

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

Timen C. ten Harkel^{a,b,*}, Guido de Jong^a, Henri A.M. Marres^b, Koen J.A.O. Ingels^b, Caroline M. Speksnijder^{c,d}, Thomas J.J. Maal^{a,c}

^a Radboud University Medical Centre, 3D Lab Radboudumc, Nijmegen 6500 HB, the Netherlands

^b Radboud University Medical Centre, Department of Otorhinolaryngology and Head and Neck Surgery, Nijmegen 6500 HB, the Netherlands

^c Radboud University Medical Centre, Department of Oral and Maxillofacial Surgery, Nijmegen 6500 HB, the Netherlands

^d University Medical Center Utrecht, Utrecht University, Department of Oral and Maxillofacial Surgery, Utrecht 3508 GA, the Netherlands

ARTICLE INFO

Keywords:

Deep learning
Facial paralysis
Machine learning
Medical imaging
Convolutional neural network
Sunnybrook facial grading system

ABSTRACT

Purpose: In order to assess the severity and the progression of a unilateral peripheral facial palsy 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 facial palsy patients based on the SFGS using a convolutional neural network.

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

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 Sunnybrook score, 0.45 for the resting symmetry subscore, 0.89 for the symmetry of voluntary movement subscore, and 0.77 for the synkinesis subscore.

Conclusions: 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 e-Health environment, since the model used 2D images captured from a video recording.

1. Introduction

The partial or complete loss of facial function associated with a 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. E.g. patients with a complete Bell's palsy have an overall recovery rate of 50 to 60 %, whilst patients with an incomplete Bell's palsy 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 Sunnybrook score ranging 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

* Corresponding author at: Radboud University Medical Centre, 3D Lab Radboudumc, Nijmegen 6500 HB, the Netherlands.

E-mail address: Timen.tenHarkel@radboudumc.nl (T.C. ten Harkel).

<https://doi.org/10.1016/j.amjoto.2023.103810>

Received 1 December 2022;

Available online 25 February 2023

0196-0709/© 2023 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

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 scoring 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.

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 facial palsy [13–29]. However, these studies

consisted of either small cohorts with <30 subjects, analyzed 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 Sunnybrook score by itself does not differentiate which area of the face is affected by the facial palsy it is crucial all 13 individual components of the SFGS are scored during follow-up. By adhering to the original SFGS the resulting Sunnybrook 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 scoring 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 unilateral 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 sub-components (resting symmetry, symmetry of voluntary movement, and synkinesis), and the composite Sunnybrook score of the SFGS (Table 1).

2. Material & methods

2.1. 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 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 facial palsy. 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.

2.2. 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 PFP patients [31,32]. The RealSense captured 30 frames per second at an approximate distance of 35 cm to the patient. The RealSense simultaneously captured a color 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 color images were used as input for the CNN. All Sunnybrook poses were captured in a single recording.

2.3. Pre-processing

Two frames were selected for each of the Sunnybrook poses; the

Table 1

Overview of the 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) ^a		
Eye	0–1	0
Cheek (nasolabial 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

^a Multiple answers in the Sunnybrook Facial Grading System can result in the same score of the individual elements [7].

starting frame was at the initiation of the Sunnybrook pose, whilst the maximum frame was selected at the maximum exertion of the Sunnybrook pose (Fig. 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 color image. The cropping centered 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 input for the CNN. Due to potential rotation between the start of the Sunnybrook 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 $112 \times 112 \times 3$, by calculating the absolute difference between the start and maximum frame for each individual color channel, creating a difference image between the starting and maximum frame (Fig. 1). All images were normalized, resulting in pixels values ranging from 0 to 1 for each input image.

2.4. 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 Fig. 2. The input consisted of the starting frame and the maximum frame, each with a size of $112 \times 112 \times 3$ pixels. The difference image, created during the preprocessing step and consisting of $112 \times 112 \times 3$ 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 0.8–1.2) and random rotation (range -20 – 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.

2.5. 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 Sunnybrook 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 Sunnybrook 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 $4 \times$ (length of training dataset / batch size) [36].

2.6. Analysis

The performance of the CNN was determined by comparing the predicted Sunnybrook scores of the models with the Sunnybrook scores

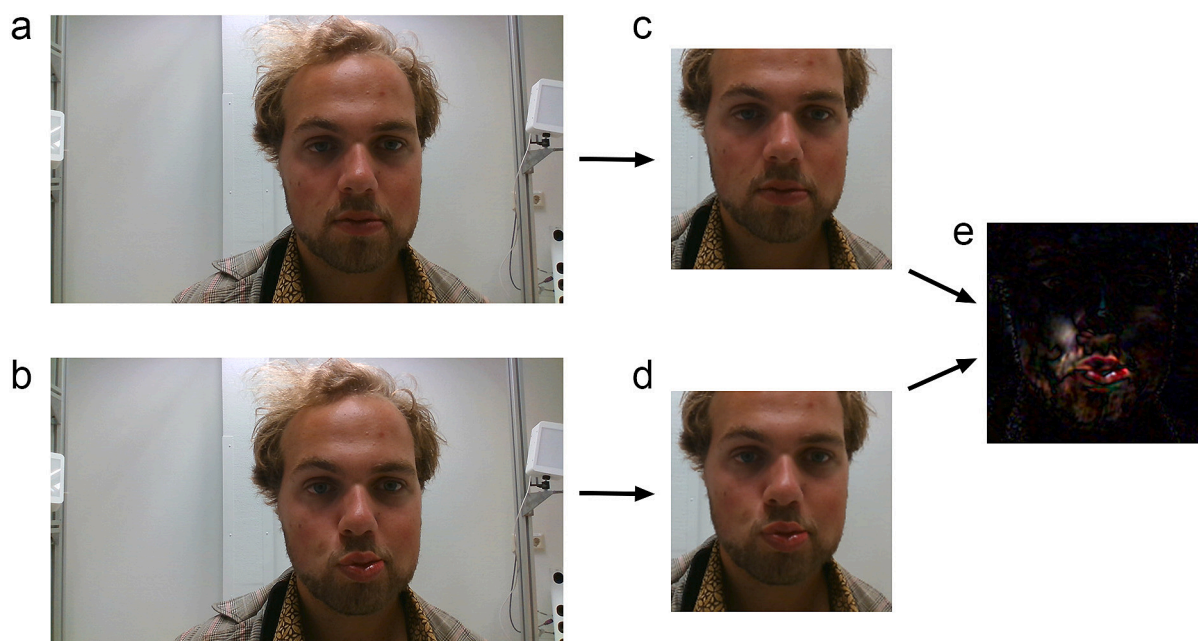


Fig. 1. Pre-processing of the RealSense recordings to optimize the input for the convolutional neural network, using the Sunnybrook pose “pucker” as an example. The starting frame (a) was selected at the initiation of the Sunnybrook pose, whilst the maximum frame (b) was selected at the maximum exertion of the Sunnybrook pose. The original images (a & b) were cropped to a $112 \times 112 \times 3$ 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 $112 \times 112 \times 3$ pixel image.

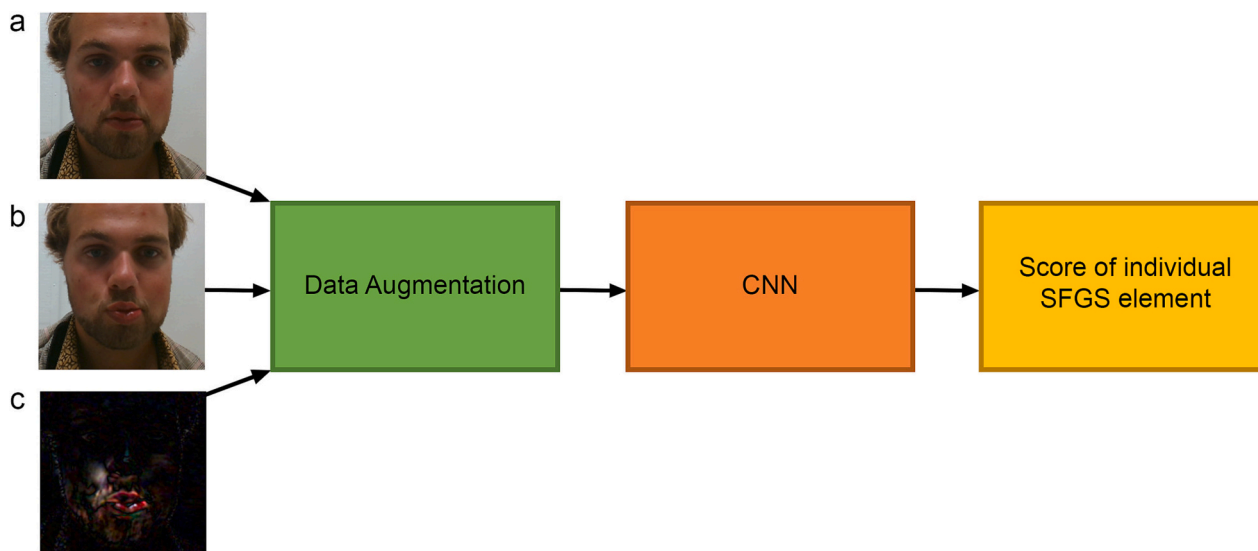


Fig. 2. Simplified overview of the training of the convolutional neural network (CNN) with the Sunnybrook pose “pucker” used as an example. The complete CNN model is shown in [Appendix A](#). The input consisted of the cropped frame during the initiation of the Sunnybrook pose (a), the frame during the maximum exertion of the Sunnybrook 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.

as graded by the experienced human observers during the patient consultation as described in the [Section 2.1](#). 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 Sunnybrook 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 Sunnybrook 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 Sunnybrook score and the columns represented the predicted Sunnybrook 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 Sunnybrook score and the predicted Sunnybrook 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 Sunnybrook 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].

3. Results

3.1. Population

A total of 116 unilateral PFP patients and 9 healthy subjects were included in this study during the period of August 2018 and November 2020. The PFP patients consisted of 49 men and 67 women, with an average age of 53 years (± 16) 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 years (± 17) ranging from 27 to 77 years.

3.2. 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 small 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.

3.3. Inter-rater reliability of the SFGS subscores and composite score

The subscores of the resting symmetry, symmetry of voluntary

Table 2

Inter-rater reliability of the individual SGFS components between the 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.

movement, synkinesis and the composite Sunnybrook 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 Section 3.2. 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 Sunnybrook all showed a good agreement.

4. Discussion

This study investigated the automated grading of patients with a unilateral 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 was 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 color 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,

Table 3

Inter-rater reliability of the SGFS subscores and composite score between the 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 Sunnybrook	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.

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 current study [45]. In this particular study the learning curve of inexperienced human observers was assessed when grading 100 PFP patients based on the SFGS. In this section the inter-rater reliability of the human observers after grading 50 PFP patients is compared to the inter-rater reliability of the CNN model from 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 PFP patients [45]. 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 versus 0.56 and the lip pucker 0.67 versus 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 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 [45]. More specifically, the CNN reaches a comparable inter-rater reliability after inexperienced human observers have graded 50 PFP patients.

The inter-rater reliability of the automated scoring 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 more 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 number of selected landmarks [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 [46]. 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 [47]. 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 (Tables 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 PFP patients using the SFGS.

5. Conclusion

This study investigated the automated grading of patients with a unilateral 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 wide variety of settings such as online consults in an e-Health 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 whom graded 50 PFP patients [45]. 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.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

CRedit authorship contribution statement

Timen C. ten Harkel: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing – original draft, Writing – review & editing, Visualization. **Guido de Jong:** Methodology, Software, Validation, Formal analysis, Writing – review & editing. **Henri A.M. Marres:** Writing – review & editing, Resources, Supervision. **Koen J.A.O. Ingels:** Conceptualization, Methodology, Validation, Investigation, Data curation, Writing – original draft, Writing – review & editing, Visualization, Resources, Supervision. **Caroline M. Speksnijder:** Conceptualization, Methodology, Validation, Formal analysis, Writing – original draft, Writing – review & editing, Visualization, Supervision. **Thomas J.J. Maal:** Conceptualization, Methodology, Validation, Writing – original draft, Writing – review & editing, Visualization, Resources, Supervision.

Declaration of competing interest

The authors declare that they have no competing interest.

Appendix A. Supplementary data

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)

(continued on next page)

(continued)

Layer	Type	Shape
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)

References

- [1] Kleiss LJ, Hohman MH, Susarla SM, HAM Marres, Hadlock TA. Health-related quality of life in 794 patients with a peripheral facial palsy using the FaCE Scale: a retrospective cohort study. *Clin Otolaryngol* 2015. <https://doi.org/10.1111/coa.12434>. n/a-n/a.
- [2] Ho AL, Scott AM, Klassen AF, Cano SJ, Pusic AL, Van Laeken N. Measuring quality of life and patient satisfaction in facial paralysis patients: a systematic review of patient-reported outcome measures. *Plast Reconstr Surg* 2012;130:91–9. <https://doi.org/10.1097/PRS.0b013e318254b08d>.
- [3] Coulson SE, O'dwyer NJ, Adams RD, Croxson GR. Expression of emotion and quality of life after facial nerve paralysis. *Otol Neurotol* 2004;25:1014–9.
- [4] Peitersen E. Bell's palsy: the spontaneous course of 2,500 peripheral facial nerve palsies of different etiologies. *Acta Otolaryngol Suppl* 2002;4–30. <https://doi.org/10.1080/000164802760370736>.
- [5] Samsudin WSW, Sundaraj K. Evaluation and grading systems of facial paralysis for facial rehabilitation. *J Phys Ther Sci* 2013;25:515–9. <https://doi.org/10.1589/jpts.25.515>.
- [6] Fattah AY, Gurusinge ADR, Gavilan J, Hadlock TA, Marcus JR, Marres H, et al. Facial nerve grading instruments: systematic review of the literature and suggestion for uniformity. *Plast Reconstr Surg* 2015;135:569–79. <https://doi.org/10.1097/PRS.0000000000000905>.
- [7] Ross BG, Fradet G, Nedzelski JM. Development of a sensitive clinical facial grading system. *Otolaryngol Neck Surg* 1996;114:380–6.
- [8] van Veen MM, Bruins TE, Artan M, Werker PMN, Dijkstra PU. Learning curve using the sunnysbrook facial grading system in assessing facial palsy: an observational study in 100 patients. *Clin Otolaryngol* 2020;45:823–6. <https://doi.org/10.1111/coa.13574>.
- [9] Lecun Y, Bengio Y, Hinton G. Deep learning. *Nature* 2015;521:436–44. <https://doi.org/10.1038/nature14539>.
- [10] Su Z, Liang B, Shi F, Gelfond J, Šegalo S, Wang J, et al. Deep learning-based facial image analysis in medical research: a systematic review protocol. *BMJ Open* 2021;11:e047549.
- [11] Liu Q, Zhang N, Yang W, Wang S, Cui Z, Chen X, et al. A review of image recognition with deep convolutional neural network. In: *Int. Conf. Intell. Comput. Springer*; 2017. p. 69–80.
- [12] Shen D, Wu G, Suk H-I. Deep learning in medical image analysis. *Annu Rev Biomed Eng* 2017;19:221–48.
- [13] Bur AM, Shew M, New J. Artificial intelligence for the otolaryngologist: a state of the art review. *Otolaryngol - Head Neck Surg (United States)* 2019;160:603–11. <https://doi.org/10.1177/0194599819827507>.
- [14] Guarini DL, Yunusova Y, Taati B, Dusseldorp JR, Mohan S, Tavares J, et al. Toward an automatic system for computer-aided assessment in facial palsy. *ArXiv* 2019;22:42–9. <https://doi.org/10.1089/fpsam.2019.29000.gua>.
- [15] GSJ Hsu, Chang MH. Deep hybrid network for automatic quantitative analysis of facial paralysis. In: *Proc AVSS 2018 - 2018 15th IEEE Int Conf Adv Video Signal-Based Surveill*; 2019. p. 1–7. <https://doi.org/10.1109/AVSS.2018.8639156>.
- [16] Mothes O, Modersohn L, Volk GF, Klingner C, Witte OW, Schlattmann P, et al. Automated objective and marker-free facial grading using photographs of patients with facial palsy. *Eur Arch Oto-Rhino-Laryngology* 2019. <https://doi.org/10.1007/s00405-019-05647-7>.
- [17] Zhuang Y, Uribe O, McDonald M, Yin X, Parikh D, Southerland A, et al. F-DIT-V: an automated video classification tool for facial weakness detection. *IEEE EMBS Int Conf Biomed Heal Informatics* 2019;2019:1–4. <https://doi.org/10.1109/bhi.2019.8834563>.
- [18] Guarini DL, Dusseldorp J, Hadlock TA, Jowett N. A machine learning approach for automated facial measurements in facial palsy. *JAMA Facial Plast Surg* 2018;20:335–7. <https://doi.org/10.1001/jamafacial.2018.0030>.
- [19] Guo Z, Dan G, Xiang J, Wang J, Yang W, Ding H, et al. An unobtrusive computerized assessment framework for unilateral peripheral facial paralysis. *IEEE J Biomed Heal Informatics* 2018;22:835–41. <https://doi.org/10.1109/JBHI.2017.2707588>.
- [20] GSJ Hsu, Huang WF, Kang JH. Hierarchical network for facial palsy detection. In: *IEEE Comput Soc Conf Comput Vis Pattern Recognit Work* 2018; 2018. p. 693–9. <https://doi.org/10.1109/CVPRW.2018.00100>.
- [21] Jiang Z, Dai W, Wang W, Wang W. A cloud-based training and evaluation system for facial paralysis rehabilitation. In: *Proc - IEEE 16th Int Conf Ind Informatics, INDIN 2018*; 2018. p. 701–6. <https://doi.org/10.1109/INDIN.2018.8471934>.
- [22] Sajid M, Shafique T, Baig MJA, Riaz I, Amin S, Manzoor S. Automatic grading of palsy using asymmetrical facial features: a study complemented by new solutions. *Symmetry (Basel)* 2018;10. <https://doi.org/10.3390/sym10070242>.
- [23] Song A, Wu Z, Ding X, Hu Q, Di X. Neurologist standard classification of facial nerve paralysis with deep neural networks. *Futur Internet* 2018;10:111. <https://doi.org/10.3390/fi10110111>.
- [24] Guo Z, Shen M, Duan L, Zhou Y, Xiang J, Ding H, et al. Deep assessment process: objective assessment process for unilateral peripheral facial paralysis via deep convolutional neural network. *Proc - Int Symp Biomed Imaging* 2017;135–8. <https://doi.org/10.1109/ISBI.2017.7950486>.
- [25] Wang T, Zhang S, Dong J, Liu LL, Yu H, Dong J, et al. Automatic evaluation of the degree of facial nerve paralysis. *Multimed Tools Appl* 2016;75:11893–908. <https://doi.org/10.1007/s11042-015-2696-0>.
- [26] Kim HS, Kim SY, Kim YH, Park KS. A smartphone-based automatic diagnosis system for facial nerve palsy. *Sensors (Switzerland)* 2015;15:26756–68. <https://doi.org/10.3390/s151026756>.
- [27] Azoulou O, Ater Y, Gersi L, Glassner Y, Bryt O, Halperin D. Mobile application for diagnosis of facial palsy. *Proc. 2nd Int. Conf. Mob. Inf. Technol. Med* 2014;1–4.
- [28] Wang T, Dong J, Sun X, Zhang S, Wang S. Automatic recognition of facial movement for paralyzed face. *Biomed Mater Eng* 2014;24:2751–60. <https://doi.org/10.3233/BME-141093>.
- [29] Tan JR, Coulson S, Keep M. Face-to-face versus video assessment of facial paralysis: implications for telemedicine. *J Med Internet Res* 2019;21. <https://doi.org/10.2196/11109>. e11109–e11109.
- [30] Jirawatnotai S, Jomkoh P, Voravitvet TY, Tirakotai W, Somboonsap N. Computerized sunnysbrook facial grading scale (SBface) application for facial paralysis evaluation. *Arch Plast Surg* 2021;48:269–77. <https://doi.org/10.5999/aps.2020.01844>.
- [31] ten Harkel TC, Speksnijder CM, van der Heijden F, Beurskens CHG, Ingels KJAO, Maal TJJ. Depth accuracy of the RealSense F200: low-cost 4D facial imaging. *Sci Rep* 2017;7:16263. <https://doi.org/10.1038/s41598-017-16608-7>.
- [32] ten Harkel TC, Vinayahalingam S, Ingels KJAO, Berge SJ, Maal TJJ, Speksnijder CM. Reliability and agreement of 3D anthropometric measurements in facial palsy patients using a low-cost 4D imaging system. *IEEE Trans Neural Syst Rehabil Eng* 2020;28:1817–24. <https://doi.org/10.1109/TNSRE.2020.3007532>.
- [33] Van Der Walt S, Schönberger JL, Nunez-Iglesias J, Bouloune F, Warner JD, Yager N, et al. Scikit-image: image processing in python. *PeerJ* 2014;2014:e453. <https://doi.org/10.7717/peerj.453>.

- [34] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. In: 3rd Int Conf Learn Represent ICLR 2015 - Conf Track Proc; 2015. p. 1–14.
- [35] Prechelt L. Early stopping - but when? In: Montavon G, Orr GB, Müller K-R, editors. Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics). 7700 LECTU. Berlin, Heidelberg: Springer Berlin Heidelberg; 2012. p. 53–67. https://doi.org/10.1007/978-3-642-35289-8_5.
- [36] Smith LN. Cyclical learning rates for training neural networks. In: Proc - 2017 IEEE Winter Conf Appl Comput Vision, WACV 2017; 2017. p. 464–72. <https://doi.org/10.1109/WACV.2017.58>.
- [37] Cohen J. Weighted kappa: nominal scale agreement provision for scaled disagreement or partial credit. *Psychol Bull* 1968;70:213.
- [38] McHugh ML. Interrater reliability: the kappa statistic. *Biochem Med* 2012;22: 276–82.
- [39] Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. *Psychol Bull* 1979;86:420.
- [40] Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *J Chiropr Med* 2016;15:155–63. <https://doi.org/10.1016/j.jcm.2016.02.012>.
- [41] Volk GF, Schaefer RA, Thielker J, Modersohn L, Mothes O, Nduka CC, et al. Reliability of grading of facial palsy using a video tutorial with synchronous video recording. *Laryngoscope* 2018. <https://doi.org/10.1002/lary.27739>.
- [42] Gaudin RA, Robinson M, Banks CA, Baiungo J, Jowett N, Hadlock TA. Emerging vs time-tested methods of facial grading among patients with facial paralysis. *JAMA Facial Plast Surg* 2016;18:251–7. <https://doi.org/10.1001/jamafacial.2016.0025>.
- [43] Neely JG, Cherian NG, Dickerson CB, Nedzelski JM. Sunnybrook facial grading system: reliability and criteria for grading. *Laryngoscope* 2010;120:1038–45.
- [44] Coulson SE, Croxson GR, Adams RD, O'Dwyer NJ. 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* 2005;132:543–9. <https://doi.org/10.1016/j.otohns.2005.01.027>.
- [45] van Veen MM, Bruins TE, Artan M, Werker PMNN, Dijkstra PU. Learning curve using the sunnybrook facial grading system in assessing facial palsy: an observational study in 100 patients. *Clin Otolaryngol* 2020;45:823–6. <https://doi.org/10.1111/coa.13574>.
- [46] Hochreiter S, Schmidhuber J. Long short-term memory. *Neural Comput* 1997;9: 1735–80.
- [47] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. *Adv Neural Inf Process Syst* 2017;30.