSense camera, and an ROI was selected to make the comparison between the RealSense and the 3D reference image possible. The ROI included key areas of the face, such as the eyes and mouth. To make the ROI consistent for all patients, the area around the pronasale was selected within a patient specific radius (Figure 3c). This radius was determined by the distance between the pronasale and the exocanthion, to include the eye region. To prevent cropping of the eye region, 10% was added to the

determined radius. The ROI used in this study was relatively conservative, to make sure a similar ROI could be selected in between patients. The average depth accuracy potentially could have improved since the excluded areas immediate to the current ROI were areas with low curvatures. However, the point cloud was cut off at the lateral sides of the face (Figure 1c) since these areas were positioned more perpendicular to the camera. The exact position of this cut-off changed in between patients. Therefore, a conservative ROI was chosen to ensure a more consistent ROI selection across patients. This ROI can be increased by using multiple synchronized RealSense cameras positioned at different angles in the measurement setup.

Additionally, during the processing of the data it was necessary to apply a registration between the RealSense and the 3D reference image, since the two images were captured with two separate imaging devices, resulting in a different location in space (Figure 4a). With the implementation of the Procrustes and ICP registration, it was possible to match the point clouds semi-automatically. However, the final ICP registration could find a sub-optimal matching in a local optimum, resulting in a lower depth accuracy [34]. In addition, another important limitation to this study, is that the clinical reference standard was only able to capture static 3D images. Therefore, only a single frame of the RealSense recording could be used in the final analysis for each SFGS pose, while there are 27 RealSense frames available each second. Future studies would benefit from the use of a professional 4D system as the reference clinical standard. However, in this study, six frames were extracted from each RealSense recording, capturing the accuracy of the system over multiple time points. All six SFGS poses reported an accuracy within a range of 1.46 to 1.53 mm, showing the consistency of the camera for various facial movements over time, in a single recording.

In conclusion, this study has shown that the RealSense can provide reliable and accurate depth data when capturing the face at rest and when performing five voluntary movements based on the SFGS, in a cohort of 34 patients with a PFP. Therefore, a similar accuracy of the RealSense point cloud can be expected when analysing the different SFGS poses, when an automated SFGS is implemented. Additionally, it has been shown submillimetre information is lost in the RealSense point cloud, especially noticeable in areas with higher curvature, which will need to be taken into account in an automated grading system. However, larger deviations will be possible to capture, especially at a closer distance to the camera, where the highest depth accuracy of 1.07 mm was achieved at a distance of 35 cm. Due to the correlation between camera distance and depth accuracy for systems such as the RealSense [33], it will be essential to keep track of the patient to camera distance in clinical applications. One aspect that needs to be included in future research is the influence of the lighting in the room on the accuracy of the RealSense. Although

this study investigated the imaging of patients with a PFP with the RealSense, there are numerous applications for a portable 3D and 4D imaging system such as the RealSense. With the emerging interest in the use of telehealth in tasks such as health monitoring, diagnostics, and performing consults, there is still room to increase the use of 3D and 4D imaging in telehealth [35–42]. Overall, when considering the portability, low-costs, and availability of the RealSense, the camera is a viable option for 3D and 4D imaging in telehealth, where the RealSense is especially suited for close range imaging. However, when submillimetre accuracy is required for the clinical application, more professional setups are still recommended to be used.

REFERENCES

1. Knoops, P. G. M. et al. Comparison of three-dimensional scanner systems for craniomaxillofacial imaging. J. Plast. Reconstr. Aesthetic Surg. 70, 441–449 (2017).
2. Hallac, R. R., Feng, J., Kane, A. A. & Seaward, J. R. Dynamic facial asymmetry in patients with repaired cleft lip using 4D imaging (video stereophotogrammetry). J. Cranio-Maxillofacial Surg. 45, 8–12 (2017).
3. Shujaat, S. et al. The clinical application of three-dimensional motion capture (4D): A novel approach to quantify the dynamics of facial animations. Int. J. Oral Maxillofac. Surg. 43, 907–916 (2014).
4. Popat, H., Richmond, S., Benedikt, L., Marshall, D. & Rosin, P. L. Quantitative analysis of facial movement - A review of three-dimensional imaging techniques. Comput. Med. Imaging Graph. 33, 377–383 (2009).
5. Al-Anezi, T. et al. A new method for automatic tracking of facial landmarks in 3D motion captured images (4D). Int. J. Oral Maxillofac. Surg. 42, 9–18 (2013).
6. Tzou, C. H. J. et al. Comparison of three-dimensional surface-imaging systems. J. Plast. Reconstr. Aesthetic Surg. 67, 489–497 (2014).
7. Bauer, S. et al. Real-Time Range Imaging in Health Care: A Survey. in Time-of-Flight and Depth Imaging. Sensors, Algorithms, and Applications: Dagstuhl 2012 Seminar on Time-of-Flight Imaging and GCPR 2013 Workshop on Imaging New Modalities (eds. Grzegorzek, M., Theobalt, C., Koch, R. & Kolb, A.) 228–254 (Springer Berlin Heidelberg, Berlin, Heidelberg, 2013). doi:10.1007/978-3-642-44964-2_11.
8. Intel® RealSense™ Data Ranges. https://software.intel.com/en-us/articles/intel-realsense-data-ranges (2016).
9. Valkenburg, R. J. & McIvor, A. M. Accurate 3D measurement using a structured light system. Image Vis. Comput. 16, 99–110 (1998).
10. Fattah, A. Y. et al. Facial Nerve Grading Instruments: Systematic Review of the Literature and Suggestion for Uniformity. Plast. Reconstr. Surg. 135, 569–579 (2015).
11. Kanerva, M., Poussa, T. & Pitkäranta, A. Sunnybrook and House-Brackmann Facial Grading Systems: Intrarater repeatability and interrater agreement. Otolaryngol. - Head Neck Surg. 135, 865–871 (2006).
12. Niziol, R., Henry, F. P., Leckenby, J. I. & Grobbelaar, A. O. Is there an ideal outcome scoring system for facial reanimation surgery? A review of current methods and suggestions for future publications. J. Plast. Reconstr. Aesthetic Surg. 68, 447–456 (2015).
13. Samsudin, W. S. W. & Sundaraj, K. Evaluation and Grading Systems of Facial Paralysis for Facial Rehabilitation. J. Phys. Ther. Sci. 25, 515–519 (2013).
14. Ross, B. G., Fradet, G. & Nedzelski, J. M. Development of a sensitive clinical facial grading system. Otolaryngol. neck Surg. 114, 380–386 (1996).

15. Coulson, S. E., Croxson, G. R. & Gilleard, W. L. Quantification of the three-dimensional displacement of normal facial movement. Ann. Otol. Rhinol. Laryngol. 109, 478–483 (2000).
16. Rusu, R. B. & Cousins, S. 3D is here: Point Cloud Library (PCL). in 2011 IEEE International Conference on Robotics and Automation 1–4 (2011). doi:10.1109/ICRA.2011.5980567.
17. Jacobson, A. et al. libigl: A simple C++ geometry processing library. http://libigl.github.io/libigl/ (2014).
18. Holz, D., Ichim, A. E., Tombari, F., Rusu, R. B. & Behnke, S. Registration with the Point Cloud Library PCL: A modular framework for aligning in 3-d. IEEE Robot. Autom. Mag. 22, 110–124 (2015).
19. Lilliefors, H. W. On the Kolmogorov-Smirnov Test for Normality with Mean and Variance Unknown. J. Am. Stat. Assoc. 62, 399–402 (1967).
20. Levene, H. Robust tests for equality of variances. Contrib. to Probab. Stat. 1, 278–292 (1960).
21. Hamza-Lup, F. G., Farrar, S. & Leon, E. Patient specific 3D surfaces for interactive medical planning and training. in Proceedings of the 20th International Conference on 3D Web Technology - Web3D '15 107–113 (ACM Press, New York, New York, USA, 2015). doi:10.1145/2775292.2775294.
22. Hernandez, M., Jongmoo Choi & Medioni, G. Laser scan quality 3-D face modeling using a low-cost depth camera. in Signal Processing Conference (EUSIPCO), 2012 Proceedings of the 20th European 1995–1999 (2012).
23. Anasosalu, P. K., Thomas, D. & Sugimoto, A. Compact and accurate 3-D face modeling using an RGB-D camera: Let's open the door to 3-D video conference. in Proceedings of the IEEE International Conference on Computer Vision 67–74 (2013). doi:10.1109/ICCVW.2013.16.
24. Berretti, S., Pala, P. & Del Bimbo, A. Increasing 3D Resolution of Kinect Faces. in European Conference on Computer Vision (ECCV) 639–653 (2014). doi:10.1007/978-3-319-16178-5.
25. Hernandez, M., Choi, J. & Medioni, G. Near laser-scan quality 3-D face reconstruction from a low-quality depth stream. Image Vis. Comput. 36, 61–69 (2015).
26. Lübbers, H.-T., Medinger, L., Kruse, A., Grätz, K. W. & Matthews, F. Precision and accuracy of the 3dMD photogrammetric system in craniomaxillofacial application. J. Craniofac. Surg. 21, 763–767 (2010).
27. Maal, T. J. J. et al. Variation of the face in rest using 3D stereophotogrammetry. Int. J. Oral Maxillofac. Surg. 40, 1252–1257 (2011).
28. Dindaroğlu, F., Kutlu, P., Duran, G. S., Görgülü, S. & Aslan, E. Accuracy and reliability of 3D stereophotogrammetry: A comparison to direct anthropometry and 2D photogrammetry. Angle Orthod. 86, 487–494 (2016).

29. Boehnen, C. & Flynn, P. Accuracy of 3D scanning technologies in a face scanning scenario. in 3-D Digital Imaging and Modeling, 2005. 3DIM 2005. Fifth International Conference on 310–317 (IEEE, 2005). doi:10.1109/3DIM.2005.13.
30. Maal, T. J. J. et al. The accuracy of matching three-dimensional photographs with skin surfaces derived from cone-beam computed tomography. Int. J. Oral Maxillofac. Surg. 37, 641–646 (2008).
31. Maal, T. J. J. et al. Variation of the face in rest using 3D stereophotogrammetry. Int. J. Oral Maxillofac. Surg. 40, 1252–1257 (2011).
32. Coulson, S. E., Croxson, G. R. & Gilleard, W. L. Quantification of the three-dimensional displacement of normal facial movement. Ann. Otol. Rhinol. Laryngol. 109, 478–483 (2000).
33. Pöhlmann, S. T. L., Harkness, E. F., Taylor, C. J. & Astley, S. M. Evaluation of Kinect 3D Sensor for Healthcare Imaging. J. Med. Biol. Eng. 857–870 (2016) doi:10.1007/s40846-016-0184-2.
34. Besl, P. & McKay, N. A Method for Registration of 3-D Shapes. IEEE Trans. Pattern Anal. Mach. Intell. 14, 239–256 (1992).
35. Kvedar, J., Coye, M. J. & Everett, W. Connected health: A review of technologies and strategies to improve patient care with telemedicine and telehealth. Health Aff. 33, 194–199 (2014).
36. Zheng, Y., Head, B. A. & Schapmire, T. J. A Systematic Review of Telehealth in Palliative Care: Caregiver Outcomes. Telemed. J. e-health 22, 1–7 (2016).
37. Regina Molini-Avejonas, D., Rondon-Melo, S., de La Higuera Amato, C. A. & Samelli, A. G. A systematic review of the use of telehealth in speech, language and hearing sciences. J. Telemed. Telecare 21, 367–376 (2015).
38. Weinstein, R. S. et al. Telemedicine, Telehealth, and Mobile Health Applications That Work: Opportunities and Barriers. J. Med. 127, 183–187 (2014).
39. Sood, A. et al. The Role of Telemedicine in Wound Care. Plast. Reconstr. Surg. 138, 248S-256S (2016).
40. AlDossary, S., Martin-Khan, M. G., Bradford, N. K. & Smith, A. C. A systematic review of the methodologies used to evaluate telemedicine service initiatives in hospital facilities. Int. J. Med. Inform. 97, 171–194 (2017).
41. Kruse, C. S., Bouffard, S., Dougherty, M. & Parro, J. S. Telemedicine Use in Rural Native American Communities in the Era of the ACA: a Systematic Literature Review. J. Med. Syst. 40, 145 (2016).
42. Klaassen, B., van Beijnum, B. J. F. F. & Hermens, H. J. Usability in telemedicine systems—A literature survey. Int. J. Med. Inform. 93, 57–69 (2016).

CHAPTER 3

VALIDATION OF THE DEPTH ACCURACY, 3D LANDMARK PLACEMENT, AND 3D ANTHROPOMETRIC MEASUREMENTS OF THE REALSENSE D415

CHAPTER 3

PART 1: RELIABILITY AND AGREEMENT OF 3D ANTHROPOMETRIC MEASUREMENTS IN FACIAL PALSY PATIENTS USING A LOW-COST 4D IMAGING SYSTEM

Published as: Timen C. ten Harkel, Shankeeth Vinayahalingam, Koen J.A.O. Ingels, Stefaan J. Bergé, Thomas J.J. Maal & Caroline M. Speksnijder, Reliability and Agreement of 3D Anthropometric Measurements in Facial Palsy Patients Using a Low-Cost 4D Imaging System. IEEE Transactions on Neural Systems and Rehabilitation Engineering 28, 1817–1824 (2020).

DOI: 10.1109/TNSRE.2020.3007532

ABSTRACT

The reliability (precision) and agreement (accuracy) of anthropometric measurements based on manually placed 3D landmarks using the RealSense D415 were investigated in this study. Thirty patients with a unilateral peripheral facial palsy, were recorded simultaneously with the RealSense and a professional 3dMD imaging system, with their face in neutral (resting) position. First the RealSense depth accuracy was determined. Subsequently, two observers placed 14 facial landmarks on the 3dMD and RealSense image, assessing the distance between landmark placement. The respective intra- and inter-rater Euclidean distance between the landmark placements was 0.84 ± 0.58 mm and 1.00 ± 0.70 mm for the 3dMD landmarks and 1.32 ± 1.27 mm and 1.62 ± 1.42 mm for the RealSense landmarks. From these landmarks 14 anthropometric measurements were derived. The intra- and inter-rater measurements had an overall reliability of 0.95 (0.87 to 0.98) and 0.93 (0.85 to 0.97) for the 3dMD measurements, and 0.83 (0.70 to 0.91) and 0.80 (0.64 to 0.89) for the RealSense measurements, respectively, expressed as the intra-class correlation coefficient. Determined by the Bland-Altman analysis, the agreement between the RealSense measurements and 3dMD measurements was on average -0.90 mm (-4.04 to 2.24) and -0.89 mm (-4.65 to 2.86) for intra- and inter-rater agreement, respectively. Based on the reported reliability and agreement of the RealSense measurements, the RealSense D415 can be considered as a viable option to perform objective 3D anthropomorphic measurements on the face in a neutral position, where a low-cost and portable camera is required.

INTRODUCTION

A unilateral peripheral facial palsy (PFP), with idiopathic PFP as the most common cause, is caused by a lesion in the 7th cranial nerve, with an estimated incidence rate between 11 to 40 per 100,000 people per annum [1]. The recovery rate of idiopathic PFP patients depends on multiple factors such as the severity of the initial PFP, the time until the start of the recovery, age, the presence of a normal taste or the stapedius reflex [2]. Patients with a complete PFP have a recovery rate between 50 to 61% compared to a recovery rate of 94 to 99% for patients with an incomplete PFP, which accounts for 70% of idiopathic PFP patients [1,2]. Although there is no standard for its quantitative measurement, a major category of asymmetry measurements is based on anthropometric measurements [3–5].

Traditionally, anthropometric measurements are acquired through direct measurements with a calliper or through the analysis of two-dimensional (2D) photography to determine linear distances between facial landmarks. Currently, three-dimensional (3D) stereophotogrammetry images are frequently used for anthropometric measurements due to a fast capture time in less than one second and being a non-invasive and non-ionizing imaging method [6–14]. 3D images add the possibility to analyse angles, surface area and volume. A common disadvantage of 3D stereophotogrammetry systems is the price and size of these systems, which might be a barrier for the implementation in clinics [15]. Additionally, these system properties make it infeasible to perform the 3D anthropomorphic measurements in a telemedicine setting, which potentially could be used to monitor the rehabilitation of the patient at home.

Therefore, this study determined the reliability and agreement of anthropometric measurements based on manually placed 3D landmarks in patients with a PFP with the face in neutral (resting) position using a portable low-cost 4D imaging system, the RealSense D415. The outcome of this study is intended to be used as a foundation for the implementation of the RealSense in a clinical or telemedicine setting, such as the objective assessment of facial asymmetry. Implementation of such an objective assessment is out of scope for this study. Subsequently, it is important to define the terms reliability and agreement, where reliability is defined as the consistency of results when a measurement is repeated (precision) and the agreement is defined as how close a measurement is to the gold standard (accuracy) [16].

Four objectives were defined to determine reliability and agreement of the RealSense measurements, whilst using a professional stereophotogrammetry imaging system, the 3dMD system, as a reference. The first objective was to determine the depth accuracy of the RealSense 3D images, to establish possible scaling issues or large discrepancies with

the 3dMD reference image (1). Secondly, the intra- and inter-rater Euclidean distance of the manual placement for 3D landmarks were assessed for both the RealSense and 3dMD images, to determine the potential influence of a lower image quality of a patient on the placement of manual landmarks (2). Thirdly, the intra- and inter-rater reliability (precision) of anthropometric measurements based on the manually placed landmarks were determined for both the RealSense and 3dMD images (3). Finally, the intra- and inter-rater agreement (accuracy) of the RealSense measurements was assessed by using 3dMD measurements as the gold standard (4).

MATERIALS & METHODS

Population

Between August 2018 and April 2019 patients with varying degrees of a unilateral PFP were included in this study, where the healthy side of the face acted as the reference for the measurements. Exclusion criteria were the presence of bilateral PFP and an age younger than 16 years. This study was conducted in compliance with the World Medical Association Declaration of Helsinki on medical research ethics and was approved by the Ethics Committee of the Radboudumc (2015-1829). All subjects provided their written informed consent for participation in this study. In addition, patients shown in this study provided a written informed consent for the use of their images in scientific publications.

Image acquisition

For each patient, the face in neutral pose was simultaneously captured with the RealSense D415 (Intel, Santa Clara, USA) and the two pod 3dMD system (3dMDface, 3dMD, Atlanta, USA). The RealSense recorded with 30 frames per second at a colour resolution of 1920 × 1080 pixels and a depth resolution of 1280 × 720 pixels, at an approximate distance of 35 cm to the patient. Due to the continuous development of the stability of the RealSense software, the latest stable version of the RealSense Software Development Kit (SDK) and camera firmware at the date of the recording were used. During the RealSense recordings, a static 3D image was captured with the 3dMD system, which acted as the clinical reference image. Due to the static nature of the 3dMD image, the RealSense was used as a 3D system as well, instead of using the full 4D capabilities of the RealSense. Recordings were performed in a windowless room used for daily clinical 3D imaging, illuminated with overhead fluorescent lighting.

Depth accuracy of the 3D image

The RealSense depth accuracy was calculated in accordance with the method outlined in previous work [17]. Briefly, a single RealSense depth image was exported from the recording at the time the static 3dMD image was taken. In this study, a

temporal filter and spatial filter were applied to the depth data, by using the built-in filters from the RealSense SDK set to default values. Further processing of the RealSense depth image was performed with the Point Cloud Library [18]. First, a region of interest was selected from the RealSense 3D image by cropping the face with a sphere, with the centre of the sphere placed at the pronasale. The radius of the sphere was determined by the distance of the exocanthion and pronasale. This radius was increased by 10% to fully include the eye region of the face. Remaining noise in the cropped RealSense images were removed with a statistical outlier filter [18]. The cropped RealSense images and 3dMD images were initially roughly aligned using the Procrustes algorithm, followed by a refined registration with the Iterative Closest Point (ICP) algorithm [19]. No scaling or reflection were applied during Procrustes and ICP registration. A distance map was calculated between the cropped RealSense depth image and the 3dMD image, based on the Euclidean distance between the vertices of the RealSense image and the 3dMD image. From this distance map, consisting of absolute values, the average Euclidean distance and standard deviation were calculated.

Facial landmark selection and placement

To determine the reliability of landmark placement, 14 facial landmarks (4 midline and 5 bilateral) were manually placed on the RealSense and 3dMD images (see Figure 1). The 14 landmarks were selected on their potential ability to describe the asymmetry of facial movement, such as the ability to close the eyes, smile, lifting the nose, lifting the eyebrows or pucker of the lips. Additionally, only clearly defined landmarks based on a standardized nomenclature of Caple & Stephan were selected to increase the accuracy and reproducibility of the landmark placement [20]. Landmarks based on the corneal apex were modified since the 3dMD system is not able to capture highly reflective surfaces, causing image artefacts at and around the eye region [15]. Therefore, the palpebrale superius was defined as the most inferior intersection of the upper eyelid with a vertical line through the palpebrale inferius. Next, the superciliare centralis was defined as the superior most intersection of the eyebrow with a vertical line through the palpebrale inferius and palpebrale superius to find the cross section of the eyebrow. In order to determine the inter-rater reliability, the 14 landmarks were manually placed by two experienced observers (TtH and SV) on the 3D surface of both the RealSense image and the 3dMD images for each patient using 3ds Max Studio 2018 (Autodesk, New York, NY, USA). The RealSense and 3dMD images were shown in a random order whereas the RealSense and 3dMD images could alternate. However, the sequence of 3D images was exactly the same for the two observers. The first observer (TtH) placed all the landmarks a second time after three weeks to determine the intra-rater reliability.

Intra- and inter-rater reliability and agreement of anthropometric measurements based on facial landmarks

A total of 14 anthropometric measurements were derived from the manually placed landmarks discussed in the previous paragraph by calculating the Euclidean distance between two selected landmarks (see Figure 1). First, the intra- and inter-rater reliability were determined separately for both the RealSense and 3dMD measurements. Subsequently, the average location of each landmark was calculated based on the landmarks from the first and second session. From this average landmark location, the intra- and inter-rater agreement of the RealSense measurements were determined by using the 3dMD measurements as the gold standard.

Statistical analysis

Statistical analysis was performed using IBM SPSS Statistics, Version 25 (IBM Corp., Armonk, NY, USA). A p-value of <0.05 was considered as statistically significant. All data were tested for a normal distribution using skewness-kurtosis tests [21]. The statistical analysis of landmark placement and anthropometric measurements were separated into the healthy and palsy side of the face. Subsequently, the intra- and inter-rater distance of the manual landmark placement was determined by calculating the Euclidean distance for each individual landmark between session one and two from observer TtH and between observer TtH and SV, for both the RealSense and 3dMD landmarks. The Wilcoxon signed-rank test was used for intra- and inter-observer measurements, to determine the significance between the intra- and inter-observer landmark placement due to a non-normal distribution of the landmark placement [22]. Anthropometric outcomes were normally distributed and therefore the reliability of the anthropometric measurements was determined by intra-class correlation coefficient (ICC) estimates and their 95% confidence interval (CI). The intra-rater reliability was calculated using a single-rater, absolute agreement, two-way mixed effects model. The inter-rater reliability was calculated using a single-rater, absolute agreement, two-way random effects model [23]. An ICC of <0.5 was considered as poor, 0.50 to 0.75 as fair, 0.75 to 0.90 as good, and 0.90 to 1.00 as excellent [24]. Agreement of the anthropometric measurements between the RealSense and the 3dMD measurements was assessed by the Bland-Altman method [25]. The Bland-Altman analysis was based on the mean systematic difference between the RealSense and 3dMD anthropometric measurements as well as on the upper and lower limits of agreement (LoA), which span 95% of all observations. The LoA was represented as a percentage of the 3dMD measurement.

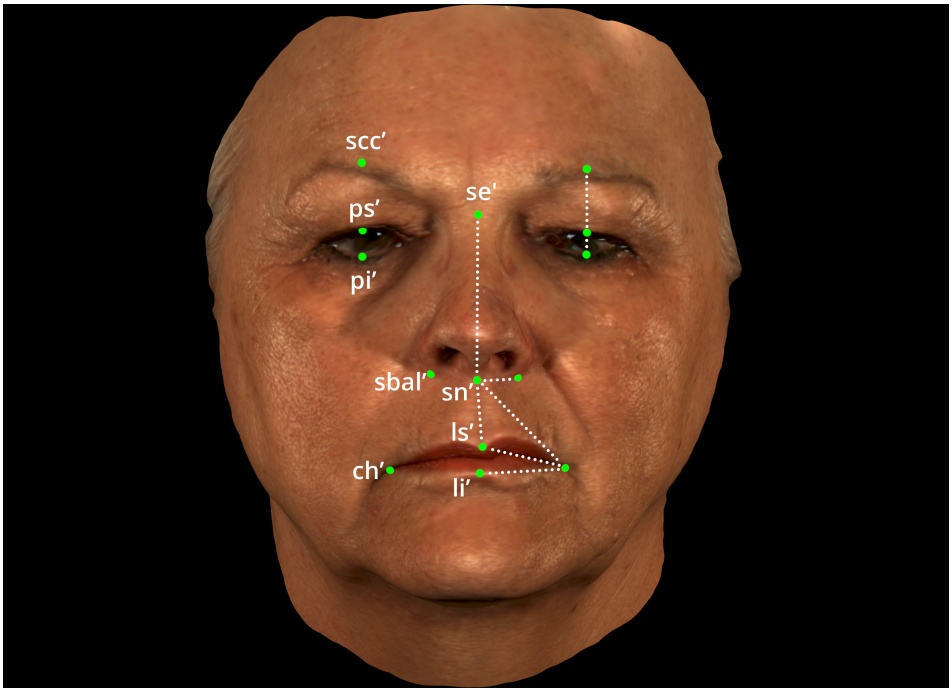

Figure 1. Manually placed landmarks on a face in a neutral (resting) position. The 14 manually placed 3D landmarks are shown as green markers (enlarged for visualization purposes) on the 3dMD image. The midline landmarks consisted of the sellion (se'), subnasale (sn'), labiale superius (ls'), and labiale inferius (li'). The bilateral landmarks consisted of the superciliare centralis (scc'), palpebrale superius (ps'), palpebrale inferius (pi'), subalare (sbal') and cheilion (ch'). The anthropometric measurements are indicated by the dotted white lines (only shown on the left side of the face for visualization purposes) and consisted of ps'-scc', ps'-pi', sbal'-sn', ch'-sn', ch'-ls', ch'-li', sn'-se', and sn'-ls'. Definitions of the landmarks are based on Caple & Stephan with scc' and ps' having modified definitions [16].

RESULTS

Population

A total of 30 patients with a PFP were included in this study, consisting of 11 men and 19 women, with an average age of 57 ± 13 years ranging from 30 to 87 years old.

Depth accuracy of the 3D image

The RealSense depth accuracy was determined by the average Euclidean distance from the RealSense image to the 3dMD image, where an average Euclidean distance of 0.97 ± 0.07 mm was found.

Table 1. The average (absolute) Euclidean distance including the standard deviation for the intra- and inter-rater landmark placement for the 3dMD and RealSense D415 measurements.

	3dMD		RealSense D415	
Landmark	**Intra-rater (mm)**	**Inter-rater (mm)**	**Intra-rater (mm)**	**Inter-rater (mm)**
ps' (p)	0.75 ± 0.46	0.98 ± 0.56	0.78 ± 0.41†	1.04 ± 0.50
ps'(h)	0.84 ± 0.55	0.81 ± 0.52	0.96 ± 0.47	1.16 ± 0.78
pi' (p)	0.77 ± 0.44	0.93 ± 0.46	0.66 ± 0.34	1.15 ± 0.57†
pi'(h)	0.69 ± 0.42	0.73 ± 0.41	0.79 ± 0.35	0.88 ± 0.52
scc'(p)	0.93 ± 0.57	1.27 ± 0.71	1.31 ± 0.67*	1.76 ± 1.28
scc'(h)	1.03 ± 0.45	1.29 ± 0.74	1.34 ± 0.75	1.90 ± 0.82*
sbal'(p)	0.72 ± 0.57	0.79 ± 0.51	2.25 ± 2.53*	2.88 ± 2.20*†
sbal'(h)	0.77 ± 0.39	0.84 ± 0.49	2.00 ± 2.26*	2.13 ± 2.41*
ch'(p)	0.74 ± 0.53	0.98 ± 0.95	1.24 ± 0.98*	1.76 ± 1.59*
ch'(h)	0.81 ± 0.61	1.20 ± 0.96	1.63 ± 1.19*	1.74 ± 1.01*
se'	1.05 ± 0.69	1.21 ± 0.80	1.55 ± 1.54	1.49 ± 1.18
sn'	0.80 ± 0.45	0.98 ± 0.53	1.12 ± 0.70	1.56 ± 1.09*
ls'	0.51 ± 0.50	0.57 ± 0.41	0.98 ± 0.61*	1.32 ± 1.70*
li'	1.38 ± 0.90	1.39 ± 0.90	1.84 ± 1.11	1.94 ± 1.50
Overall	**0.84 ± 0.58**	**1.00 ± 0.70**	**1.32 ± 1.27**	**1.62 ± 1.42**

The bilateral landmarks are grouped by the palsy side of the face (p) and the healthy side of the face (h). See Figure 1. For landmark abbreviations.

*Statistically significant difference between the placement of the RealSense D415 landmark and their respective 3dMD landmark ($p < 0.05$).

Intra- and inter-rater Euclidean distance of landmark placement

The average Euclidean distance was determined for the intra- and inter-rater landmark placement of both the RealSense and the 3dMD image as shown in Table 1, with a respective intra- and inter-rater distance of 0.84 ± 0.58 mm and 1.00 ± 0.70 mm for the 3dMD landmarks and 1.32 ± 1.27 mm and 1.62 ± 1.42 mm for the RealSense landmarks. Additionally, Table 1 marks the landmarks that showed a statistically significant difference between the RealSense and their 3dMD counterpart. When comparing the landmarks from the palsy and healthy side, a significant difference was found for the palpebrale superius for the intra-rater landmark placement on the RealSense image, and the palpebrale inferius and subalare for the inter-rater landmark placement on the RealSense image, also indicated on Table 1.

Table 2. The intra- and inter-rater reliability of the anthropometric measurements of the 3dMD and the RealSense D415.

	3dMD						RealSense					
	Intra-rater			Inter-rater			Intra-rater			Inter-rater		
Measurement	**ICC**	**Length (mm)**	**AD (mm)**	**ICC**	**Length (mm)**	**AD (mm)**	**ICC**	**Length (mm)**	**AD (mm)**	**ICC**	**Length (mm)**	**AD (mm)**
ps' – scc' (p)	0.98 (0.96 – 0.99)	19.8	0.4	0.97 (0.91 – 0.99)	20.0	0.6	0.93 (0.85 – 0.97)	19.8	0.8	0.93 (0.85 – 0.96)	19.7	0.8
ps' – scc' (h)	0.97 (0.94 – 0.99)	20.7	0.6	0.97 (0.94 – 0.99)	20.6	0.6	0.96 (0.91 – 0.98)	20.0	0.7	0.94 (0.87 – 0.97)	19.9	0.9
ps' – pi' (p)	0.97 (0.92 – 0.99)	9.1	0.4	0.97 (0.95 – 0.99)	8.9	0.4	0.96 (0.92 – 0.98)	8.6	0.5	0.95 (0.89 – 0.97)	8.7	0.6
ps' – pi' (h)	0.93 (0.85 – 0.96)	9.7	0.4	0.89 (0.78 – 0.95)	9.7	0.4	0.86 (0.73 – 0.93)	9.3	0.5	0.76 (0.55 – 0.88)	9.2	0.7
sbal' – sn' (p)	0.84 (0.62 – 0.93)	17.2	0.8	0.88 (0.76 – 0.94)	16.8	0.7	0.38 (0.05 – 0.65)	15.7	1.8	0.54 (0.18 – 0.76)	16.0	2.0
sbal' – sn' (h)	0.90 (0.65 – 0.96)	17.8	0.7	0.88 (0.77 – 0.94)	17.5	0.7	0.43 (0.11 – 0.68)	15.8	1.3	0.27 (0.00 – 0.56)	15.9	1.9
ch' – sn' (p)	0.96 (0.92 – 0.98)	44.3	0.8	0.91 (0.78 – 0.96)	43.7	1.2	0.88 (0.75 – 0.95)	43.0	1.5	0.87 (0.74 – 0.94)	42.3	1.9
ch' – sn' (h)	0.96 (0.91 – 0.98)	44.2	0.9	0.91 (0.74 – 0.96)	43.7	1.3	0.86 (0.74 – 0.93)	43.1	1.7	0.82 (0.66 – 0.91)	42.8	2.0
ch' – ls' (p)	0.97 (0.93 – 0.99)	35.9	0.7	0.96 (0.91 – 0.98)	35.5	0.9	0.94 (0.88 – 0.97)	34.8	1.2	0.84 (0.65 – 0.93)	34.1	1.9
ch' – ls' (h)	0.95 (0.90 – 0.98)	35.5	0.9	0.93 (0.86 – 0.97)	35.3	1.0	0.85 (0.71 – 0.92)	34.8	1.6	0.84 (0.68 – 0.92)	34.7	1.9
ch' – li' (p)	0.93 (0.86 – 0.97)	33.1	1.2	0.93 (0.85 – 0.96)	32.8	1.3	0.85 (0.71 – 0.93)	32.1	1.7	0.82 (0.66 – 0.91)	31.9	1.9
ch' – li' (h)	0.94 (0.88 – 0.97)	34.4	1.2	0.94 (0.87 – 0.97)	34.3	1.2	0.87 (0.74 – 0.94)	33.3	1.7	0.87 (0.74 – 0.93)	33.1	1.8
sn' – se'	0.97 (0.94 – 0.99)	53.8	1.0	0.97 (0.88 – 0.99)	54.2	1.1	0.92 (0.84 – 0.96)	53.5	1.4	0.93 (0.86 – 0.97)	53.3	1.5
sn' – ls'	0.97 (0.95 – 0.99)	18.3	0.5	0.95 (0.87 – 0.98)	18.1	0.6	0.94 (0.81 – 0.97)	17.2	0.9	0.81 (0.64 – 0.90)	17.1	1.4
Overall	**0.95 (0.87 – 0.98)**	**28.1**	**0.7**	**0.93 (0.85 – 0.97)**	**27.9**	**0.9**	**0.83 (0.70 – 0.91)**	**27.2**	**1.2**	**0.80 (0.64 – 0.89)**	**27.1**	**1.5**

The reliability is represented as the intra-class correlation coefficient (ICC), including the lower and upper bound of the 95% confidence interval. Furthermore, the average length and the absolute difference (AD) between the observer measurements are shown. The measurements are grouped by the palsy side of the face (p) and the healthy side of the face (h). See Figure 1 for landmark abbreviations.

Table 3. Numeric representations of the Bland-Altman analysis for the intra- and inter-rater agreement of the anthropometric measurements.

	Intra-rater agreement		Inter-rater agreement	
Measurement	**Mean difference (mm)**	**LoA (%)**	**Mean difference (mm)**	**LoA (%)**
ps' – scc' (p)	-0.03 (-2.10 – 2.04)	10.9	-0.32 (-3.47 – 2.83)	15.8
ps' – scc' (h)	-0.69 (-3.01 – 1.63)	10.9	-0.71 (-3.79 – 2.37)	14.2
ps' – pi' (p)	-0.44 (-2.30 – 1.42)	23.5	-0.25 (-2.30 – 1.80)	24.8
ps' – pi' (h)	-0.43 (-2.04 – 1.18)	17.5	-0.50 (-2.17 – 1.17)	18.8
sbal' – sn' (p)	-1.41 (-4.98 – 2.16)	21.6	-0.79 (-4.99 – 3.41)	25.6
sbal' – sn' (h)	-1.99 (-5.87 – 1.88)	22.5	-1.57 (-5.23 – 2.10)	22.1
ch' – sn' (p)	-1.23 (-5.41 – 2.95)	9.4	-1.43 (-6.36 – 3.51)	11.9
ch' – sn' (h)	-1.12 (-5.25 – 3.00)	9.4	-0.89 (-5.25 – 3.46)	9.8
ch' – ls' (p)	-1.08 (-5.49 – 3.34)	12.5	-1.34 (-7.41 – 4.72)	18.5
ch' – ls' (h)	-0.72 (-4.23 – 2.79)	9.8	-0.61 (-5.33 – 4.10)	13.5
ch' – li' (p)	-1.01 (-4.95 – 2.93)	12.2	-0.97 (-6.71 – 4.77)	19.2
ch' – li' (h)	-1.05 (-4.33 – 2.23)	9.9	-1.14 (-5.41 – 3.13)	12.6
sn' – se'	-0.29 (-3.49 – 2.92)	6.4	-0.91 (-3.80 – 1.97)	5.3
sn' – ls'	-1.10 (-3.12 – 0.92)	13.1	-1.06 (-2.87 – 0.75)	11.5
Overall	**-0.90 (-4.04 – 2.24)**	**13.6**	**-0.89 (-4.65 – 2.86)**	**16.0**

The data shows the mean systematic difference between the RealSense D415 and 3dMD measurements (mm) including the limits of agreement (LoA). Besides the LoA range in mm, the LoA are represented as the percentage (%) of the measured 3dMD distance. The anthropometric measurements are grouped by the palsy side of the face (p) and the healthy side of the face (h). See Figure 1 for landmark abbreviations.

Intra- and inter-rater reliability of anthropometric measurements

Intra- and inter-rater reliability (precision) was determined for the anthropometric measurements expressed as the ICC, as shown in Table 2, with an overall ICC of 0.95 and 0.93 for the intra- and inter-rater reliability of the 3dMD and 0.83 and 0.80 for the RealSense. When categorizing the ICC scores, all 3dMD anthropometric measurements for both the intra- and inter-raters are in the good to excellent category compared to 85.7% for the RealSense.

Intra- and inter-rater agreement of anthropometric measurements

A summary of the Bland-Altman analysis is shown in Table 3, where the intra- and inter-rater RealSense measurements were on average 0.90 mm and 0.89 mm lower compared to the 3dMD measurements, respectively. The LoA ranged on average from -4.04 to 2.24 mm and from -4.65 to 2.86 mm for the intra and inter-rater agreement, respectively.

DISCUSSION

Depth accuracy of the 3D image

This study assessed the reliability (precision) and agreement (accuracy) of anthropometric measurements for the RealSense D415 based on 3D images of 30 patients with a PFP. The first objective of this study was to determine the overall depth accuracy of the RealSense, expressed as the average Euclidean distance, using the 3dMD system as the gold standard where a depth accuracy of 0.97 ± 0.07 mm was found. In our previous study, a similar method was used to determine the depth accuracy of the RealSense F200, a predecessor of the RealSense D415, where an average depth accuracy of 1.48 ± 0.22 mm was found for patients with a PFP [17]. The higher depth accuracy of the RealSense D415, represented by a lower average Euclidean distance, can partially be explained by the lower recording distance of 35 cm, which is associated with a higher depth accuracy [17]. Additionally, the default temporal and spatial filters from the RealSense SDK were applied to the raw image data, which could positively influence the depth accuracy by removing temporal and spatial noise. In contrast, this study used overhead fluorescent lighting which reflected a more realistic representation of a patient's home situation, compared to the professional diffuse lighting setup used in our previous study [17]. Overall, the depth accuracy results indicated that there were no major scaling issues with the RealSense D415 depth data and that the general curvature of the face was successfully captured, with potentially a slightly higher depth accuracy compared to the RealSense F200 [17]. However, the average Euclidean distance is a general analysis of depth accuracy which does not specifically determines the accuracy of certain regions such as the mouth or the nose.

Intra- and inter-rater Euclidean distance of landmark placement

The second objective was to assess the average Euclidean distance of the intra- and inter-rater manual landmark placement on the RealSense D415 and the 3dMD images. The 3dMD landmarks acted as a reference with an average intra- and inter-rater landmark distance of 0.84 ± 0.58 mm and 1.00 ± 0.70 mm, respectively. Previous literature does not describe the Euclidean distance of manual 3D landmark placement of patients with a PFP. However, this distance is reported for healthy subjects and the average intra- and inter-rater distances range from 0.76 mm to 1.32 mm (intra) and 0.88 mm to 1.42 mm (inter) [26–30], which is consistent with the results from this study. The reported average landmark distance highly depends on the selected landmarks. Therefore, the distances of individual landmarks were also compared to the existing literature, although not all studies reported the individual distance for each landmark as used in this study. The palpebrale inferius, cheilion, sellion, subnasale, and labiale superius fell within a range

of 0.2 mm compared to the reported distances in the literature, whereas the palpebrale superius was found to be more than 0.5 mm accurate in this study [26–30]. Therefore, we consider the 3dMD landmark placement of the two observers in this study representative from what can be expected in a practical situation and can be used as a realistic reference for the RealSense landmark placement.

As landmarks can be directly placed on the RealSense image, in theory, the landmark placement could be equally as reliable as the landmark placement on the 3dMD image. However, the distance between the manual landmarks for the RealSense was on average 1.32 mm for the intra- and 1.62 mm for the inter-rater landmark placement, which falls around the 95th percentile of landmark placement compared to professional cameras such as the 3dMD [26–30]. Notably, a statistically significant higher Euclidean distance was found for the superciliare centralis, subalare, subnasale, cheilion, and labiale superius on the RealSense image compared to the 3dMD image. The first possible explanation for the higher landmark distance is the lack of depth data around certain regions not visible to the RealSense camera, due to the use of a single camera. An example is the alar region, where the nose blocks a direct view to the alar region, making the localization of the subalare unreliable. Hence the highest landmark distance was found for the subalare on the RealSense image. Secondly, the landmark distance is influenced by the colour quality of the RealSense image, since a subset of the landmarks are based on certain colour transitions, such as the labiale superius, labiale inferius or superciliare centralis. Due to the relatively higher age of the subjects in this study, the cheilion was often surrounded by skin folds causing shaded areas. The shaded areas made it harder to identify the location where the upper and lower vermillion border met, which explains the relatively higher inter-rater distance on the 3dMD image. When comparing the healthy and palsy side of the face, only significant differences were found for the RealSense landmarks. However, the placement of the subalare was found to be unreliable due to the lack of depth information, which is the likely cause of the significant difference instead of the presence of the PFP. Furthermore, the intra-rater palpebrale superius and inter-rater palpebrale inferius were found to be significantly different. However, these differences are not present for all the palpebrale superius and inferius landmarks, and in one case the healthy side had a lower landmark distance whilst in the other case, the palsy side had a lower distance. Therefore, it is not clear whether this difference is caused by the presence of the PFP or overall difficulty of selecting these landmarks. In general, the results from the Euclidean distances of the landmarks indicate that the lower depth and colour quality of the RealSense increase the difficulty of selecting the same landmark between multiple sessions or observers compared to the 3dMD images.

Intra- and inter-rater reliability of anthropometric measurements

The third objective was to determine the intra- and inter-rater reliability (precision) of the anthropometric measurements based on the manually placed landmarks as shown in Table 2. It is essential to determine whether the 3dMD measurements can be used as a reference for the RealSense measurements. Previous studies have not reported the reliability of anthropometric measurements for patients with a PFP. Instead, previous work used 2D landmarks for their analysis or focused on analysing motion [31–34]. Hence, only an indirect comparison could be made with anthropometric measurement studies based on healthy subjects [7,14,35–37]. The 3dMD system has a reported average absolute difference between observers of 0.8 mm ranging from 0.5 to 1.2 mm for anthropometric measurements [7]. If this difference is expressed as the percentage of the measurement the average difference is 4.1% with a range of 0.5 to 11.0%. Using the Di3D, another professional stereophotogrammetry system, a similar absolute difference of 0.87 ± 0.56 mm and 1.64 ± 1.08% has been reported [14]. Finally, an absolute difference of 0.99 ± 0.93 mm has been found using the Cyberware 3030RGB laser scanner [35]. These results were comparable to the 3dMD measurements performed in this study with an average absolute difference of 0.7 mm with a range of 0.4 to 1.2 mm and an average percentage difference of 2.9% with a range of 1.7 to 4.9% for the intra-rater measurements, respectively. The inter-rater measurements are within a similar range with an average absolute difference of 0.9 mm from a range to 0.4 to 1.3 mm and a percentage of 3.3% with a range from 2.1 to 4.2%. Similarly, the reported average intra-rater ICC ranges from 0.97 to 1.00 and from 0.83 to 0.99, which was comparable to this study where the average ICC ranged from 0.84 to 0.99 [14,37]. The reported average intra-rater ICC ranged from 0.70 to 0.98 and from 0.73 to 0.98, where this study has found an ICC in the range of 0.85 to 0.97 [36,37]. Since the absolute differences and the ICC reliability were both in a similar range as previously reported results, the 3dMD anthropometric measurements were considered to be representative results.

Therefore, a comparison could be made with the RealSense measurements. The RealSense measurements showed a relatively high ICC for the majority of the measurements, with 85.7% being in the good or excellent category, although all RealSense ICC values were lower compared to their relative 3dMD ICC score [24]. The 95% CI had a considerably wider range compared to the 3dMD measurements, indicating a lower reliability. A lower reliability was expected for the RealSense measurements, due to the overall higher distance between the individual landmark placement as explored in objective two. This was evident in the subalare to subnasale measurement, where the subalare landmark placement was found to be unreliable for the RealSense due to the lack of depth data, resulting in the lowest ICC value overall. When comparing the palsy side to the healthy side

the majority of the measurements had an ICC difference below 0.03. The remaining largest difference was between the palpebrale superius and inferius measurement. However, the ICC for the healthy side of the face was already lower for the 3dMD measurement. This result seemed to be amplified in the RealSense measurements. Despite the overall lower reliability of the RealSense measurements compared to the 3dMD it is important to determine the clinical impact of the lower reliability of the RealSense measurements.

Intra- and inter-rater agreement of anthropometric measurements

The fourth objective of this paper assessed the agreement (accuracy) of anthropometric measurements. In an ideal situation, all intra- and inter-rater measurements for both the 3dMD and the RealSense would be similar. Therefore, an initial estimate of agreement was made by comparing the length of the 3dMD and RealSense measurements as shown in Table 2. After subtracting the measured 3dMD length from the RealSense length for each individual measurement, an average underestimation of 0.91 mm and 0.89 was found for the intra- and inter-rater RealSense measurements, respectively. The Bland-Altman analysis confirmed these results with an overall underestimation of 0.90 mm and 0.89 mm for the intra- and inter-observer measurements, respectively. Furthermore, the average Bland-Altman analysis showed that in a clinical setting it is expected that 95% of the measurements will be less than 4.04 mm and 4.65 mm difference compared to the 3dMD measurements for the intra- and inter-rater, respectively. When expressed as a percentage of the original 3dMD measurement, this equates up to 95% of the measurements having a discrepancy of less than 13.6% and 16.0%, respectively. Most notably, the measurement between the palpebrale superius and inferius have an average LoA of 21.5%. Although the actual LoA was relatively low, ranging from -2.2 to 1.4 mm, the original measurement had a relatively short length of 9.2 mm. This caused the lower average depth accuracy of 0.97 mm from the RealSense image to have a higher impact on the overall measurement. Therefore, it is important to be aware that in a clinical implementation, 95% of the RealSense measurements based on the neutral face are expected to be within a range of 13.6 and 16.0% of the 3dMD measurements for the intra- and inter-rater measurements, respectively. However, this percentage most likely will increase when measuring short distances closer to the overall depth accuracy of the RealSense image.

Future research

As far as the authors are aware, this is the first study to use images from the RealSense D415 to assess anthropometric measurements. Therefore, this study focused on the manual 3D landmark placement of 14 selected landmarks on a neutral (resting) face from 30 patients with a PFP. This leaves multiple areas of research yet to be explored.

The current research primarily used the 3D capabilities of the RealSense camera by analysing the face in a neutral position. However, in order to effectively assess a PFP, multiple facial poses should be analysed [3–5]. Although the required facial poses will depend on the chosen grading scale, it is clear that the analysis of a single neutral pose would not be sufficient in assessing the severity of a PFP [3–5]. Therefore, the reliability and agreement of the anthropometric measurements based on different facial poses should be investigated. This could be realized by recording multiple facial poses with a 4D camera such as the RealSense D415. The 4D data could offer additional quantification methods for the assessment of the severity of a PFP, on top of the 3D landmarks and anthropometric measurements defined in this paper [3–5].

This study used the manual landmark placement on the 3dMD images as the gold standard since the reliability and agreement of manual landmark placement on high quality 3D images has been extensively researched [26–30]. Therefore, the reliability and agreement of the landmark placement and the anthropometric measurements found in this study could be directly compared to the existing literature. However, in order to minimize the effort required to perform the anthropometric measurements, the implementation of automatic landmark detection algorithms could be an interesting topic for future research [3–5]. Additionally, the RealSense camera simultaneously captured 2D colour images and 3D depth images in this study. This would make it possible for future research to compare the reliability and agreement of anthropometric measurements based on either the 2D or 3D landmarks [12,13,38]. Depending on the implementation of the objective assessment of the PFP, additional landmarks may be required on top of the 14 selected landmarks defined in this study. The results from this study could be used as an indication of expected reliability and agreement of the landmark placement and anthropometric measurements.

CONCLUSION

This study has assessed the reliability (precision) and agreement (accuracy) of anthropometric measurements based on manually placed 3D landmarks using the RealSense D415 within a population of patients with a PFP in a neutral (resting) position. This research can be used as a foundation for the implementation of the RealSense in a clinical or telemedicine setting, such as the objective assessment of facial asymmetry. Based on the reported reliability and agreement of the RealSense measurements, the RealSense D415 can be considered as a viable option to perform objective 3D anthropomorphic measurements on the neutral face in a clinical or telemedicine setting, where a low-cost and portable camera is required.

REFERENCES

1. McCaul, J. A. et al. Evidence based management of Bell's palsy. Br. J. Oral Maxillofac. Surg. 52, 387–391 (2014).
2. Peitersen, E. Bell's Palsy: The Spontaneous Course of 2,500 Peripheral Facial Nerve Palsies of Different Etiologies. Acta Otolaryngol. Suppl. 4–30 (2002) doi:10.1080/000164802760370736.
3. Samsudin, W. S. W. & Sundaraj, K. Image processing on facial paralysis for facial rehabilitation system: A review. Proc. - 2012 IEEE Int. Conf. Control Syst. Comput. Eng. ICCSCE 2012 259–263 (2012) doi:10.1109/ICCSCE.2012.6487152.
4. Revenaugh, P. C. et al. Use of Objective Metrics in Dynamic Facial Reanimation A Systematic Review. JAMA Facial Plast. Surg. 20, 501–508 (2018).
5. Niziol, R., Henry, F. P., Leckenby, J. I. & Grobbelaar, A. O. Is there an ideal outcome scoring system for facial reanimation surgery? A review of current methods and suggestions for future publications. J. Plast. Reconstr. Aesthetic Surg. 68, 447–456 (2015).
6. Aldridge, K., Boyadjiev, S. A., Capone, G. T., DeLeon, V. B. & Richtsmeier, J. T. Precision and error of three-dimensional phenotypic measures acquired from 3dMD photogrammetric images. Am. J. Med. Genet. 138 A, 247–253 (2005).
7. Wong, J. Y. et al. Validity and reliability of craniofacial anthropometric measurement of 3D digital photogrammetric images. Cleft Palate-Craniofacial J. 45, 232–239 (2008).
8. Weinberg, S. M. et al. Anthropometric Precision and Accuracy of Digital Three-Dimensional Photogrammetry. J. Craniofac. Surg. 17, 477–483 (2006).
9. Schimmel, M. et al. Distances between facial landmarks can be measured accurately with a new digital 3-dimensional video system. Am. J. Orthod. Dentofac. Orthop. 137, 1–10 (2010).
10. Aynechi, N., Larson, B. E., Leon-Salazar, V. & Beiraghi, S. Accuracy and precision of a 3D anthropometric facial analysis with and without landmark labeling before image acquisition. Angle Orthod. 81, 245–252 (2011).
11. Hong, C. et al. Evaluation of the 3dMDface system as a tool for soft tissue analysis. Orthod. Craniofacial Res. 20, 119–124 (2017).
12. Dindaroğlu, F., Kutlu, P., Duran, G. S., Görgülü, S. & Aslan, E. Accuracy and reliability of 3D stereophotogrammetry: A comparison to direct anthropometry and 2D photogrammetry. Angle Orthod. 86, 487–494 (2016).
13. Ghoddousi, H., Edler, R., Haers, P., Wertheim, D. & Greenhill, D. Comparison of three methods of facial measurement. Int. J. Oral Maxillofac. Surg. 36, 250–258 (2007).
14. Fourie, Z., Damstra, J., Gerrits, P. O. & Ren, Y. Evaluation of anthropometric accuracy and reliability using different three-dimensional scanning systems. Forensic Sci. Int. 207, 127–134 (2011).

15. Heike, C. L., Upson, K., Stuhaug, E. & Weinberg, S. M. 3D digital stereophotogrammetry: A practical guide to facial image acquisition. Head Face Med. 6, 1–11 (2010).
16. Zaki, R. Validation of Instrument Measuring Continuous Variable in Medicine. in Advances in Statistical Methodologies and Their Application to Real Problems 217–237 (IntechOpen, 2017). doi:10.5772/66151.
17. ten Harkel, T. C. et al. Depth accuracy of the RealSense F200: Low-cost 4D facial imaging. Sci. Rep. 7, 16263 (2017).
18. Rusu, R. B. & Cousins, S. 3D is here: Point Cloud Library (PCL). 2011 IEEE Int. Conf. Robot. Autom. 1–4 (2011) doi:10.1109/ICRA.2011.5980567.
19. Segal, A. V, Haehnel, D. & Thrun, S. Generalized-ICP. Proc. Robot. Sci. Syst. 2, 4 (2009).
20. Caple, J. & Stephan, C. N. A standardized nomenclature for craniofacial and facial anthropometry. Int. J. Legal Med. (2015) doi:10.1007/s00414-015-1292-1.
21. Lilliefors, H. W. On the Kolmogorov-Smirnov Test for Normality with Mean and Variance Unknown. J. Am. Stat. Assoc. 62, 399–402 (1967).
22. Wilcoxon, F. Individual Comparisons by Ranking Methods. Biometrics Bull. 1, 80–83 (1945).
23. McGraw, K. O. & Wong, S. P. Forming Inferences about Some Intraclass Correlation Coefficients. Psychol. Methods 1, 30–46 (1996).
24. Koo, T. K. & Li, M. Y. A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. J. Chiropr. Med. 15, 155–63 (2016).
25. Bland, J. M. & Altman, D. G. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 327, 307–310 (1986).
26. Plooij, J. M. et al. Evaluation of reproducibility and reliability of 3D soft tissue analysis using 3D stereophotogrammetry. Int. J. Oral Maxillofac. Surg. 38, 267–73 (2009).
27. Lin, H., Zhu, P., Lin, Y., Zheng, Y. & Xu, Y. Reliability and Reproducibility of Landmarks on Three-Dimensional Soft-Tissue Cephalometrics Using Different Placement Methods. Plast. Reconstr. Surg. 134, 102e-110e (2014).
28. Toma, A. M., Zhurov, A., Playle, R., Ong, E. & Richmond, S. Reproducibility of facial soft tissue landmarks on 3D laser-scanned facial images. Orthod. Craniofacial Res. 12, 33–42 (2009).
29. Fagertun, J. et al. 3D facial landmarks: Inter-operator variability of manual annotation. BMC Med. Imaging 14, 1–9 (2014).
30. Baysal, A., Sahan, A. O., Ozturk, M. A. & Uysal, T. Reproducibility and reliability of three-dimensional soft tissue landmark identification using three-dimensional stereophotogrammetry. Angle Orthod. 86, 1004–1009 (2016).
31. Neely, J. G., Wang, K. X., Shapland, C. A., Sehizadeh, A. & Wang, A. Computerized Objective Measurement of Facial Motion. Otol. Neurotol. 31, 1488–1492 (2010).
32. Mehta, R. P., Zhang, S. & Hadlock, T. A. Novel 3-D video for quantification of facial movement. Otolaryngol. - Head Neck Surg. 138, 468–472 (2008).

33. Hadlock, T. A. & Urban, L. S. Toward a universal, automated facial measurement tool in facial reanimation. Arch. Facial Plast. Surg. 14, 277–282 (2012).
34. Shujaat, S. et al. The clinical application of three-dimensional motion capture (4D): A novel approach to quantify the dynamics of facial animations. Int. J. Oral Maxillofac. Surg. 43, 907–916 (2014).
35. Ramieri, G. A. et al. Reconstruction of facial morphology from laser scanned data. Part I: Reliability of the technique. Dentomaxillofacial Radiol. 35, 158–164 (2006).
36. Ceinos, R., Tardivo, D., Bertrand, M. F. & Lupi-Pegurier, L. Inter- and Intra-Operator Reliability of Facial and Dental Measurements Using 3D-Stereophotogrammetry. J. Esthet. Restor. Dent. 28, 178–189 (2016).
37. Othman, S. A., Majawit, L. P., Hassan, W. N. W., Wey, M. C. & Razi, R. M. Anthropometric study of three-dimensional facial morphology in Malay adults. PLoS One 11, 1–15 (2016).
38. Caple, J. & Stephan, C. N. A standardized nomenclature for craniofacial and facial anthropometry. Int. J. Legal Med. 130, 863–879 (2016).

CHAPTER 3

PART 2: RELIABILITY AND AGREEMENT OF 3D ANTHROPOMETRIC MEASUREMENTS DURING THE VOLUNTARY MOVEMENTS OF THE SUNNYBROOK FACIAL GRADING SYSTEM

Authors: Timen C. ten Harkel, Shankeeth Vinayahalingam, Thomas J.J. Maal, Henri A.M. Marres, Caroline M. Speksnijder & Koen J.A.O. Ingels

ABSTRACT

Background

This study determined the reliability (precision) and agreement (accuracy) of 14 anthropometric measurements during 5 voluntary movements of the Sunnybrook Facial Grading System (SFGS).

Materials & Methods

Thirty patients with a unilateral peripheral facial palsy (PFP) were recorded simultaneously with the 3dMD system (gold standard) and a low-cost RealSense D415 camera. Measurements were derived from 14 manually placed landmarks on the 3dMD and RealSense 3D images at maximum exertion for each of the voluntary movements.

Results

The depth accuracy of the RealSense ranged between 0.95 mm and 1.01 mm during the voluntary movements. The reliability of the landmark placement ranged from 0.87 mm to 1.02 mm (intra-rater) and 1.00 mm to 1.25 mm (inter-rater) for the 3dMD image. For the RealSense image this ranged from 1.04 mm to 1.28 mm (intra-rater) and from 1.39 mm to 1.57 mm (inter-rater). The reliability of the anthropometric measurements, expressed as the intra-class correlation coefficient, ranged from 0.95 to 0.97 (intra-rater) and from 0.93 to 0.96 (inter-rater) for the 3dMD measurements and from 0.90 to 0.95 (intra-rater) and from 0.87 to 0.92 (inter-rater) for the RealSense measurements. The agreement of the anthropometric measurements resulted in an underestimation of -1.47 mm to -1.02 mm (intra-rater) and -1.34 mm to -1.02 mm (inter-rater) of the RealSense measurements compared to the 3dMD.

Conclusion

Firstly, the reliability of the 3D landmark placement and anthropometric measurements was similar for patients with a PFP performing the voluntary movements of the SFGS compared to healthy subjects at rest, when using the high quality 3dMD images. Secondly, the 3D anthropometric measurements on the RealSense images showed a relatively consistent reliability and agreement during the voluntary movements of the SFGS based on the patients with a PFP. Therefore, the RealSense can be considered as a viable option to perform objective measurements in case a low-cost and portable camera is required, e.g. in an eHealth environment.

INTRODUCTION

Peripheral facial palsy (PFP) is caused by a lesion at or below the facial nucleus of the seventh cranial nerve, resulting in a partial or complete impairment of the ipsilateral mimic muscles. PFP has an estimated incidence rate of between 11 to 40 per 100,000 people per annum, with the majority of cases classified as an idiopathic PFP [1]. In order to assess the severity of a PFP during diagnostic workup and follow-up there are multiple subjective diagnostic tools available, such as the Sunnybrook Facial Grading System (SFGS), House Brackmann, and Sydney score [2,3]. Additionally, objective measurements can be used to determine asymmetry measurements for the grading of a PFP, where a major class of objective measurements are based on anthropometric measurements [2–4].

Anthropometry is the scientific method to measure dimensions of the human body. Traditionally, this method utilized callipers to measure the distance between two facial landmarks [5]. The calliper has mostly been replaced by digital measurements based on 2D images, 2D videos, static 3D images, or 3D videos (also called 4D imaging) [2–5]. One of the advantages of 3D imaging is the ability to include depth and express the anthropometric measurements in real world coordinates [6]. A 4D imaging system might be preferred over a 3D imaging system since certain aspects of a PFP are only visible during certain facial poses. Preferably, the 4D imaging system would be limited in size, complexity, and cost, making the implementation of a 4D analysis more accessible. However, these properties could result in a lower image quality and depth accuracy, which could influence the accuracy of the landmark placement and the anthropometric measurements.

In previous work, the reliability and agreement of anthropometric measurements were determined using such a low-cost portable 4D imaging system, the RealSense D415 [7]. This study analysed a population of patients with a PFP with the face at rest. However, it is important to determine the effect on the reliability and agreement of the anthropometric measurements during multiple facial poses. One of the major subjective grading systems for PFP is the SFGS [8]. The SFGS includes the scoring of the facial symmetry at rest, during five voluntary movements and the scoring of synkinesis. The five voluntary movements of the SFGS consist of raising the eyebrows, gently closing the eyes, open mouth smiling, snarling (raising the nasal ala), and puckering the lips [9]. Due to the extensive use of the SFGS and the variety of voluntary movements, the SFGS provides a good representation of clinically relevant voluntary movements to assess a PFP in clinical practice.

Therefore, the aim of this study is to determine the impact of the five voluntary movements of the SFGS on the reliability and agreement of the anthropometric measurements, based on manually placed 3D landmarks on patients with a PFP, using the RealSense D415. The results of this study will be compared to the research analysing the face at rest to determine the potential influence of the voluntary movements on the reliability and agreement of the anthropometric measurements during clinical measurements [7]. In order to compare the results to the research analysing the face at rest, the following four objectives were adapted using the previous study protocol using the same professional 3dMD imaging system as the gold standard, with a reported depth accuracy of 0.20 to 0.25 mm [7,10–12]. The first objective was to determine the depth accuracy of the RealSense 3D images for each of the voluntary movements (1). Secondly, the intra- and inter-rater reliability (precision) of the manual 3D landmark placement were assessed for the voluntary movements for both the RealSense and 3dMD images (2). Thirdly, the intra- and inter-rater reliability of anthropometric measurements based on the manually placed landmarks were determined for the voluntary movements for both the RealSense and 3dMD images (3). Finally, the intra- and inter-rater agreement (accuracy) of the RealSense measurements was assessed for each of the voluntary movements by using 3dMD measurements as the gold standard (4).

MATERIALS & METHODS

Population

Between August 2018 and April 2019, all patients with a unilateral PFP seen during a multidisciplinary consultation at the Department of Otorhinolaryngology of the Radboudumc, were eligible for participation in this study. Exclusion criteria were the presence of epilepsy and an age younger than 16 years. Approval of this study was authorized by the Ethics Committee of the Radboudumc (2015-1829) and was conducted in compliance with the World Medical Association Declaration of Helsinki on medical research ethics. All subjects provided a written informed consent for the participation in this study. Additionally, patients shown in this study provided a written informed consent for the use of their images.

Image acquisition

The image acquisition consisted of recording the five voluntary movements based on the SFGS (i.e., forehead wrinkle, gentle eye closure, open mouth smile, snarl, and lip pucker). These voluntary movements were captured simultaneously with the two-pod 3dMD system (3dMDface, 3dMD, Atlanta, USA) and the RealSense D415 (Intel, Santa Clara, USA). The RealSense D415 recorded with 30 frames per second at an approximate distance

of 35 cm to the patient, with a colour resolution of 1920 × 1080 pixels and a depth resolution of 1280 × 720 pixels. A static 3D image was captured with the 3dMD system at maximum exertion of each voluntary movement, which acted as the clinical reference image.

Objective 1: Depth accuracy of the 3D image

The method to determine the RealSense depth accuracy was described in previous work, which consisted of exporting the 3D images from the RealSense recording at the time of maximum exertion of one of the five voluntary movements, which was at the same time of the static 3dMD image [7]. The default temporal and spatial filters from the RealSense Software Development Kit were applied to the RealSense depth images. The RealSense depth image was cropped by a sphere placed on the pronasale, with a radius determined by the distance between the exocanthion and pronasale. In order to fully include the eye region of the face, this radius was increased by 10%. Remaining noise was removed from the RealSense image with a statistical outlier filter [13]. Registration between the RealSense image and the 3dMD image was performed with the Procrustes algorithm followed by the Iterative Closest Point algorithm without applying scaling or reflection [14]. A distance map was calculated between the RealSense depth image and the 3dMD image for each of the voluntary movements. From this distance map, the average Euclidean distance was calculated, which only consisted of absolute values.

Objective 2: Intra- and inter-rater reliability of landmark placement

During this study, a total of 14 key facial landmarks (4 midline and 5 bilateral), selected on the ability to track movement during the voluntary movements, were placed on the RealSense and 3dMD images for each of the five voluntary movements (Figure 1). The landmark definitions were based on Caple & Stephan with alterations described in our earlier research due to limitations of the 3dMD system to capture reflective surfaces [7,15]. The alterations included the palpebrale superius (ps'), which was defined as the most inferior intersection of the upper eyelid with a vertical line through the palpebrale inferius (pi').The superciliare centralis (scc') was defined as the superior most intersection of the eyebrow with a vertical line through pi' and ps' to find the cross section of the eyebrow [7]. The landmarks were manually placed by two experienced observers (TtH and SV) on the 3D surface of both the RealSense image and the 3dMD images on all the five voluntary movements for each individual patient, in order to determine the intra-rater reliability. Observer TtH placed all the landmarks a second time after three weeks, to determine the intra-observer reliability. The RealSense and 3dMD images were shown in a random order whereas the RealSense and 3dMD images could alternate, with the exact same sequence for the two observers.

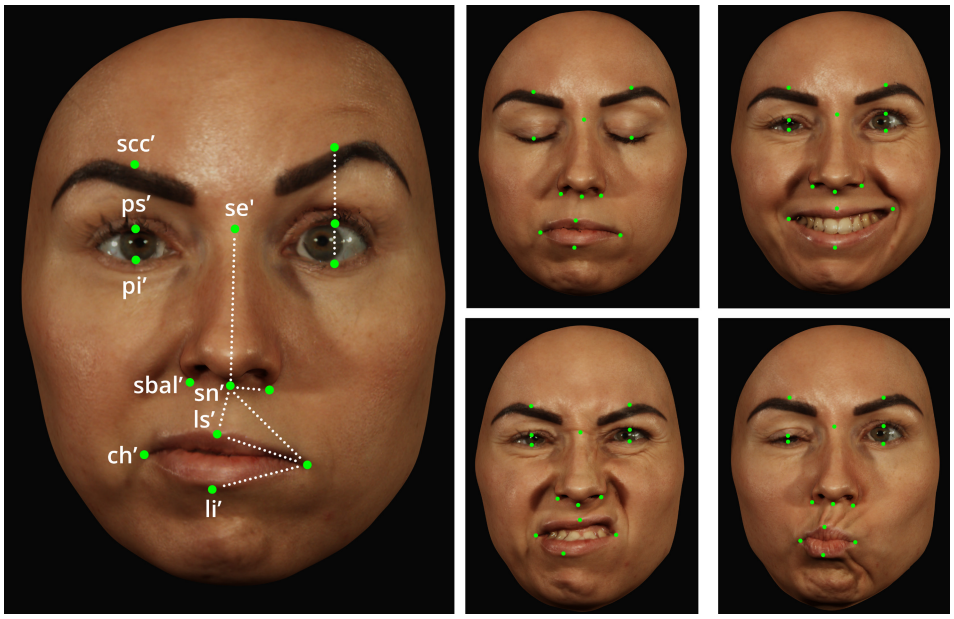

Figure 1. Manually placed landmarks at the moment of maximum exertion of the five voluntary movements of the Sunnybrook Facial Grading System (SFGS). The 14 manually placed 3D landmarks are shown as green dots on the 3dMD image, consisting of 4 midline landmarks and 10 bilateral landmarks, with the landmark definitions based on Caple & Stephan with scc' and ps' having modified definitions as defined in the baseline study of the face at rest [6,7]. The midline landmarks consisted of the sellion (se'), subnasale (sn'), labiale superius (ls'), and labiale inferius (li'). The bilateral landmarks consisted of the superciliare centralis (scc'), palpebrale superius (ps'), palpebrale inferius (pi'), subalare (sbal') and cheilion (ch'). The anthropometric measurements are indicated by the dotted white lines as shown on the left side of the face. The measurements consisted of ps'–scc', ps'–pi', sbal'–sn', ch'–sn', ch'–ls', ch'–li', sn'–se', and sn'–ls'.

Objective 3 & 4: Intra- and inter-rater reliability and agreement of anthropometric measurements

A total of 14 anthropometric measurements were determined by calculating the Euclidean distance between a combination of 2 out of the 14 manually placed landmarks (Figure 1). The intra- and inter-rater reliability was determined separately for all 14 anthropometric measurements for each of the voluntary movements for both the RealSense D415 and 3dMD images.

Statistical analysis

The statistical analysis was performed for each of the voluntary movements individually in order to compare the results to the face at rest. The analysis was split up between the healthy and palsy side of each patient for the bilateral landmarks and measurements and the intra- and inter-rater Euclidean distance was determined for both the RealSense and 3dMD landmarks individually. Due to the non-normal distribution of the landmark placement the Wilcoxon signed-rank test was used to

Table 1. The average depth accuracy of the RealSense D415 for each voluntary movement of the Sunnybrook Facial Grading System (SFGS). The depth accuracy is expressed as the average Euclidean distances between the RealSense D415 surface to the 3dMD surface.

Voluntary movement of the SFGS	Depth accuracy (mm)
Forehead wrinkle	0.96 ± 0.07
Gentle eye closure	0.95 ± 0.07
Open mouth smile	1.01 ± 0.10
Snarl	0.99 ± 0.10
Lip pucker	0.97 ± 0.08

determine the reliability of the intra- and inter-observer landmark placement [16]. The anthropometric measurements were normally distributed, and the reliability of these measurements was determined by the intra-class correlation coefficient (ICC) estimates and their 95% confidence interval (CI). An ICC of <0.5 was considered as poor, 0.50 to 0.75 as fair, 0.75 to 0.90 as good, and 0.90 to 1.00 as excellent [17]. The intra-rater reliability was calculated using a single-rater, absolute agreement, two-way mixed effects model. The inter-rater reliability was calculated using a single-rater, absolute agreement, two-way random effects model [18]. Finally, the agreement of the anthropometric measurements was determined by the Bland-Altman method [19], based on the mean systematic difference between the RealSense and 3dMD anthropometric measurements. The analysis also included the upper and lower limit of agreement (LoA), which span 95% of all observations, represented as a percentage of the 3dMD measurement.

RESULTS

Population

This study included the recordings of 30 patients with a PFP, which consisted of 11 men and 19 women, with an average age of 57 ± 13 years ranging from 30 to 87 years. This population was identical to the population included in the baseline study analysing the face at rest and the voluntary movements were performed immediately after the recording of the face at rest [7].

Objective 1: Depth accuracy of the 3D image

The RealSense depth accuracy was determined for all individual voluntary movements by calculating the average Euclidean distance from the RealSense image to the 3dMD image. The average Euclidean distance ranged between 0.95 mm and 1.01 mm (Table 1), with the best and worst depth accuracy found for the gentle eye closure and open mouth smile, respectively.

Table 2. The average Euclidean distance including the standard deviation for the intra- and inter-rater landmark placement for each voluntary movement of the Sunnybrook Facial Grading System (SFGS) based on the 3dMD and RealSense D415 images.

	3dMD		RealSense D415	
Voluntary movement of the SFGS	**Intra-rater (mm)**	**Inter-rater (mm)**	**Intra-rater (mm)**	**Inter-rater (mm)**
Forehead wrinkle	0.87 ± 0.57	1.00 ± 0.80	1.15 ± 0.76	1.46 ± 1.06
Gentle eye closure	1.01 ± 0.71	1.25 ± 0.94	1.09 ± 0.70	1.39 ± 0.90
Open mouth smile	1.02 ± 0.78	1.14 ± 0.88	1.04 ± 0.68	1.41 ± 1.08
Snarl	0.91 ± 0.63	1.14 ± 0.80	1.28 ± 0.98	1.57 ± 1.09
Lip pucker	0.90 ± 0.58	1.06 ± 0.66	1.06 ± 0.62	1.37 ± 0.92

Table 3. Overview of landmarks with a statistically significant difference between the RealSense D415 landmark placement and their respective 3dMD landmark ($p < 0.05$).

Voluntary movement of the SFGS	**Intra-rater**	**Inter-rater**
Forehead wrinkle	scc' (h), sbal' (p & h), sn', ls'	scc' (p & h), sbal' (p & h), ch' (p & h), sn', ls', li'
Gentle eye closure	sbal' (p & h), ch' (p), sn', ls'	sbal' (h), ch' (p & h), sn', ls'
Open mouth smile	-	sbal' (p & h), se', sn'
Snarl	pi' (h), scc' (p), sbal' (p & h), ch' (p & h), sn', ls'	scc' (h), sbal (h'), se', ls', li'
Lip pucker	ch' (p), sn', ls'	scc' (h), sbal' (h), sn'

The bilateral landmarks are indicated by either the palsy side (p) and the healthy side (h) of the face. See Figure 1 for the landmark abbreviations. SFGS = Sunnybrook Facial Grading System.

Objective 2: Intra- and inter-rater reliability of landmark placement

The intra- and inter-rater reliability of the manual landmark placement was determined for the 14 landmarks as shown in Figure 1. Table 2 shows the average Euclidean distance for the intra- and inter-rater landmarks based on the individual voluntary movements for both the 3dMD and RealSense image separately. The following five landmarks were found to have a statistically significant difference between the healthy and the palsy side of the face: (1) ps' during the inter-rater 3dMD landmark placement of the lip pucker, (2) ps' during the intra-rater placement of the RealSense landmark of the forehead wrinkle, (3 & 4), ps' and pi' during the inter-rater placement of the RealSense landmark of the gentle eye closure, and (5) ps' during the inter-rater placement of the RealSense landmark of the lip pucker. Additionally, the landmarks resulting in a significant difference between the 3dMD and RealSense placement are shown in Table 3. A notable result in the reliability of the RealSense landmark placement was the subalare (sbal'), with an average intra- and inter-rater Euclidean distance of 1.22 mm and 1.46 mm during all voluntary movements.

Table 4. The intra- and inter-rater reliability of the anthropometric measurements for each voluntary movement of the Sunnybrook Facial Grading System (SFGS) based on the 3dMD measurements.

	3dMD measurements					
	Intra-rater			Inter-rater		
Voluntary movement of the SFGS	ICC	Length (mm)	AD (mm)	ICC	Length (mm)	AD (mm)
Forehead wrinkle	0.95 (0.88 – 0.98)	28.1	0.8	0.93 (0.85 – 0.97)	28.2	0.8
Gentle eye closure	0.95 (0.91 – 0.98)	26.9	0.7	0.94 (0.87 – 0.97)	26.9	0.8
Open mouth smile	0.95 (0.88 – 0.98)	28.9	0.8	0.93 (0.87 – 0.97)	28.9	1.0
Snarl	0.96 (0.91 – 0.98)	26.7	0.7	0.94 (0.88 – 0.97)	26.7	0.9
Lip pucker	0.97 (0.91 – 0.99)	24.7	0.7	0.96 (0.92 – 0.98)	24.8	0.7

The reliability is represented as the intra-class correlation coefficient (ICC), including the lower and upper bound of the 95% confidence interval. Furthermore, the average length and the absolute difference (AD) between the observer measurements are shown.

Table 5. The intra- and inter-rater reliability of the anthropometric measurements for each voluntary movement of the Sunnybrook Facial Grading System (SFGS) based on the RealSense D415 measurements.

	RealSense D415 measurements					
	Intra-rater			Inter-rater		
Voluntary movement of the SFGS	ICC	Length (mm)	AD (mm)	ICC	Length (mm)	AD (mm)
Forehead wrinkle	0.90 (0.80 – 0.95)	26.6	1.1	0.87 (0.69 – 0.94)	26.8	1.3
Gentle eye closure	0.93 (0.87 – 0.97)	25.6	0.9	0.91 (0.80 – 0.96)	25.8	1.1
Open mouth smile	0.95 (0.90 – 0.98)	27.8	0.8	0.89 (0.78 – 0.94)	27.8	1.2
Snarl	0.92 (0.84 – 0.96)	25.6	1.1	0.89 (0.79 – 0.94)	25.7	1.4
Lip pucker	0.94 (0.87 – 0.97)	23.7	0.9	0.92 (0.82 – 0.96)	23.7	1.1

The reliability is represented as the intra-class correlation coefficient (ICC), including the lower and upper bound of the 95% confidence interval. Furthermore, the average length and the absolute difference (AD) between the observer measurements are shown.

Objective 3: Intra- and inter-rater reliability of anthropometric measurements

The intra- and inter-rater reliability (precision) was determined for the 14 anthropometric measurements as described in Figure 1 for each of the voluntary movements and separately for both the 3dMD and RealSense image. The reliability was expressed as the ICC and Table 4 shows the average ICC, measured length, and absolute distance of the 14 anthropometric 3dMD measurements for each voluntary movement. Table 5 shows these results for the RealSense D415 measurements.

Table 6. Numeric representations of the Bland-Altman analysis for the intra- and inter-rater agreement of the anthropometric measurements for each voluntary movement of the Sunnybrook Facial Grading System (SFGS).

	Intra-rater agreement		Inter-rater agreement	
Voluntary movement of the SFGS	**Mean difference (mm)**	**LoA (%)**	**Mean difference (mm)**	**LoA (%)**
Forehead wrinkle	-1.47 (-4.88 – 1.95)	14.5 (13.3)	-1.34 (-4.98 – 2.30)	15.1 (14.0)
Gentle eye closure	-1.36 (-4.81 – 2.10)	29.1 (14.2)	-1.17 (-4.62 – 2.29)	34.2 (14.2)
Open mouth smile	-1.07 (-4.68 – 2.54)	16.5 (14.2)	-1.16 (-4.81 – 2.49)	16.0 (14.3)
Snarl	-1.02 (-5.02 – 2.97)	24.3 (16.3)	-1.02 (-5.18 – 3.14)	25.4 (17.1)
Lip pucker	-1.08 (-4.44 – 2.28)	19.4 (13.9)	-1.06 (-4.41 – 2.28)	19.3 (13.9)

The mean systematic difference between the RealSense D415 and 3dMD measurements (mm) including the limit of agreement (LoA). Additionally, the LoA is represented as the percentage (%) of the measured 3dMD distance. The LoA percentage in between brackets is the LoA percentage excluding the measurement ps' – pi'.

Objective 4: Intra- and inter-rater agreement of anthropometric measurements

The intra- and inter-rater reliability (accuracy) was determined for the 14 anthropometric measurements for each of the voluntary movements. The numeric representation of the Bland-Altman analysis including the LoA is shown in Table 6 for the average of the 14 measurements for each of the voluntary movements. The lowest average agreement was found during the intra-rater measurements of the forehead wrinkle (-1.47 mm) and gentle eye closure (-1.36 mm). The agreement was mostly affected by the measurements involving ch', with an average agreement of -2.05 mm for the palsy side and -2.7 mm for the healthy side of the face for the intra-rater agreement. Additionally, Table 6 shows the LoA expressed as the percentage of the 3dMD measurement. In addition, the LoA percentage without the ps' – pi' measurement is shown in between brackets, due to the major increase in LoA percentage when measuring lower distances without impacting the agreement. For example, during the gentle eye closure, the intra-rater LoA percentage was 118.7% whilst having a higher agreement compared to the other voluntary movements for the ps' – pi' measurement.

DISCUSSION

This study determined the impact of the five voluntary movements on the reliability and agreement of anthropometric measurements, based on manually placed 3D landmarks on patients with a PFP, using the RealSense D415. The results from this study can act as a reference when implementing anthropometric measurements using a low cost-camera in a clinical setting or when further exploring the 3D and 4D

capabilities of the RealSense. This study extends previous work, where the reliability and agreement of the anthropometric measurements was determined for the face at rest [7]. In order to compare data between the studies, the material and methods were kept as consistent as possible between the two studies. For example, the objectives, population of patients with a PFP, depth accuracy calculations, selection of landmarks, anthropometric measurements, observers, and statistical analysis, did not change between the studies.

Objective 1: Depth accuracy of the 3D image

The first objective of this study was to determine the overall depth accuracy of the RealSense for each of the voluntary movements using the 3dMD image as the gold standard. The depth accuracy ranged between 0.95 and 1.01 mm for the voluntary movements (Table 1), which was in the same range as the face at rest (0.97 mm) [7]. These results correspond with the depth accuracy of a previous generation RealSense camera, the RealSense F200, which was not significantly influenced by the voluntary movements [20]. However, there is still a major difference between the depth accuracy of the RealSense D415 and the 3dMD system with a respective depth accuracy of 0.20 to 0.25 mm for healthy subjects [10–12].

Objective 2: Intra- and inter-rater reliability of landmark placement

Reliability of 3dMD landmark placement

The second objective assessed the intra- and inter-rater reliability (precision) of the manual 3D landmark placement for the voluntary movements for both the 3dMD and RealSense images. The landmarks of the 3dMD are discussed first, in order to determine the reliability of the landmark placement in an ideal scenario with a high-quality 3D image. The baseline study analysing the face at rest, reported an average intra- and inter-rater distance of 0.84 mm and 1.00 mm, respectively for the 3dMD landmarks [7]. All voluntary movements were found to have a higher average landmark distance, as shown in Table 2.

The landmarks in the eye region played a crucial role in the decreased landmark reliability. During the (partial) eye closure it was difficult to determine the location of pi', since the inferior border of the lower eyelid could be blocked by the upper eyelid or eyelashes. The reliability of ps' and scc' were indirectly affected as well since these landmarks were derived from pi'. Due to synkinesis, eye closure could also occur during other voluntary movements [21]. During the lip pucker this resulted in a single significant difference of the inter-rater landmark placement of pi' between the healthy and palsy side of the face. The landmark placement was not significantly influenced by the palsy and healthy side of the face for the remaining 98% of the bilateral landmarks.

No studies were found reporting the reliability of landmark placement on patients with a PFP using high-quality stereophotogrammetry images. Therefore, the landmark placement was compared to studies based on healthy subjects with the face at rest, with the reliability of the landmark placement ranging from 0.76 mm to 1.32 mm (intra-rater) and 0.88 mm to 1.42 mm (inter-rater) landmark distances [22–26]. These results indicate that the reliability of landmark placement is not negatively impacted for patients with a PFP during the voluntary movements compared to healthy subjects at rest, when using high quality 3D imaging systems such as the 3dMD. Therefore, the reliability of the 3dMD landmark placement during the voluntary movements was found to be sufficient to act as a reference for the RealSense landmark placement.

Reliability of RealSense D415 landmark placement

The RealSense landmark placement showed an increase in landmark reliability during the voluntary movements, as shown in Table 2, compared to the face at rest, with an average intra- and inter-rater distance of 1.32 mm and 1.62 mm [7]. Additionally, the direct comparison between the 3dMD and RealSense landmark placement resulted in less statistical differences compared to the face at rest (Table 3), which also indicated a more reliable landmark placement. When comparing the landmark placement between the healthy and palsy side of the face of the RealSense image, 94% of the bilateral landmarks did not find a significant difference between the healthy and palsy side of the face.

The increase in reliability for the RealSense landmarks placement during the voluntary movements was mainly caused by the subalare (sbal'), with an average intra- and inter-rater Euclidean distance of 1.22 mm and 1.46 mm compared to a respective distance of 2.13 mm and 2.51 mm during the baseline study [7]. The difficulty of the sbal' placement during the face at rest was caused by the lack of depth data around the noise region [7]. During landmark placement of the voluntary movements, it was noted the depth data was present more often around the subalare region due to a slight upward rotation of the face of the patients, making landmark placement more reliable.

Objective 3: Intra- and inter-rater reliability of anthropometric measurements

Reliability of 3dMD anthropometric measurements

The third objective determined the intra- and inter-rater reliability (precision) of the anthropometric measurements based on the voluntary movements. The baseline study found an excellent intra- and inter-rater ICC of 0.95 and 0.93, respectively for the anthropometric measurements based on the 3dMD image [7]. All measurements based

on the voluntary movements had either a similar or better ICC including the lower and upper boundary of the confidence interval (Table 4), with an average intra- and inter-rater reliability of 0.96 and 0.94, respectively. Therefore, the reliability of the voluntary movement measurements was considered as excellent. The absolute difference of intra-raters ranged between 0.7 mm and 0.8 mm for the voluntary movements (Table 4), which was the same as the 0.7 mm absolute difference at rest [7]. The inter-rater absolute difference ranged between 0.7 mm and 1.0 mm compared to the 0.9 mm with the face at rest.

Apart from the baseline study, no other studies were found analysing 3D anthropometric measurements based on patients with a PFP [7]. Instead, previous work focused on the analysis of motion or was based on 2D landmarks [27–30]. Hence an indirect comparison was made with studies based on healthy subjects [31–35]. The average ICC for healthy subjects ranged from 0.83 to 1.00 (intra-rater) and from 0.70 to 0.98 (inter-rater) [33–35]. The average absolute differences ranged from 0.80 mm to 0.99 mm for anthropometric measurements for healthy subjects [31,32,35]. Since both the ICC and the absolute differences of the voluntary movements were in a similar range as the baseline study with the face at rest and were in the highest range for the healthy subjects, the 3dMD anthropometric measurements were considered to be reliable results to act as the reference values.

Reliability of RealSense D415 anthropometric measurements

The reliability of the RealSense voluntary movement measurements increased compared to the face at rest, as seen with the 3dMD measurements. The average intra- and inter-rater ICC of the voluntary movements was 0.93 and 0.90, respectively (Table 5) and therefore excellent, compared to a good ICC of 0.83 and 0.80 for the face at rest, respectively [7]. These results were confirmed by an overall lower absolute difference between observers. The intra-rater absolute difference improved from 1.2 mm to 1.0 mm between the face at rest and the voluntary movements, respectively (Table 5). The inter-rater absolute difference saw a similar improvement from 1.5 mm to 1.2 mm, respectively.

The reliability of the RealSense measurements was within the range of the average ICC for healthy subjects ranging from 0.83 to 1.00 (intra-rater) and from 0.70 to 0.98 (inter-rater) [33–35]. However, a higher ICC was found during the 3dMD measurements and the absolute difference of the intra-rater measurements for the RealSense was 1.2 mm, which fell outside of the reported range for healthy subject of 0.8 mm to 1.0 mm [31,32,35]. Therefore, a lower reliability of the RealSense measurements should still be expected compared to the 3dMD measurements.

Objective 4: Intra- and inter-rater agreement of anthropometric measurements
The fourth objective assessed the agreement (accuracy) of anthropometric measurements using the 3dMD measurements as the gold standard. The Bland-Altman analysis was used to determine the differences between the 3dMD measurements and the RealSense measurements, which would be zero in the case of perfect agreement. In the baseline study based on the face at rest, an average underestimation of -0.90 mm and -0.89 mm was found for the intra- and inter-rater RealSense measurements compared to the 3dMD measurements [7]. During the voluntary movements, this underestimation increased to an average of -1.21 mm and -1.16 mm for the intra- and inter-rater measurements, respectively (Table 6).

A difference between the RealSense and 3dMD measurements was expected even in case landmarks were placed on exactly the same location, due to the RealSense depth accuracy of 0.98 mm for the voluntary movements (Table 1). However, the depth accuracy did not seem to be the main cause of the difference in agreement, since the voluntary movements with the lowest agreement, the forehead wrinkle and gentle eye closure, were found to have the highest depth accuracy (Table 1 & 6). The lowest agreement was found for measurements involving ch' on the healthy side of the face during the forehead wrinkle and gentle eye closure. These two voluntary movements should not affect the mouth region significantly for the healthy side of the face compared to the face at rest. These results indicate the lower agreement was not caused by the voluntary movements or the PFP, but by the overall difficulty to determine ch'. The average age of the population in this study was relatively high, increasing the amount of skin folds around ch'. This made it harder to identify the location where the upper and lower vermillion border met and caused a higher variation in the measurements between the 3dMD and RealSense measurements.

The limit of agreement (LoA), showed the expected percentage difference between the RealSense and 3dMD measurement in 95% of the cases. A major increase in LoA percentage was seen during the ps' – pi' measurement with an intra-rater LoA percentage of 118.7% for the gentle eye closure compared to 20.5% for the face at rest [7]. However, the agreement of ps' – pi' was higher during the gentle eye closure indicating the increase of the LoA percentage was caused by the relatively short distance measured during the eye closure. When excluding ps' – pi' from the results, the LoA percentage changed to 14.4% and 14.7% for the intra- and inter-rater measurements, respectively (Table 6). This is in the same range as the face at rest, with the overall LoA percentage of 12.4% and 15.0%, respectively.

Combining the results of the Bland-Altman analysis and the LoA percentage, the overall agreement between the RealSense and 3dMD measurements remained relatively consistent between the voluntary movements and the face at rest. Therefore, 95% of the RealSense measurements are expected to be at least within -5.0 mm and 3.0 mm of the 3dMD measurements during the voluntary movements. This would be a 14% difference compared to the 3dDM measurements. It is clear from these results that submillimetre accuracy should not be expected when using the RealSense to perform anthropometric measurement in a clinical setting. However, the clinical application will determine whether the agreement of the RealSense measurements is within reasonable limits and sufficient for the required task.

Future research

To our knowledge this is the first study to use the RealSense D415 camera to assess the reliability and agreement of anthropometric measurements during the voluntary movements in patients with a PFP. The research extended previous research based on the face at rest and followed a similar study design [7]. Therefore, the scope of this research was mainly determined by the baseline study. This leaves multiple areas of research to be explored.

First of all, the depth accuracy of the RealSense was determined based on the entire surface of the face. During this study, the depth accuracy remained stable during the voluntary movements and therefore differences found in the reliability and agreement of the anthropometric measurements were most likely not caused by differences in the underlying 3D depth data. However, the face could be divided in multiple regions to determine the effect on the depth accuracy off specific regions. The main focus of the current study was based on the analysis of the manually placed 3D landmarks and their derived anthropometric measurements. However, the use of the complete 3D and 4D data, as captured by the RealSense, could be of additional value in the assessment of a PFP [2–4,8]. Since this current study has shown a relatively consistent agreement and reliability of the anthropometric measurements, the inclusion of more 3D and 4D data could be a viable option in future research.

Due to the (partial) eye closure during the voluntary movements the landmark placement of pi', ps', and scc' became more challenging. The overall effect on the reliability and agreement of the derived anthropometric measurements from these landmarks was minimized since the observers were measuring the same distance (e.g., zero during eye closure). However, in order to improve the reliability of the landmark placement the location of pi', ps', and scc' could be based on the centre location of the endocanthion and exocanthion, as the endocanthion and exocanthion are less likely to be blocked during the eye closure.

In contrast, ch' had a more significant impact on the reliability and agreement of the anthropometric measurements. The data indicate this was not caused by the voluntary movements, but the overall difficulty of ch' placement due to the presence of skin folds and shaded areas mainly present on the RealSense image. However, the reliability of the ch' placement could potentially be increased by using the 4D data of the RealSense recording by tracking the landmarks over time and compare changes between frames.

The current study analysed the voluntary movements of the SFGS, where other grading systems use (slightly) different poses to determine the degree of the PFP [2–4]. Additionally, a selection of 14 landmarks and anthropometric measurements were analysed on a total of 30 patients with a PFP. Therefore, it might be desirable to expand this research to include more facial poses, landmarks, and patients.

Finally, this study can act as a foundation for the implementation of clinical measurements or the automatic assessment of a PFP using the RealSense D415 [2–4,8]. Due to the low-cost and portability of the RealSense, these measurements could be implemented in an eHealth environment or in circumstances where a professional 4D camera would be too bulky or expensive. The requirements of the clinical implementation will determine whether the reliability and agreement of the anthropometric measurements are sufficient for the specific clinical implementation.

CONCLUSION

This study has assessed the reliability (precision) and agreement (accuracy) of anthropometric measurements of the five voluntary movements of the SFGS, based on manually placed 3D landmarks on patients with a PFP using the RealSense D415. First, it was found that the reliability of the landmark placement and anthropometric measurements were similar for the patients with a PFP performing the voluntary movements compared to healthy subjects at rest, when using high quality 3D images such as the 3dMD images [22–26,31–35]. Additionally, the reliability and agreement of the 3D landmark placement and anthropometric measurements during the voluntary movements on the RealSense images were relatively consistent. Therefore, the RealSense D415 can be considered as a viable option to perform objective measurements in case a low-cost and portable camera is required, e.g. in an eHealth environment.

REFERENCES

1. McCaul, J. A. et al. Evidence based management of Bell's palsy. Br. J. Oral Maxillofac. Surg. 52, 387–391 (2014).
2. Samsudin, W. S. W. & Sundaraj, K. Image processing on facial paralysis for facial rehabilitation system: A review. Proc. - 2012 IEEE Int. Conf. Control Syst. Comput. Eng. ICCSCE 2012 259–263 (2012) doi:10.1109/ICCSCE.2012.6487152.
3. Revenaugh, P. C. et al. Use of Objective Metrics in Dynamic Facial Reanimation A Systematic Review. JAMA Facial Plast. Surg. 20, 501–508 (2018).
4. Niziol, R., Henry, F. P., Leckenby, J. I. & Grobbelaar, A. O. Is there an ideal outcome scoring system for facial reanimation surgery? A review of current methods and suggestions for future publications. J. Plast. Reconstr. Aesthetic Surg. 68, 447–456 (2015).
5. Thierens, L. A. M., De Roo, N. M. C., De Pauw, G. A. M. & Brusselaers, N. Assessment modalities of non-ionizing three-dimensional images for the quantification of facial morphology, symmetry, and appearance in cleft lip and palate: a systematic review. Int. J. Oral Maxillofac. Surg. 47, 1095–1105 (2018).
6. Caple, J. & Stephan, C. N. A standardized nomenclature for craniofacial and facial anthropometry. Int. J. Legal Med. (2015) doi:10.1007/s00414-015-1292-1.
7. ten Harkel, T. C. et al. Reliability and Agreement of 3D Anthropometric Measurements in Facial Palsy Patients Using a Low-Cost 4D Imaging System. IEEE Trans. Neural Syst. Rehabil. Eng. 28, 1817–1824 (2020).
8. Fattah, A. Y. et al. Facial Nerve Grading Instruments: Systematic Review of the Literature and Suggestion for Uniformity. Plast. Reconstr. Surg. 135, 569–579 (2015).
9. Ross, B. G., Fradet, G. & Nedzelski, J. M. Development of a sensitive clinical facial grading system. Otolaryngol. neck Surg. 114, 380–386 (1996).
10. Lübbers, H.-T., Medinger, L., Kruse, A., Grätz, K. W. & Matthews, F. Precision and accuracy of the 3dMD photogrammetric system in craniomaxillofacial application. J. Craniofac. Surg. 21, 763–767 (2010).
11. Maal, T. J. J. et al. Variation of the face in rest using 3D stereophotogrammetry. Int. J. Oral Maxillofac. Surg. 40, 1252–1257 (2011).
12. Dindaroğlu, F., Kutlu, P., Duran, G. S., Görgülü, S. & Aslan, E. Accuracy and reliability of 3D stereophotogrammetry: A comparison to direct anthropometry and 2D photogrammetry. Angle Orthod. 86, 487–494 (2016).
13. Rusu, R. B. & Cousins, S. 3D is here: Point Cloud Library (PCL). 2011 IEEE Int. Conf. Robot. Autom. 1–4 (2011) doi:10.1109/ICRA.2011.5980567.
14. Segal, A. V, Haehnel, D. & Thrun, S. Generalized-ICP. Proc. Robot. Sci. Syst. 2, 4 (2009).
15. Caple, J. & Stephan, C. N. A standardized nomenclature for craniofacial and facial anthropometry. Int. J. Legal Med. 130, 863–879 (2016).

16. Wilcoxon, F. Individual Comparisons by Ranking Methods. Biometrics Bull. 1, 80–83 (1945).
17. Koo, T. K. & Li, M. Y. A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. J. Chiropr. Med. 15, 155–63 (2016).
18. McGraw, K. O. & Wong, S. P. Forming Inferences about Some Intraclass Correlation Coefficients. Psychol. Methods 1, 30–46 (1996).
19. Bland, J. M. & Altman, D. G. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 327, 307–310 (1986).
20. ten Harkel, T. C. et al. Depth accuracy of the RealSense F200: Low-cost 4D facial imaging. Sci. Rep. 7, 16263 (2017).
21. Beurskens, C. H. G., Oosterhof, J. & Nijhuis-van der Sanden, M. W. G. Frequency and location of synkineses in patients with peripheral facial nerve paresis. Otol. Neurotol. 31, 671–5 (2010).
22. Plooij, J. M. et al. Evaluation of reproducibility and reliability of 3D soft tissue analysis using 3D stereophotogrammetry. Int. J. Oral Maxillofac. Surg. 38, 267–73 (2009).
23. Lin, H., Zhu, P., Lin, Y., Zheng, Y. & Xu, Y. Reliability and Reproducibility of Landmarks on Three-Dimensional Soft-Tissue Cephalometrics Using Different Placement Methods. Plast. Reconstr. Surg. 134, 102e-110e (2014).
24. Toma, A. M., Zhurov, A., Playle, R., Ong, E. & Richmond, S. Reproducibility of facial soft tissue landmarks on 3D laser-scanned facial images. Orthod. Craniofacial Res. 12, 33–42 (2009).
25. Fagertun, J. et al. 3D facial landmarks: Inter-operator variability of manual annotation. BMC Med. Imaging 14, 1–9 (2014).
26. Baysal, A., Sahan, A. O., Ozturk, M. A. & Uysal, T. Reproducibility and reliability of three-dimensional soft tissue landmark identification using three-dimensional stereophotogrammetry. Angle Orthod. 86, 1004–1009 (2016).
27. Neely, J. G., Wang, K. X., Shapland, C. A., Sehizadeh, A. & Wang, A. Computerized Objective Measurement of Facial Motion. Otol. Neurotol. 31, 1488–1492 (2010).
28. Mehta, R. P., Zhang, S. & Hadlock, T. A. Novel 3-D video for quantification of facial movement. Otolaryngol. - Head Neck Surg. 138, 468–472 (2008).
29. Hadlock, T. A. & Urban, L. S. Toward a universal, automated facial measurement tool in facial reanimation. Arch. Facial Plast. Surg. 14, 277–282 (2012).
30. Shujaat, S. et al. The clinical application of three-dimensional motion capture (4D): A novel approach to quantify the dynamics of facial animations. Int. J. Oral Maxillofac. Surg. 43, 907–916 (2014).
31. Wong, J. Y. et al. Validity and reliability of craniofacial anthropometric measurement of 3D digital photogrammetric images. Cleft Palate-Craniofacial J. 45, 232–239 (2008).
32. Ramieri, G. A. et al. Reconstruction of facial morphology from laser scanned data. Part I: Reliability of the technique. Dentomaxillofacial Radiol. 35, 158–164 (2006).

33. Ceinos, R., Tardivo, D., Bertrand, M. F. & Lupi-Pegurier, L. Inter- and Intra-Operator Reliability of Facial and Dental Measurements Using 3D-Stereophotogrammetry. J. Esthet. Restor. Dent. 28, 178–189 (2016).
34. Othman, S. A., Majawit, L. P., Hassan, W. N. W., Wey, M. C. & Razi, R. M. Anthropometric study of three-dimensional facial morphology in Malay adults. PLoS One 11, 1–15 (2016).
35. Fourie, Z., Damstra, J., Gerrits, P. O. & Ren, Y. Evaluation of anthropometric accuracy and reliability using different three-dimensional scanning systems. Forensic Sci. Int. 207, 127–134 (2011).

CHAPTER 4

AUTOMATIC GRADING OF PATIENTS WITH A UNILATERAL FACIAL PALSY BASED ON THE SUNNYBROOK FACIAL GRADING SYSTEM: A DEEP LEARNING STUDY BASED ON A CONVOLUTIONAL NEURAL NETWORK

Published as: Timen C. ten Harkel, Guido de Jong, Henri A.M. Marres, Koen J.A.O. Ingels, Caroline M. Speksnijder &Thomas J.J. Maal. Automatic grading of patients with a unilateral facial paralysis based on the Sunnybrook Facial Grading System – A deep learning study based on a convolutional neural network. American Journal of Otolaryngology 44, 103810 (2023).

DOI: 10.1016/j.amjoto.2023.103810

ABSTRACT

Purpose

In order to assess the severity and the progression of a unilateral peripheral facial palsy (PFP) the Sunnybrook Facial Grading System (SFGS) is a well-established grading system due to its clinical relevance, sensitivity, and robust measuring method. However, training is required in order to achieve a high inter-rater reliability. This study investigated the automated grading of patients with a PFP based on the SFGS using a convolutional neural network.

Materials & Methods

A total of 116 patients with a unilateral PFP and 9 healthy subjects were recorded performing the SFGS poses. A separate model was trained for each of the 13 elements of the SFGS and then used to calculate the SFGS subscores and composite score. The performance of the automated grading system was compared to three clinicians experienced in the grading of a PFP.

Results

The inter-rater reliability of the convolutional neural network was within the range of human observers, with an average intra-class correlation coefficient of 0.87 for the composite SFGS score, 0.45 for the resting symmetry subscore, 0.89 for the symmetry of voluntary movement subscore, and 0.77 for the synkinesis subscore.

Conclusion

This study showed the potential of the automated SFGS to be implemented in a clinical setting. The automated grading system adhered to the original SFGS, which makes the implementation and interpretation of the automated grading more straightforward. The automated system can be implemented in numerous settings such as online consults in an eHealth environment, since the model used 2D images captured from a video recording.

INTRODUCTION

The partial or complete loss of facial function associated with a unilateral peripheral facial palsy (PFP) can have a significant impact on the physical, social, and emotional quality of life, due to the potential inability to blink, to eat and drink, or to communicate both verbally and non-verbally [1–3]. The cause and severity of the initial PFP has a major impact on the expected recovery rate. For example, patients with a complete idiopathic PFP have an overall recovery rate of 50 to 60%, whilst patients with an incomplete idiopathic PFP have a recovery rate of 95 to 99% [4].

In order to assess the severity and the progression of the PFP, multiple grading systems exist, such as the House-Brackmann scale, Sunnybrook Facial Grading System (SFGS), and eFACE [5,6]. One of the recommended and well-established grading systems is the SFGS due to its clinical relevance, sensitivity, and robust measuring method [6]. The SFGS is a weighted grading system where the composite SFGS score ranges from 0 to 100 [7]. A score of 0 indicates a complete flaccid unilateral facial paralysis (without synkinesis) and a score of 100 indicates normal functioning of the mimic muscles. The SFGS assesses 13 individual elements and are grouped into three subcomponents; the resting symmetry (3 elements), symmetry of voluntary movement (5 elements), and synkinesis (5 elements). A complete breakdown of the SFGS is shown in Table 1.

Despite the clinical relevance and sensitive measurements of the SFGS, there are certain disadvantages using a subjective grading system. First of all, training is required in order to achieve a high reliability between observers [8]. Additionally, the grading of the PFP is most commonly performed during consultation of the patients, where an increase in grading frequency is not always possible due to time constraints in the clinic and also due to travel distances of patients. These limitations could be alleviated by the automation of grading of a PFP based on the SFGS. The automated system would remove the learning curve of the SFGS and make the SFGS more accessible for e.g., researchers, students, clinicians in training, or other untrained co-workers. This automated system could then potentially be used during online consults in an eHealth environment. Ideally, the automated grading system would be so user-friendly it could be used by the patient at home without any assistance. This would enable more frequent monitoring of the rehabilitation process of the patient without increasing the workload of clinicians.

Table 1. Overview of the Sunnybrook Facial Grading System (SFGS) assessing 13 individual elements during the resting symmetry (3 elements), symmetry of voluntary movement (5 elements) and synkinesis of the facial muscles (5 elements).

SFGS component	Score range (discrete values)	Score for healthy subjects
Resting symmetry (RS)*		
Eye	0 – 1	0
Cheek (naso-labial fold)	0 – 2	0
Mouth	0 – 1	0
Symmetry of Voluntary Movement (SVM)		
Forehead wrinkle	1 – 5	5
Gentle eye closure	1 – 5	5
Open mouth smile	1 – 5	5
Snarl	1 – 5	5
Lip pucker	1 – 5	5
Synkinesis (SK)		
Forehead wrinkle	0 – 4	0
Gentle eye closure	0 – 4	0
Open mouth smile	0 – 4	0
Snarl	0 – 4	0
Lip pucker	0 – 4	0
Subscore SFGS components		
RS subscore (sum RS x 5)	0 – 20	0
SVM subscore (sum SVM x 4)	20 – 100	100
SK subscore (sum SK)	0 – 20	0
Composite score		
SVM subscore - RS subscore - SK subscore	0 – 100	100

*Multiple answers in the SFGS can result in the same score for the individual elements [7].

Deep learning has shown great results in the automation of image based recognition and classification tasks [9–12]. A subtype of deep learning, the convolutional neural network (CNN), is particularly suitable for image based classification and is able to surpass the human-level performance in recognition and classification tasks [10,11]. Therefore, an automated SFGS based on a CNN has the potential to exceed the reliability compared to human observers. In order to achieve this accuracy, the CNN model is usually trained on a large amount of input data. This training process can take a long time and will sometimes require expensive hardware. However, once the training phase of model has been finished, the execution of the model generally can be performed within milliseconds on relatively affordable electronic devices such as smartphones, laptops, and desktops [9–12], which is ideal for the implementation in a clinical setting.

Deep learning has been applied for studies investigating the automation of the grading of a PFP [13–29]. However, these studies consisted of either small cohorts with less than 30 subjects, analysed only the composite score of the SFGS, or focused on different grading systems such as the House-Brackmann scale or eFace [13–30]. Since the composite SFGS score by itself does not differentiate which area of the face is affected by the PFP it is crucial all 13 individual components of the SFGS are scored during follow-up. By adhering to the original SFGS the resulting SFGS scores are easy to interpret for clinicians familiar with the SFGS. Additionally, all previous research about the clinical relevance and reliability of the SFGS would remain valid. This would make the automated grading system more straightforward to implement in daily clinical practice or in an eHealth environment.

Therefore, this prospective study investigated the automated grading of patients with a PFP based on the SFGS using a CNN. The long-term goal of the automated SFGS grading system would be to create a user-friendly system that can be used by the patient at home without any assistance, whilst ideally exceeding the inter-rater reliability of human observers. However, the scope of this study was first to determine the feasibility of an automated SFGS grading system based on a CNN. Therefore, the objective of this study was to determine the inter-rater reliability of the automated SFGS based on a CNN compared to human observers, experienced in the grading of the SFGS, for all 13 individual components of the SFGS. Additionally, the scoring key of the SFGS was used to determine the inter-rater reliability of the three subcomponents (resting symmetry, symmetry of voluntary movement, and synkinesis), and the composite SFGS score (Table 1).

MATERIALS & METHODS

Population

Patients seen during facial palsy consultation at the Department of Otorhinolaryngology of the Radboudumc were included in this study during the period of August 2018 and November 2020, independent of etiology of the unilateral PFP. Additionally, healthy subjects were allowed to participate in this study to act as reference measurements. The subjects were graded during patient consultation according to the SFGS by three clinicians experienced in SFGS grading. The team consisted of an otorhinolaryngologist, a plastic surgeon, and a physical therapist, all experienced for many years in diagnosis and treatment of a PFP. The observers were present in the same room and discussion between the observers was allowed, as was standard clinical practice during the consultation. Approval of this study was authorized by the Ethics Committee of the Radboudumc (2015-1829) and was conducted in compliance with the World Medical Association Declaration of Helsinki on medical research ethics. Each subject provided a written informed consent for the participation in this study and subjects shown in this study provided a written informed consent for the use of their images.

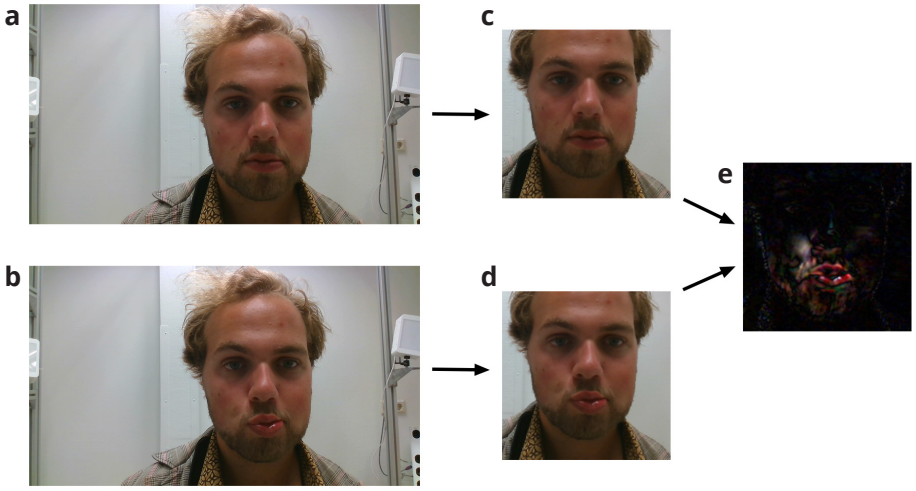

Figure 1. Pre-processing of the RealSense recordings to optimize the input for the convolutional neural network, using the Sunnybrook Facial Grading System (SFGS) pose "pucker" as an example. The starting frame (a) was selected at the initiation of the SFGS pose, whilst the maximum frame (b) was selected at the maximum exertion of the SFGS pose. The original images (a & b) were cropped to a 112 × 112 × 3 pixel image (c & d) based on manually placed landmarks on the on the left and right exocanthion (not shown). Image registration was applied between the cropped starting frame (c) and maximum frame (d) to correct for potential movement between the frames. Finally, the difference image (e) was calculated between the cropped starting frame (c) and maximum frame (d), resulting in a 112 × 112 × 3 pixel image.

Image acquisition

Image acquisition consisted of recording the six poses based on the SFGS, i.e., neutral, forehead wrinkle, gentle eye closure, open mouth smile, snarl, and lip pucker. Recordings were performed with the RealSense D415 (Intel, Santa Clara, USA), used for previous studies recording patients with a PFP [31,32]. The RealSense captured 30 frames per second at an approximate distance of 35 cm to the patient. The RealSense simultaneously captured a colour recording with a resolution of 1920 × 1080 pixels and a depth recording with a resolution of 1280 × 720 pixels. During this study only the 2D colour images were used as input for the CNN. All SFGS poses were captured in a single recording.

Pre-processing

Two frames were selected for each of the SFGS poses; the starting frame was at the initiation of the SFGS pose, whilst the maximum frame was selected at the maximum exertion of the SFGS pose (Figure 1). This resulted in 12 selected frames per subject. On each of the 12 selected frames, landmarks were placed on the left and right exocanthion, which were used for cropping the image to a 112 × 112 pixel colour image. The cropping centred the face of the subject and removed the majority of the background of the image. The 112 × 112 resolution was also required in order to make the image suitable as an

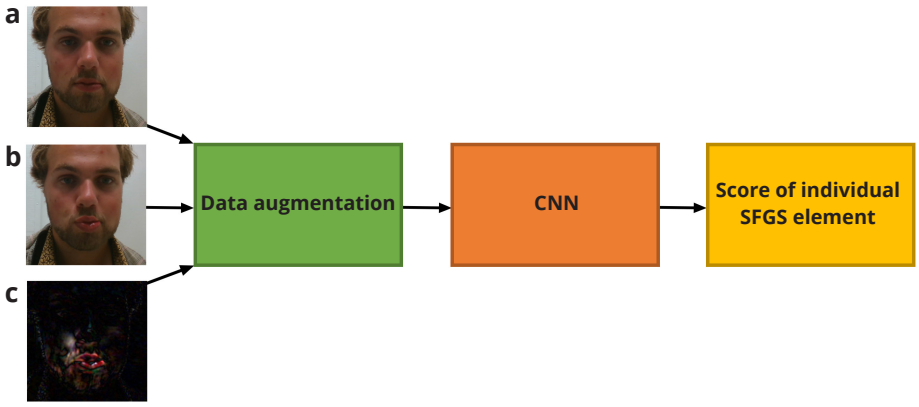

Figure 2. Simplified overview of the training of the convolutional neural network (CNN) with the Sunnybrook Facial Grading System (SFGS) pose "pucker" used as an example. The complete CNN model is shown in Appendix A. The input consisted of the cropped frame during the initiation of the SFGS pose (a), the frame during the maximum exertion of the SFGS pose (b), and the difference image between the starting and maximum frame (c). The difference image was only added during the dynamic SFGS poses during the symmetry of voluntary movement and synkinesis. Due to the relatively small cohort size data augmentation, early stopping, dropout, batch normalization, and Gaussian noise was used during training to prevent overfitting. The predicted scores of the individual elements of the SFGS were converted from a continuous scale to the respective nominal score of the SFGS as shown in Table 1. E.g., the final output score for the symmetry of voluntary movement ranged from discrete values from 1 to 5.

input for the CNN. Due to potential rotation between the start of the SFGS pose and the maximum exertion, image registration was applied between the starting and maximum frame using optical flow registration [33]. Finally, a third image was created, with a matching resolution of 112 × 112 × 3, by calculating the absolute difference between the start and maximum frame for each individual colour channel, creating a difference image between the starting and maximum frame (Figure 1). All images were normalized, resulting in pixels values ranging from 0 to 1 for each input image.

Architecture

The CNN architecture was based on CNN configuration D as described by Simonyan & Zisserman consisting of 16 weight layers, with 13 convolution layers and 3 fully connected layers [34]. Due to the relatively small cohort size multiple alterations were made to the architecture, resulting in the CNN as shown in Appendix A, with a simplified overview shown in Figure 2. The input consisted of the starting frame and the maximum frame, each with a size of 112 × 112 × 3 pixels. The difference image, created during the preprocessing step and consisting of 112 × 112 × 3 pixels, was added as a third input during the dynamic components of the SFGS, i.e., the symmetry of voluntary movement and synkinesis. The input layer was followed by three data augmentation layers; a

random horizontal flip, random zoom (range factor 0.8 to 1.2) and random rotation (range -20 to 20 degrees), which was only activated during training of the model. The data augmentation was followed by the CNN, with a kernel size reduced by a factor of four compared to CNN configuration D from Simonyan & Zisserman [34]. Additionally, a kernel and bias constraint with a maximum norm value of three was added and each maxpool layer was preceded by a batch normalization layer. The fully connected layers consisted of 1024 nodes. Dropout layers (p = 0.5) and batch normalization (momentum = 0.95) layers were added after each fully connected layer. A linear activation function was used for the output layer, followed by a Gaussian noise layer ($\sigma = 0.1$) for further regularization. Finally, the logcosh loss function was used in combination with the Adam optimizer.

Training

Each of the 13 elements of the SFGS were trained separately based on the output labels as determined by the three experienced observers. As the composite SFGS score is calculated from the 13 SFGS elements, the training and testing groups were kept consistent between the 13 SFGS elements. E.g., the trained CNN model of the symmetry of voluntary movement of the pucker was based on exactly the same training and testing group as the model of the synkinesis of the gentle eye closure. A stratified k-fold was applied during training, which divided the dataset into five folds, using 80% of the subjects for training during each fold. This meant the CNN model was trained and tested five times, where the testing data always consisted of completely new subjects during each fold (20% of the subjects per fold). The stratified k-fold was based on the composite SFGS score to promote a fair distribution of the subjects. Data augmentation was set to a random zoom factor ranging from 0.8 to 1.2 and a random rotation factor ranging between -20 to 20 degrees. Early stopping was used with a patience of 1500 epochs and a batch size of 32 was used [35]. A cyclic triangular learning rate was applied with a base learning rate of 1e-8, a max learning rate of 1e-3, and a step size of 4× (length of training dataset / batch size) [36].

Analysis

The performance of the CNN was determined by comparing the predicted SFGS scores of the models with the SFGS scores as graded by the experienced human observers during the patient consultation as described in the section Population. The CNN was trained for five different combinations of subjects, as determined by the stratified k-folds. The analysis as described below was repeated for each of the five folds. Due to the linear output of the CNN model, the predicted scores of the individual elements of the SFGS were converted from a continuous scale to the

respective nominal score of the SFGS as shown in Table 1. E.g., the final output score for the symmetry of voluntary movement ranged from discrete values from 1 to 5. Additionally, the predicted scores were capped at the minimum and maximum score of each individual SFGS. From these 13 individual scores the subscores of the 3 SFGS components and the composite SFGS score were calculated according to the scoring key of the SFGS (Table 1). Therefore, no separate CNN models were trained to calculate the subscores of the SFGS components and the composite SFGS score.

The individual scores of the resting symmetry, symmetry of voluntary movement, and the synkinesis were based on a nominal scale. In order to determine the agreement between the CNN model and the experienced human observers, confusion matrices were made. The confusion matrix visualized the performance of a classification model where the row represented the actual SFGS score, and the columns represented the predicted SFGS score. The cells of the matrix displayed the frequency for that particular combination. E.g., in case of perfect agreement all outcomes are on the diagonal of the confusion matrix, since the actual SFGS score and the predicted SFGS score are the same score. From the confusion matrices the quadratic weighted Cohen's Kappa was calculated to determine the inter-rater reliability between the predicted values of the model and the observers [37]. A Cohen's Kappa lower than 0.20 was considered as having no agreement, 0.21 to 0.39 a minimal agreement, 0.40 to 0.59 a weak agreement, 0.60 to 0.79 a moderate agreement, 0.80 to 0.90 a strong agreement, and 0.91 to 1.00 an almost perfect agreement [38].

Due to the continuous values of the SFGS components and the composite SFGS score, the inter-rater reliability of the total SFGS scores was expressed as the intra-class correlation coefficient (ICC, type 2,1) [39]. An ICC of <0.5 was considered as poor, 0.50 to 0.75 as fair, 0.75 to 0.90 as good, and 0.90 to 1.00 as excellent [40].

RESULTS

Population

A total of 116 patients with a PFP and 9 healthy subjects were included in this study during the period of August 2018 and November 2020. The patients with a PFP consisted of 49 men and 67 women, with an average age of 53 ± 16 years ranging from 18 to 88 years. The side of paralysis was equally distributed (50 / 50 % r / l). The 9 healthy subjects consisted of 3 men and 6 women, with an average age of 56 ± 17 years ranging from 27 to 77 years.

Table 2. Inter-rater reliability of the individual Sunnybrook Facial Grading System (SFGS) components between the convolutional neural network (CNN) model and the experienced observers.

SFGS component	Inter-rater reliability training data	Inter-rater reliability testing data
Resting symmetry		
Eye	0.32 (0.03 - 0.73)	0.37 (0.00 - 0.75)
Cheek (naso-labial fold)	0.22 (-0.22 - 0.61)	0.29 (0.17 - 0.46)
Mouth	0.41 (0.03 - 0.88)	0.47 (0.34 - 0.60)
Symmetry of Voluntary Movement		
Forehead wrinkle	0.84 (0.73 - 0.89)	0.84 (0.81 - 0.90)
Gentle eye closure	0.74 (0.53 - 0.91)	0.79 (0.66 - 0.92)
Open mouth smile	0.86 (0.83 - 0.90)	0.81 (0.76 - 0.84)
Snarl	0.70 (0.10 - 0.88)	0.65 (0.32 - 0.79)
Lip pucker	0.61 (0.35 - 0.80)	0.63 (0.47 - 0.86)
Synkinesis		
Forehead wrinkle	0.69 (0.54 - 0.91)	0.69 (0.56 - 0.84)
Gentle eye closure	0.64 (0.36 - 0.79)	0.56 (0.27 - 0.76)
Open mouth smile	0.37 (0.20 - 0.50)	0.54 (0.35 - 0.71)
Snarl	0.17 (-0.07 - 0.34)	0.36 (0.30 - 0.51)
Lip pucker	0.84 (0.80 - 0.89)	0.77 (0.71 - 0.90)

The mean inter-rater reliability is expressed as the quadratic weighted Cohen's Kappa and is shown for both the training and testing data. The values in between brackets show the range of Kappa values for the five k-folds.

Inter-rater reliability of the individual SFGS elements

Table 2 shows the inter-rater reliability between the predicted CNN scores and the experienced observers for each of the 13 elements of the SFGS (Table 1), expressed as the quadratic weighted Cohen's Kappa. The mean inter-rater reliability was determined for five different combinations of subjects, as determined by the stratified k-folds. The range of inter-rater reliability found for these k-folds are shown in between brackets in Table 2. The CNN model was first trained on 100 subjects and then tested on 25 subjects in order to determine the inter-rater reliability of the CNN. Both the inter-rater reliability for the training and testing data is shown in Table 2 to determine potential overfitting during the training process of the CNN model. The data from Table 2 indicates no overfitting occurred due to the relatively minor differences between the inter-rater reliability of the testing and training data. When looking at the test data for the resting symmetry elements a minimal agreement was found between the predicted CNN scores and the experienced observers. The elements of the symmetry of voluntary movement mostly showed a moderate to strong agreement, whilst the synkinesis elements ranged from a minimal to moderate agreement.

Table 3. Inter-rater reliability of the Sunnybrook Facial Grading System (SFGS) subscores and composite score between the convolutional neural network (CNN) model and the experienced observers.

SFGS component	Inter-rater reliability training data	Inter-rater reliability testing data
Resting symmetry subscore	0.39 (0.13 - 0.58)	0.45 (0.35 - 0.58)
Symmetry of voluntary movement subscore	0.90 (0.85 - 0.94)	0.89 (0.86 - 0.94)
Synkinesis subscore	0.75 (0.71 - 0.79)	0.77 (0.72 - 0.85)
Composite score	0.87 (0.79 - 0.91)	0.87 (0.79 - 0.93)

The mean inter-rater reliability is expressed as the intra-class correlation coefficient (ICC, type 2,1) and is shown for both the training and testing data. The values in between brackets show the range of ICC values for the five k-folds.

Inter-rater reliability of the SFGS subscores and composite score

The subscores of the resting symmetry, symmetry of voluntary movement, synkinesis and the composite SFGS score were calculated according to the scoring key of the SFGS (Table 1) and were derived from the individual SFGS components as determined in the previous section. The inter-rater reliability between the total scores of the CNN and the experienced observers is expressed as the ICC (type 2,1) and is shown in Table 3. The total score of the resting symmetry showed a poor agreement, whereas the symmetry of voluntary movement, synkinesis, and composite SFGS all showed a good agreement.

DISCUSSION

This study investigated the automated grading of patients with a PFP based on the SFGS using a CNN. A separate CNN model was trained for each of the 13 elements of the SFGS consisting of the resting symmetry (3 elements), symmetry of voluntary movement (5 elements) and synkinesis (5 elements). The training and testing data were kept consistent throughout the individual elements of the SFGS, in order to calculate the total scores of the SFGS using the associated scoring key (Table 1). By adhering to the original SFGS, the results found in this study can be compared to previous research about the clinical relevance and reliability of the SFGS. Additionally, the CNN model used two colour 2D frames as an input, which could potentially be captured by any available 2D camera such as a smartphone camera or a (laptop) webcam. This would make the implementation of the automated SFGS into daily clinical practice more straightforward. This would also allow for a user-friendly implementation of the automated SFGS grading system that could be used by the patient at home without any assistance. However, before implementing the automated SFGS in the clinic, the inter-rater reliability of the automated SFGS scores need to be compared to the expected inter-rater reliability between human observers.
Multiple studies investigated the inter-rater reliability between human observers based on the SFGS, but not all studies used the same statistical analysis or included all the

individual elements from the SFGS [6,8,29,41–44]. However, these studies did find a predominantly minimal to weak agreement for the individual components of the resting symmetry, a moderate agreement for the symmetry of voluntary movement, and a weak to moderate agreement for the synkinesis. Existing literature investigating the inter-rater reliability between human observers of the subcomponents of the SFGS, predominantly showed a fair agreement for the resting symmetry, where the voluntary movements and synkinesis showed a good agreement, and the composite SFGS score showed a good to excellent agreement [6,8,29,41–44]. The results shown in this current study indicate that the average inter-rater reliability of the CNN model falls within the expected ranges of human observers (Tables 2 & 3) and therefore performed similarly to human observers. This provides a first good indication the automated SFGS would be suitable to implement in a clinical setting.

After the general comparisons with existing reliability studies, a more direct comparison could be made with an inter-rater reliability study between human observers using the same methods as used in this current study [8]. In this particular study the learning curve of inexperienced human observers was assessed when grading 100 patients with a PFP based on the SFGS. In this section, we compare the inter-rater reliability of human observers, who evaluated 50 patients with a PFP, to the inter-rater reliability of the CNN model used in this current study. The largest differences were found for the resting symmetry where the CNN model had a lower quadratic Cohen's Kappa compared to the human observers for all individual elements. This was also reflected with a high range of the inter-rater reliability between folds for the CNN model (Table 2). There are multiple factors that could contribute to the lower inter-rater reliability. The resting symmetry is the most difficult component of the SFGS, and previous studies reported a wide range of inter-rater reliability, and the CNN models still falls within this range [6,8,29,41–44]. However, the CNN model was trained on a relatively small cohort of 100 subjects and tested on 25 subjects. Considering the difficulty of grading the resting symmetry, an increase in cohort size could benefit the inter-rater reliability of the CNN model [34]. In contrast, the inter-rater reliability of the symmetry of voluntary movement and synkinesis was on par or exceeded the human observers after grading 50 patients with a PFP [8]. For the symmetry of voluntary movement, the snarl showed the largest difference between the human observers and the CNN model, with a respective quadratic Cohen's Kappa of 0.77 and 0.65. One fold of the CNN model found a quadratic Cohen's Kappa of 0.32, lowering the overall agreement. This was most likely caused by a batch of difficult subjects to score in that particular fold and not due to the architecture of the CNN model. Especially since the CNN performed better on the forehead wrinkle with a quadratic Cohen's Kappa of 0.84 compared to 0.75 for the human observers. During the synkinesis the largest differences were found for the gentle eye closure and lip pucker. The gentle eye close

found a quadratic Cohen's Kappa of 0.73 vs. 0.56 and the lip pucker 0.67 vs. 0.77, for the human observers and CNN model, respectively. Therefore, the CNN model and human observers seem to be balanced in grading the synkinesis. The inter-rater reliability of the subcomponents and the composite score were all within an ICC range of 0.02, except for the symmetry of voluntary movement where the CNN outperformed the human observers with a respective ICC of 0.89 vs. 0.85. Overall, this comparison confirms that the CNN performs similar to human observers [8]. More specifically, the CNN reaches a comparable inter-rater reliability after inexperienced human observers have graded 50 patients with a PFP.

The inter-rater reliability of the automated grading system could potentially be further improved by changing the deep learning architecture. The subjects were recorded with the RealSense D415, which simultaneously captured 2D and 3D images [31]. The 3D depth data would be able to add additional details about changes in the facial structure during the training of the model. Alternatively, specific (3D) facial landmarks could be added to focus on a select number of regions [32]. The current study used the neutral frame and frame of maximum exertion as a training input, whereas a Long Short-Term Memory (LSTM) deep learning network could provide more temporal information during the training of the model [45]. Another alternative deep learning network would be the Vision Transformer (ViT) model, which is less dependent on the spatial dependency of the regions of interest [46]. However, ViT models generally require large databases for training.

In general, deep learning models improve their accuracy by increasing the size of the training dataset, independent of the specific chosen deep learning architecture [34]. This is also the case for the CNN model used in this study, where a larger database would show more variations of a PFP. However, the impact of the cohort size was reduced by applying a high dropout rate, data augmentation, batch normalization, early stopping, and noise layers during training of the model (Appendix A). This resulted in relatively minor differences between the inter-rater reliability of the training data and testing data (Table 2 & 3), which indicates overfitting was minimized in this study. Additionally, the robustness of the CNN architecture was tested by applying five stratified k-folds, thereby making efficient use of the cohort, and using all 125 subjects in the testing of the CNN model during the five folds. The CNN model performed well with the different sets of subjects when taking into consideration that certain parts of the SFGS are relatively difficult to grade for human observers as well [6,8,29,41–44]. A potential limitation to improve the inter-rater reliability of the CNN might be the inter-rater reliability of human observers. This study used the average of three experienced observers during clinical consultation as was clinical standard practice, allowing discussion between the observers,

which made the SFGS used in this study less biased towards a single observer. However, it could be valuable to re-evaluate discrepancies between the human observers and the CNN, especially when the CNN approaches or exceeds the inter-rater reliability of human observers. This could result in the CNN achieving a higher inter-rater reliability compared to experienced human observers for the grading of patients with a PFP using the SFGS.

CONCLUSION

This study investigated the automated grading of patients with a PFP based on the SFGS in a cohort of 125 subjects consisting of 116 patients and 9 healthy subjects. This automated grading system can make the SFGS more accessible for researchers, students, clinicians in training, or other untrained co-workers, by removing the learning curve associated with the SFGS [34]. The implemented CNN model adhered to the original SFGS, which makes the implementation and interpretation of the automated grading more straightforward in a clinical setting. Additionally, the automated SFGS can be implemented in a wide variety of settings such as online consults in an eHealth environment, since the CNN is based on 2D images captured from a video recording. This would allow image capture devices such as smartphones or laptop webcams to be used as an input for the CNN model. The inter-rater reliability of the CNN found in this study was within the expected ranges of human observers [6,8,29,41–44]. More specifically, the CNN achieved a similar inter-rater reliability as human observers who graded 50 patients with a PFP [8]. However, the inter-rater reliability of the automated SFGS can potentially exceed the reliability of human observers by increasing the size of the cohort used to train the CNN model [34]. Therefore, this study showed the potential of the automated SFGS based on the CNN as a first step towards a user-friendly automated grading system that can be used by the patient at home.

REFERENCES

1. Kleiss, I. J., Hohman, M. H., Susarla, S. M., Marres, H. A. M. & Hadlock, T. A. Health-related quality of life in 794 patients with a peripheral facial palsy using the FaCE Scale: A retrospective cohort study. Clin. Otolaryngol. 40, 651–656 (2015).
2. Ho, A. L. et al. Measuring quality of life and patient satisfaction in facial paralysis patients: a systematic review of patient-reported outcome measures. Plast. Reconstr. Surg. 130, 91–9 (2012).
3. Coulson, S. E., O'dwyer, N. J., Adams, R. D. & Croxson, G. R. Expression of emotion and quality of life after facial nerve paralysis. Otol. Neurotol. 25, 1014–1019 (2004).
4. Peitersen, E. Bell's Palsy: The Spontaneous Course of 2,500 Peripheral Facial Nerve Palsies of Different Etiologies. Acta Otolaryngol. Suppl. 4–30 (2002) doi:10.1080/000164802760370736.
5. Samsudin, W. S. W. & Sundaraj, K. Evaluation and Grading Systems of Facial Paralysis for Facial Rehabilitation. J. Phys. Ther. Sci. 25, 515–519 (2013).
6. Fattah, A. Y. et al. Facial Nerve Grading Instruments: Systematic Review of the Literature and Suggestion for Uniformity. Plast. Reconstr. Surg. 135, 569–579 (2015).
7. Ross, B. G., Fradet, G. & Nedzelski, J. M. Development of a sensitive clinical facial grading system. Otolaryngol. - Head Neck Surg. 114, 380–386 (1996).
8. van Veen, M. M., Bruins, T. E., Artan, M., Werker, P. M. N. & Dijkstra, P. U. Learning curve using the Sunnybrook Facial Grading System in assessing facial palsy: An observational study in 100 patients. Clin. Otolaryngol. 45, 823–826 (2020).
9. Lecun, Y., Bengio, Y. & Hinton, G. Deep learning. Nature 521, 436–444 (2015).
10. Su, Z. et al. Deep learning-based facial image analysis in medical research: a systematic review protocol. BMJ Open 11, e047549 (2021).
11. Liu, Q. et al. A review of image recognition with deep convolutional neural network. in International onference on intelligent computing 69–80 (Springer, 2017).
12. Shen, D., Wu, G. & Suk, H.-I. Deep learning in medical image analysis. Annu. Rev. Biomed. Eng. 19, 221–248 (2017).
13. Bur, A. M., Shew, M. & New, J. Artificial Intelligence for the Otolaryngologist: A State of the Art Review. Otolaryngol. - Head Neck Surg. (United States) 160, 603–611 (2019).
14. Guarin, D. L. et al. Toward an Automatic System for Computer-Aided Assessment in Facial Palsy. Facial Plast. Surg. aesthetic Med. 22, 42–49 (2020).
15. Hsu, G. S. J. & Chang, M. H. Deep Hybrid Network for Automatic Quantitative Analysis of Facial Paralysis. Proc. AVSS 2018 - 2018 15th IEEE Int. Conf. Adv. Video Signal-Based Surveill. 1–7 (2019) doi:10.1109/AVSS.2018.8639156.
16. Mothes, O. et al. Automated objective and marker-free facial grading using photographs of patients with facial palsy. Eur. Arch. Oto-Rhino-Laryngology (2019) doi:10.1007/s00405-019-05647-7.

17. Zhuang, Y. et al. F-DIT-V: An Automated Video Classification Tool for Facial Weakness Detection. 2019 IEEE EMBS Int. Conf. Biomed. Heal. Informatics 1–4 (2019) doi:10.1109/bhi.2019.8834563.
18. Guarin, D. L., Dusseldorp, J., Hadlock, T. A. & Jowett, N. A Machine Learning Approach for Automated Facial Measurements in Facial Palsy. JAMA Facial Plast. Surg. 20, 335–337 (2018).
19. Guo, Z. et al. An unobtrusive computerized assessment framework for unilateral peripheral facial paralysis. IEEE J. Biomed. Heal. Informatics 22, 835–841 (2018).
20. Hsu, G. S. J., Huang, W. F. & Kang, J. H. Hierarchical network for facial palsy detection. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Work. 2018-June, 693–699 (2018).
21. Jiang, Z., Dai, W., Wang, W. & Wang, W. A Cloud-Based Training and Evaluation System for Facial Paralysis Rehabilitation. Proc. - IEEE 16th Int. Conf. Ind. Informatics, INDIN 2018 701–706 (2018) doi:10.1109/INDIN.2018.8471934.
22. Sajid, M. et al. Automatic grading of palsy using asymmetrical facial features: A study complemented by new solutions. Symmetry (Basel). 10, (2018).
23. Song, A., Wu, Z., Ding, X., Hu, Q. & Di, X. Neurologist Standard Classification of Facial Nerve Paralysis with Deep Neural Networks. Futur. Internet 10, 111 (2018).
24. Guo, Z. et al. Deep assessment process: Objective assessment process for unilateral peripheral facial paralysis via deep convolutional neural network. Proc. - Int. Symp. Biomed. Imaging 135–138 (2017) doi:10.1109/ISBI.2017.7950486.
25. Wang, T. et al. Automatic evaluation of the degree of facial nerve paralysis. Multimed. Tools Appl. 75, 11893–11908 (2016).
26. Kim, H. S., Kim, S. Y., Kim, Y. H. & Park, K. S. A smartphone-based automatic diagnosis system for facial nerve palsy. Sensors (Switzerland) 15, 26756–26768 (2015).
27. Azoulay, O. et al. Mobile Application for Diagnosis of Facial Palsy. in Proc. 2nd Int. Conf. Mobile Inf. Technol. Med (2014).
28. Wang, T., Dong, J., Sun, X., Zhang, S. & Wang, S. Automatic recognition of facial movement for paralyzed face. Biomed. Mater. Eng. 24, 2751–2760 (2014).
29. Tan, J. R., Coulson, S. & Keep, M. Face-to-Face Versus Video Assessment of Facial Paralysis: Implications for Telemedicine. J. Med. Internet Res. 21, e11109–e11109 (2019).
30. Jirawatnotai, S., Jomkoh, P., Voravitvet, T. Y., Tirakotai, W. & Somboonsap, N. Computerized Sunnybrook facial grading scale (SBface) application for facial paralysis evaluation. Arch. Plast. Surg. 48, 269–277 (2021).
31. ten Harkel, T. C. et al. Depth accuracy of the RealSense F200: Low-cost 4D facial imaging. Sci. Rep. 7, 16263 (2017).
32. ten Harkel, T. C. et al. Reliability and Agreement of 3D Anthropometric Measurements in Facial Palsy Patients Using a Low-Cost 4D Imaging System. IEEE Trans. Neural Syst.

Rehabil. Eng. 28, 1817–1824 (2020).
33. Van Der Walt, S. et al. Scikit-image: Image processing in Python. PeerJ 2014, e453 (2014).
34. Simonyan, K. & Zisserman, A. Very deep convolutional networks for large-scale image recognition. 3rd Int. Conf. Learn. Represent. ICLR 2015 - Conf. Track Proc. 1–14 (2015).
35. Prechelt, L. Early stopping - But when? in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) (eds. Montavon, G., Orr, G. B. & Müller, K.-R.) vol. 7700 LECTU 53–67 (Springer Berlin Heidelberg, Berlin, Heidelberg, 2012).
36. Smith, L. N. Cyclical learning rates for training neural networks. Proc. - 2017 IEEE Winter Conf. Appl. Comput. Vision, WACV 2017 464–472 (2017) doi:10.1109/WACV.2017.58.
37. Cohen, J. Weighted kappa: nominal scale agreement provision for scaled disagreement or partial credit. Psychol. Bull. 70, 213 (1968).
38. McHugh, M. L. Interrater reliability: the kappa statistic. Biochem. medica 22, 276–282 (2012).
39. Shrout, P. E. & Fleiss, J. L. Intraclass correlations: uses in assessing rater reliability. Psychol. Bull. 86, 420 (1979).
40. Koo, T. K. & Li, M. Y. A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. J. Chiropr. Med. 15, 155–63 (2016).
41. Volk, G. F. et al. Reliability of grading of facial palsy using a video tutorial with synchronous video recording. Laryngoscope (2018) doi:10.1002/lary.27739.
42. Gaudin, R. A. et al. Emerging vs time-tested methods of facial grading among patients with facial paralysis. JAMA Facial Plast. Surg. 18, 251–257 (2016).
43. Neely, J. G., Cherian, N. G., Dickerson, C. B. & Nedzelski, J. M. Sunnybrook facial grading system: reliability and criteria for grading. Laryngoscope 120, 1038–1045 (2010).
44. Coulson, S. E., Croxson, G. R., Adams, R. D. & O'Dwyer, N. J. Reliability of the "Sydney," "Sunnybrook," and "House Brackmann" facial grading systems to assess voluntary movement and synkinesis after facial nerve paralysis. Otolaryngol. - Head Neck Surg. 132, 543–549 (2005).
45. Hochreiter, S. & Schmidhuber, J. Long short-term memory. Neural Comput. 9, 1735–1780 (1997).
46. Vaswani, A. et al. Attention is all you need. Adv. Neural Inf. Process. Syst. 30, (2017).

APPENDIX A

Table 1. Full overview of the model architecture used for the convolutional neural network (CNN).

Layer	Type	Shape
0	InputLayer	[(None, 112, 112, 3)]
1	InputLayer	[(None, 112, 112, 3)]
2	InputLayer	[(None, 112, 112, 3)]
3	Concatenate	(None, 112, 112, 9)
4	RandomFlip	(None, 112, 112, 9)
5	RandomRotation	(None, 112, 112, 9)
6	RandomZoom	(None, 112, 112, 9)
7	Conv2D	(None, 112, 112, 16)
8	Conv2D	(None, 112, 112, 16)
9	BatchNormalization	(None, 112, 112, 16)
10	MaxPooling2D	(None, 56, 56, 16)
11	Conv2D	(None, 56, 56, 32)
12	Conv2D	(None, 56, 56, 32)
13	BatchNormalization	(None, 56, 56, 32)
14	MaxPooling2D	(None, 28, 28, 32)
15	Conv2D	(None, 28, 28, 64)
16	Conv2D	(None, 28, 28, 64)
17	Conv2D	(None, 28, 28, 64)
18	BatchNormalization	(None, 28, 28, 64)
19	MaxPooling2D	(None, 14, 14, 64)
20	Conv2D	(None, 14, 14, 128)
21	Conv2D	(None, 14, 14, 128)
22	Conv2D	(None, 14, 14, 128)
23	BatchNormalization	(None, 14, 14, 128)
24	MaxPooling2D	(None, 7, 7, 128)
25	Conv2D	(None, 7, 7, 128)
26	Conv2D	(None, 7, 7, 128)
27	Conv2D	(None, 7, 7, 128)
28	BatchNormalization	(None, 7, 7, 128)
29	MaxPooling2D	(None, 3, 3, 128)
30	Flatten	(None, 1152)
31	Dense	(None, 1024)
32	BatchNormalization	(None, 1024)
33	Dropout	(None, 1024)
34	Dense	(None, 1024)
35	BatchNormalization	(None, 1024)
36	Dropout	(None, 1024)
37	Dense	(None, 1)
38	GaussianNoise	(None, 1)

CHAPTER 5

OPTIMIZATION OF THE AUTOMATED SUNNYBROOK FACIAL GRADING SYSTEM: IMPROVING THE RELIABILITY OF A DEEP LEARNING NETWORK WITH FACIAL LANDMARKS

Published as: Timen C. ten Harkel, Freek Bielevelt, Henri A.M. Marres, Koen J.A.O. Ingels, Thomas J.J. Maal & Caroline M. Speksnijder. Optimization of the automated Sunnybrook Facial Grading System – Improving the reliability of a deep learning network with facial landmarks. European Annals of Otorhinolaryngology, Head and Neck Diseases 142, 5–10 (2025).

DOI: 10.1016/j.anorl.2024.07.005

ABSTRACT

Objective

The Sunnybrook Facial Grading System (SFGS) is a well-established grading system to assess the severity and progression of a unilateral facial palsy (PFP). The automation of the SFGS makes the SFGS more accessible for researchers, students, clinicians in training, or other untrained co-workers and could be implemented in an eHealth environment. This study investigated the impact on the reliability of the automated SFGS by adding a facial landmark layer in a previously developed convolutional neural network (CNN).

Materials & Methods

An existing dataset of 116 patients with a unilateral PFP and 9 healthy subjects performing the SFGS poses was used to train a CNN with a newly added facial landmark layer. A separate model was trained for each of the 13 elements of the SFGS and then used to calculate the SFGS subscores and composite score. The intra-class coefficient of the automated grading system was calculated based on three clinicians experienced in the grading of a PFP.

Results

The inter-rater reliability of the CNN with the additional facial landmarks increased in performance for all composite scores compared to the previous model. The intra-class coefficient for the composite SFGS score increased from 0.87 to 0.91, the resting symmetry subscore increased from 0.45 to 0.62, the symmetry of voluntary movement subscore increased from 0.89 to 0.92, and the synkinesis subscore increased from 0.75 to 0.78.

Conclusion

The integration of a facial landmark layer into the CNN showed a clear improvement in the reliability of the automated SFGS, reaching a performance level comparable to human observers. These results were attained without increasing the dataset underscoring the impact of incorporating facial landmarks into a CNN. These findings indicate that the automated SFGS with facial landmarks is a reliable tool for assessing patients with a PFP and is applicable in an eHealth environment.

INTRODUCTION

There are several diagnostic tools available to determine the severity of a unilateral peripheral facial palsy (PFP), such as the House-Brackmann scale, Sunnybrook Facial Grading System (SFGS), and eFACE [1–3]. Ideally, the diagnostic tool has a high reliability and validity with a low learning curve and is fast to use in a clinical setting. One of these diagnostic tools, the SFGS, has been shown to be a robust, sensitive, and clinically relevant method to grade and assess a PFP [2]. The SFGS comprises of 13 elements that evaluate different aspects of a PFP across the following six poses; neutral, forehead wrinkle, gentle eye closure, open mouth smile, snarl, and lip pucker. These elements are grouped into three subcomponents: the resting symmetry (3 elements), symmetry of voluntary movement (5 elements), and synkinesis (5 elements). An overview of the SFGS is shown in Table 1. Each of these elements can evolve over time, either in conjunction with or independently from one another. For this reason, the SFGS is a robust method to monitor the development of a PFP over time.

Although the SFGS is a widely used diagnostic tool, there is a learning curve associated with the SFGS and there might be limited time within or outside the clinic to assess all the elements [4]. Hence, an automated SFGS was developed to make the SFGS more accessible for researchers, students, clinicians in training, or other untrained co-workers [5]. Our long-term goal is to develop a user-friendly system that can be used by the patient at home without any assistance, whilst ideally exceeding the inter-rater reliability of human observers. Therefore, the automated SFGS was designed to be relatively inexpensive, portable, non-invasive, and fast, to make the automated system a low barrier of entry in clinical practice. In addition, the automated system generated the same output as the manual SFGS to keep the clinical relevance, validation, and experience gained over the years with the SFGS. This was done by implementing a type of deep learning network, a convolutional neural network (CNN), which has been widely used for image-based recognition and image classification tasks [6–9]. The CNN analysed multiple 2D colour images of patients with a PFP and produced scores for all the 13 individual elements of the SFGS [5]. The analysed images consisted of the face at rest just before the initiation of the SFGS pose, the moment of maximum exertion, and a difference image which was the absolute difference between the two previously selected images. The three subscores and composite SFGS score were calculated from the 13 scores generated by the CNN models, replicating the process of the manual SFGS. The reliability of the automated SFGS was determined by comparing the automated score with the score of three human observers experienced in the grading of the SFGS. The automated SFGS managed to achieve an inter-rater reliability within the expected ranges of human observers albeit at the lower end of the reported range [2,4,5,10–15]. Since the CNN is based on 2D

Table 1. Overview of the Sunnybrook Facial Grading System (SFGS) assessing 13 individual elements during the resting symmetry (3 elements), symmetry of voluntary movement (5 elements) and synkinesis of the facial muscles (5 elements).

SFGS component	Score range (discrete values)	Score for healthy subjects
Resting symmetry (RS)*		
Eye	0 – 1	0
Cheek (naso-labial fold)	0 – 2	0
Mouth	0 – 1	0
Symmetry of Voluntary Movement (SVM)		
Forehead wrinkle	1 – 5	5
Gentle eye closure	1 – 5	5
Open mouth smile	1 – 5	5
Snarl	1 – 5	5
Lip pucker	1 – 5	5
Synkinesis (SK)		
Forehead wrinkle	0 – 4	0
Gentle eye closure	0 – 4	0
Open mouth smile	0 – 4	0
Snarl	0 – 4	0
Lip pucker	0 – 4	0
Subscore SFGS components		
RS subscore (sum RS x 5)	0 – 20	0
SVM subscore (sum SVM x 4)	20 – 100	100
SK subscore (sum SK)	0 – 20	0
Composite score		
SVM subscore – RS subscore – SK subscore	0 – 100	100

*Multiple answers in the SFGS can result in the same score for the individual elements [3].

images captured from a video recording, the automated SFGS can be implemented in a diverse range of environments, such as online consultations in an eHealth environment. Additionally, by adhering to the SFGS the implementation and interpretation of the automated SFGS is more straightforward in a clinical setting.

One of the limitations of our previous study was the cohort size consisting of 125 subjects [5]. However, in certain (clinical) settings there may be no option to increase the cohort size. Therefore, this study investigated the impact on the reliability of the automated SFGS by adding a facial landmark layer to the CNN, instead of significantly increasing the cohort size. In the original CNN model, a difference image was added as an input to highlight areas of movement. During the absence of movement this layer

would deactivate. The additional facial landmark layer could potentially increase the reliability of the model by indicating regions of interest for the CNN even in the absence of motion. In order to determine the effect of adding a facial landmark layer to the CNN, the image dataset, processing, and training were kept as consistent as possible to our previous work [5].

MATERIALS & METHODS

Dataset

To train the CNN, this study utilized the identical dataset as described in our previous work [5]. The dataset consisted of 116 patients with a unilateral PFP and 9 healthy subjects performing the six SFGS poses, recorded with the RealSense D415 (Intel, Santa Clara, USA) during the period of August 2018 and November 2020. The patients with a PFP were seen during facial palsy consultation at the Department of Otorhinolaryngology of the Radboudumc and were included independent of etiology of the PFP. The included patients consisted of 49 men and 67 women, with an average age of 53 ± 16 years ranging from 18 to 88 years. The composite SFGS score for the patients with a PFP ranged from 0 to 91 with an average score of 41. The side of the PFP was equally distributed (50/50% r/l). The 9 healthy subjects consisted of 3 men and 6 women, with an average age of 56 ± 17 years ranging from 27 to 77 years.

The dataset contained the SFGS scores as graded by three clinicians (an otorhinolaryngologist, a plastic surgeon, and a physiotherapist) all with multiple years of experience applying the SFGS and in the diagnosis and treatment of patients with PFP. The clinicians were present within the same room and discussion between the observers was allowed, as was standard clinical practice during the consultation.

Approval of this study was authorized by the Ethics Committee of the Radboudumc (2015-1829). This study was conducted in compliance with the World Medical Association Declaration of Helsinki on medical research ethics. Each subject provided a written informed consent for the participation in this study and subjects shown in this study provided a written informed consent for the use of their images.

Landmark placement

Thirteen facial landmarks were manually placed by observer FB on the original 2D colour images (1920 × 1080 pixels) using the definitions of Caple & Stephan: superciliare (bilateral), palpebrale superius (bilateral), palpebrale inferius (bilateral), subalare (bilateral), subnasale, cheilion (bilateral), labiale superius, and labiale inferius (Figure 1a & b) [16].

All landmarks were placed on both the start frame, i.e., neutral position of the face, and frame of maximum exertion for each SFGS pose and subject. From these landmark positions, a colour image of 1920 × 1080 pixels was generated with the landmark coordinates set to a value of 255 for each colour channel and all other pixels set to a value of 0.

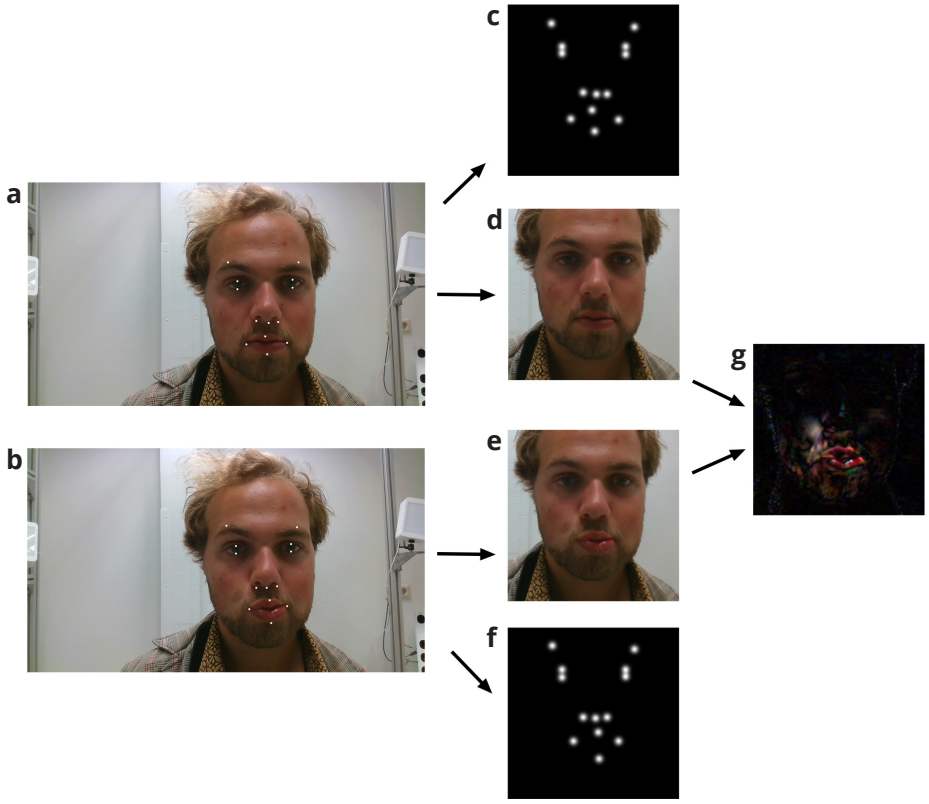

Figure 1. Pre-processing of the input images for the convolutional neural network, using the Sunnybrook Facial Grading System (SFGS) pose "pucker" as an example. The starting frame (a) and the maximum frame (b) were pre-selected in the dataset [5]. The 13 manually placed landmarks were superimposed and enlarged on the original images for visualization purposes (a & b). The original images were cropped to a 112 × 112 × 3 pixel image with image registration applied between the starting frame and frame of maximum exertion (d & e). The same cropping and image registration were applied to the landmark images, with an additional Gaussian filter, resulting in 112 × 112 × 3 pixel images (c & f).Finally, the difference image (g) was calculated between the cropped starting frame (d) and maximum frame (e), resulting in a 112 × 112 × 3 pixel image.

Pre-processing

This study followed the exact same pre-processing steps as outlined in our previous work to generate identical input images during the training of the CNN with the additional landmark images (Figure 1c – g) [5]. The pre-processing pipeline consisted of cropping the 2D colour images to a 112 × 112 pixel image and transforming the image using optical flow registration. A difference image with a matching resolution of 112 × 112 × 3 pixels was generated by calculating the absolute difference between the start (Figure 1d) and maximum frame (Figure 1e) for each individual colour channel, which highlighted the movement between frames (Figure 1g). All images were normalized, resulting in pixel values ranging from 0 to 1 for each input image.

Before the existing pre-processing pipeline was applied to the landmark images, a Gaussian filter ($\sigma = 3$) was applied to the landmark images. To keep the images perfectly aligned, the same cropping and the optical flow registration from the colour images were applied to the landmark images (Figure 1c & f). Lastly, the landmark images were normalized, resulting in a maximum pixel value of 1.

Architecture

A total of four inputs were given to the CNN: the starting frame image (Figure 1d), the landmark image of the starting frame (Figure 1c), the maximum frame image (Figure 1e), and the landmark image of the maximum frame (Figure 1f), each with a size of 112 × 112 × 3 pixels. The difference image was added as a fifth input representing the dynamic components of the SFGS, i.e., the symmetry of voluntary movement and synkinesis (Figure 1g). The subsequent layers in the CNN were adapted from our previous work using the same training parameters [5]. The input layer was followed by three data augmentation layers which were only activated during training of the model, consisting of a random horizontal flip, random zoom (range factor 0.8 to 1.2), and random rotation (range -20 to 20 degrees). The data augmentation was followed by 16 weight layers, with 13 convolution layers with 5 maxpool layers and 5 batch normalization layers [5]. This was followed by 3 fully connected layers with 1024 nodes, each with a dropout layer ($p = 0.5$) and a batch normalization layer (momentum = 0.95). A linear activation function was applied to produce a linear output, which was followed by a Gaussian noise layer ($\sigma = 0.1$). During training, the logcosh loss function was used to measure the difference between the predicted and actual values and the Adam optimizer was used to minimize the logcosh loss function. A complete overview of the CNN architecture is shown in Appendix B.

Training

All training parameters were kept consistent with our previous work, using the identical five different testing and training groups using a stratified k-fold based on the composite SFGS score [5]. The CNN model was trained and tested five times, where the testing data always consisted of 25 completely new subjects during each fold (20% of the subjects per fold). The training and testing groups were kept consistent between the 13 SFGS elements. A cyclic triangular learning rate was applied with a base learning rate of 1e-8, a max learning rate of 1e-3, and a step size of 4 (length of training dataset / batch size) [17]. During the training process a batch size of 32 was used and early stopping was implemented with a patience level set to 1500 epochs [18].

Analysis

The performance of the CNN was determined by comparing the predicted SFGS scores of the models with the SFGS scores as graded by the experienced human observers. The analysis was repeated for each of the five folds. The predicted scores of the individual elements of the SFGS were converted from a continuous scale to the respective nominal score of the SFGS as shown in Table 1, with the predicted scores capped at the minimum and maximum score of each individual element. From these 13 individual scores the subscores of the 3 SFGS components and the composite SFGS score were calculated according to the scoring key of the SFGS (Table 1). The inter-rater reliability between the CNN model and the human observers was determined by the intra-class correlation coefficient (ICC, type 2,1) for the SFGS components and the composite SFGS score [19].

The inter-rater reliability for the individual SFGS elements was determined both by the ICC (type 2,1) and by the quadratic weighted Cohen's Kappa [20]. An ICC of <0.5 was considered poor, 0.50 to 0.75 fair, 0.75 to 0.90 good, and 0.90 to 1.00 excellent [21]. A Cohen's Kappa lower than 0.20 was considered as having no agreement, 0.21 to 0.39 a minimal agreement, 0.40 to 0.59 a weak agreement, 0.60 to 0.79 a moderate agreement, 0.80 to 0.90 a strong agreement, and 0.91 to 1.00 an almost perfect agreement [22].

RESULTS

The inter-rater reliability between the CNN and the experienced observers is shown for each of the 13 elements of the SFGS, the SFGS subscores, and the composite SFGS score, in Table 2. Both the ICC (type 2,1) and quadratic weighted Cohen's Kappa were calculated for the 13 elements of the SFGS. All ICC values were higher compared to the Cohen's Kappa, with an average difference between the two reliability scores of 0.008. Table 2 only shows the Cohen's Kappa for the 13 elements with it being the most conservative score.

Table 2. Inter-rater reliability between the convolutional neural network (CNN) model and the experienced observers.

SFGS component	Inter-rater reliability training data	Inter-rater reliability testing data
Resting symmetry		
Eye	0.34 (0.00 – 0.61)	0.41 (0.15 – 0.64)
Cheek (naso-labial fold)	0.47 (0.34 – 0.62)	0.52 (0.34 – 0.69)
Mouth	0.70 (0.40 – 0.96)	0.67 (0.56 – 0.84)
Symmetry of Voluntary Movement		
Forehead wrinkle	0.88 (0.83 – 0.92)	0.89 (0.85 – 0.94)
Gentle eye closure	0.81 (0.75 – 0.86)	0.81 (0.69 – 0.91)
Open mouth smile	0.88 (0.82 – 0.92)	0.86 (0.80 – 0.89)
Snarl	0.69 (0.45 – 0.86)	0.71 (0.55 – 0.85)
Lip Pucker	0.72 (0.53 – 0.83)	0.68 (0.61 – 0.74)
Synkinesis		
Forehead wrinkle	0.69 (0.51 – 0.88)	0.74 (0.59 – 0.84)
Gentle eye closure	0.61 (0.31 – 0.82)	0.59 (0.46 – 0.86)
Open mouth smile	0.43 (0.34 – 0.56)	0.46 (0.34 – 0.58)
Snarl	0.49 (0.18 – 0.81)	0.49 (0.40 – 0.67)
Lip pucker	0.78 (0.61 – 0.88)	0.75 (0.68 – 0.78)
Resting symmetry subscore	0.64 (0.52 – 0.81)	0.62 (0.42 – 0.76)
Symmetry of voluntary movement subscore	0.92 (0.90 – 0.93)	0.92 (0.88 – 0.94)
Synkinesis subscore	0.79 (0.68 – 0.92)	0.78 (0.71 – 0.85)
Composite score	0.91 (0.89 – 0.94)	0.91 (0.88 – 0.94)

The inter-rater reliability is expressed as the quadratic weighted Cohen's Kappa for the individual Sunnybrook Facial Grading System (SFGS) elements and as the intra-class correlation coefficient (ICC, type 2,1) for the SFGS subscores and composite score. The mean inter-rater reliability was calculated from the five different k-folds. The values in between brackets show the inter-rater reliability range for the five k-folds.

The inter-rater reliability for the SFGS subscores and the composite SFGS score all fell within the same ICC category when comparing the testing and training data (Table 2). More specifically, the symmetry subscore showed a fair agreement, the symmetry of voluntary movement subscore showed an excellent agreement, the synkinesis subscore showed a good agreement, and the composite score showed an excellent agreement.

The majority of the inter-rater reliability for the individual SFGS elements were classified within the same category for both the training and testing data. Only the eye at rest and gentle eye closure during synkinesis deviated with one category. The elements of the resting symmetry showed a minimal to moderate agreement, the elements of s voluntary movement showed a moderate or strong agreement, and the synkinesis elements showed a weak to moderate agreement.

DISCUSSION

This study investigated the impact on the reliability of the automated SFGS by adding a facial landmark layer to a previously developed CNN [5]. The automated SFGS makes the SFGS more accessible for researchers, students, clinicians in training, or other untrained co-workers and could be implemented in settings such as online consults in an eHealth environment. The previous version of the automated SFGS achieved a similar inter-rater reliability as human observers who graded 50 patients with a PFP, which potentially leaves room for improvement [4,5]. As it is not always feasible to increase the cohort size in certain (clinical) settings, this study aimed to improve the reliability of the automated SFGS by adding a facial landmark layer to the training process. To our knowledge this is the first study to have investigated the impact of integrating a facial landmark layer to an existing CNN network. Previous studies have either added a landmark layer as a parallel model or added the landmark layer in a deeper layer of the model [23,24]. As the implementation is based on a separate landmark layer, it would be possible to add this layer to existing, validated models used for different applications, without changing the underlying architecture, making the implementation more broadly applicable.

To compare the two CNN models, as many potential confounding variables were kept consistent between the two studies. Hence, the same dataset was used, consisting of 116 patients with a PFP and 9 healthy subjects with their associated SFGS scores. All the pre-processing pipelines were kept the same, resulting in identical input images for the training of the CNN, with the only difference being the added landmark images to the input (Figure 1 c & f). The CNN architecture was kept consistent including the training parameters, and during the training of the CNN model the same training and testing groups were used. Finally, the same statistical analysis was applied to determine the reliability of the CNN models vs. the experienced human observers. One additional ICC analysis was calculated for the individual SFGS elements, as both the quadratic weighted Cohen's Kappa and ICC have been used to calculate the reliability for the individual SFGS elements [2,4,5,10–15]. The Cohen's Kappa and ICC deviated with an average of 0.008 with the Cohen's Kappa being the most conservative score and always being lower than the ICC. Therefore, the reliability of the individual elements was based on the quadratic weighted Cohen's Kappa and was considered a good approximation of the ICC.

By using the same dataset certain limitations were inherited from the previous study. For example, the cohort size of 125 subjects is relatively limited, despite this being the largest cohort used for the automation of the SFGS to date [25]. A larger dataset would show more variations of a PFP, which generally is beneficial for the reliability of CNN

models [6–9]. However, in the case of the SFGS there are no public databases available which include all the SFGS poses required for the training and testing of the model, highlighting the potential benefit of increasing the reliability of existing models based on smaller datasets [25–27]. To counteract some of the limitations of the cohort size a stratified k-fold was implemented which allows for a better estimation of the reliability of a CNN model in a smaller dataset [28]. Another aspect of the cohort should be highlighted as the current cohort included 9 healthy subjects out of the 125 subjects, which might indicate an unbalanced dataset. However, as the symmetry of voluntary movement or synkinesis can consist of five potential scores (Table 1) an even distribution would result in 25 subjects in each category considering the dataset of 125 subjects. However, a PFP can affect specific regions of the face, where it is possible for a patient to have an identical score as a healthy individual for the unaffected regions, acting as a healthy control. In case of the synkinesis, the eyebrow lift is expected to affect roughly 20% of the individuals, where 80% of the patients would act as a healthy reference [29]. Therefore, the addition of healthy subjects was relatively limited in the dataset, to prevent a bias towards healthy scores, although the ideal ratio of healthy subjects was not investigated in this study.

When comparing the reliability of the testing dataset between the old and new CNN model with facial landmarks, all subscores and the composite score showed an increase in reliability for the new CNN model (Table 2) [5]. More specifically, the composite SFGS score increased from good (0.87) to excellent (0.91), the resting symmetry subscore increased from poor (0.45) to fair (0.62), and the symmetry of voluntary movement subscore increased from good (0.89) to excellent (0.92) (Table 2) [5]. Only the synkinesis subscore showed the same agreement, a good agreement, for both the old model (0.77) and the new model (0.78). The improvement in subscores and composite score originated from the higher reliability of the individual SFGS elements where 12 out of the 13 elements showed an improvement. When calculating the average Cohen's Kappa for each category the resting symmetry elements increased from 0.38 to 0.53, the symmetry of voluntary movement elements increased from 0.74 to 0.79, and the synkinesis elements increased from 0.58 to 0.66 (Table 2) [5]. These results combined indicate a clear improvement of the reliability of the automated SFGS when adding a facial landmark layer to the CNN.

The performance of the automated SFGS can also be compared to the reliability of human observers, based on historic research, with a total of 5 studies determining the reliability of the 13 individual SFGS elements [4,13,30–32], with 2 studies not including the symmetry at rest [14,15], 3 studies only reporting the subscores and composite scores [10,33,34], and 1 study only reporting the composite score [12]. The reliability reported between human observers showed a relatively high range, albeit on the higher end of

agreement, with the ICC ranging from 0.81 to 1.00 with an average of 0.91 for the SFGS composite score [4,10,13–15,30–34]. This metric corresponds with the ICC of 0.91 for the composite score found in this study, indicating that the automated SFGS is grading as well as human observers (Table 2).

Cabrol et al. attempted to reconcile the range of reliability in SFGS grading with a study based on 20 patients with a PFP with a wide variety of SFGS composite scores, rated by 31 health professionals involved in the management of patients with PFP with different clinical backgrounds and not trained in the SFGS [15]. The new CNN model with the facial landmarks layer outperformed the average reliability of the 31 health professionals for all the subscores and the composite score, with a respective resting symmetry subscores of 0.62 vs. 0.55, a symmetry of voluntary movement subscore of 0.92 vs. 0.84, a synkinesis subscore of 0.78 vs. 0.48, and a composite score of 0.91 vs. 0.85 (Table 2) [15]. The new CNN was comparable to the best performing health professionals, the ENT specialists, with a relative resting symmetry subscore of 0.62 vs. 0.70, a symmetry of voluntary movement subscore of 0.92 vs. 0.89, a synkinesis subscore of 0.78 vs. 0.45, and a composite score of 0.91 vs. 0.89 (Table 2) [15]. These results support the conclusion that the new CNN model is grading at the level of human observers.

The old CNN model was compared to the SFGS learning curve of novice observers during the grading of 100 patients with a PFP and the old model performed at the same reliability as human observers after grading 50 patients with a PFP [4,5]. The reliability of the new CNN model saw an improvement with a comparable reliability of human observers after grading 70 to 100 patients, with an average difference in Cohen's Kappa of 0.00 when comparing all individual SFGS elements (Table 2) [4]. The SFGS learning curve stabilized for the human observers after grading 70 patients, which indicates that the automated SFGS is on a similar level as novice observers after an extensive 7-week training and feedback program [15].

The new CNN with the facial landmark layer improved the reliability of the automated SFGS, but the inner workings of CNNs are not easily interpretable [6–9]. However, we hypothesize that the facial landmarks functioned as guidance for the CNN focusing on key areas of the face, which worked in conjunction with the difference image already present in the previous CNN model highlighting areas of movement (Figure 1g). A potential downside of the facial landmark layer was the possibility of overfitting, where the model would become overly specialized, and was unable to generalize well to new data, resulting in a lower reliability for the testing data [6–9]. However, the difference between the Cohen's Kappa of the training and testing data for the 13 SFGS elements was on average 0.03 for the old CNN model, whilst this was 0.01 for the new CNN

model, indicating decreased overfitting between training and testing data (Table 2) [5]. Additionally, all training data and testing data fell inside the same agreement category. Hence, it is unlikely the increased reliability of the new CNN model was a result of overfitting.

To further minimize overfitting of the data, a specific set of clearly defined landmarks was selected to track key areas for the SFGS poses [16]. The superciliare tracked the eyebrow lift, the palpebrale superius and inferius tracked the gentle eye closure and the resting symmetry of the eye, the subalare and subnasale tracked the snarl. Lastly, the cheilion and labiale superius and inferius tracked the open mouth smile, pucker, and resting symmetry of the mouth. The resting symmetry of the cheek is missing from this overview as there was no landmark available in this region that could be placed with a high reliability [16]. Nonetheless, the quadratic Cohen's Kappa increased from 0.29 to 0.52 for the resting symmetry of the cheek (Table 2) [5]. A potential explanation would be that the difference image would show a low amount of movement as the face was at rest, making it increasingly difficult for the CNN to find an area to focus on. The facial landmarks would still indicate key areas of the face and might be more efficient in guiding the areas of interest for the CNN. An improvement in reliability would be expected for the remaining poses at rest, which was indeed the case as the quadratic Cohen's Kappa for resting symmetry of the eye increased from 0.37 to 0.41 and the resting symmetry of mouth increased from 0.47 to 0.67 (Table 2) [5].

Finally, to minimize errors introduced by the landmark placement, the landmarks were placed manually [16]. The impact of possible inaccurate landmark placement was further reduced as a Gaussian filter was applied to the landmarks with the initial purpose of preventing overfitting (Figure 1c & f). However, the manual placement of landmarks is an undesirable step as the automated SFGS should make the SFGS more accessible. As this study has shown the benefit of adding a landmark layer to the CNN, the automation of the landmark placement should be investigated in future research, with the ultimate goal to develop a fully automatic and highly reliable SFGS that is easily applicable to both health care providers and patients with a PFP.

CONCLUSION

The integration of a facial landmark layer into the CNN showed a clear improvement in the reliability of the automated SFGS, reaching a performance level comparable to human observers. These results were attained without increasing the dataset underscoring the impact of incorporating facial landmarks into a CNN. These findings indicate that the automated SFGS with facial landmarks is a reliable tool for assessing patients with a PFP.

REFERENCES

1. Samsudin, W. S. W. & Sundaraj, K. Evaluation and Grading Systems of Facial Paralysis for Facial Rehabilitation. J. Phys. Ther. Sci. 25, 515–519 (2013).
2. Fattah, A. Y. et al. Facial Nerve Grading Instruments: Systematic Review of the Literature and Suggestion for Uniformity. Plast. Reconstr. Surg. 135, 569–579 (2015).
3. Ross, B. G., Fradet, G. & Nedzelski, J. M. Development of a sensitive clinical facial grading system. Otolaryngol. - Head Neck Surg. 114, 380–386 (1996).
4. van Veen, M. M., Bruins, T. E., Artan, M., Werker, P. M. N. & Dijkstra, P. U. Learning curve using the Sunnybrook Facial Grading System in assessing facial palsy: An observational study in 100 patients. Clin. Otolaryngol. 45, 823–826 (2020).
5. ten Harkel, T. C. et al. Automatic grading of patients with a unilateral facial paralysis based on the Sunnybrook Facial Grading System - A deep learning study based on a convolutional neural network. Am. J. Otolaryngol. 44, 103810 (2023).
6. Lecun, Y., Bengio, Y. & Hinton, G. Deep learning. Nature 521, 436–444 (2015).
7. Su, Z. et al. Deep learning-based facial image analysis in medical research: a systematic review protocol. BMJ Open 11, e047549 (2021).
8. Liu, Q. et al. A review of image recognition with deep convolutional neural network. in International onference on intelligent computing 69–80 (Springer, 2017).
9. Shen, D., Wu, G. & Suk, H.-I. Deep learning in medical image analysis. Annu. Rev. Biomed. Eng. 19, 221–248 (2017).
10. Volk, G. F. et al. Reliability of grading of facial palsy using a video tutorial with synchronous video recording. Laryngoscope (2018) doi:10.1002/lary.27739.
11. Gaudin, R. A. et al. Emerging vs time-tested methods of facial grading among patients with facial paralysis. JAMA Facial Plast. Surg. 18, 251–257 (2016).
12. Neely, J. G., Cherian, N. G., Dickerson, C. B. & Nedzelski, J. M. Sunnybrook facial grading system: reliability and criteria for grading. Laryngoscope 120, 1038–1045 (2010).
13. Tan, J. R., Coulson, S. & Keep, M. Face-to-Face Versus Video Assessment of Facial Paralysis: Implications for Telemedicine. J. Med. Internet Res. 21, e11109–e11109 (2019).
14. Coulson, S. E., Croxson, G. R., Adams, R. D. & O'Dwyer, N. J. Reliability of the "Sydney," "Sunnybrook," and "House Brackmann" facial grading systems to assess voluntary movement and synkinesis after facial nerve paralysis. Otolaryngol. - Head Neck Surg. 132, 543–549 (2005).
15. Cabrol, C. et al. Sunnybrook Facial Grading System: Intra-rater and Inter-rater Variabilities. Otol. Neurotol. 42, 1089–1094 (2021).
16. Caple, J. & Stephan, C. N. A standardized nomenclature for craniofacial and facial anthropometry. Int. J. Legal Med. 130, 863–879 (2016).
17. Smith, L. N. Cyclical learning rates for training neural networks. Proc. - 2017 IEEE

Winter Conf. Appl. Comput. Vision, WACV 2017 464–472 (2017) doi:10.1109/WACV.2017.58.

18. Prechelt, L. Early stopping - But when? in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) (eds. Montavon, G., Orr, G. B. & Müller, K.-R.) vol. 7700 LECTU 53–67 (Springer Berlin Heidelberg, Berlin, Heidelberg, 2012).
19. Shrout, P. E. & Fleiss, J. L. Intraclass correlations: uses in assessing rater reliability. Psychol. Bull. 86, 420 (1979).
20. Cohen, J. Weighted kappa: nominal scale agreement provision for scaled disagreement or partial credit. Psychol. Bull. 70, 213 (1968).
21. Koo, T. K. & Li, M. Y. A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. J. Chiropr. Med. 15, 155–63 (2016).
22. McHugh, M. L. Interrater reliability: the kappa statistic. Biochem. medica 22, 276–282 (2012).
23. Hu, M., Wang, H., Wang, X., Yang, J. & Wang, R. Video facial emotion recognition based on local enhanced motion history image and CNN-CTSLSTM networks. J. Vis. Commun. Image Represent. 59, 176–185 (2019).
24. Kollias, D. & Zafeiriou, S. Exploiting Multi-CNN Features in CNN-RNN Based Dimensional Emotion Recognition on the OMG in-the-Wild Dataset. IEEE Trans. Affect. Comput. 12, 595–606 (2021).
25. Jirawatnotai, S., Jomkoh, P., Voravitvet, T. Y., Tirakotai, W. & Somboonsap, N. Computerized Sunnybrook facial grading scale (SBface) application for facial paralysis evaluation. Arch. Plast. Surg. 48, 269–277 (2021).
26. Guarin, D. L. et al. Toward an Automatic System for Computer-Aided Assessment in Facial Palsy. Facial Plast. Surg. aesthetic Med. 22, 42–49 (2020).
27. Xia, Y. et al. AFLFP: A Database With Annotated Facial Landmarks for Facial Palsy. IEEE Trans. Comput. Soc. Syst. 10, 1975–1985 (2023).
28. Vabalas, A., Gowen, E., Poliakoff, E. & Casson, A. J. Machine learning algorithm validation with a limited sample size. PLoS One 14, 1–20 (2019).
29. Beurskens, C. H. G., Oosterhof, J. & Nijhuis-van der Sanden, M. W. G. Frequency and location of synkineses in patients with peripheral facial nerve paresis. Otol. Neurotol. 31, 671–5 (2010).
30. Waubant, A., Franco-Vidal, V. & Ribadeau Dumas, A. Validation of a French version of the Sunnybrook facial grading system. Eur. Ann. Otorhinolaryngol. Head Neck Dis. 139, 119–124 (2022).
31. Pavese, C. et al. Validation of the Italian version of the Sunnybrook Facial Grading System. Neurol. Sci. 34, 457–463 (2013).
32. Kayhan, F. T., Zurakowski, D. & Rauch, S. D. Toronto facial grading system: Interobserver reliability. Otolaryngol. - Head Neck Surg. 122, 212–215 (2000).

33. Hu, W. L., Ross, B. & Nedzelski, J. Reliability of the Sunnybrook Facial Grading System by Novice Users. J. Otolaryngol. 30, 208–211 (2001).
34. Kanerva, M., Poussa, T. & Pitkäranta, A. Sunnybrook and House-Brackmann Facial Grading Systems: Intrarater repeatability and interrater agreement. Otolaryngol. - Head Neck Surg. 135, 865–871 (2006).

APPENDIX B

Table 1. Full overview of the model architecture used for the convolutional neural network (CNN).

Layer	Type	Shape
0	InputLayer	[(None, 112, 112, 3)]
1	InputLayer	[(None, 112, 112, 3)]
2	InputLayer	[(None, 112, 112, 3)]
3	InputLayer	[(None, 112, 112, 3)]
4	InputLayer	[(None, 112, 112, 3)]
5	Concatenate	(None, 112, 112, 15)
6	RandomFlip	(None, 112, 112, 15)
7	RandomRotation	(None, 112, 112, 15)
8	RandomZoom	(None, 112, 112, 15)
9	Conv2D	(None, 112, 112, 16)
10	Conv2D	(None, 112, 112, 16)
11	BatchNormalization	(None, 112, 112, 16)
12	MaxPooling2D	(None, 56, 56, 16)
13	Conv2D	(None, 56, 56, 32)
14	Conv2D	(None, 56, 56, 32)
15	BatchNormalization	(None, 56, 56, 32)
16	MaxPooling2D	(None, 28, 28, 32)
17	Conv2D	(None, 28, 28, 64)
18	Conv2D	(None, 28, 28, 64)
19	Conv2D	(None, 28, 28, 64)
20	BatchNormalization	(None, 28, 28, 64)
21	MaxPooling2D	(None, 14, 14, 64)
22	Conv2D	(None, 14, 14, 128)
23	Conv2D	(None, 14, 14, 128)
24	Conv2D	(None, 14, 14, 128)
25	BatchNormalization	(None, 14, 14, 128)
26	MaxPooling2D	(None, 7, 7, 128)
27	Conv2D	(None, 7, 7, 128)
28	Conv2D	(None, 7, 7, 128)
29	Conv2D	(None, 7, 7, 128)
30	BatchNormalization	(None, 7, 7, 128)
31	MaxPooling2D	(None, 3, 3, 128)
32	Flatten	(None, 1152)
33	Dense	(None, 1024)
34	BatchNormalization	(None, 1024)
35	Dropout	(None, 1024)
36	Dense	(None, 1024)
37	BatchNormalization	(None, 1024)
38	Dropout	(None, 1024)
39	Dense	(None, 1)
40	GaussianNoise	(None, 1)

CHAPTER 6

GENERAL DISCUSSION & FUTURE PERSPECTIVES

INTRODUCTION

The assessment of the severity of a unilateral peripheral facial palsy (PFP) is a crucial step in the treatment and monitoring of a PFP. One of the recommended grading systems to determine the severity of a PFP is the Sunnybrook Facial Grading System (SFGS) due to its validity, reproducibility, and responsiveness [1–6]. However, there is a learning curve associated with the SFGS, which will require the time of a trained observer to grade the PFP [7]. This might make the SFGS less accessible for researchers, students, clinicians in training, or other untrained co-workers and patients. In turn, this places limits on how frequently the PFP can be assessed. The automation of the SFGS could help alleviate these issues and potentially even exceed the reliability of human observers performing the SFGS manually.

Therefore, this thesis investigated the automation of the SFGS with the long-term aim to develop a user-friendly system that could be used by the patient at home without any assistance. The automated SFGS should be reliable, cost effective, portable, fast, and intuitive to use, in order to make the automated system a low barrier of entry in clinical practice. The use of non-invasive three-dimensional (3D) video imaging was preferred for the automation of the SFGS due to the complexity of the human face and the changes to the surface during the dynamic poses of the SFGS. The implementation of a professional 3D video (4D) imaging system would not be feasible due to its size and cost. Therefore, the Intel RealSense™ F200 and RealSense™ D415 (Intel®, Santa Clara, USA) were investigated for the implementation of the automated SFGS. The first part of this thesis validated the depth data of the RealSense cameras, including derived 3D landmarks and anthropometric measurements. The second part of the thesis implemented an automated SFGS based on recordings of the RealSense D415.

The research questions related to these two topics are stated in the section *"Research questions"* in **Chapter 1**.

The next section of this chapter, *"Discussion of the research questions"*, will provide a general overview of the methodology and results of the presented studies and will answer the stated research questions. The discussion of the research questions is followed by the discussion of the future perspectives for the automation of the SFGS.

DISCUSSION OF THE RESEARCH QUESTIONS

The introduction of new hardware in a clinical setting brings certain challenges with it, as was the case with the RealSense F200. Considering the size and price difference between the RealSense F200 ($100 USD) and a professional imaging system commonly used in clinical settings often with a cost of tens of thousands of US dollars, a difference in depth accuracy was expected [8,9]. This difference could affect the translation of the digital data into real world coordinates and subsequent data analysis. Therefore, **Chapter 2** answered the research question: "*What is the depth accuracy of the RealSense F200 during the SFGS poses?*" As the automated SFGS is intended to be used for patients with a PFP, the depth accuracy was determined in a cohort of 34 patients with a PFP. The subjects were simultaneously recorded with the RealSense F200 and the clinically validated 3dMD system (3dMDface, 3dMD, Atlanta, USA) whilst performing the six SFGS poses, which includes the face at rest. The depth accuracy was determined by comparing the depth images of these two systems during maximum exertion of the SFGS poses. Due to the ipsilateral nature of the PFP, it was also possible to compare the depth accuracy between the healthy and palsy side of the face. The results showed that the RealSense F200 average depth accuracy was 1.48 mm for the face at rest and 1.49 mm during the voluntary movements of the SFGS, where the SFGS poses did not significantly influence the RealSense depth accuracy. In addition, the depth accuracy was not significantly affected by the PFP (1.48 mm), compared to the healthy side of the face (1.46 mm). However, the distance of the patients to the RealSense F200 was shown to affect the accuracy of the system, where the best depth accuracy of 1.07 mm was measured at a distance of 35 cm. To put these results in perspective, the 3dMD system has a reported depth accuracy of around 0.25 mm for the face at rest [8,10–13]. Therefore, the average depth accuracy of the RealSense F200 was roughly six times worse compared to the 3dMD. Although no other research was found regarding the close range imaging of human subjects based on the RealSense F200, similar low-cost portable cameras have reported depth accuracies in the same millimetre range as the RealSense F200 [14–16]. These results gave a good first indication of the expected depth accuracy of the RealSense F200. However, a more detailed analysis could be achieved by segmenting the face in predefined areas to determine the role of each area in the depth inaccuracy [17]. In addition, the recording of the subjects was performed in a windowless room with diffuse lighting, which might have overestimated the depth accuracy compared to a home monitoring situation [14,18,19]. Finally, the 3dMD setup captured static 3D images and could only compare the moment of maximum exertion during the SFGS poses. A 4D imaging setup could determine the depth accuracy of multiple RealSense F200 frames as long as the reference system would not interfere with the structured light pattern of the RealSense F200 [15]. These professional 4D systems have become more attainable, but

their use is still mostly limited to dedicated healthcare centres. While it was anticipated that the RealSense F200 would have a lower overall depth accuracy compared to the professional 3dMD system, this research has provided quantifiable measurements of its depth accuracy when imaging patients with PFP in a clinical setting.

A second generation of RealSense 4D cameras was released in 2018 with significant improvements to the software and hardware to warrant a change from the RealSense F200 to the RealSense D415 [9,14,18,19]. Therefore, **Chapter 3** validated multiple aspects of the RealSense D415 where the chapter was divided into two parts. The first part analysed 30 patients with a PFP at rest, whilst the second part analysed the exact same patient population during the voluntary movements of the SFGS. In this discussion both parts from **Chapter 3** are discussed as one. As the depth accuracy of the RealSense D415 was unknown in a clinical setting the following research question was first answered: *"What is the depth accuracy of the RealSense D415 during the SFGS poses?"* A similar methodology was used as in **Chapter 2** where the 3dMD system was used as the gold standard to determine the depth accuracy of the RealSense F200. However, the diffuse lights were removed from the measurement setup to represent a more realistic home scenario. Furthermore, the recording distance to the patient was maintained more consistently at approximately 35 cm. The results showed an average depth accuracy for the RealSense D415 of 0.97 mm for the face at rest and 0.98 mm for the voluntary movements. This was around 50 percent better compared to the RealSense F200 but could partially be explained by the lower average recording distance. Other studies have reported similar depth accuracies of the RealSense D415 ranging from 0.5 mm to 2.0 mm at distances at or lower than 35 cm [14,16,19–22]. Although these studies analysed either flat planes or inanimate objects instead of human subjects, the results do confirm a similar depth accuracy of the RealSense D415. Considering the depth accuracy of the RealSense D415 and the additional software and hardware improvements to the camera, the RealSense D415 was considered a suitable successor to the RealSense F200 [23].

After the analysis of the overall depth accuracy of the RealSense D415, **Chapter 3** further explored the potential influence of the lower colour and depth image quality of the RealSense D415 based on the same patient cohort used to determine the depth accuracy. First, the reliability of manual placement of 3D landmarks was investigated, where the landmarks could be used for the implementation of anthropometric measurements or act as features for the automation of the SFGS. Therefore, the following research question was stated: *"What is the reliability of 3D landmark placement on RealSense D415 images during the SFGS poses?"* To answer this question two observers placed 14 facial landmarks on the RealSense D415 and 3dMD depth images at moment

of maximum exertion of the SFGS poses to determine the inter-rater reliability. The first observer repeated the landmark placement three weeks after the first session to determine the intra-rater reliability. The reliability of the landmark placement first needed to be determined on the 3dMD images, as there was no reference data available for the voluntary movements of the SFGS in a cohort of patients with a PFP. The reliability of the 3dMD landmark placement could then be used as a baseline for the landmark placement on the RealSense D415. The average landmark reliability for the face at rest was 0.92 mm for the 3dMD images, which was considered within the same range as the landmark placement for healthy subjects at rest [8,12,24–29]. The average landmark reliability for the voluntary movements was 1.03 mm. This indicated a trend towards a slightly lower reliability of the landmark placement during the voluntary movements, but the reliability still fell within the range for healthy subjects at rest. As the selection of individual landmarks might bias the overall average reliability, individual landmarks were compared where possible and also fell within the range of the healthy subjects at rest. Therefore, the 3dMD landmark placement was considered a good reference for the RealSense D415 landmark placement. The landmark placement on the RealSense D415 images reported a lower average reliability of 1.47 mm for the face at rest. This resulted in multiple landmarks having a significant difference in reliability compared to their 3dMD counterpart. During the voluntary movements, the average reliability improved to 1.28 mm, causing a lower number of landmarks to have a significant difference between the RealSense D415 and 3dMD landmark placement. The landmarks around the nose region played a major role of the improved reliability during the voluntary movements. It was noted that the patients had their head tilted slightly more backwards during the voluntary movements resulting in a better coverage of the depth data around the nose region. Therefore, head positioning should be an important consideration during the recording of subjects. However, the lack of depth data in certain regions is unfortunately an inherent limitation of using a compact stereo vision camera at close range, due to the limited spacing between the two internal cameras [30]. This could be resolved by adding a second RealSense D415 during the recording, but would negatively impact the cost, complexity, and portability of the automated SFGS, and was not considered a feasible solution. Another factor that influenced the overall decrease in landmark placement reliability was due to the lower colour quality of the RealSense D415, as a subset of the landmarks were placed according to certain colour transitions. Overall, the reliability of the landmark placement was lower on the RealSense D415 images compared to the 3dMD images. This study was the first to quantify the exact impact on the manual landmark placement using a lower quality 4D camera for all SFGS poses in a population of patients with a PFP. By analysing the landmark placement for both the RealSense and 3dMD images the validation can be applied in clinical settings where either a professional system or a consumer grade depth camera is used.

From the 14 manually placed 3D landmarks it was possible to derive clinically relevant anthropometric measurements. Although the landmark placement and anthropometric measurements are closely related, it is possible for the anthropometric measurements to be inaccurate even with perfect landmark placement. These discrepancies can be caused by inaccuracies in the underlying depth data or by basing the landmark placement on different features due to variations in colour quality. Therefore, a third research question was stated in **Chapter 3**: *"What is the reliability and agreement of 3D anthropometric measurements on RealSense D415 images during the SFGS poses?"* A total of 14 anthropometric measurements were derived from the 14 manual landmarks placed on both the RealSense D415 and 3dMD images during the SFGS poses. The first part of the research question addressed the reliability of the 3D anthropometric measurements, where the measurements based on the 3dMD landmarks showed an excellent reliability for all the SFGS poses. The average intra-class correlation coefficient (ICC) was 0.94 for the measurement for the face at rest and the average ICC for the voluntary movements was 0.95. This led to the interesting observation that although the 3D landmark placement was less reliable for the voluntary movements, as discussed in the previous paragraph, the derived anthropometric measurements became more reliable during the voluntary movements. Existing literature has reported a similar ICC range for anthropometric measurements based on high quality depth images in cohorts of healthy subjects at rest [10,28,31–38]. Therefore, all the 3dMD anthropometric measurements based on the SFGS poses were considered to be excellent reference values for the RealSense D415 measurements. The RealSense D415 measurements showed good reliability for the face at rest with an average ICC of 0.82 and increased to an excellent reliability for the voluntary movements with an average ICC of 0.91. This increase could be due to the combination of the higher reliability in the 3D landmark placement, as determined in the previous paragraph, and the increased reliability during the voluntary movements as seen on the 3dMD images. Despite this improvement, the average reliability of the RealSense D415 anthropometric measurements fell outside the reported range based on healthy subjects using high quality depth images [10,28,31–38]. The second part of the research question investigated the agreement of the anthropometric measurements of the RealSense D415 compared to the 3dMD measurements. The RealSense D415 measurements showed an average underestimation of -0.90 mm for the face at rest and -1.12 mm for the voluntary movements compared to the 3dMD measurements, where 95% of the measurement are expected to be within a 5 mm error of the 3dMD measurements. Due to the differences in depth accuracy between the RealSense D415 and 3dMD depth images, a measurement error was expected. However, the voluntary movement with the highest depth accuracy, showed the lowest overall agreement. This indicated that the landmark placement itself played a larger role in the lower agreement

of the measurements, which could be caused by more difficult feature detection due to a lower colour and depth quality of the RealSense D415. In comparison, the 3dMD anthropometric measurements can achieve submillimetre accuracy compared to direct anthropometric measurements [10,28,31–38]. The 3dMD measurements during the voluntary movements were not compared to direct anthropometric measurements, which would be infeasible for patients with a PFP. However, the analysis of the depth accuracy, landmark placement, and anthropometric measurements in **Chapter 3** indicated that all SFGS poses were found to be in the same range as healthy subjects at rest for the 3dMD images. Therefore, the influence of the voluntary movements was expected to be limited in this analysis. It was clear however, that submillimetre accuracy should not be expected during anthropometric measurements based on the RealSense D415.

After gaining a better understanding of the data generated by the RealSense cameras a first version of the automated SFGS was implemented in **Chapter 4**. A traditional machine learning approach would use features such as landmarks and anthropometric measurements for the automation of image processing tasks. However, there has been a shift towards deep learning implementations, which remove the need for manual feature selection, where the convolutional neural network (CNN) is especially suited for image feature selection [39–46]. This led to the following research question: *"What is the reliability of an automated SFGS grading system based on a CNN compared to human observers?"* The training and testing of the CNN model were based on a dataset consisting of 116 patients and 9 healthy subjects performing the SFGS poses recorded with the RealSense D415. All subjects were graded according to the SFGS by multiple observers experienced in the SFGS grading of patients with a PFP. From these recordings three two-dimensional (2D) colour images were used as the input data to train a separate CNN model for each of the 13 individual SFGS elements. The input images consisted of the face at rest just before the start of the SFGS pose, the moment of maximum exertion during the SFGS pose, and a difference image which was the absolute difference between the two previously selected images. The three subscores and composite SFGS score were calculated from the 13 element scores generated by the 13 CNN models, replicating the process of the manual SFGS. The reliability of the automated SFGS was determined by comparing the automated score with the score of the human observers. The inter-rater reliability of the automated SFGS, including the individual SFGS elements, fell within the reported range of human observer reliability, albeit at the lower end of the reported range [3,7,47–52]. More specifically, the results showed an average ICC of 0.87 for the composite SFGS score, 0.45 for the resting symmetry subscore, 0.89 for the symmetry of voluntary movement subscore, and 0.77 for the

synkinesis subscore. Although there have been efforts to automate the grading of a PFP, existing research either investigated other grading systems than the SFGS, used small cohorts with less than 30 subjects, or only analysed the composite score of the SFGS, making a direct comparison unfeasible or unreliable [53,54,63–72,55,73,74,56–62]. The first implementation of the automated SFGS showed promising results, but there were certain limitations in the methodology. The used database of 125 subjects was relatively small for the implementation of a CNN model [39,41,42,45,75,76]. Therefore, an existing CNN architecture was reduced in size to minimize the complexity of the CNN model [77]. In addition, the difference image used as a third input was intended to highlight areas of movement between the frame of rest and the frame of maximum exertion. Ideally, these features would automatically be detected by the CNN, but this was challenging for the small dataset, hence the implementation of the difference image. Another downside of the smaller dataset was that the same data was used for both model validation and testing, which could lead to an overestimation of the reliability of the model [78]. Therefore, a stratified k-fold was implemented to better estimate the inter-rater reliability of the model [79]. This meant that the model was trained 5 times, where each fold used a unique set of 25 subjects in the testing set. The impact of the small cohort size was further reduced by using data augmentation, dropout layers, noise layers, batch normalization, and early stopping during the training of the CNN model [46,80–82]. All these preventative measures resulted in a similar inter-rater reliability between the training and testing set, indicating that overfitting was minimized. Therefore, the results presented in this chapter were considered a good indication of the reliability of the automated SFGS, showing the potential of the clinical implementation of the automated SFGS.

With a first iteration of the automated SFGS, **Chapter 5** investigated if the reliability of the CNN model could be improved without increasing the size of the dataset. In the original CNN model, a difference layer was added as an input to highlight areas of movement. During the absence of movement this layer would deactivate. Therefore, the hypothesis was that an input layer with facial landmarks could improve the reliability by indicating regions of interest for the CNN even in the absence of motion. This led to the following research question: *"What is the impact on the reliability of the automated SFGS by adding a facial landmark layer to the CNN?"* In order to compare the two CNN models, as many potentially confounding variables were kept consistent between **Chapter 4** and **Chapter 5**. Hence, the same dataset was used, consisting of the 116 patients with a PFP and 9 healthy subjects with their corresponding SFGS scores. All the pre-processing pipelines were kept the same, resulting in identical input images for the training of the CNN, with the only difference being the added landmark image to the input of the model. The landmark image

consisted of 13 manually placed facial landmarks with a Gaussian filter applied to the image. The addition of the landmark layer resulted in a clear improvement in the reliability of the automated SFGS, reaching the level of experienced observers with an excellent agreement for the composite score [3,7,47–50]. More specifically, the ICC for the composite SFGS score increased from 0.87 to 0.91, the resting symmetry subscore increased from 0.45 to 0.62, the symmetry of voluntary movement subscore increased from 0.89 to 0.92, and the synkinesis subscore increased from 0.75 to 0.78. The increase in reliability was unlikely to be caused by overfitting, as the overall difference in inter-rater reliability between the training and testing dataset decreased for the CNN model with the landmark layer. While the improved reliability of the CNN model with the facial landmarks is beneficial, there is a clear trade-off. One of the main advantages of deep learning is the elimination of manual feature selection, and introducing manual features increases the overhead for automating the SFGS [43,76]. For example, in this study the features were based on manually placed landmarks due to their reliability [31,34,83–86]. However, the manual landmark placement would not be feasible for the automated SFGS due to the time constraint involved. In turn, this would require the implementation of automated landmark placement, which will need to be validated. The impact of the potential errors from automated landmark placement might be reduced due to the use of a Gaussian filter on the landmark image but still needs to be investigated. In addition, the feature selection is expected to impact the reliability, where this study used 13 facial landmarks important during the movement of the SFGS poses, based on the 14 landmarks introduced in **Chapter 3**. Although the additional landmark layer would require further research, the increase in reliability might justify the inclusion, especially if it enables an earlier implementation of the automated SFGS in a clinical setting.

FUTURE PERSPECTIVES

This thesis investigated the automation of the SFGS with the long-term aim to develop a user-friendly system that can be used by the patient at home without any assistance. The focus for this thesis was the validation and implementation of an automated SFGS using a relatively affordable 4D imaging system in a clinical setting. This meant certain areas of research were out of scope for this thesis, such as the regulatory requirements for the automated SFGS. However, these unexplored topics will be important for the long-term implementation of the automated SFGS in a clinical or non-clinical setting. To this end we will first discuss possible future improvements regarding the topics more closely related to this thesis, followed by a broader overview of areas of research that are required for a successful implementation of the automated SFGS, finishing with an overview of the possible short-term and long-term clinical impact of the automated SFGS.

One of the first technical improvement that come to mind for future work would be the implementation of the RealSense depth data, as validated in **Chapter 2** and **Chapter 3**, in the automated SFGS [44]. As previously discussed, the depth data and the 3D landmarks and anthropometric measurements based on the RealSense recordings could play an important role in improving reliability of automated SFGS, due to the complex and dynamic features of the face during the SFGS poses [11,12,30,87]. However, in this thesis it was chosen to first determine a baseline for the reliability of the automated SFGS based on the 2D images of the RealSense recordings. From this baseline it will be possible to determine the impact of the depth data on the reliability of the automated SFGS. A valid outcome could be that the 2D data is sufficient to achieve a high reliability for the automation of the SFGS. This would have the advantage of lowering the barrier of entry in a home implementation of the automated SFGS. However, this trade-off can only be made in case the impact of the depth data is determined. Therefore, this thesis captured and validated the depth data of a relatively affordable 4D camera from the start. It should be noted that although the RealSense F200 and D415 were used in this thesis to capture the colour and depth images, these cameras are not the only portable and low-cost 4D imaging solutions available [15,16,18]. The RealSense cameras are intended as a representation of depth cameras with lower performance compared to professional systems. Technological advancements could make 4D cameras more affordable making it more likely patients will already have a suitable depth camera at home, such as the integrated depth cameras used for facial authentication in smartphones and laptops [88,89]. Although it would be preferred to support all these cameras to lower the barrier of entry of the automated SFGS, the usage of multiple depth cameras complicates the initial implementation and validation of the automated SFGS. Therefore, the inclusion of multiple depth cameras remains to be implemented in future work.

A major challenge in this thesis was the relatively small cohort size used for the automation of the SFGS due to the time required for the inclusion of the patients with a PFP in a single centre study design. A larger dataset would show more variations of the PFP and would typically be beneficial for the reliability of the CNN model [41,45,46,76]. Although multiple mitigations were implemented during the training of the CNN model due to the small dataset, as discussed in **Chapter 4** and **Chapter 5**, not all recommended techniques were implemented. Most notably, it was initially attempted to use pre-existing CNN architectures that have been successfully implemented in related image automation tasks. This would allow the usage of a technique called transfer learning, where a much larger dataset is used to train the network for a related image classification task [41,75,80]. However, in unpublished work the implementation of these pre-trained networks did not result in a reliability comparable to human observers for the automated SFGS. Due to the potential impact of a successful implementation of transfer learning it would be recommended to revisit this technique in future work. Another limitation of the single study centre design is the relatively limited variation in demographic and recording circumstances represented by a single clinical centre, which might overestimate the reliability of the model when deployed in other clinical centres. Therefore, future work could benefit from a multinational multi-centre study, which would allow for a much faster growth of the dataset with a more varied demographic and recording circumstances. A larger available dataset could have further implications on the optimization of the existing CNN model. For example, **Chapter 4** and **Chapter 5** added a difference image and a landmark image to the CNN to increase the reliability of the CNN. However, with the training on a larger dataset, these features are expected to be detected by the CNN itself, resulting in a simplified model. Other improvements to the existing CNN model could include the implementation of a more complex CNN model, where all SFGS elements are determined simultaneously in a single model, which could take the relationship between different SFGS elements into account. Apart from optimizing the existing CNN model, it might be beneficial to implement completely new deep learning architectures or pre-processing pipelines in future work, especially considering the rapidly evolving field of deep learning. This could include the usage of depth data as discussed in the previous paragraph with 3D CNNs, addition of temporal data in recurrent neural networks, or the usage of attention mechanisms such as visual transformer networks, or the focus on hierarchical relationships between features using capsule networks [40,41,44,76,90–93].

There are also areas of research that were not explored in this thesis which need to be considered for the long-term implementation of the automated SFGS, such as regulatory requirements, workflow optimization including the development of a graphical user interface (GUI), multi-language support, integration with existing hospital systems, and logistics and financing of the distribution of software and cameras. To determine the

best path forward for the automated SFGS a multidisciplinary team is required, including the input from a representative patient population. Therefore, the following overview should be considered as a rough potential guideline and is expected to change based on the input from the multidisciplinary team. Although the home implementation of the automated SFGS is the final aim of this project, there are serious complexities regarding data management, camera availability, and overall workflow. This makes the initial implementation of the automated SFGS in a multi-centre study an interesting option, which allows for a more controlled environment and less financial and logistical challenges compared to a home implementation. Another advantage of a multi-centre study would be the limited number of staff interacting with the automated system, which will allow for more elaborate instructions at the start of the development of the automated SFGS as these instructions do not need to be repeated for each new patient. This would make the staff better suited to deal with (minor) challenges encountered during the development stage. During this process, the final aim of a home implementation should be leading in design choices, where a significant effort should be placed in the ease of use of the automated SFGS. To this extent the clinical stage can act as the beta development to optimize the workflow of the automated SFGS until untrained staff is able to use the automated SFGS. Workflow concepts that can be considered are the development of a single GUI that will include all the steps of the recording process, such as the positioning of the patient, start of the recording, capturing of all SFGS poses, and any required manual processing steps until the final SFGS score is calculated. Due to the usage of a depth camera the positioning of the patient could be indicated on screen, where audiovisual markers could be used to direct the patient into the right position. After the positioning of the patient, the GUI could indicate which SFGS pose should be recorded, with the option to randomize the order of SFGS poses or to perform repeated measurements, where the final design choices will depend on the input of the multidisciplinary team.

From this initial development of the automated SFGS there are two major paths that can be taken. The first is to use the existing workflow to create a fully regulated software package for use in dedicated clinics, where the manual SFGS is completely replaced by the automated SFGS. This software package will be considered a medical device in many countries, which will have specific regulatory requirements, such as the implementation of a quality management system which will include topics such as the documentation of control and records, data management, software and product design, risk management, whilst following common good manufacturing processes [94–96]. In addition, there are specific legal frameworks that should be taken into account. For example, in the European Union the General Data Protection Regulation (GDPR), the Medical Device Regulation (MDR), the European Health Data Space (EDHS), and the European Artificial (AI) Intelligence Act will play an important

role [96]. The process of creating a fully regulated software package is expected to take a long time with a significant associated cost but would allow the usage of the automated SFGS as a standalone application in clinics where the manual SFGS can be fully replaced. However, the final aim of the automated SFGS is the ability to implement a user-friendly system that can be used by the patient at home without any assistance. Therefore, the second path forward would focus on the home implementation of the automated SFGS, which would most likely include all the requirements of the application in a clinical setting but is expected to face more challenges. Certain questions might arise such as whether patients will be required to login to be able to access the software and decrypt potential raw footage and SFGS scores available on the local device. Is the raw footage going to be stored locally at all and is the patient allowed to view the footage or SFGS scores? Is there going to be a direct connection to hospital servers to update the medical records with the automated SFGS score and does this include any raw data that can be used to further train the deep learning model? In case the RealSense cameras are used for the recordings, how would these be financed and what are the logistics for lending and retrieving these cameras? Will there be support for multiple depth cameras to alleviate some of these issues or are there going to be multiple parallel deep learning models that can accept 2D or depth data, each with their own caveats? From these questions it is clear that there is a significant amount of research required by the multidisciplinary team during the implementation of the automated SFGS in a home setting. However, the potential benefits of a home implementation of the automated SFGS should be taken into consideration as it could offer a more personalized rehabilitation process for the patient. For example, the patient would be able to select their preferred time to take the SFGS measurement, without the need to travel or being restricted to regular business hours. Depending on the preference of the patient, the automated SFGS could also provide a more objective overview of the rehabilitation process by comparing the video recordings and the corresponding SFGS score over time. The home implementation of the automated SFGS could also have a positive impact for clinicians and researchers and thereby benefiting patients. With a home implementation of the SFGS it would become easier to increase the frequency of the SFGS measurements. This could help to closely monitor changes in the PFP and determine the impact of any modifications to the treatment. Over time, this could even facilitate the development of a deep learning model designed to monitor the rehabilitation process of the patient and the identification of patients in need of closer supervision. This model might even be able to identify certain patterns in the data to individualize and improve the treatment of the patient. These actions would be performed in close collaboration with researchers, where the researchers would benefit from having access to reliable data input, with a higher temporal frequency, with SFGS scores directly comparable to studies implemented

elsewhere. It might even be possible to make the SFGS more sensitive to changes in the PFP. Currently, the individual elements of the symmetry of voluntary movement and synkinesis are based on an ordinal scale. E.g., there is no differentiation between a score of 2.8 and 3.2 for an individual element as both scores will result in a score of 3. A deep learning model could create a continuous scale of the individual SFGS elements identifying smaller changes of the PFP. However, these changes to the SFGS should be approached with care, as the benefits of using the existing SFGS would potentially be lost, such as the existing validation of the SFGS or the direct comparison to studies not using the adjusted automated SFGS.

Having discussed a broad range of potential improvements to the automated SFGS, we will now focus on the short-term and long-term recommendations for the future steps of the automated SFGS and the associated clinical impact. A first recommendation would be to add the patients recorded after November 2020 to the dataset, which was the inclusion cutoff used in this thesis. Adding these recordings to the dataset will give a more accurate indication of the performance of the automated SFGS where a separate validation and testing set can be used during the training of the model. As these recordings are already available this could be implemented in a relatively short timeframe. This would also be a good opportunity to evaluate the existing CNN model against a deep learning model incorporating the depth data and potentially revisiting the usage of transfer learning. In case the reliability of the CNN model is comparable to or exceeding the higher end of the reliability of human observers, the automated SFGS could immediately be implemented in a research setting where the outcome of the SFGS does not directly influence the treatment plan of the patient, and where the patient cohort is based on a similar patient demographic as used during the training of the CNN model. From here the focus can shift towards optimizing the workflow of the automated SFGS as discussed in the previous paragraphs, preferably in preparation of a multinational multi-centre study. The multi-centre study would result in a dataset of order of magnitude bigger than the current dataset and should be achievable in the medium-term. With a dataset this size it might be an opportune moment to revisit other deep learning techniques as discussed in the future perspectives. With a broader demographic included in the multi-centre study, the automated SFGS could be used as a multi-national research tool, which would allow for a more reliable comparison between studies incorporating the SFGS, without limitations on the availability of observers experienced in the SFGS. Whilst the multi-centre study is running, the remaining requirements for a validated and regulated automated SFGS should be investigated and implemented in the long-term. From this point on the manual SFGS can be completely replaced, which will allow monitoring at a much higher frequency without significantly impacting the workload for the clinical staff with a potentially higher reliability compared to experienced human observers.

REFERENCES

1. Kim, S. J. & Lee, H. Y. Acute Peripheral Facial Palsy: Recent Guidelines and a Systematic Review of the Literature. J. Korean Med. Sci. 35, e245 (2020).
2. Samsudin, W. S. W. & Sundaraj, K. Evaluation and Grading Systems of Facial Paralysis for Facial Rehabilitation. J. Phys. Ther. Sci. 25, 515–519 (2013).
3. Fattah, A. Y. et al. Facial nerve grading instruments: Systematic Review of the Literature and Suggestion for Uniformity. Plast. Reconstr. Surg. 135, 569–579 (2015).
4. Niziol, R., Henry, F. P., Leckenby, J. I. & Grobbelaar, A. O. Is there an ideal outcome scoring system for facial reanimation surgery? A review of current methods and suggestions for future publications. J. Plast. Reconstr. Aesthetic Surg. 68, 447–456 (2015).
5. Ross, B. G., Fradet, G. & Nedzelski, J. M. Development of a sensitive clinical facial grading system. Otolaryngol. neck Surg. 114, 380–386 (1996).
6. Berner, J. E., Kamalathevan, P., Kyriazidis, I. & Nduka, C. Facial synkinesis outcome measures: A systematic review of the available grading systems and a Delphi study to identify the steps towards a consensus. J. Plast. Reconstr. Aesthetic Surg. 72, 946–963 (2019).
7. van Veen, M. M., Bruins, T. E., Artan, M., Werker, P. M. N. & Dijkstra, P. U. Learning curve using the Sunnybrook Facial Grading System in assessing facial palsy: An observational study in 100 patients. Clin. Otolaryngol. 45, 823–826 (2020).
8. Liu, J. et al. Accuracy of 3-dimensional stereophotogrammetry: Comparison of the 3dMD and Bellus3D facial scanning systems with one another and with direct anthropometry. Am. J. Orthod. Dentofac. Orthop. 160, 862–871 (2021).
9. Siena, F. L., Byrom, B., Watts, P. & Breedon, P. Utilising the Intel RealSense Camera for Measuring Health Outcomes in Clinical Research. J. Med. Syst. 42, 53 (2018).
10. Dindaroğlu, F., Kutlu, P., Duran, G. S., Görgülü, S. & Aslan, E. Accuracy and reliability of 3D stereophotogrammetry: A comparison to direct anthropometry and 2D photogrammetry. Angle Orthod. 86, 487–494 (2016).
11. Maal, T. J. J. et al. Variation of the face in rest using 3D stereophotogrammetry. Int. J. Oral Maxillofac. Surg. 40, 1252–1257 (2011).
12. Maal, T. J. J. et al. Registration of 3-Dimensional Facial Photographs for Clinical Use. J. Oral Maxillofac. Surg. 68, 2391–2401 (2010).
13. Schipper, J. A. M. et al. Reliability and validity of handheld structured light scanners and a static stereophotogrammetry system in facial three-dimensional surface imaging. Sci. Rep. 14, 8172 (2024).
14. Servi, M. et al. Metrological Characterization and Comparison of D415, D455, L515 RealSense Devices in the Close Range. Sensors 21, (2021).
15. Halmetschlager-Funek, G., Suchi, M., Kampel, M. & Vincze, M. An empirical evaluation

of ten depth cameras: Bias, precision, lateral noise, different lighting conditions and materials, and multiple sensor setups in indoor environments. IEEE Robot. Autom. Mag. 26, 67–77 (2019).
16. Burger, L. et al. Comparative evaluation of three commercially available markerless depth sensors for close-range use in surgical simulation. Int. J. Comput. Assist. Radiol. Surg. 18, 1109–1118 (2023).
17. Vallen, H. et al. Three-dimensional stereophotogrammetry measurement of facial asymmetry in patients with congenital muscular torticollis: a non-invasive method. Int. J. Oral Maxillofac. Surg. 50, 835–842 (2021).
18. da Silva Neto, J. G., da Lima Silva, P. J., Figueredo, F., Teixeira, J. M. X. N. & Teichrieb, V. Comparison of RGB-D sensors for 3D reconstruction. in 2020 22nd Symposium on Virtual and Augmented Reality (SVR) 252–261 (2020). doi:10.1109/SVR51698.2020.00046.
19. Carfagni, M. et al. Metrological and Critical Characterization of the Intel D415 Stereo Depth Camera. Sensors 19, (2019).
20. Bajzik, J., Koniar, D., Hargas, L., Volak, J. & Janisova, S. Depth sensor selection for specific application. 13th Int. Conf. ELEKTRO 2020, ELEKTRO 2020 - Proc. 2020-May, (2020).
21. Gutta, V., Lemaire, E. D., Baddour, N. & Fallavollita, P. A Comparison of Depth Sensors for 3D Object Surface Reconstruction. C. Proc. 42, 4–7 (2019).
22. Chen, R., Xu, J. & Zhang, S. Comparative study on 3D optical sensors for short range applications. Opt. Lasers Eng. 149, 106763 (2022).
23. Curto, E. & Araujo, H. An Experimental Assessment of Depth Estimation in Transparent and Translucent Scenes for Intel RealSense D415, SR305 and L515. Sensors 22, (2022).
24. Toma, A. M., Zhurov, A., Playle, R., Ong, E. & Richmond, S. Reproducibility of facial soft tissue landmarks on 3D laser-scanned facial images. Orthod. Craniofacial Res. 12, 33–42 (2009).
25. Lin, H., Zhu, P., Lin, Y., Zheng, Y. & Xu, Y. Reliability and Reproducibility of Landmarks on Three-Dimensional Soft-Tissue Cephalometrics Using Different Placement Methods. Plast. Reconstr. Surg. 134, (2014).
26. Baysal, A., Sahan, A. O., Ozturk, M. A. & Uysal, T. Reproducibility and reliability of three-dimensional soft tissue landmark identification using three-dimensional stereophotogrammetry. Angle Orthod. 86, 1004–1009 (2016).
27. Fagertun, J. et al. 3D facial landmarks: Inter-operator variability of manual annotation. BMC Med. Imaging 14, 1–9 (2014).
28. Düppe, K., Becker, M. & Schönmeyr, B. Evaluation of Facial Anthropometry Using Three-Dimensional Photogrammetry and Direct Measuring Techniques. J. Craniofac. Surg. 29, (2018).

29. Petrides, G. A. et al. Introduction of a Low-Cost and Automated Four-Dimensional Assessment System of the Face. Plast. Reconstr. Surg. 150, 639e-643e (2022).
30. Heike, C. L., Upson, K., Stuhaug, E. & Weinberg, S. M. 3D digital stereophotogrammetry: a practical guide to facial image acquisition. Head Face Med. 6, 18 (2010).
31. Hobbs-Murphy, K., Olmedo-Nockideneh, I., Brazile, W. J., Morris, K. & Rosecrance, J. Intra-rater and inter-rater reliability of 3D facial measurements. Appl. Ergon. 116, 104218 (2024).
32. Fourie, Z., Damstra, J., Gerrits, P. O. & Ren, Y. Evaluation of anthropometric accuracy and reliability using different three-dimensional scanning systems. Forensic Sci. Int. 207, 127–134 (2011).
33. Koban, K. C. et al. 3D Anthropometric Facial Imaging - A comparison of different 3D scanners. Facial Plast. Surg. Clin. North Am. 30, 149–158 (2022).
34. Ayaz, I. et al. Accuracy and reliability of 2-dimensional photography versus 3-dimensional soft tissue imaging. Imaging Sci Dent 50, 15–22 (2020).
35. Ramieri, G. A. et al. Reconstruction of facial morphology from laser scanned data. Part I: Reliability of the technique. Dentomaxillofacial Radiol. 35, 158–164 (2006).
36. Wong, J. Y. et al. Validity and reliability of craniofacial anthropometric measurement of 3D digital photogrammetric images. Cleft Palate-Craniofacial J. 45, 232–239 (2008).
37. Ceinos, R., Tardivo, D., Bertrand, M. F. & Lupi-Pegurier, L. Inter- and Intra-Operator Reliability of Facial and Dental Measurements Using 3D-Stereophotogrammetry. J. Esthet. Restor. Dent. 28, 178–189 (2016).
38. Othman, S. A., Majawit, L. P., Wan Hassan, W. N., Wey, M. C. & Razi, R. M. Anthropometric study of three-dimensional facial morphology in Malay adults. PLoS One 11, 1–15 (2016).
39. Taghizadeh, M. & Chalechale, A. A comprehensive and systematic review on classical and deep learning based region proposal algorithms. Expert Syst. Appl. 189, 116105 (2022).
40. Wang, Y. et al. A systematic review on affective computing: emotion models, databases, and recent advances. Inf. Fusion 83–84, 19–52 (2022).
41. Alzubaidi, L. et al. Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. J. Big Data 8, 53 (2021).
42. Suganyadevi, S., Seethalakshmi, V. & Balasamy, K. A review on deep learning in medical image analysis. Int. J. Multimed. Inf. Retr. 11, 19–38 (2022).
43. Sharifani, K. & Amini, M. Machine Learning and Deep Learning: A Review of Methods and Applications. World Inf. Technol. Eng. J. 10, 3897–3904 (2023).
44. Singh, S. P. et al. 3D Deep Learning on Medical Images: A Review. Sensors 20, (2020).
45. Li, F. & Du, Y. Introduction: A Brief History of Deep Learning and Its Applications in Power Systems. in Deep Learning for Power System Applications: Case Studies Linking Artificial Intelligence and Power Systems 1–13 (Springer International

Publishing, Cham, 2024). doi:10.1007/978-3-031-45357-1_1.
46. Krizhevsky, A., Sutskever, I. & Hinton, G. E. ImageNet Classification with Deep Convolutional Neural Networks. in Advances in Neural Information Processing Systems (eds. Pereira, F., Burges, C. J., Bottou, L. & Weinberger, K. Q.) vol. 25 (Curran Associates, Inc., 2012).
47. Cabrol, C. et al. Sunnybrook Facial Grading System: Intra-rater and Inter-rater Variabilities. Otol. Neurotol. 42, (2021).
48. Coulson, S. E., Croxson, G. R., Adams, R. D. & O'Dwyer, N. J. Reliability of the 'Sydney,' 'Sunnybrook,' and 'House Brackmann' facial grading systems to assess voluntary movement and synkinesis after facial nerve paralysis. Otolaryngol. - Head Neck Surg. 132, 543–549 (2005).
49. Neely, J. G., Cherian, N. G., Dickerson, C. B. & Nedzelski, J. M. Sunnybrook facial grading system: reliability and criteria for grading. Laryngoscope 120, 1038–1045 (2010).
50. Volk, G. F. et al. Reliability of grading of facial palsy using a video tutorial with synchronous video recording. Laryngoscope 129, 2274–2279 (2019).
51. Gaudin, R. A. et al. Emerging vs time-tested methods of facial grading among patients with facial paralysis. JAMA Facial Plast. Surg. 18, 251–257 (2016).
52. Tan, J. R., Coulson, S. & Keep, M. Face-to-face versus video assessment of facial paralysis: Implications for telemedicine. J. Med. Internet Res. 21, (2019).
53. Wang, T., Dong, J., Sun, X., Zhang, S. & Wang, S. Automatic recognition of facial movement for paralyzed face. Biomed. Mater. Eng. 24, 2751–2760 (2014).
54. Wang, T. et al. Automatic evaluation of the degree of facial nerve paralysis. Multimed. Tools Appl. 75, 11893–11908 (2016).
55. Kim, H. S., Kim, S. Y., Kim, Y. H. & Park, K. S. A smartphone-based automatic diagnosis system for facial nerve palsy. Sensors (Switzerland) 15, 26756–26768 (2015).
56. Guo, Z. et al. An unobtrusive computerized assessment framework for unilateral peripheral facial paralysis. IEEE J. Biomed. Heal. Informatics 22, 835–841 (2018).
57. Guo, Z. et al. Deep assessment process: Objective assessment process for unilateral peripheral facial paralysis via deep convolutional neural network. Proc. - Int. Symp. Biomed. Imaging 135–138 (2017) doi:10.1109/ISBI.2017.7950486.
58. Sajid, M. et al. Automatic grading of palsy using asymmetrical facial features: A study complemented by new solutions. Symmetry (Basel). 10, (2018).
59. Song, A., Wu, Z., Ding, X., Hu, Q. & Di, X. Neurologist Standard Classification of Facial Nerve Paralysis with Deep Neural Networks. Futur. Internet 10, 111 (2018).
60. Hsu, G. S. J., Huang, W. F. & Kang, J. H. Hierarchical network for facial palsy detection. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Work. 2018-June, 693–699 (2018).
61. Hsu, G. S. J. & Chang, M. H. Deep Hybrid Network for Automatic Quantitative Analysis of Facial Paralysis. Proc. AVSS 2018 - 2018 15th IEEE Int. Conf. Adv. Video Signal-

Based Surveill. 1–7 (2019) doi:10.1109/AVSS.2018.8639156.
62. Bur, A. M., Shew, M. & New, J. Artificial Intelligence for the Otolaryngologist: A State of the Art Review. Otolaryngol. - Head Neck Surg. (United States) 160, 603–611 (2019).
63. Mothes, O. et al. Automated objective and marker-free facial grading using photographs of patients with facial palsy. Eur. Arch. Oto-Rhino-Laryngology (2019) doi:10.1007/s00405-019-05647-7.
64. Zhuang, Y. et al. F-DIT-V: An Automated Video Classification Tool for Facial Weakness Detection. 2019 IEEE EMBS Int. Conf. Biomed. Heal. Informatics 1–4 (2019) doi:10.1109/bhi.2019.8834563.
65. Jirawatnotai, S., Jomkoh, P., Voravitvet, T. Y., Tirakotai, W. & Somboonsap, N. Computerized Sunnybrook facial grading scale (SBface) application for facial paralysis evaluation. Arch. Plast. Surg. 48, 269–277 (2021).
66. Guarin, D. L. et al. Toward an Automatic System for Computer-Aided Assessment in Facial Palsy. Facial Plast. Surg. aesthetic Med. 22, 42–49 (2020).
67. Mitchell, D. T., Allen, D. Z., Greives, M. R. & Nguyen, P. D. A Critical Assessment and Review of Artificial Intelligence in Facial Paralysis Analysis: Uncovering the Truth. FACE 2, 200–207 (2021).
68. Alagha, M. A., Ayoub, A., Morley, S. & Ju, X. Objective grading facial paralysis severity using a dynamic 3D stereo photogrammetry imaging system. Opt. Lasers Eng. 150, 106876 (2022).
69. Ojha, P. T. et al. Validation of a New Graphic Facial Nerve Grading System: FAME Scale. Ann. Indian Acad. Neurol. 25, (2022).
70. Knoedler, L. et al. A Ready-to-Use Grading Tool for Facial Palsy Examiners—Automated Grading System in Facial Palsy Patients Made Easy. J. Pers. Med. 12, (2022).
71. Alagha, M. A., Ju, X., Morley, S. & Ayoub, A. F. Mathematical Validation of the Modified Sunnybrook Facial Grading System Using Four-dimensional Imaging. J. Plast. Reconstr. Surg. 2, 77–88 (2023).
72. Raj, A. et al. Automatic and Objective Facial Palsy Grading Index Prediction Using Deep Feature Regression. in Medical Image Understanding and Analysis (eds. Papież, B. W., Namburete, A. I. L., Yaqub, M. & Noble, J. A.) 253–266 (Springer International Publishing, Cham, 2020).
73. Jiang, Z., Dai, W., Wang, W. & Wang, W. A Cloud-Based Training and Evaluation System for Facial Paralysis Rehabilitation. Proc. - IEEE 16th Int. Conf. Ind. Informatics, INDIN 2018 701–706 (2018) doi:10.1109/INDIN.2018.8471934.
74. Ali, W. et al. A Transfer Learning Approach for Facial Paralysis Severity Detection. IEEE Access 11, 127492–127508 (2023).
75. Wang, J., Zhu, H., Wang, S.-H. & Zhang, Y.-D. A Review of Deep Learning on Medical Image Analysis. Mob. Networks Appl. 26, 351–380 (2021).
76. Wang, S. et al. Machine/Deep Learning for Software Engineering: A Systematic

Literature Review. IEEE Trans. Softw. Eng. 49, 1188–1231 (2023).
77. Simonyan, K. & Zisserman, A. Very deep convolutional networks for large-scale image recognition. 3rd Int. Conf. Learn. Represent. ICLR 2015 - Conf. Track Proc. 1–14 (2015).
78. Eelbode, T., Sinonquel, P., Maes, F. & Bisschops, R. Pitfalls in training and validation of deep learning systems. Best Pract. Res. Clin. Gastroenterol. 52–53, 101712 (2021).
79. Vabalas, A., Gowen, E., Poliakoff, E. & Casson, A. J. Machine learning algorithm validation with a limited sample size. PLoS One 14, 1–20 (2019).
80. Bansal, M. A., Sharma, D. R. & Kathuria, D. M. A Systematic Review on Data Scarcity Problem in Deep Learning: Solution and Applications. ACM Comput. Surv. 54, (2022).
81. Li, H. et al. Research on Overfitting of Deep Learning. in 2019 15th International Conference on Computational Intelligence and Security (CIS) 78–81 (2019). doi:10.1109/CIS.2019.00025.
82. Bejani, M. M. & Ghatee, M. A systematic review on overfitting control in shallow and deep neural networks. Artif. Intell. Rev. 54, 6391–6438 (2021).
83. Caple, J. & Stephan, C. N. A standardized nomenclature for craniofacial and facial anthropometry. Int. J. Legal Med. 130, 863–879 (2016).
84. Bodini, M. A Review of Facial Landmark Extraction in 2D Images and Videos Using Deep Learning. Big Data Cogn. Comput. 3, (2019).
85. Khabarlak, K. & Koriashkina, L. Fast Facial Landmark Detection and Applications: A Survey. J. Comput. Sci. Technol. 22, e02 (2022).
86. Flores, M. R. P. et al. Comparative Assessment of a Novel Photo-Anthropometric Landmark-Positioning Approach for the Analysis of Facial Structures on Two-Dimensional Images. J. Forensic Sci. 64, 828–838 (2019).
87. Knoops, P. G. M. et al. Comparison of three-dimensional scanner systems for craniomaxillofacial imaging. J. Plast. Reconstr. Aesthetic Surg. 70, 441–449 (2017).
88. Blahnik, V. & Schindelbeck, O. Smartphone imaging technology and its applications. Adv. Opt. Technol. 10, 145–232 (2021).
89. Hunt, B., Ruiz, A. J. & Pogue, B. W. Smartphone-based imaging systems for medical applications: a critical review. J. Biomed. Opt. 26, 40902 (2021).
90. Arnab, A. et al. ViViT: A Video Vision Transformer. Proc. IEEE Int. Conf. Comput. Vis. 6816–6826 (2021) doi:10.1109/ICCV48922.2021.00676.
91. Han, K. et al. A Survey on Vision Transformer. IEEE Trans. Pattern Anal. Mach. Intell. 45, 87–110 (2023).
92. Liu, Z. et al. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. in 2021 IEEE/CVF International Conference on Computer Vision (ICCV) 9992–10002 (IEEE Computer Society, Los Alamitos, CA, USA, 2021). doi:10.1109/ICCV48922.2021.00986.
93. Maurício, J., Domingues, I. & Bernardino, J. Comparing Vision Transformers and

Convolutional Neural Networks for Image Classification: A Literature Review. Appl. Sci. 13, (2023).
94. Muehlematter, U. J., Daniore, P. & Vokinger, K. N. Approval of artificial intelligence and machine learning-based medical devices in the USA and Europe (2015–20): a comparative analysis. Lancet Digit. Heal. 3, e195–e203 (2021).
95. Petersen, E. et al. Responsible and Regulatory Conform Machine Learning for Medicine: A Survey of Challenges and Solutions. IEEE Access 10, 58375–58418 (2022).
96. Pecchia, L., Maccaro, A., Matarrese, M. A. G., Folkvord, F. & Fico, G. Artificial Intelligence, data protection and medical device regulations: squaring the circle with a historical perspective in Europe. Health Technol. (Berl). 14, 663–670 (2024).

CHAPTER 7

APPENDICES

SUMMARY

The human face plays an integral part of verbal and non-verbal communication in everyday life, such as during the expression of emotions and thoughts. A disturbance in these functions can cause major effects to the physical, social, and emotional quality of life for affected individuals. One of these conditions that can affect facial functioning is a unilateral peripheral facial palsy (PFP). The PFP causes a partial or complete loss in the facial muscle functionality on a single side of the face. There are numerous factors which can affect the facial nerve functioning, such as trauma, herpes zoster, or diabetic mellitus. However, the majority of PFP cases are classified as idiopathic facial palsy, which means that the exact cause is unknown. The treatment and expected recovery rate of a PFP will, among other factors, depend on the severity of the PFP. Therefore, it is crucial to assess the severity of the PFP and monitor this over time. There are several grading systems available to determine the severity of a PFP. One of the recommended and well-established grading systems is the Sunnybrook Facial Grading System (SFGS). This recommendation is due to the high reliability, sensitivity to changes in the severity of the PFP, and the clinical relevance of the SFGS. The SFGS achieves these properties by assessing the facial nerve function of the muscles most important for facial expression. The SFGS assesses six different facial poses which consist of the face at rest as well as five voluntary movements; the forehead wrinkle, gentle eye closure, open mouth smile, snarl (raising of the nostrils), and lip pucker. A total of 13 elements are individually assessed and grouped into three subcomponents; the face at rest, the symmetry of voluntary movement, and synkinesis. Synkinesis refers to the involuntary activation of a part of the face, during a deliberate movement of a different part of the face. Each of these three subcomponents results in a subscore, which together determine the composite SFGS score. The composite SFGS score is a point scale ranging from 0 to 100, where the score of 0 indicates a complete PFP and a score of 100 indicates normal functioning of the facial muscles. A more detailed overview of the SFGS is given in **Chapter 1**. Although the SFGS is a recommended grading system to determine the severity of a PFP, the SFGS is a subjective grading system, which is influenced by the individual input of the human observer. Therefore, the grading will depend on the experience of the human observer due to the learning curve associated with the SFGS, which may bias the SFGS grading. This learning curve makes the SFGS less accessible or even inaccessible for researchers, students, clinicians in training, other untrained co-workers, or in the usage for the home monitoring of patients. Therefore, this thesis investigated the automation of the SFGS with the long-term aim to develop a user-friendly system that could be used by the patient at home without any assistance. In an ideal situation, this automated system would be more reliable than human observers experienced in the grading of PFPs, using the SFGS. A more detailed background and rationale for this aim is presented in **Chapter 1**.

In order to automatically determine the SFGS score a certain input will be required for the automated system. As the manual SFGS is based on visual examination, the choice was made to use a non-invasive imaging technology. Due to the complex nature of the face and the potential changes in the facial surface during the execution of the SFGS poses, the usage of a series of three-dimensional (3D) images could be of additional value for the automated SFGS. To make the automated system applicable for home use, the 3D video camera (4D) should be relatively inexpensive, portable, and easy to use. A 4D camera that met these requirements was the RealSense F200 (Intel, Santa Clara, USA). The RealSense F200 can simultaneously capture a 2D colour video and a 3D depth video in the form factor of a webcam with a price of $100 USD. Professional systems, such as the clinically validated 3dMD system (3dMDface, 3dMD, Atlanta, USA), often have a cost of tens of thousands of dollars, which most likely will cause a quality difference between the RealSense F200 and a professional system. Therefore, **Chapter 2** determined the depth accuracy of the RealSense F200 in a clinical setting by recording 34 patients with a PFP performing the six SFGS poses. Each patient was simultaneously recorded by the RealSense F200 and the 3dMD system. The depth accuracy of the RealSense F200 was determined at moment of maximum exertion for each of the SFGS poses by comparing the depth images from the RealSense F200 to the depth images from the 3dMD system. This analysis started with the alignment of the two depth images to match their position and rotation in 3D space. After this alignment, the depth accuracy of the RealSense F200 image was determined by calculating the distance from each point of the RealSense depth image to the closest point on the 3dMD image. The RealSense F200 depth image typically consists of thousands of points and this distance calculation was performed for each individual point. From these measurements an average depth accuracy was calculated for each patient and SFGS pose separately. This resulted in an average depth accuracy of 1.48 mm for the face at rest and 1.49 mm during the voluntary movement of the SFGS for the RealSense F200. It was statistically determined that the SFGS poses did not significantly influence the depth accuracy of the RealSense F200. When compared to the reported depth accuracy for the 3dMD system which ranged from 0.20 mm to 0.25 mm for the face at rest for healthy individuals, there is a clear difference in depth accuracy between the two systems. This result was not unexpected due to the difference in cost, size, and complexity of the two systems. Taken these factors in account, the RealSense F200 provided relatively reliable and accurate depth data when recording a range of facial movements and was considered a viable option as a portable and low-cost 4D camera for the automated SFGS.

After the first generation of RealSense cameras, a successor of the RealSense F200 was released, the RealSense D415. Due to significant improvements to the software and hardware of the camera the decision was made to switch to the RealSense D415. This meant the depth accuracy of the RealSense D415 needed to be determined.

However, the automated SFGS might not require the entire depth image of the face. Instead, the tracking of 3D facial landmarks might be sufficient. Additionally, it would be possible to perform facial measurements, also called anthropometric measurements, based on these 3D landmarks. Therefore, **Chapter 3** validated the RealSense D415 on three different topics; the depth accuracy of the entire depth image, the reliability of 3D landmark placement, and the reliability and agreement of 3D anthropometric measurements. The reliability is defined as the consistency of results when a measurement is repeated. The agreement is defined as how close a measurement is to the gold standard; in this case these are the measurements performed on the 3dMD image. The validation of each of the three topics was based on the same population of 30 patients with a PFP performing the six SFGS poses, simultaneously recorded with the RealSense D415 and 3dMD system. The methodology to determine the depth accuracy of the RealSense D415 was consistent with the methodology as described in **Chapter 2** in order to compare the depth accuracy of the RealSense F200. The facial landmarks were manually placed by two human observers on both the RealSense D415 and 3dMD depth images at the moment of maximum exertion of each SFGS pose. The average depth accuracy of the RealSense D415 was 0.97 mm for the face at rest and 0.98 mm for the voluntary movements. This depth accuracy improved compared to the RealSense F200, making the RealSense D415 a viable successor for the automated SFGS. The average reliability of the 3dMD landmark placement was 0.92 mm for the face at rest and 1.03 mm for the voluntary movements, which was within the expected range compared to the landmark placement on healthy subjects at rest. This indicated that the 3dMD results could be used as reference values for the RealSense D415. The reliability of the RealSense D415 landmark placement declined to 1.47 mm for the face at rest and to 1.28 mm during the voluntary movements, due to the reduced depth and colour quality of the RealSense compared to the 3dMD system. This effect was also visible in the anthropometric measurements, where the average reliability of the RealSense D415 measurements fell outside the reported range based on healthy subjects recorded in rest with a professional 3D system. This resulted in an average underestimation of the anthropometric measurements of -0.90 mm for the face at rest and -1.12 mm for the voluntary movements compared to the 3dMD measurements, where 95% of the RealSense D415 measurements were within a 5 mm error of the 3dMD measurements. In comparison, the 3dMD anthropometric measurements can achieve submillimetre accuracy compared to direct anthropometric measurements. Therefore, there is a clear difference between the accuracy of the RealSense D415 and the 3dMD system, where the specific clinical application will determine whether the performance of the RealSense D415 is sufficient. Overall, the results were considered as reasonable when the size and cost of the RealSense D415 were taken into account.

After the validation of the RealSense D415 depth accuracy and the 3D facial landmark placement with their derived anthropometric measurements, a first automated SFGS was implemented in **Chapter 4**. The automated SFGS was based on a convolutional neural network (CNN), which is a type of deep learning network especially suited for the automation of image processing tasks. One of the common methods to train a CNN is by presenting a large amount of input data with a known output. During the training process of the CNN model, the input data is used to identify patterns in the dataset and generalize this knowledge. This allows the trained model to generate the desired output when unseen data from different situations is introduced. This translated into using the 2D video recordings from the RealSense D415 as an input for the automated SFGS, with the known SFGS score as an output. The SFGS score was determined by three clinicians experienced in the grading of patients with a PFP based on the SFGS. The CNN was trained on the RealSense D415 recordings from 116 patients with a PFP and 9 healthy subjects performing the SFGS poses. Each of the 13 SFGS elements was trained on a separate CNN, resulting in 13 scores for the SFGS elements. From these elements the SFGS subscores and SFGS composite score were calculated, which is the same process when performing the SFGS manually. The reliability of the CNN was determined by comparing the automated SFGS score to the SFGS score from the three experienced human observers. The reliability was expressed as the intra-class correlation coefficient (ICC), which represents the reliability with a value between 0 and 1. An ICC value close to 1 would indicate a high reliability between the scores of the automated SFGS and the human observers, which would be the desirable result. The reliability for each SFGS element, subscore and composite score were determined, where an average ICC of 0.87 was found for the composite SFGS which is considered a good agreement. The reported ICC of the SFGS composite score for human observers ranges between 0.81 to 1.00 with an average of 0.91. This meant that the reliability of the automated SFGS was similar to human observers manually performing the SFGS.

The first implementation of the automated SFGS showed promising results and **Chapter 5** explored a further optimization of the CNN model by using manually placed 2D facial landmarks as an additional input. These landmarks could help focus on the relevant regions of the face during the training of the CNN. To compare the results to the first automated SFGS as many variables were kept consistent with **Chapter 4**. For example, the exact same dataset with 125 subjects and their respective SFGS grades as determined by the three expert human observers were used during training of the CNN. The main difference compared to the old CNN model were the additional fourth and fifth input, consisting of 2D images of the facial landmarks during rest and maximum exertion of the SFGS pose. The analysis of the reliability was kept consistent with the

previous chapter as well, which compared the automated SFGS to the human observers, expressed as the ICC. The addition of the input with the 2D facial landmarks resulted in an increase in reliability of the automated SFGS, where the overall ICC for the composite SFGS increased from 0.87 to 0.91, which is considered an excellent agreement. This meant the automated SFGS was grading with a similar reliability as experienced human observers, without increasing the size of the underlying dataset. Apart from improving the reliability for the automated SFGS, these optimizations show the potential impact for clinical applications where a limited sized dataset is available.

The results presented in this thesis are a first step towards the implementation of an automated SFGS in a clinical setting. However, there are areas of research that have not been addressed in this thesis. First of all, not all available data from the RealSense recordings were used in the development of the automated SFGS. This thesis has implemented a baseline model based on 2D data to show the potential of a deep learning network. The depth data, 3D landmarks, and 3D anthropometric measurements from the RealSense D415, which have been validated in this thesis, could potentially improve the reliability of the automated SFGS. As it can take a significant amount of time to build a reasonable sized dataset, it was preferred to immediately capture 4D recordings, despite not using these data in the current automated SFGS. There are also areas of research that should be further explored with a multi-disciplinary team before the automated SFGS can be implemented in daily clinical practise. For example, the regulatory requirements and relevant legal frameworks should be taken into account before the deployment of the automated SFGS, especially when the automated SFGS influences the treatment plan of the patient. This most likely will require the further validation of the reliability of the automated SFGS with the inclusion of more subjects. The expansion of the dataset would preferable be done in different settings, such as in an eHealth environment or in other clinical centres. This requires the development of an easy-to-use interface, where the required manual steps are minimized. A larger dataset would make the further development of the underlying techniques of the automated SFGS an interesting option, where new CNN or deep learning architectures could be implemented, with the goal to further increase the reliability of the automated SFGS. In combination with the work presented in this thesis these efforts could result in the implementation of an automated SFGS in research, daily clinical practice, and by the patient at home in an eHealth environment, with a higher reliability compared to human observers.

SAMENVATTING

Het menselijk aangezicht speelt een integrale rol tijdens verbale en non-verbale communicatie in het dagelijks leven, zoals tijdens de expressie van emoties en gedachten. Een verstoring in deze functies kan grote gevolgen veroorzaken in de fysieke, sociale en emotionele kwaliteit in het leven van de aangedane individuen. Eén van de condities die het functioneren van het gezicht kan aantasten is een unilaterale perifere aangezichtsverlamming (PAV). De PAV zorgt voor een gedeeltelijk of volledig verlies van de functie van de aangezichtsspieren aan een enkele zijde van het gezicht. Er zijn meerdere factoren die de werking van de gezichtszenuw kunnen aantasten, zoals een trauma, herpes zoster, of diabetes mellitus. Echter, de meerderheid van de PAV gevallen wordt geclassificeerd als een idiopathische aangezichtsverlamming, wat inhoudt dat de exacte oorzaak onbekend is. De behandeling en het verwachte herstel van een PAV hangt onder andere af van de ernst van de PAV. Daarom is het cruciaal om de ernst van de PAV te bepalen en deze over de tijd te monitoren. Er zijn meerdere graderingsystemen beschikbaar om de ernst van de PAV te bepalen. Eén van de aanbevolen en gevestigde graderingsystemen is de Sunnybrook Facial Grading System (SFGS). Deze aanbeveling is te danken aan de hoge betrouwbaarheid, gevoeligheid voor veranderingen in de ernst van de PAV, en de klinische relevantie van de SFGS. De SFGS behaalt deze eigenschappen door het beoordelen van de spieren die het belangrijkst zijn voor gezichtsuitdrukking om zo het functioneren van de aangezichtszenuw (n. facialis) in kaart te brengen. De SFGS behaalt deze eigenschappen door zes verschillende mimische posities van het aangezicht te beoordelen. De mimische posities bestaan uit het gezicht in rust en vijf vrijwillige bewegingen: het optrekken van de wenkbrauwen, het rustig sluiten van de ogen, een glimlach met open mond, snauwen (het optrekken van de neusvleugels), en het tuiten van de lippen. Er worden in totaal 13 individuele elementen beoordeeld die gegroepeerd zijn binnen drie onderdelen: het aangezicht in rust, de symmetrie van beweging en de synkinese. Synkinese verwijst hier naar de onvrijwillige activatie van een deel van de mimische spieren, tijdens een bewuste beweging van een ander deel van de mimische spieren. Elk van deze drie onderdelen resulteert in een gedeeltelijke score, welke samen de totale SFGS score bepalen. De totale SFGS score is een puntenschaal welke tussen een waarde van 0 en 100 ligt, waarbij de score van 0 een volledige PAV aangeeft en een score van 100 een normale aangezichtsfunctie. Een uitgebreid overzicht van de SFGS kan worden gevonden in **Hoofdstuk 1**. Ondanks dat de SFGS een aanbevolen graderingsysteem is om de ernst van een PAV te bepalen, wordt de SFGS beïnvloed door de individuele input van de menselijke beoordelaar aangezien de SFGS een subjectief graderingsysteem is. Daarom hangt de kwaliteit van de gradering af van de ervaring van de menselijke beoordelaar door de bijbehorende leercurve van de SFGS, welke de SFGS gradering kan vertekenen. Deze leercurve maakt de SFGS minder toegankelijk of zelfs

ontoegankelijk voor onderzoekers, studenten, clinici in opleiding, andere ongetrainde collega's of in het gebruik van de thuismonitoring van patiënten. Daarom onderzoekt dit proefschrift de automatisering van de SFGS met het lange-termijn doel om een gebruikersvriendelijk systeem te ontwikkelen welke door de patiënt thuis gebruikt kan worden zonder enige externe hulp. Dit geautomatiseerd systeem zou idealiter een hogere betrouwbaarheid hebben ten opzichte van menselijke beoordelaars ervaren in het graderen van de PAV met behulp van de SFGS. Een meer gedetailleerde achtergrond en redenatie voor dit doel wordt uiteengezet in **Hoofdstuk 1**.

Om de SFGS score automatisch te kunnen bepalen zal er een bepaalde vorm van input nodig zijn voor het automatische systeem. Aangezien de handmatige SFGS is gebaseerd op visuele inspectie, werd ervoor gekozen om een non-invasieve beeldvormingstechnologie te gebruiken. Door de complexe aard van het aangezicht en de mogelijke veranderingen in het gezichtsoppervlak tijdens het uitvoeren van de vrijwillige SFGS bewegingen, zou het gebruik van een reeks driedimensionale (3D) beelden een toegevoegde waarde kunnen hebben voor de geautomatiseerde SFGS. Om het automatische systeem toepasbaar te maken voor thuisgebruik, zou de 3D videocamera (4D) relatief goedkoop, draagbaar en makkelijk in gebruik moeten zijn. Een 4D camera die aan deze eisen voldeed was de RealSense F200 (Intel, Santa Clara, USA). De RealSense F200 is in staat om tegelijkertijd een 2D kleuren video en een 3D diepte video op te nemen, terwijl de RealSense F200 het formaat heeft van een webcam met een prijs van $100 USD. Professionele systemen, zoals het klinisch gevalideerde 3dMD systeem (3dMDface, 3dMD, Atlanta, USA), hebben vaak een prijs van tienduizenden dollars, waardoor er naar alle waarschijnlijkheid een kwaliteitsverschil zal zitten tussen de RealSense F200 en een professioneel systeem. Daarom bepaalde **Hoofdstuk 2** de diepte accuraatheid van de RealSense F200 in een klinische omgeving, door een video opname te maken met de RealSense F200 van 34 patiënten met een PAV tijdens het uitvoeren van de zes SFGS mimische posities. Elke patiënt werd tegelijkertijd opgenomen door de RealSense F200 en het 3dMD systeem. De diepteaccuraatheid van de RealSense F200 werd bepaald op het moment van maximale uitslag van iedere SFGS mimische positie, door de dieptebeelden van de RealSense F200 te vergelijken met de dieptebeelden van het 3dMD systeem. Deze analyse begon met het uitlijnen van de twee dieptebeelden om de positie en rotatie overeen te laten komen in de 3D ruimte. Na deze uitlijning werd de diepteaccuraatheid van de RealSense F200 bepaald door de afstand van elk punt van het RealSense dieptebeeld naar het dichtstbijzijnde punt van het 3dMD dieptebeeld uit te rekenen. Het RealSense F200 dieptebeeld bestaat typisch gezien uit duizenden punten, en deze afstandsberekening werd bepaald voor elk individueel punt. Van deze metingen werd er een gemiddelde diepteaccuraatheid bepaald voor elke afzonderlijke patiënt en SFGS mimische positie. Dit resulteerde in een gemiddelde diepteaccuraatheid van 1,48 mm voor het gezicht in rust en 1,49 mm

tijdens de vrijwillige bewegingen van de SFGS voor de RealSense F200. Uit de statistische analyse bleek dat de SFGS mimische posities geen significante invloed hadden op de diepteaccuraatheid van de RealSense F200. Wanneer de accuraatheid werd vergeleken met de gerapporteerde diepteaccuraatheid van het 3dMD systeem, welke tussen de 0,20 mm en 0,25 mm lag voor het gezicht in rust bij gezonde individuen, was er een duidelijk verschil te zien in de diepteaccuraatheid tussen de twee systemen. Dit resultaat was niet onverwacht gezien het verschil in kosten, formaat en complexiteit van de twee systemen. Rekening houdend met deze factoren voorzag de RealSense F200 van een relatief betrouwbare en accurate diepte data tijdens het opnemen van een scala aan aangezichtsbewegingen en werd daarom gezien als een goede optie voor een draagbare en goedkope 4D camera voor de automatisering van de SFGS.

Na de eerste generatie van RealSense camera's werd er een opvolger van de RealSense F200 uitgebracht, de RealSense D415. Door significante verbeteringen aan de hardware en software, werd er besloten om de overstap te maken naar de RealSense D415. Dit betekende dat de diepteaccuraatheid van de RealSense D415 bepaald moest worden. Echter, de geautomatiseerde SFGS heeft mogelijk niet het volledige dieptebeeld van het aangezicht nodig. In plaats daarvan zou het mogelijk voldoende zijn om specifieke oriëntatiepunten in het gezicht te volgen, oftewel facial landmarks. Daarnaast zou het mogelijk zijn om metingen in het gezicht uit te voeren, ook wel antropometrische metingen genoemd, gebaseerd op deze 3D facial landmarks. Daarom valideerde **Hoofdstuk 3** de RealSense D415 op drie verschillende onderwerpen: de diepteaccuraatheid van het gehele dieptebeeld, de betrouwbaarheid van de 3D facial landmarkplaatsing, en de betrouwbaarheid en overeenkomst van de 3D antropometrische metingen. De betrouwbaarheid is gedefinieerd als de consistentie van de uitkomst wanneer een meting wordt herhaald. De overeenkomst is gedefinieerd als hoe dicht een meting bij de gouden standaard ligt, in dit geval de metingen uitgevoerd op het 3dMD dieptebeeld. De validatie van deze drie onderwerpen was gebaseerd op dezelfde populatie van 30 PAV patiënten welke de zes SFGS mimische posities uitvoerden, tegelijkertijd opgenomen met de RealSense D415 en het 3dMD systeem. De methodologie om de diepteaccuraatheid van de RealSense D415 te bepalen was consistent met de methodologie zoals beschreven in **Hoofdstuk 2** om de vergelijking met de diepteaccuraatheid van de RealSense F200 te kunnen maken. De facial landmarks werden handmatig geplaatst door twee menselijke beoordelaars op de RealSense en 3dMD dieptebeelden op het moment van maximale uitslag van iedere SFGS mimische positie. De gemiddelde diepteaccuraatheid voor de RealSense D415 was 0,97 mm voor het gezicht in rust en 0,98 mm voor de vrijwillige bewegingen. Deze diepteaccuraatheid was beter dan die van de RealSense F200, wat de RealSense D415 een geschikte opvolger maakte voor de automatische SFGS. De gemiddelde betrouwbaarheid voor

de 3dMD landmarkplaatsing was 0,92 mm voor het gezicht in rust en 1,03 mm voor de vrijwillige bewegingen, wat binnen het verwachte bereik was ten opzichte van de landmarkplaatsing bij de gezonde personen in rust. Dit toonde aan dat de resultaten van de 3dMD gebruikt konden worden als referentiewaardes voor de RealSense D415. De betrouwbaarheid van de RealSense D415 landmarkplaatsing nam af tot 1,47 mm voor het gezicht in rust en tot 1,28 mm tijdens de vrijwillige bewegingen, veroorzaakt door de lagere diepte- en kleurkwaliteit van het RealSense D415 systeem ten opzichte van het 3dMD systeem. Dit effect was ook zichtbaar tijdens de antropometrische metingen, waar de gemiddelde betrouwbaarheid van de RealSense D415 metingen buiten het gerapporteerde bereik viel op metingen van gezonde personen in rust, opgenomen met een professioneel 3D systeem. Dit zorgde voor een gemiddelde onderschatting van de antropometrische metingen van -0,90 mm voor het gezicht in rust en -1,12 mm voor de vrijwillige bewegingen ten opzichte van de 3dMD metingen, waar 95% van de RealSense D415 metingen binnen een 5 mm foutmarge vielen van de 3dMD metingen. Ter vergelijking, de 3dMD antropometrische metingen kunnen een accuraatheid onder één millimeter behalen ten opzichte van directe antropometrische metingen. Daarom is er een duidelijk verschil tussen de accuraatheid van de RealSense D415 en het 3dMD systeem, waar de specifieke klinische toepassing zal bepalen of de prestaties van de RealSense D415 voldoende zijn. Over het geheel genomen werden de resultaten als redelijk gezien, wanneer er rekening werd gehouden met het formaat en de kosten van de RealSense D415.

Na de validatie van de diepteaccuraatheid van de RealSense D415 en de plaatsing van de 3D facial landmarks en de afgeleide antropometrische metingen, werd er een eerste geautomatiseerde SFGS geïmplementeerd in **Hoofdstuk 4**. De geautomatiseerde SFGS was gebaseerd op een convolutional neural network (CNN), wat een type deep learning netwerk is, specifiek geschikt voor het automatiseren van taken gerelateerd aan beeldverwerking. Eén van de gebruikelijke methodes om een CNN te trainen, is door het aanbieden van een grote hoeveelheid input data met een bekende output. Tijdens het trainingsproces van het CNN model wordt de input data gebruikt om patronen in de dataset te identificeren en om deze kennis te generaliseren. Dit maakt het mogelijk voor het getrainde model om de gewenste output te genereren wanneer nieuwe data uit verschillende situaties wordt geïntroduceerd. Dit vertaalde zich in het gebruik van 2D video opnames van de RealSense D415 voor de input van de geautomatiseerde SFGS, met de SFGS score als de output. De SFGS score was bepaald door drie clinici ervaren in het graderen van patiënten met een PAV, gebaseerd op de SFGS. De CNN was getraind op opnames van de RealSense D415, bestaande uit 116 patiënten met een PAV en 9 gezonde personen die de SFGS mimische posities uitvoerden. Elk van de 13 individuele SFGS elementen werd getraind op een aparte CNN, resulterend in de 13 scores voor de

13 verschillende SFGS elementen. Vanuit deze 13 scores konden de gedeeltelijke SFGS scores en de totale SFGS score worden uitgerekend, wat hetzelfde proces is wanneer de SFGS handmatig wordt uitgevoerd. De betrouwbaarheid van de CNN werd bepaald door de automatisch bepaalde SFGS score te vergelijken met de SFGS score van de drie ervaren menselijke beoordelaars. De betrouwbaarheid was uitgedrukt als het intra-klasse correlatiecoëfficiënt (ICC), wat de betrouwbaarheid weergeeft met een waarde tussen 0 en 1. Een ICC waarde dichtbij 1 geeft een hoge betrouwbaarheid aan tussen de geautomatiseerde SFGS en de menselijke beoordelaars, wat de gewenste uitkomst zou zijn. De betrouwbaarheid van ieder SFGS element, gedeeltelijke score en de totaalscore werden bepaald, waarbij een gemiddelde ICC van 0,87 werd gevonden voor de totale SFGS score, wat wordt gezien als een goede overeenkomst. De gerapporteerde ICC voor menselijke beoordelaars ligt tussen 0,81 en 1,00 met een gemiddelde ICC van 0,91. Dit betekende dat de betrouwbaarheid van de geautomatiseerde SFGS vergelijkbaar was met de menselijke beoordelaars die de SFGS handmatig bepalen.

De eerste implementatie van de geautomatiseerde SFGS liet veelbelovende resultaten zien en **Hoofdstuk 5** onderzocht een verdere optimalisatie van het CNN model door het gebruik van handmatig geplaatste 2D facial landmarks als extra input. Deze landmarks zouden kunnen helpen bij het focussen op de relevante regio's van het gezicht tijdens de training van de CNN. Om de resultaten te kunnen vergelijken met de eerste geautomatiseerde SFGS, werden er zoveel mogelijk variabelen gelijk gehouden met **Hoofdstuk 4**. Er werd bijvoorbeeld exact dezelfde dataset gebruikt tijdens het trainen van de CNN, bestaande uit de 125 proefpersonen met de SFGS score zoals bepaald door de drie menselijke beoordelaars. Het grootste verschil met het oude CNN model waren de extra vierde en vijfde input, bestaande uit een 2D afbeelding van de handmatig geplaatste facial landmarks voor het gezicht in rust en tijdens de maximale uitslag van de mimische positie. De analyse van de betrouwbaarheid was ook consistent gehouden met het vorige hoofdstuk, waarbij de geautomatiseerde SFGS werd vergeleken met de menselijke beoordelaars, uitgedrukt als de ICC. De toevoeging van de input met 2D facial landmarks zorgde voor een toename in de betrouwbaarheid van de geautomatiseerde SFGS, waarbij de ICC van de totale SFGS score toenam van 0,87 naar 0,91, wat wordt gezien als een excellente overeenkomst. Dit betekende dat de geautomatiseerde SFGS gradeerde met een vergelijkbare betrouwbaarheid als ervaren menselijke beoordelaars, zonder het uitbreiden van de onderliggende dataset. Naast het toenemen van de betrouwbaarheid van de geautomatiseerde SFGS, laat deze optimalisatie de potentiële impact zien voor klinische applicaties waar een beperkte dataset beschikbaar is.

De gepresenteerde resultaten in dit proefschrift zijn een eerste stap naar de implementatie van een geautomatiseerde SFGS in een klinische setting. Er zijn echter bepaalde onderzoeksgebieden die niet zijn behandeld tijdens dit proefschrift. Ten eerste, niet alle beschikbare data van de RealSense opnames zijn gebruikt tijdens de ontwikkeling van de geautomatiseerde SFGS. Dit proefschrift heeft een basismodel geïmplementeerd gebaseerd op 2D data om de potentie van een deep learning netwerk te laten zien. De diepte data, 3D landmarks en 3D antropometrische metingen van de RealSense D415, die zijn gevalideerd in dit proefschrift, zouden de betrouwbaarheid van de geautomatiseerde SFGS mogelijk verder kunnen verbeteren. Aangezien het een significante hoeveelheid tijd kan kosten om een dataset van redelijke omvang op te bouwen was er de voorkeur om direct 4D opnames te maken, ondanks dat deze data nog niet wordt gebruikt in de huidige geautomatiseerde SFGS. Er zijn ook gebieden van onderzoek die nog verder moeten worden verkend met een multidisciplinair team voordat de geautomatiseerde SFGS kan worden geïmplementeerd in de dagelijkse klinische praktijk. Er moet bijvoorbeeld rekening worden gehouden met de wettelijke vereisten en relevante wettelijke kaders voordat de geautomatiseerde SFGS ingezet kan worden, met name wanneer de geautomatiseerde SFGS het behandelplan van de patiënt beïnvloedt. Dit zal naar alle waarschijnlijkheid een verdere validatie van de betrouwbaarheid van de geautomatiseerde SFGS vereisen, met de inclusie van meer proefpersonen. De uitbreiding van de dataset zou bij voorkeur worden gedaan in verschillende omgevingen, zoals in een eHealth omgeving of in andere klinische centra. Dit maakt het nodig om een gebruiksvriendelijke interface te ontwikkelen waarbij de benodigde handmatige stappen zijn geminimaliseerd. Een grotere dataset zou voor de verdere ontwikkeling van de onderliggende techniek van de geautomatiseerde SFGS interessant zijn, waarin nieuwe CNN of deep learning architecturen geïmplementeerd kunnen worden, met het doel om de betrouwbaarheid van de geautomatiseerde SFGS te verbeteren. In combinatie met het werk gepresenteerd in dit proefschrift, kunnen deze inspanningen resulteren in de implementatie van een geautomatiseerde SFGS in onderzoek, dagelijks klinisch gebruik en door de patiënt thuis in een eHealth omgeving, met een hogere betrouwbaarheid ten opzichte van menselijke beoordelaars.

RESEARCH DATA MANAGEMENT

Ethics and privacy

This thesis is based on the results of human studies, which were conducted in compliance with the World Medical Association Declaration of Helsinki on medical research ethics. The institutional ethical review committee CMO Radboudumc, Nijmegen, the Netherlands has given approval to conduct these studies (CMO Radboudumc dossier number: 2015-1829). The privacy of the participants in these studies was warranted by the use of pseudonymization. The pseudonymization key was stored on a secured Radboudumc network drive that was only accessible to members of the project who needed access to it because of their role within the project. The pseudonymization key was stored separately from the research data. Informed consent was obtained from participants to collect and process their data for this research project. The sensitivity and confidentiality of the raw qualitative data makes sharing of the data without compromising confidentiality and privacy impossible. Therefore, specific optional consent for the sharing of the identifiable data in scientific and/ or general publications was retrieved from the participants. Any subjects shown in the scientific publications provided the written informed consent for the use of their images. However, none of the participants provided consent of the sharing of the raw data in data repositories.

Data collection and storage

Data for Chapter 2 to 5 were obtained through recordings with the RealSense™ F200 and Intel RealSense™ D415 (Intel®, Santa Clara, USA) and the 3dMD system (3dMDface, 3dMD, Atlanta, USA). Data from Chapter 2 to 5 were stored and analysed on the department server and are only accessible by project members working at the Radboudumc. These secure storage options safeguard the availability, integrity and confidentiality of the data. Paper (hardcopy) data is stored in cabinets on the department. The data will be saved for 15 years after termination of the respective studies. Using these patient data in future research is only possible after a renewed permission by the patient as recorded in the informed consent.

Table 1. Overview of the details where the data and research documentation for each chapter can be found on the Radboud Data Repository (RDR).

Chapter	Data Acquisition Collection
2	RDR, DOI:10.34973/6gq6-pm78
3 Part 1	RDR, DOI:10.34973/6gq6-pm78
3 Part 2	RDR, DOI:10.34973/6gq6-pm78
4	RDR, DOI:10.34973/6gq6-pm78
5	RDR, DOI:10.34973/6gq6-pm78

Data sharing according to the FAIR principles

Chapter 2, 4, and 5 are published open access. No consent has been given by the subjects for sharing the raw data in (restricted) online data repositories and anonymization of the raw data is not possible. Therefore, the raw data are stored on a secured Radboudumc network drive that was only accessible to members of the project who needed access to it because of their role within the project. The pseudonymized data underlying Chapters 2, 3 Part 1 & 2, 4 and 5 are published on the Radboud Data Repository in a closed access Data Acquisition Collection and will remain available for at least 15 years after termination of the studies (see Table 1 for a more detailed overview). Data were made reusable by adding sufficient documentation and by using preferred and sustainable data formats where possible, such as .txt files and .csv files.

ACKNOWLEDGEMENTS

Dit proefschrift is tot stand gekomen door de hulp van velen en daarom wil ik graag iedereen bedanken die hier een bijdrage aan heeft geleverd. Allereerst wil ik alle patiënten en vrijwilligers bedanken voor hun deelname aan dit onderzoek, aangezien dit onderzoek begint en eindigt bij de patiënt. Jullie verhalen over de aangezichtsverlamming en de feedback over het onderzoek hebben voor veel motivatie gezorgd bij het uitvoeren van dit proefschrift. Hopelijk kan dit onderzoek een onderdeel uitmaken bij het verbeteren van het rehabilitatieproces voor de aangezichtsverlamming.

Prof. dr. Maal, beste Thomas, volgens mij heb je in 2014 een iets te positieve indruk achtergelaten als stagebegeleider van het 3D Lab, waardoor we nu alweer meer dan 10 jaar hebben samengewerkt! Tijdens deze periode heb je ervoor gezorgd dat ik mijn grenzen bleef verleggen en was het tegelijkertijd mogelijk om mijn eigen pad te vinden binnen het 3D Lab. Hierdoor heb ik zoveel onvergetelijke ervaringen en kennis op kunnen doen. Het is ook heel erg fijn dat ik altijd op je kon bouwen, of dit nu op professioneel vlak was voor al mijn technische vragen, of op persoonlijk vlak, voor soms de serieuze, maar vooral leuke onderwerpen. Ik kijk dan ook terug met een gevoel van trots dat ik een onderdeel van jouw team heb kunnen uitmaken en wil je bedanken voor deze geweldige tijd!

Prof. dr. Marres, beste Henri, na het afronden van mijn afstuderen waren we bij het 3D Lab druk op zoek naar mogelijkheden hoe we het afstudeeronderzoek konden voortzetten als een PhD traject. Het was een enorme opluchting om de ondersteuning van jou en de KNO te ontvangen, waardoor dit onderzoek vervolgd kon worden. Daarnaast was het heel waardevol om jouw input te krijgen bij het opzetten van de onderzoekslijn en feedback te ontvangen bij de wetenschappelijke artikelen, waar ik altijd op je kon rekenen om de puntjes op de i te zetten (al was het in sommige gevallen nodig om eerst de ontbrekende i te plaatsen). Aan het eind van het promotietraject was het ook super fijn om nog een extra duwtje in de rug te krijgen om het proefschrift succesvol af te ronden, dus heel erg bedankt voor deze steun van begin tot eind tijdens dit promotietraject.

Dr. Ingels, beste Koen, onze eerste ontmoeting was tijdens mijn M3 stage in 2015 tijdens een facialis spreekuur. We bespraken hoe we de patiëntinclusie voor het onderzoek konden regelen. Dit was één van de spannendste momenten voor mij, aangezien de patiëntinclusie één van de belangrijkste maar ook direct lastigste punten van klinisch onderzoek is. Hier heb ik zoveel geluk gehad om jou te ontmoeten, want je bent vanaf het begin super enthousiast geweest over het onderzoek, waardoor het altijd voelde alsof ik de wind in de rug had tijdens het onderzoek. Het was ook enorm fijn om een klankbord te

hebben voor mijn klinische vragen, waar je expertise over de aangezichtsverlamming mijn verstand ver te boven gaat. Ik zou je daarom direct als (klinisch) begeleider aanbevelen voor andere studenten, maar of dat er in zit? Ik denk dat je de komende jaren, zeer verdiend, van je pensioen gaat genieten.

Dr. Speksnijder, beste Caroline, zonder jou was dit onderzoek er niet geweest, aangezien jij de oorspronkelijke bedenker was van mijn eerste stageopdracht bij het 3D Lab over de toepassing van spiegeltherapie bij een aangezichtsverlamming. Ik kan me nog goed herinneren dat ons eerste gesprek niet alleen maar over het onderzoek ging, maar dat je oprecht geïnteresseerd was in de persoon die tegenover je zat. Hierdoor zaten we al snel op dezelfde golflengte, waardoor het samenwerken me heel goed beviel. Jouw persoonlijke betrokkenheid was ook duidelijk te zien in het onderzoek, waar je altijd een drijvende kracht bent geweest om het onderzoek verder te ontwikkelen. Dit vertaalde zich in bizar snelle reacties op mijn vragen, uitgebreide feedback op artikelen en al die extra hulp die je op wetenschappelijk en persoonlijk gebied hebt gegeven. Ik weet dat je nog veel grotere plannen hebt met het onderzoek en ik hoop dat dit een eerste mooie opzet is waar in de komende jaren nog verder op gebouwd kan worden! Het zal in ieder geval niet aan jouw passie en drijfkracht liggen, waar de patiënten en medeonderzoekers zich heel gelukkig mee kunnen prijzen.

Graag wil ik de leden van de manuscriptcommissie bestaande uit **prof. dr. D.J.O. Ulrich, prof. dr. P.J. van der Wees en prof. dr. P.P.G. van Benthem** bedanken voor de tijd en toewijding die zij hebben genomen om dit proefschrift te beoordelen en voor de aanwezigheid tijdens de verdediging.

Beste Frank en Tom, mijn paranimfen. Jullie hebben beiden een grote rol gespeeld in de totstandkoming van dit proefschrift en ik ben heel dankbaar dat jullie nog steeds aan mijn zijde staan bij de afronding van deze periode. **Frank**, onze 3D Lab carrière is vrijwel tegelijkertijd begonnen, waardoor we onze (academische) vaardigheden samen konden ontwikkelen en waarbij we veel heldere momenten hebben meegemaakt! In deze periode is onze vriendschap ook flink gegroeid, waardoor het werk vaak een onderbreking was van de talloze activiteiten daarbuiten. Tijdens deze momenten heb ik je leren kennen als een echte allemansvriend die altijd klaarstaat om anderen te helpen, zelfs als dit ten koste van jezelf gaat. Hier heb ik meermaals dankbaar gebruik van gemaakt. Of het nu ging om het beklimmen van de academische trap of wanneer ik 's ochtends vroeg ergens was gestrand en een lift nodig had, ik kon altijd op je leunen. En nu sta je opnieuw voor mij klaar op dit belangrijke moment, waar ik je ontzettend dankbaar voor ben! **Tom**, ik sta er iedere keer weer versteld van hoeveel theoretische en praktische kennis jij hebt. Wanneer ik denk dat ik een beetje doorheb wat jij allemaal

kan, verbouw jij opeens je hele huis in de kortst mogelijke tijd. Met zoveel verschillende kwaliteiten heb ik geluk gehad dat we aan zoveel verschillende projecten hebben kunnen samenwerken. Wanneer nodig konden we onze kennis ook op nieuwe gebieden toepassen, zoals bij het ontcijferen van versleutelde bestanden, wat buitengewoon handig bleek te zijn. Naast de vele koppen koffie die wij uit de door jou gerepareerde koffiemachine hebben gehaald, heb je ook mijn interesse in domotica aangewakkerd en ben je een enorme inspiratiebron geweest om meer praktisch bezig te zijn, waarvoor ik je graag wil bedanken. Ik kijk nu al uit naar het volgende project dat we samen kunnen oppakken!

Beste Arico, waar we elkaar tijdens de studie regelmatig tegen het lijf liepen, raakte onze vriendschap echt in een stroomversnelling tijdens onze stage bij het 3D Lab. Binnen de kortste keren was ik praktisch één van je huisgenoten en moesten jullie me bijna elke avond de deur uitzetten. Dat was niet altijd even gemakkelijk aangezien je een echte levensgenieter bent, waardoor de frituurpan vaak nog aan het eind van de avond werd aangeslingerd onder het genot van een drankje. Dat genieten doe je volgens mij nog steeds, waardoor je tijdens mijn verdediging ook lekker op vakantie bent. Gelukkig hebben we wel vaker op een onlineverbinding moeten terugvallen, al moet je dan wel een gokje wagen dat de internetverbinding goed genoeg is. Toch zou ik je liever binnenkort een keer live willen ontmoeten om te proosten op de geweldige momenten die we samen hebben beleefd. Met een beetje geluk bespaar je de planten om je heen van dat laatste halfvolle glas!

Beste (oud)collega's van het 3D Lab, beste Anouk, Ashley, Bas, Bo, Dylan, Gert, Guido, Han, Harold, Jene, Jessica, Joost, Lars, Leanne, Luc, Merel, Rinaldo, Robin, Ruud en Tycho, in de afgelopen jaren hebben jullie niet alleen een grote impact gehad op mijn proefschrift, maar ook op zoveel meer aspecten in mijn (dagelijks) leven. Jullie hebben door de combinatie van passie voor onderzoek, humor, een relaxte sfeer, uiteenlopende interesses en een databank van enorme kennis, laten zien dat we zoveel met elkaar konden bereiken. Door deze diverse achtergronden en interesses was er altijd wel iets te leren over een nieuw onderwerp. Dit zorgde er vaak genoeg voor dat we tijd tekortkwamen tijdens de normale werkuren, dus dan zat er niets anders op om het gesprek maar bij de Aesculaaf of St. Anneke voort te zetten. Om een beetje tegengewicht te geven konden we dan toch af en toe ook nog wat wielrennen of boulderen, om dan weer te eindigen met een BBQ aan het Waalstrand. Het moge duidelijk zijn dat jullie meer dan alleen collega's zijn geweest in de afgelopen jaren en al deze ervaringen hebben nog steeds een grote (positieve) impact op mijn leven, waar ik jullie altijd dankbaar voor zal blijven!

Beste Freek, als het goed is, voelde je je al aangesproken in het vorige stuk over het 3D Lab, maar ik wilde je hier nog apart noemen gezien het werk dat wij samen hebben gedaan binnen het aangezichtsonderzoek. Je bent begonnen als M3 student bij het 3D Lab en het werd al snel duidelijk dat je een zeer gemotiveerde en zelfstandige werker bent die feedback extreem snel oppakt. Dit kwam goed uit, aangezien je in 2020 bent begonnen aan je stage waardoor we elkaar in de daaropvolgende jaren meer digitaal hebben gesproken dan ik graag zou hebben gewild. Ondanks deze tegenvaller ben ik heel erg blij dat jij een plek binnen het 3D Lab hebt kunnen vinden. Daarnaast is het super fijn dat ik mijn taken met vol vertrouwen heb kunnen overdragen aan jou, waar ik ervan overtuigd ben dat jij deze beter uitvoert dan ik ooit heb gedaan.

Beste Facialisteam, klinisch onderzoek staat of valt bij de patiëntinclusie en daarom wil ik het hele Facialisteam bedanken voor alle inzet en hulp bij dit onderzoek! In de afgelopen jaren hebben we een unieke database met 2D, 3D en 4D data voor de aangezichtsverlamming kunnen opbouwen. Ik wil jullie ook bedanken voor de open ontvangst tijdens de facialis bijeenkomsten, waardoor het makkelijk was om vragen te stellen en om feedback over het onderzoek te ontvangen. Daarnaast was het heel erg mooi om te zien hoe jullie vol inzetten op het creëren van een excellente zorg voor de patiënt. Hopelijk kan dit onderzoek in de toekomst een mooie aanvulling zijn om deze kwaliteit van zorg zo hoog te houden.

Beste collega's van de KNO & MKA, staf, aios en medewerkers van de poli, heel erg bedankt voor de gezellige sfeer op de afdelingen waar de toon goed wordt gezet wanneer je in de wandelgangen met gezang wordt ontvangen! Het was ook heel leuk om elkaar beter te leren kennen tijdens de gezellige dagjes uit en de (kerst)feestjes die nooit uit de hand liepen. Daarnaast wil ik **Stefaan** nog extra bedanken, aangezien een groot deel van mijn tijd bij het 3D Lab onlosmakelijk was verbonden aan de MKA. Deze integratie heeft voor veel kruisbestuiving gezorgd, wat een grote rol heeft gespeeld in mijn opgedane ervaringen en kennis bij het 3D Lab.

Best collega's van het UMC Utrecht MKA en 3D FaceLab, beste Celine, Florine, Harmien, Hilde, Jaron, Karlien, Maartje, Marit, Nard, Robbie, Wouter, Willem en Laura van de KNO, tijdens het opzetten van de samenwerking tussen het 3D Lab en het 3D FaceLab ging het initieel over het brengen van augmented reality naar de operatiekamer, maar dit bleek een virtuele realiteit te zijn, aangezien mijn specialisatie meer lag bij de ondersteuning op het technisch vlak. **Toine**, het was heel fijn om jou als leidraad te hebben omdat je directe en duidelijke doelstellingen stelde om het onderzoek in de juiste banen te leiden. Daarnaast was je heel erg open in je aanpak, waardoor je

altijd bereid was om in gesprek te gaan hoe bepaalde zaken het beste aangepakt konden worden. Hierin was **Marvick** ook enorm waardevol, waar jij als echte duizendpoot op zoveel verschillende aspecten input hebt geleverd voor het onderzoek. Vaak was één dag in de week te weinig om alles te bereiken wat we graag wilden en ik had volgens mij zo fulltime aan de slag kunnen gaan bij jullie. Ik zie het in ieder geval als een positief teken dat de samenwerking nog steeds succesvol wordt voortgezet en hopelijk is dit het begin van een veel langere samenwerking.

Lieve familie en vrienden, helaas zijn jullie met te grote getalen om allemaal apart te kunnen bedanken, maar wat ben ik tegelijkertijd blij dat jullie met zovelen zijn. Al die momenten die we samen hebben gedeeld tijdens het studeren, sporten, kamperen, geocachen, familiedagen, reizen en zoveel meer activiteiten die we daaromheen hebben uitgevoerd. Ik kon op jullie rekenen wanneer het nodig was, of dit nu voor of tijdens (en hopelijk na) het promotietraject was. Ondanks het feit dat ik met een groot aantal van jullie ben overgeschakeld naar digitale afspraken, hoop ik dat we in de komende jaren, waar dan ook ter wereld, nieuwe momenten samen kunnen beleven.

Dear Sylvia, Graeme, Andy, and grandma, it is difficult to express my gratitude to you, as you all have played a major part in making a new country feel like home. Although at a quick glance during passport control, it's often easy to find the Dutchy in the group, I really feel part of your family. I'm truly appreciative of the ongoing involvement and support in our life. Whether it's helping with this thesis, housing us for an indefinite period when we first moved overseas, sharing meals, surfing and holidays together, or just helping with everyday things such as babysitting, gardening, and fixing up the house, you are always there to support and encourage us. This makes the challenging moments in life, away from my original home, so much easier to deal with, for which I'm extremely thankful.

Lieve Madelein, **Wouter en Anne-Jan,** met ons viertjes thuis was er altijd wel iets te doen, al was het maar omdat we elkaar in de haren zaten. Natuurlijk wist ik dat deze vechtlust werd omgezet in hulp wanneer dat nodig was, zo ook tijdens dit proefschrift. Jullie waren altijd beschikbaar wanneer ik afleiding kon gebruiken en gelukkig hadden jullie nog wat versterking met **Nynke, Klaas, Lieke en Lynn** om te helpen wanneer er heel veel afleiding nodig was. Daarnaast was het altijd genieten om samen bij te praten over alles wat we allemaal hadden beleefd. Dit is nu wel één van de grootste uitdagingen om veel van deze ontwikkelingen op afstand mee te maken. Ik ben blij dat we tegenwoordig de mogelijkheid hebben om deze momenten

digitaal met elkaar te delen, dus blijf die foto's en video's doorsturen, maar het zal nooit een vervanging zijn om elkaar in het echt te zien. Ondanks dat ik jullie nu veel minder vaak zie dan ik zou willen, is het wel heel erg fijn om te weten dat we al die jaren zoveel tijd hebben door kunnen brengen en een speciale band hebben kunnen opbouwen waar we nu op kunnen steunen!

Lieve mam (in memoriam) en pap, wat heb ik een geluk gehad om jullie als ouders te hebben, met alle liefde en vrijheid die jullie mij hebben gegeven. Daarnaast hebben jullie mij altijd gesteund en aangemoedigd bij academische, sportieve en persoonlijke uitdagingen, waardoor ik mezelf voortdurend verder kon ontwikkelen. Hierdoor heb ik uiteindelijk stappen gezet die me aan de andere kant van de wereld hebben gebracht. Hoewel dit praktisch gezien allemaal niet zo handig is, werd het juist met veel enthousiasme ontvangen dankzij jullie onvoorwaardelijke steun. Hiermee hebben jullie beiden een grote rol gespeeld in het ontwikkelen van de vastberadenheid om tot het einde door te zetten om dit proefschrift af te ronden. Zonder jullie steun zou ik dit nooit hebben bereikt. Dat maakt het extra moeilijk dat mam er niet bij is om deze mijlpaal samen te vieren, maar het laat ook zien dat jullie altijd bij me zullen blijven, op welke manier dan ook.

Lieve Lizzie, we hebben samen al flink wat avonturen meegemaakt en wat ben ik blij dat ik jou naast me heb, zodat we het (letterlijk) vallen en opstaan samen kunnen doen. Logischerwijs heb jij ook een ontzettend grote rol gespeeld bij dit proefschrift. Met je eeuwige enthousiasme, waar ik me soms van afvraag waar je al die energie vandaan kan toveren, heb je me continu weten te motiveren. Daarnaast herinner je me er maar al te graag aan dat jij de wijzere bent van ons tweeën met je twee maanden extra levenservaring. Hierdoor was het heel waardevol om jou als sparringpartner te hebben om ideeën te bespreken. Ondertussen zijn we alweer met flink wat nieuwe avonturen gestart, waar we het ouderschap al een beetje konden oefenen met **Vita**, maar waar we aan het grootste avontuur (tot nu toe) zijn begonnen met **Peter**. Wat brengen jullie mij een plezier, liefde en geluk met z'n allen, waardoor iedere dag weer een nieuw feest (en soms een beetje een gekkenhuis) is. Hoewel het hoofdstuk van dit proefschrift nu is afgesloten, begint het hoofdstuk van ons leven nog maar net, en ik kan niet wachten om te ontdekken wat er ons nog meer te wachten staat!

LIST OF PUBLICATIONS

1. **ten Harkel, T. C.**, Speksnijder, C. M., van der Heijden, F., Beurskens, C. H. G., Ingels, K. J. A. O. & Maal, T. J. J. Depth accuracy of the RealSense F200: Low-cost 4D facial imaging. Scientific Reports 7, 16263, 2017, DOI: 10.1038/s41598-017-16608-7.
2. **ten Harkel, T. C.**, Vinayahalingam, S., Ingels, K. J. A. O., Bergé, S. J., Maal, T. J. J. & Speksnijder, C. M. Reliability and Agreement of 3D Anthropometric Measurements in Facial Palsy Patients Using a Low-Cost 4D Imaging System. IEEE Transactions on Neural Systems and Rehabilitation Engineering 28, 1817–1824, 2020, DOI: 10.1109/TNSRE.2020.3007532.
3. Baer, K., Kieser, S., Schon, B., Rajendran, K., **ten Harkel, T.C.**, Ramyar, M., Löbker, C., Bateman, C., Butler, A., Raja, A., Hooper, G., Anderson, N. & Woodfield, T. Spectral CT imaging of human osteoarthritic cartilage via quantitative assessment of glycosaminoglycan content using multiple contrast agents. APL Bioengineering 5, 26101, 2021, DOI: 10.1063/5.0035312.
4. De Kort, W. W. B., van Hout, W. M. M. T., **ten Harkel, T. C.**, van Cann, E. M. & Rosenberg, A. J. W. P. A Novel Method for Quantitative Three-Dimensional Analysis of Zygomatico-Maxillary Complex Symmetry. Journal of Craniofacial Surgery 33, 1474–1478, 2022, DOI: 10.1097/SCS.0000000000008382.
5. van Hout, W. M. M. T., de Kort, W. W. B., **ten Harkel, T. C.**, van Cann, E. M. & Rosenberg, A. J. W. P. Zygomaticomaxillary complex fracture repair with intraoperative CBCT imaging. A prospective cohort study. Journal of Cranio-Maxillofacial Surgery 50, 54–60, 2022, DOI: 10.1016/j.jcms.2021.09.009.
6. de Jongh, F. W., Sanches, E. E., Pouwels, S., **ten Harkel, T. C.** & Ingels, K. J. A. O. E-Health and telemedicine applications in plastic surgery and the treatment of facial palsy. Health Sciences Review 2, 100009, 2022, DOI: 10.1016/j.hsr.2021.100009.
7. Schutte, H., Muradin, M. S. M., Bielevelt, F., **ten Harkel, T. C.**, Speksnijder, C. M. & Rosenberg, A. J. W. P. Creating Three-Dimensional Templates of Smiling and Pouting Faces for Different Sex- and Age Groups. Journal of Clinical Medicine 11, 2022, DOI: 10.3390/jcm11247257.
8. Markodimitraki, L. M., **ten Harkel, T. C.**, Bleys, R. L. A. W., Stegeman, I. & Thomeer, H. G. X. M. Cochlear implant positioning and fixation using 3D-printed patient specific surgical guides; a cadaveric study. PLOS ONE 17, 1–12, 2022, DOI: 10.1371/journal.pone.0270517.
9. Stoop, C. C., Janssen, N. G., **ten Harkel, T. C.** & Rosenberg, A. J. W. P. A Novel and Practical Protocol for Three-Dimensional Assessment of Alveolar Cleft Grafting Procedures. The Cleft Palate Craniofacial Journal 60, 601–607, 2023, DOI: 10.1177/10556656221074210.

10. **ten Harkel, T. C.**, de Jong, G., Marres, H. A. M., Ingels, K. J. A. O., Speksnijder, C. M. & Maal, T. J. J. Automatic grading of patients with a unilateral facial paralysis based on the Sunnybrook Facial Grading System - A deep learning study based on a convolutional neural network. American Journal of Otolaryngology 44, 103810, 2023, DOI: 10.1016/j.amjoto.2023.103810.
11. Markodimitraki, L. M., **ten Harkel, T. C.**, Bennink, E., Stegeman, I. & Thomeer, H. G. X. M. A monocenter, patient-blinded, randomized, parallel-group, non-inferiority study to compare cochlear implant receiver/stimulator device fixation techniques (COMFIT) with and without drilling in adults eligible for primary cochlear implantation. Trials 24, 605, 2023, DOI: 10.1186/s13063-023-07568-7.
12. **ten Harkel, T. C.**, Bielevelt, F., Marres, H. A. M., Ingels, K. J. A. O., Maal, T. J. J. & Speksnijder, C. M. Optimization of the automated Sunnybrook Facial Grading System – Improving the reliability of a deep learning network with facial landmarks. European Annals of Otorhinolaryngology, Head and Neck Diseases 142, 5–10, 2025, DOI: 10.1016/j.anorl.2024.07.005.

PHD PORTFOLIO

PhD candidate	T.C. ten Harkel
Department	3D Lab Radboudumc Otorhinolaryngology / Head and Neck Surgery
PhD period	01/01/2017 — 07/04/2025
PhD supervisors	prof. dr. T.J.J. Maal, prof. dr. H.A.M. Marres
PhD co-supervisors	dr. K.J.A.O. Ingels, dr. C.M. Speksnijder

TRAINING ACTIVITIES	**Hours**
Courses	
Radboudumc - Introduction day (2017)	6
RIHS - Introduction course for PhD candidates (2017)	15
DGP MRI fusion training (2018)	12
Radboudumc - eBROK course (2018)	42
3ds Max course (2018)	11
Newtom VGI EVO & NNT course (2019)	4
Radboudumc - Scientific integrity (2020)	20
Seminars	
Radboudumc Research Rounds (2017 — 2018)	6
JCI meetings of the MKA (2017 — 2018)	14
Meetings of the facialis team (2017 — 2020)	17
Research meetings of the 3D Lab (2017 — 2021)	42
Other	
Presentation for MITeC Radboudumc (2018)	6
Robosculpt KNO (2018)	56
Presentation for the facial palsy expertise centre Radboudumc (2020)	6
Expertise team Facial Analysis, 3D Design and PSI, Software 3D Lab (2018 — 2020)	28

TEACHING ACTIVITIES	
Lecturing	
TMS-THK-CBCT course (2018)	22
Supervision of internships / other	
Supervision and guidance of students Technical Medicine (2017 — 2022)	420
External support for students Technical Medicine, residents MKA, residents KNO UMC Utrecht (2017 — 2022)	560
TOTAL	**1287**

ABOUT THE AUTHOR

Timen ten Harkel was born in Hefshuizen, the Netherlands, on the 31st of May 1990. After finishing secondary school in 2008, he earned his bachelor's degree in Technical Medicine at the University of Twente, in 2011. He then took a gap year to gain hands on industry experience, and travel internationally. In 2012, he began a master's of Technical Medicine at the University of Twente, specialising in Medical Imaging and Interventions. During this program, he completed (clinical) internships at the Netherlands Cancer Institute (NKI), Amsterdam UMC, Radboudumc, and the University of Otago (New Zealand). Timen received his Master of Science in 2016, after completing a thesis on the 3D facial landmark tracking of patients with a facial palsy at the 3D Lab of the Radboudumc. This master's research laid the foundation for a subsequent PhD program at the Radboudumc and Timen began his research into the automation of the Sunnybrook Facial Grading System in 2017. This research was done in collaboration between the 3D Lab and the Department of Otorhinolaryngology / Head and Neck Surgery, under the supervision of prof. dr. Thomas Maal, prof. dr. Henri Marres, dr. Koen Ingels, and dr. Caroline Speksnijder. During his PhD, Timen contributed to other research projects, performed clinical tasks, created 3D (surgical) visualizations, and supervised both masters' and medical students. He also worked as a Technical Physician at the UMC Utrecht, supporting students and surgical residents in the Department of Oral and Maxillofacial Surgery. After completing his PhD, Timen plans to continue to explore his passion for the application and implementation of (medical) technology.

Radboudumc university medical center

Radboud University