Mobile Application for Diagnosis of Facial Palsy

Ofer Azoulay M.D.[2,3], Yotam Ater[1], Lior Gersi[1], Yonatan Glassner[1], Ori Bryt[1] and Doron Halperin M.D.[2,3]

[1] Vision and Image Sciences Lab (VISL), Technion – IIT, Haifa, Israel
[2] Kaplan Medical Center, Rehovot, Israel
[3] The Hebrew University, Jerusalem, Israel

Abstract. Facial nerve palsy is common, affecting 1 out of 60 people around the world. Patients with facial movement dysfunction suffer from major functional, aesthetic and psychological disability. Grading the severity of facial palsy has prognostic and follow-up significance. Traditional grading systems, e.g. the House-Brackmann scale, are based on physician clinical observation, and are prone to inter and intra observer variability. Several objective grading systems, including Computer-Vision-based ones, have been suggested. None of them have gained wide clinical use, mainly due to practical issues. In this paper, we present an objective, learning-based, mobile application for the diagnosis and grading of facial nerve palsy. The algorithm is based on analyzing a set of 9 facial expressions, and is designed to be used on a mobile device. The application score exhibits a high correlation to subjective grading scores, and high computational efficiency.

1 Background

Facial Nerve Palsy (FNP) is a common medical condition. One out of every sixty people will suffer from peripheral facial nerve palsy [1]. FNP patients suffer from major functional, aesthetic and psychological disability [11].

The severity of FNP has a prognostic significance. Grading facial function is essential for evaluating and communicating the spontaneous course of FNP and the effect of medical or surgical treatments [1, 6]. Traditional grading systems are based on subjective clinical observation made by the treating physician. The method most commonly used and accepted by the American Academy of Otolaryngology- Head and Neck Surgery (AAO-HNS) is the House-Brackmann Scale (HBS) [6]. Facial paralysis is ranked I-VI (from normal to complete paralysis) in accordance with facial movement during predefined facial expressions performed by the patient. A second method is the Yanagihara Grading System (YGS), which is an unweighted regional system largely accepted in the Japanese literature [16]. Each of ten functions is scored from 0 to 4, giving a maximum score of 40 . Another method is the Sunnybrook Grading System (SGS), which is a weighted regional scale that measures also synkinesis regionally, to form a maximum composite score of 100 [13].

Subjective grading systems have an inherent disadvantage of being prone to intra- and inter-assessor variability [8, 17, 12–14]. Therefore, the motivation for creating an objective FNP grading system is obvious.

Our work relies on Computer Vision (CV) and Machine Learning (ML), which are prominent fields in Computer Engineering. We hereby give a concise review of them. The field of Computer Vision deals with the art of automatic image understanding, the attempt to provide machines the ability to interpret correctly the contents of images. This field is relevant to medical applications, as several medical diagnoses are done mainly or partly by observation of the patient. Clearly, capturing videos of people performing predefined facial expressions may assist in developing an objective scoring method of FNP. Extracting relevant facial features from the videos should facilitate the creation of a model of the nature and severity of the paralysis. This can be done using CV tools for detection and measurement.

A model for the severity of FNP can also be achieved by adding Machine Learning tools. Machine Learning (aka Artificial Intelligence) is a general term for a set of algorithms which allow computers to learn and infer from data. This learning procedure, in its general form, involves set of examples (marked as $\{\boldsymbol{x}_i\} \in X$), a corresponding ground-truth set of labels ($\{y_i\} \in Y$) and a learning algorithm which learns a connecting function ($\mathcal{F} : X \to Y$) between X and Y. If Y is discrete and has a finite number of classes, we call that problem a **classification** problem. If Y is continuous, we call it a **regression** problem. Eventually, every new patient is given a grade according to his extracted facial features, and an objective grading system for FNP is established.

During the last few decades several objective grading systems have been suggested, most of them are based on CV and ML methods. Wood et al. analyzed facial videos using a computer, in which measurement of distances between facial landmarks during movement by microscaling was performed

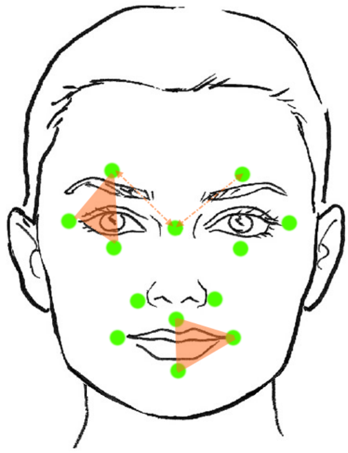

Fig. 1. Facial landmarks from which we calculate the features used by the algorithm. Triangle areas and linear distances represent typical measurements that are extracted.

[15]. Cheung and Neely presented an objective method called Computerized Facial Analysis Evaluation System (FACE) that is based on summing pixel values of facial light distortion formed during facial movement [9]. Meier-Gallati et al. used a similar principle and presented a method called Objective Scaling of Facial Nerve Function Based on Area Analysis (OSCAR) [7]. Burres et al. developed a system that measures and compares area ratios of bilateral facial polygons created by marking 11 landmarks sketched on the patient's photos during predetermined facial expressions [2]. O'Reilly et al. have recently developed the Glasgow Facial Palsy Score (GFPS) which is based on artificial neural network. A computer station performs video analysis of patients preforming five predefined facial expressions. Quantitative scores based on the HBS are obtained and displayed as a facial movement chart called "FACOGRAM" [10].

Our work most resembles the work by O'Reilly et al., as we also capture video footage of subjects, extract features and use a learning algorithm to achieve a grading for FNP. However, we have extended the number and variety of predefined facial expressions being captured, the facial features extraction process we employ is restricted to local features, and we also compare to two additional subjective grading methods. Another advantage of our work is in computational efficiency.

2 Objectives

None of the objective methods described nor others have gained wide clinical use, mainly due to practical issues. They are time-consuming, difficult to use or require special equipment or techniques.

Indeed, the Facial Nerve Disorders Committee of the AAO-HNS concluded in 2009: "Further improvement of quantification of facial nerve function will require an objective rating scale... Additional advances in motion analysis software are expected and, when refined, should allow widespread use of consistent, repeatable, objective scoring" [14].

With this mission, we set off to develop a mobile application for the diagnosis and grading of facial nerve palsy. A time-efficient mobile application will provide physicians as well as patients a standardized, validated diagnostic tool at all sites.

3 Methods

The presented research is an outcome of a clinical observational study together with algorithmic development.

3.1 Clinical Research

The study was approved by the local hospital ethical committee. videos of 9 predefined facial expressions (see Table.1) were recorded from a cohort of 14 patients with various degrees of facial nerve dysfunction and 31 healthy subjects (15 women and 30 men). All adult patients with unilateral FNP were included. No clinical decisions were altered in regard to patients conventional treatments were made due to the participant in the study. Three Otolaryngologists independently graded the severity of the patients FNP according to HBS, YGS and SGS (mentioned in Sec. 1). The 13 predefined facial landmarks were labeled by simple stickers.

3.2 Algorithmic Development

We are interested in a grading algorithm to map patients to their FNP severity. The learnt examples are numeric representations of the subjects (termed features) and the labeling is their severity as assessed by treating physicians according to the subjective methods described in Sec.1. Finding the appropriate features is a critical to the learning process, since they must be informative and distinctive.

Similarly to [10], we looked for features which capture the noticeable facial asymmetry of patients, to evaluate their palsy severity. We relied on the method of [4], which used measurements between landmarks to quantify facial movement in healthy subjects. We set 13 facial landmarks, and chose 99 measurements, triangle areas and linear distances (see Fig. 1). For each one, we examined the temporal value during predetermined facial expressions (see Table.1), and normalized it by its initial value at rest. We then subtracted the right and left values, to get a temporal differences vector for each measurement. Finally, the sequence features are set as the value of the vector at frame of the maximal movement. Intuitively, these features quantify asymmetry, where low values indicate symmetry.

	Movements
1	Face at rest
2	Strong eye closure
3	Weak eye closure
4	Raised eyebrows
5	Closed mouth smile
6	Big smile
7	Puckering of lips (whistle motion)
8	Puff-up cheeks
9	Stretching down lower lip

Table 1. Facial movements used in the study, 11 features were calculated for each.

For the algorithm training and validation, we divided our cohort to training and test groups. Our algorithm must meet the requirements of low computational cost and real time response. We thus limited ourselves to simpler learning algorithms, which satisfy those requirements. We chose the well-known SVM (Support Vector Machine) classifier [3], which presents a linear computational time with respect to the number of features.

We first trained a **binary classifier** to separate patients from healthy subjects. This was done both as a proof of concept to validate our features and to be applied before a fine-grained classifier. The classifier was trained on the training group using leave-one-out cross validation, and then was examined on the test set.

We next trained a **regression model** to achieve fine-grained severity grading. For training we used the grading of 3 physicians, which used the 3 subjective methods (see Sec. 1). For each patient, we considered the ground-truth grade as the average of the physicians' clinical score, using averaging to overcome inter-observability. As before, we separated the cohort into training and test groups. The regressor was trained using Ridge regression [5], with a third degree polynomial of the same features.

3.3 The application

We developed a mobile application for the Android platform. The application provides a UI (user interface) which guides the user through the process of acquiring the required videos, and then locally applies our grading algorithm. The mobile application realizes the projects objectives by combining the ability to both diagnose and grade the severity of FNP in an easy and efficient manner on a readily available mobile platform.

The process of performing a diagnosis consists of three stages. First, the physician places labels (e.g. stickers, washable marker) on 13 predefined facial landmarks (see Fig.1). Secondly, the application guides the user through a short calibration process to validate correct identification of the labels under current lighting conditions and particular choice of labels. The final, and primary, stage consists of capturing video of the 9 predefined facial expressions. A short (approx. 10 seconds) video is captured of each facial expression, and the location of the labels is tracked in real time. Once all videos have been acquired, the application applies the grading algorithm and returns the calculated severity. The capturing process takes between five to ten minutes, and the calculation of the grade itself, once all video sequences have been captured, is less than 500 milliseconds on average. We ran our experiments on a Samsung Galaxy S3.

4 Results

Binary classification results of subjects to sick and healthy are presented in Table 2. We reach an overall accuracy of 95.5% (43/45).

	Classified patient	Classified healthy
Patient	12	2
Healthy subject	0	31

Table 2. Binary classification results.

Grading severity results on the test group are presented in Fig.2 and in Table 3 with respect to the three subjective grading methods. The mean score difference and std reflect the average error between the algorithm's grading and the ground-truth, and the standard deviation, according to the subjective method's scale. The correlation value indicates the level of consistency between the algorithm's grading and the ground-truth labeling, where the maximal possible correlation value is 1. Fig.2 represents subjective and algorithm grading of all subjects according to the three subjective methods. Healthy subjects were automatically labeled as healthy in all methods.

5 Discussion

Our algorithm demonstrates excellent binary classification and high correlation to currently used subjective FNP grading methods. Two Patients which were misclassified as healthy had a very mild disease, difficult to grade also for the physicians observing them. In terms of HBS, this is the minimal error. Our processing time takes less than half a second on a mobile device, comparing to O'Reilly et al. reported time of five minutes on a portable computer. These results indicate that the extracted features correlate well to FNP severity grading, and prove that the process we developed is feasible under the constraints of a mobile platform. Despite the small number of subjects, the algorithm exhibits stable behaviour, independent of the selection of the training and test groups. Further validation of the classifier is needed in a comparative

Scoring Method	Mean score difference (std)	Correlation (p-value)
HBS	0.12 (0.7)	0.894 (1.03e-17)
YGS	0.32 (3.16)	0.959 (5.41e-27)
SGS	1.6 (10.01)	0.94 (8.71e-24)

Table 3. Full severity grading results.

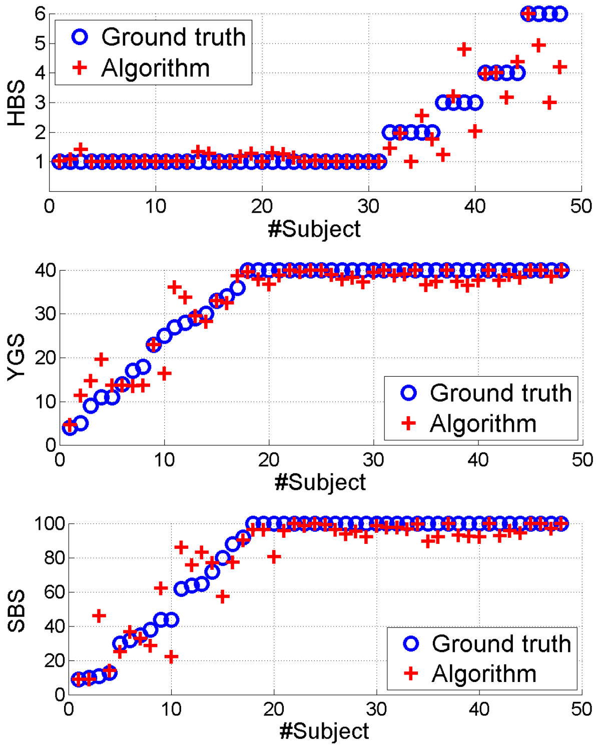

Fig. 2. Full regression results, compared to HBS (top), YGS and SBS (bottom). The X axes represent the subjects in the research, and the Y axes represent the score according to the specified method.

clinical study. Nevertheless, the algorithm is constantly improving with the growth of the dataset and the addition of features

6 Conclusion

We presented an algorithm and a mobile application for diagnosis and grading of facial nerve palsy. A user-friendly mobile application will not only allow physicians to diagnose patients at the clinic, bed side or emergency room, but for the first time will allow patients, suffering a severe psychological condition, to monitor their own condition. Furthermore, an objective and standardized scale is of great significance as a research tool for surgical or medical treatment of FNP.

Future steps in the development of the application will include: detection of facial landmarks without labels, integration of semi-supervised learning techniques to allow self-improvement of the application with usage, and the addition of separate grading of each facial area.

We believe that the progress in the fields of ML and CV holds big promise for the medical field. Such systems will be frequently integrated at health centers for more precise analysis and treatment. Our work is an important step in that direction.

7 Acknowledgments

We are grateful to Oded Cohen M.D. and Gil Zoizner M.D. from Kaplan Medical Center for clinically grading the patients in this study, and also to Johanan Erez from VISL, for coordinating this research.

References

1. Reginald F Baugh, Gregory J Basura, Lisa E Ishii, Seth R Schwartz, Caitlin Murray Drumheller, Rebecca Burkholder, Nathan A Deckard, Cindy Dawson, Colin Driscoll, M Boyd Gillespie, et al. Clinical practice guideline bells palsy. *Otolaryngology–Head and Neck Surgery*, 149(3 suppl):S1–S27, 2013.
2. SA Burres. Objective grading of facial paralysis. *The Annals of otology, rhinology, and laryngology*, 95(3 Pt 1):238–241, 1985.
3. Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.
4. Pavel Dulguerov, Desheng Wang, Thomas V Perneger, Francis Marchal, and Willy Lehmann. Videomimicography: the standards of normal revised. *Archives of Otolaryngology–Head & Neck Surgery*, 129(9):960–965, 2003.
5. Arthur E Hoerl and Robert W Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.
6. John W House and Derald E Brackmann. Facial nerve grading system. *Otolaryngology–head and neck surgery: official journal of American Academy of Otolaryngology-Head and Neck Surgery*, 93(2):146, 1985.
7. Verena Meier-Gallati, Henning Scriba, and Ugo Fisch. Objective scaling of facial nerve function based on area analysis (oscar). *Otolaryngology-Head and Neck Surgery*, 118(4):545–550, 1998.
8. George E Murty, Joseph P Diver, Peter J Kelly, GM O'Donoghue, and Patrick J Bradley. The nottingham system: objective assessment of facial nerve function in the clinic. *Otolaryngology–head and neck surgery*, 110(2):156–161, 1994.
9. John Gail Neely, Katie X Wang, Chelsea A Shapland, Ali Sehizadeh, and Amy Wang. Computerized objective measurement of facial motion: normal variation and test-retest reliability. *Otology & Neurotology*, 31(9):1488–1492, 2010.
10. Brian F O'Reilly, John J Soraghan, Stewart McGrenary, and Shu He. Objective method of assessing and presenting the house-brackmann and regional grades of facial palsy by production of a facogram. *Otology & Neurotology*, 31(3):486–491, 2010.
11. Erik Peitersen. Bell's palsy: the spontaneous course of 2,500 peripheral facial nerve palsies of different etiologies. *Acta Oto-Laryngologica*, 122(7):4–30, 2002.
12. Jürg Rickenmann, Claude Jaquenod, D Cerenko, and Ugo Fisch. Comparative value of facial nerve grading systems. *Otolaryngology–Head and Neck Surgery*, 117(4):322–325, 1997.
13. Brenda G Ross, Gaeton Fradet, and Julian M Nedzelski. Development of a sensitive clinical facial grading system. *Otolaryngology–head and neck surgery*, 114(3):380–386, 1996.
14. Jeffrey T Vrabec, Douglas D Backous, Hamid R Djalilian, Paul W Gidley, John P Leonetti, Sam J Marzo, Daniel Morrison, Matthew Ng, Mitchell J Ramsey, Barry M Schaitkin, et al. Facial nerve grading system 2.0. *Otolaryngology-Head and Neck Surgery*, 140(4):445–450, 2009.
15. David A Wood, Gordon B Hughes, Michelle Secic, and Terry L Good. Objective measurement of normal facial movement with video microscaling. *Otology & Neurotology*, 15(1):61–65, 1994.
16. Naoki Yanagihara. Grading system for evaluation of facial palsy. *Facial nerve. New York: Masson Publishing USA*, pages 41–4, 1985.
17. Thomas L Yen, Colin LW Driscoll, and Anil K Lalwani. Significance of house-brackmann facial nerve grading global score in the setting of differential facial nerve function. *Otology & neurotology*, 24(1):118–122, 2003.