A. Raj, O. Mothes, S. Sickert, G. F. Volk, O. Guntinas-Lichius, J. Denzler:
Automatic and Objective Facial Palsy Grading Index Prediction using Deep Feature Regression.
Annual Conference on Medical Image Understanding and Analysis (MIUA). 2020.


# Automatic and Objective Facial Palsy Grading Index Prediction using Deep Feature Regression

Anish Raj[1], Oliver Mothes[1], Sven Sickert[1], Gerd Fabian Volk[2], Orlando Guntinas-Lichius[2], and Joachim Denzler[1]

[1] Computer Vision Group, Friedrich Schiller University Jena, Ernst-Abbe-Platz 2, 07743 Jena, Germany

[2] Department of Otorhinolaryngology, Jena University Hospital, Am Klinikum 1, 07747 Jena, Germany

**Abstract.** One of the main reasons for a half-sided facial paralysis is a dysfunction of the facial nerve. Physicians have to assess such a unilateral facial palsy with the help of standardized grading scales to evaluate the treatment. However, such assessments are usually very subjective and they are prone to variance and inconsistency between physicians due to their varying experience. We propose an automatic non-biased method using deep features combined with a linear regression method for facial palsy grading index prediction. With an extension of the free software tool *Auto-eFace* we annotated images of facial palsy patients and healthy subjects according to a common facial palsy grading scale. In our experiments, we obtained an average grading error of 11%.

**Keywords:** Automatic Assessment · Facial Palsy · Regression · Deep Learning.

## 1 Introduction

A unilateral facial palsy essentially leads to a one-sided motor dysfunction of the facial muscles and the associated movement disorder [26]. There are multiple reasons to render such a dysfunction. A special case of a peripheral facial palsy is the idiopathic form (*Bell's paresis*) with an incidence of 20-40 / 100,000 individual [24, 26]. It involves a one-sided motor dysfunction of the facial muscles [34]. There are different levels of such a functional deficit. For both diagnosis and treatment, it is important to grade these levels as precisely as possible.

Ideally, they should be measured objectively. For that purpose, several indices and grading schemes were proposed in the past. For instance, very popular and widely used methods are the House-Brackmann grading scale [15], the Sunnybrook grading scale [27] or the newer more fine-grained eFace scaling [1]. However, in practice, they are largely dependent on the experience of the attending physician. Thus, they are prone to variance and inconsistency. Furthermore, the intra-reliability of one physician, as well as the inter-reliability of multiple physicians, is very small [37]. However, in therapy evaluation an objective assessment is in urgent need.

Inspired by this lack of objectivity, we started working on a completely different and more data-driven approach. In this paper, we introduce an automatic facial palsy measurement approach based on regression, that is less subjective and dependent on the medical staff's experience. We will show, how state-of-the-art visual learning techniques can be used to converge on this goal. Basically, the visual inspection by a physician is substituted by a visual learning system trained to recognize different grades of facial palsy in photos of a patient.

In the last decade convolutional neural networks (CNNs) [21] turned into the leading approach in most of the computer vision tasks. Areas that benefited the most are general image classification [19] and face recognition [25]. Recent approaches can make use of subtle details in images. They discriminate between thousands of general object classes or hundreds of sub-classes (e.g. animal species). Especially, works in the field of fine-grained classification [22, 30] shows the strength of deep learning approaches based on CNNs. We demonstrate, that such techniques are also able to measure subtle differences and changes in parts of a human face. Typically, a lot of data is required to train CNNs for such a special task. As a solution, a pre-trained CNN on a similar task can be used for extracting features [7]. Afterwards, these image features can be used to train a task-specific model.

The main contribution of this paper is an approach based on deep learning for an automatic and objective grading index prediction for facial palsy. Instead of training classifiers like the state-of-the-art approaches, we train single linear regressors, each for every static and dynamic eFace [1] sub-scores using deep features of a pre-trained CNN and an annotated image database of facial palsy patients. As a prerequisite an annotation phase was necessary. We extended the open software tool *Auto-eFace*[1] including the Emotricstool [11, 12] to manually grade facial palsy according to various facial palsy index grading systems including the eFace grading system.

## 2 Facial Palsy Assessment

Facial palsy occurs when the facial muscles become frail due to some form of temporary or long term damage to the facial nerve [8]. Both sides of the face have their own set of independent facial nerves. The nerves help with the coordination of facial expressions, and in various ways help regulate taste, saliva and tear production [8]. Facial palsy is classified into two categories: A *Peripheral Facial Paralysis* occurs due to a lack of functioning of the nerve in the pons region of the brainstem. It leads to adverse effects in the lower, middle and upper regions of the facial muscles on the affected side. In the case of a *Central Facial Paralysis*, there is a perturbation of the nerve in the cortical region where only the lower half of the face on one side is affected.

For grading one-sided facial palsy the eFace scale was proposed [1]. It produces an overall disfigurement score of the patient using 16 parameters that

[1] https://github.com/dguari1/Auto-eFace

Table 1: Parameters for eFace ratings [1].

| Category | Parameter no. | Parameter name | Parameter range |
|---|---|---|---|
| **Static** | R1 | Resting brow height | 0-200 |
| | R2 | Resting palpebral fissure width | 0-200 |
| | R3 | Nasolabial fold depth at rest | 0-200 |
| | R4 | Nasolabial fold orientation at rest | 0-200 |
| | R5 | Oral commissure position at rest | 0-200 |
| **Dynamic** | M1 | Brow elevation | 0-100 |
| | M2 | Palpebral fissure narrowing during gentle eye closure | 0-100 |
| | M3 | Palpebral fissure narrowing during full eye closure | 0-100 |
| | M4 | Oral commissure movement with smile | 0-100 |
| | M5 | Nasolabial fold depth with smile | 0-200 |
| | M6 | Nasolabial fold orientation with smile | 0-200 |
| | M7 | Lower lip movement with 'EEEE' sound | 0-100 |
| **Synkinesis** | S1 | Ocular synkinesis | 0-100 |
| | S2 | Midfacial synkinesis | 0-100 |
| | S3 | Mentalis dimpling | 0-100 |
| | S4 | Platysmal synkinesis | 0-100 |

are adjusted by a physician. These parameters are split into three categories *Static*, *Dynamic* and *Synkinesis*. An overview can be found in Tab. 1. For parameters that lie in the range between 0 and 200, the score of 100 indicates a balanced position while 0 or 200 indicates extreme malposition. When it comes to the parameters that lie between 0 and 100, the score of 0 represents extreme paralysis and 100 indicates balanced position or absence of synkinesis. Finally, the sub-score for each category is the average of the values in the respective categories.

## 3 Related Work

There have been multiple works on facial palsy analysis with the help of visual recognition systems and software tools. The authors of [11] propose Emotrics that uses the work of [17] to automatically localize facial keypoints (*landmarks*) in frontal facial images. It is trained on a database of only healthy faces [28] and can be used to compute distances and angular measurements between the landmarks. However, landmark localization is prone to errors for patients with strong facial asymmetry. As a consequence, the user has to manually modify the location of several landmarks. In [12] the authors manually annotated face images of facial palsy patients to train a suitable landmark localization model. For Emotrics to work, at least one iris must be visible in the image. Furthermore, there can not

be head tilt or otherwise, their proposed method fails. However, if landmark localization is improved for palsy patients, Emotrics can produce accurate facial distances and angular measurements that pertain to facial paralysis. Typical examples of such measurements are smile excursion, smile symmetry, and eyebrow symmetry. Additionally, it can be used to compare two images and hence, to keep track of the patients' progress before and after an intervention. In contrast to our method, the approach does not predict a commonly used facial palsy grading index.

In [2] the authors propose an approach that uses facial images to discriminate between normal and facial palsy patients. Additionally, the facial paralysis type (peripheral or central) can also be distinguished. They are able to classify the degree of severity based on the House-Brackmann scale [15]. The authors show how the iris can be extracted and facial landmarks are detected. They employ a combination of local active contour model [20] and Daugman's algorithm [5]. The authors stress the importance of iris segmentation for finding the difference between a healthy and paralyzed side. However, the House-Brackmann scale is not suitable for a fine-grained facial palsy assessment as it contains only a few discrete grades. In contrast, the authors of [23] propose to use active appearance models (AAM) [4] to grade facial paralysis with Stennert scale [32] in addition to House-Brackmann scale. A pre-trained AAM model is fitted on nine images for each patient. Afterwards, the distance between landmarks, fitting parameters from the AAM and the predicted action units [13] are fed as features into a random decision forest [3]. For the House-Brackmann scale, the authors of [23] achieve about 80%, whereas the more fine-grained Stennert score reaches an accuracy of 72% for the rest sub-scores and 66% for the motion sub-scores.

In another work [10], Microsoft's Kinect (v2) is used to grade facial paralysis. The authors propose a new facial palsy grading system for resting symmetry and voluntary movements (e.g. raising eyebrows, closing eyes or smiling). Their system is also able to predict grades for the traditional grading systems described in Sec. 2. The process involves detecting facial landmarks as 3-D coordinates in real-time using Microsoft's Kinect. To calculate the asymmetry of the facial regions the ratios of distances between corresponding landmarks and a common reference point on the two sides of the face are used. Furthermore, their previous resting symmetry assessment module [9] is extended by adding gamma correction, eye area and mouth slope features. However, it is to be noted that they did not test their system on facial palsy patients, but on healthy subjects only.

In this paper, we tackle the problem of objective facial palsy assessment as a linear regression problem by using the fine-scaled eFace grading index [1], which has 16 ordinal fine-grained grading scales for the rest face and facial motions. In contrast to the described state-of-the-art approaches, a facial palsy index classification method is not suitable regarding the fine-grained ordinal sub-scales of the eFace-scale. Hence, an approach based on regression is more appropriate for our automatic prediction task. Besides the eFace grading scale, our approach can also be adapted to predict the other facial palsy scales mentioned in Sec. 1.

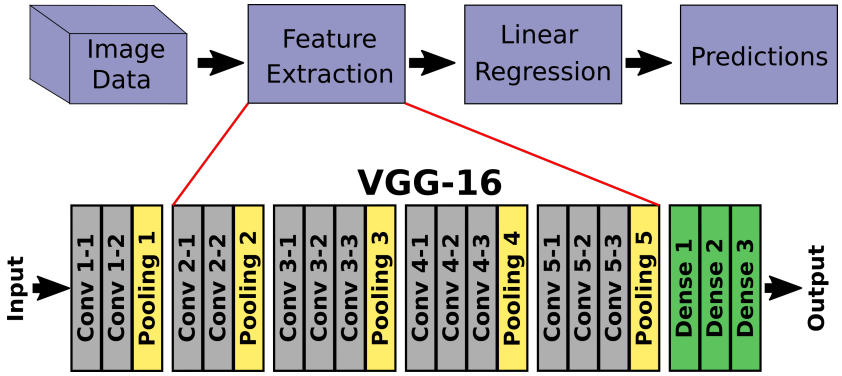


Fig. 1: Image features are extracted from different layers of a pre-trained CNN excerpt (e.g. VGG-16 [31]). These features are then used as input for a linear regression model, that generates the facial palsy score predictions. We use an $\epsilon$-SVR for the regression task.

## 4 Automated Facial Palsy Grading

We now describe our objective automated facial palsy index grading pipeline using a deep learning approach. It includes our extension of the *Auto-eFace* software tool [12] for facial palsy grading index annotation, as it is an important first step to be able to learn sophisticated models. An overview of our whole machine learning pipeline is visualized in Fig. 1.

### 4.1 Facial Palsy Annotation

For facial palsy annotation, we extended the software tool *Auto-eFace*, which was developed for automatic facial asymmetry measurements with the included Emotrics tool [11, 12]. It originally features facial landmark correction and facial asymmetry measurements. In addition, our extension *JAuto-eFace* provides facial palsy assessment based on the grading scales eFace [1], House-Brackmann [15], Stennert [32] and Sunnybrook [27]. Furthermore, we added the functionality to annotate image recording series of two different grading recording standards with 9 and 12 images [29], respectively. The annotator can select an image data folder which contains the recorded images of the patient. All facial palsy scales can be assessed by bars within the corresponding scaling limits, including all the sub-scores. Additionally, after each interaction, the tool saves all rated values, automatically.

### 4.2 Pre-processing

Image pre-processing is performed in mainly three steps. First, the face region is extracted from the images, as background information is not required in the

following steps. To achieve this, a face detector provided by the *Dlib* framework [18] is used. Afterwards, the face images are converted into a $n \times n$ pixel square shape. This step is necessary to train deep learning models that require square images. To reshape initial images into a square while maintaining the aspect ratio, a suitable amount of zero paddings is applied at the image borders. The index scores are pre-processed by rescaling them to the range of $[0, 1]$.

### 4.3 Deep Feature Extraction

To be able to apply linear regression, it is most important to extract meaningful image features. Deep neural networks can learn such features during training time, which can then be used for the task at hand [7]. For the training of such neural networks, a large amount of data is necessary. However, when enough data is not available, it is possible to exploit a pre-trained CNN. Such a network is trained for another task using data of the same or a related image domain [7]. Thus, we can extract deep image features by simply exploiting the activation of layers from this pre-trained CNN. These are the features we use for our regression task. Figure 1 illustrates our proposed machine learning pipeline for automatic facial palsy grading index prediction. It includes the architecture of a VGG-16 [31] model with respective convolutional, pooling and dense layers.

### 4.4 Facial Palsy Index Regression

The extracted deep features are now used to learn a linear regression model for facial palsy index prediction. Specifically, we train an *$\epsilon$-support vector regressor* ($\epsilon$-SVR) as introduced by [35]. It uses the given training data $\{(x_1, y_1), \ldots, (x_L, y_L)\} \subset \mathcal{X} \times \mathbb{R}$, where $x_i \in \mathbb{R}^d$ denotes the extracted deep features and $y_i \in \mathbb{R}$ the corresponding facial palsy score. The goal is to find a hyperplane $f(x) = \langle w, x \rangle + b$. Additionally, a maximum deviation of $\epsilon$ from the ground-truth target value $y$ on average over for all training data is enforced. Given that vector $w$ is perpendicular to the hyperplane $f(x)$, it is sufficient to minimize the norm of $w$, i.e. $||w||^2 = \langle w, w \rangle$. The interested reader is referred to [35] for a detailed derivation. For each facial palsy score, an individual single linear regressor is learned, with $\epsilon = 0.1$. We employed SVR as other regression methods (*e.g.*, random forests or a different kernel) did not perform well in our preliminary experiments. As described in Sec. 4.2 all palsy scores in the pre-processing step are normalized. Thus, in the last step, an inverted scaling is applied to obtain the absolute score values.

## 5 Experiments

In the following, we evaluate our proposed regression approach for predicting facial palsy grading scores. After establishing the experimental setup, we demonstrate automatic eFace prediction for subjects with and without facial palsy.


Fig. 2: During the creation of the dataset, subjects have performed 9 different facial movements: neutral face (1), brow elevation (2), gentle closed eyes (3), squeezed eyes (4), wrinkled nose (5), smile with closed mouth (6), show teeth (7), pursing lips (8) and lowered mouth corners (9) [23].

### 5.1 Datasets and Metrics

**Datasets.** The facial palsy dataset (D1) used in our experiments was recorded by the Department of Otorhinolaryngology of the University Hospital Jena, Germany. Its medical purpose is to evaluate the success of a *Hypoglossus Fazialis Jump Anastomose* performed on facial palsy patients. In this surgery, a part of the nerve of the tongue (hypoglossal nerve) is connected to the facial nerve, to induce reinnervation of the facial muscles [36]. The dataset consists of 2D RGB image series of 52 multi-ethnical patients (22 women, 30 men) of different ages which are recorded before and after the mentioned surgery. Each image series contains nine standardized images of the patient's frontal face. They include a neutral facial expression and eight different facial movements. An example of these nine expression images can be found in Fig. 2. The very same image recording protocol is applied in our second dataset (D2), which includes 28 adult healthy subjects (14 women, 14 men) without facial palsy. This dataset serves as a control study.

**Annotation.** For facial palsy index annotation, we extended the free software tool *Auto-eFace* [11] (see Sec. 4.1) to manually grade different facial palsy grading indices including eFace scale [1] described in Sec. 2. The eFace rest sub-scores R1-R5 are annotated by using the rest image (1), while the eFace motion sub-scores are annotated by using different images, which can be seen in Tab. 3. Please note that synkinesis sub-scores can only be assessed by using video data, where the motion can be clearly seen. Thus, the synkinesis scores S1-S4 are not annotated in both datasets. The facial palsy dataset (D1) was assessed by a medical expert. In contrast, the dataset of healthy subjects (D2) has not been annotated by an expert. We set all facial palsy grading scales for those subjects to the values of a healthy person.

**Evaluation Metrics.** For evaluating the automatic eFace sub-score prediction, we compared the ground-truth annotations with the predictions of our regression models. We achieve this by calculating the mean absolute error (MAE) of the grading score. The MAE is defined as

$$MAE = \frac{\sum_{i=1}^{n} |y_i - \hat{y}_i|}{n}, \tag{1}$$

Table 2: The combined MAE (± std dev) of all the eFace sub-scores including static and dynamic parameters from the VGG-faces model is decreasing with increasing depth of the CNN-layer from which features are extracted.

| **Layer** | **Conv2-2** | **Conv3-3** | **Conv4-3** | **Conv5-3** | **Pooling5** |
|---|---|---|---|---|---|
| **MAE** | $27.44 \pm 6.92$ | $25.96 \pm 6.67$ | $24.39 \pm 6.59$ | $\mathbf{22.43} \pm 5.77$ | $22.45 \pm 5.82$ |

where $y_i$ are ground truth values, $\hat{y}_i$ the predicted values and $n$ is the total number of observations. The data is split into train and test set to run an 11-fold cross-validation. Hence, no recording series of a patient is in both train set and test set at the same time.

### 5.2 Automatic eFace Prediction

**Setup.** For evaluation, the introduced datasets D1 and D2 are combined and split for 11-fold cross-validation. Thus, there are images of facial palsy patients and healthy subjects in the train data splits. Note, for evaluation of the test splits both the datasets D1 and D2 are used. Furthermore, healthy subjects are evaluated in a control study.

For the individual eFace sub-score predictions we use different images of each recording series regarding the performed motions. We used the rest image (1) for predicting the eFace rest sub-scores R1-R5. For prediction of the eFace motion sub-scores M1-M7, different images (2,3,4,6,7) are utilized, which can be seen in Tab. 3. As the synkinesis sub-scores S1-S4 are not labelled in the facial palsy dataset (D1), we have neglected them in this experiment.

For feature extraction, described in Sec. 4.3, two different pre-trained CNN models with a VGG-16 architecture [31] are used. Both CNNs are trained on completely different data with varying tasks. The VGG-faces model [25] is pre-trained on facial data for face recognition. For comparison, we use the exact same VGG-16 architecture trained for an object recognition task using *Imagenet* dataset [6]. Before extracting features by exploiting the layer activations of the CNN models, pre-processing is applied to all images (see Sec. 4.2). Specifically, all images are scaled to a shape of $224 \times 224$ pixels to match the input shape for the VGG-16 model.

**Layer Influence.** To decide which CNN layer has the most descriptive features for the eFace sub-score prediction, in a first experiment, we extract the features of various layers of the VGG-faces model. For each eFace score as well as for each feature set of a chosen layer an individual linear regressor is trained. They are all evaluated using MAE. Afterwards, for each layer, a combined evaluation is applied by calculating the average of all sub-score errors.

Results of this experiment are summarized in Tab. 2. It shows the combined MAEs of the different chosen layers of the VGG-faces model. As can be seen,

Table 3: The linear regressors exploiting the extracted features of the VGG-faces model lead on average to an MAE of 22.44. In comparison, the VGG model trained on the *Imagenet* dataset [6] reaches an average MAE of 24.34. The errors are indicated in standard deviations.

| **Param.** | **Image** | **VGG-faces** | **VGG-imagenet** |
|---|---|---|---|
| R1 | 1 | **13.17** ± 13.54 | 15.16 ± 14.38 |
| R2 | 1 | **30.16** ± 23.98 | 36.24 ± 28.70 |
| R3 | 1 | **23.91** ± 22.15 | 25.49 ± 20.98 |
| R4 | 1 | **25.60** ± 19.82 | 26.24 ± 21.87 |
| R5 | 1 | 23.32 ± 19.35 | **23.20** ± 19.62 |
| M1 | 2 | **31.22** ± 17.88 | 31.86 ± 19.55 |
| M2 | 3 | **19.07** ± 17.83 | 23.67 ± 19.73 |
| M3 | 4 | **15.65** ± 15.51 | 16.81 ± 15.43 |
| M4 | 6 | 24.98 ± 18.47 | **23.79** ± 16.99 |
| M5 | 6 | **28.48** ± 24.00 | 28.67 ± 23.48 |
| M6 | 6 | **17.47** ± 12.28 | 25.15 ± 20.83 |
| M7 | 7 | 16.24 ± 14.20 | **15.74** ± 12.42 |

the deeper Conv5-3 and Pooling5 layers of VGG-faces have nearly similar MAE. Results suggest that the features of later layers represent the facial information better. As a consequence, we have chosen the Conv5-3 layer for the feature extraction process for our further experiments.

**Pre-Training.** In our next experiment, we compare different weights of the VGG architecture, which in turn provides different features for the linear regressor model. For comparing different pre-trained VGG models, Tab. 3 shows the MAE and standard deviations based on a 11-fold cross-validation. It shows results for both the predicted static as well as dynamic eFace sub-scores.

With an average MAE of 22.44, the features of the VGG-faces model perform slightly better than the features of the VGG model trained on *Imagenet*. The latter model reaches an average MAE of 24.34. In other words, the regressors trained with features of VGG-faces achieve an error of 11%, while the regressors trained with *Imagenet* features have an error of 12%. It should not come as a surprise that features generated from purely facial data perform better. However, it is worth noting that the difference is not very large. A model trained with non-facial images of high diversity from 1000 different object classes can also extract descriptive features for faces. Since there has not been any previous work regarding automated eFace scores prediction, we do not have a baseline to compare our results to. However, these results can act as a baseline for future works.

In Fig. 3 we illustrate our results of the static and dynamic eFace scores more detailed as box plots indicating standard deviations and outliers. It can be

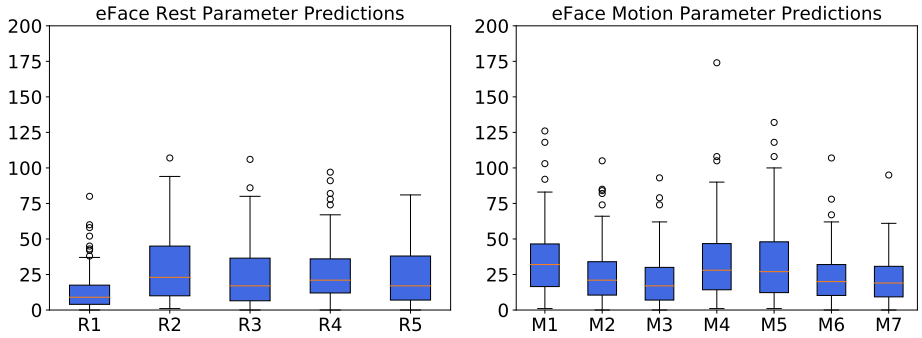


Fig. 3: The comparison of the single linear regressors trained for each eFace sub-score show different absolute eFace prediction errors. For instance, the brow height sub-score in the rest image (1) performs quite well with an MAE of 13.17, while the brow elevation sub-score of motion image (2) performs only with an MAE of 31.22.

easily seen that the single linear regressors trained for each eFace sub-score show different absolute eFace prediction errors. Overall the static sub-scores perform with an MAE of 23.23, while the dynamic sub-scores perform with an MAE of 21.87. This difference can be explained by the fact that even for an expert it is easier to evaluate a eFace sub-score in a motion image than in the rest image.

As there are more advanced architectures for CNNs available that perform better on *ImageNet* dataset we like to mention that we also trained regressors using such architectures. For instance, on a ResNet-50 model [14], we achieve an average MAE of 28.90 by extracting features from activations of Resnet-50s last convolutional layer. It suggests that a deeper network does not necessarily produce more representative features for our task. Table 4 shows the average MAE for some recent CNN architectures pre-trained on different datasets.

### 5.3 Control Study

To validate our results in a control study, we used dataset D2 including 28 healthy subjects. For that study, we only evaluate the healthy subjects of the test splits. We obtain an average MAE over all eFace sub-scores of 12.82, where the average MAE of the static sub-scores is 9.76. The average MAE of the dynamic sub-scores is 15.01. For our case study, this range of errors is acceptable, since we can not assume that subjects have a completely symmetric face. This assumption is supported, for instance, by measurements of the Emotrics tool [11, 12] in *Auto-eFace*. With the Emotrics tool we measured the distances between eyebrows and the related eye pupils of all subjects in dataset D2 using the rest image (1). On average the left and the right distances differed by $5.93 \pm 4.62$ %.

Table 4: Linear regressors exploiting features from the last convolutional layer of recent network architectures pre-trained on *Imagenet* dataset [6] do not achieve the average accuracy as the VGG-faces model based on VGG architecture pre-trained on facial data [25]. The errors are indicated using standard deviations.

| **Architecture** | **Pre-trained Dataset** | **Average MAE** |
|---|---|---|
| **ResNet-50** [14] | Imagenet [6] | $28.90 \pm 6.62$ |
| **Mobilenet** [16] | Imagenet [6] | $23.94 \pm 6.70$ |
| **EfficientNet-B4** [33] | Imagenet [6] | $26.99 \pm 7.38$ |
| **VGG-16** [31] | Imagenet [6] | $24.34 \pm 6.32$ |
| **VGG-faces** [25] | Faces [25] | $\mathbf{22.44 \pm 6.02}$ |

## 6 Conclusions

In this paper, we proposed a machine learning approach for automatic and objective facial palsy grading index prediction. As a prerequisite, we extended an existing annotation tool to work for various facial palsy grading scales. In our experiments, we were able to demonstrate that the eFace grading scale can be predicted by our approach with an average MAE of 22.43. Moreover, we found that deeper networks like ResNet-50 do not provide more suitable features for our application. They contain much more parameters than a typical VGG-16 model when fully connected layers are excluded. Additionally, we verified our results in a control study by applying our method to healthy subjects with an average MAE of 12.82.

Our machine learning approach could potentially also be applied to other ordinal interval facial palsy scales like House-Brackmann [15] or Sunnybrook [27]. Furthermore, our dataset with 52 facial palsy patients is quite small, We expect to further reduce the MAE given more annoated data during training. In addition, data augmentation techniques like image flipping would support these efforts. It completely removes a potential left/right bias for facial palsy cases in the training data.

We believe that our proposed data-driven approach is a step towards more objective grading. The system can help physicians prepare a better treatment plan for each patient, while, a patient can use a smartphone-based app at home to track the progress of their therapy. However, one point of criticism could be that by subjectively annotating the data by one medical expert, the model also makes subjective predictions. To avoid this, we suggest that several experts annotate the same data and, thus, averaged grades are used for training. This prevents the subjective decision of an individual and averages the assessment. Besides, an individual expert becomes a weighting based on its annotation experience. Simultaneously, an inter-rater evaluation is possible and we can compare the

model accuracy with the human assessment performance. Those are points for further work and are not yet implemented in our approach.

## Acknowledgments

The research was supported by grant DE 735/15-1 and GU 463/12-1 of the German Research Foundation (DFG). We also gratefully acknowledge the support of NVIDIA Corporation with the donation of Titan Xp GPUs used for this research.

## References

1. Banks, C.A., Bhama, P.K., Park, J., Hadlock, C.R., Hadlock, T.A.: Clinician-graded electronic facial paralysis assessment: the eface. Plastic and Reconstructive Surgery **136**(2), 223e–230e (2015)
2. Barbosa, J., Lee, K., Lee, S., Lodhi, B., Cho, J.G., Seo, W.K., Kang, J.: Efficient quantitative assessment of facial paralysis using iris segmentation and active contour-based key points detection with hybrid classifier. BMC Medical Imaging **16**(1), 23 (Mar 2016)
3. Breiman, L.: Random forests. Machine Learning **45**(1), 5–32 (2001)
4. Cootes, T.F., Edwards, G.J., Taylor, C.J.: Active appearance models. IEEE Transactions on Pattern Analysis & Machine Intelligence **23**(6), 681–685 (Jun 2001)
5. Daugman, J.: How iris recognition works. In: The Essential Guide to Image Processing, pp. 715–739. Elsevier (2009)
6. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 248–255. IEEE (2009)
7. Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E., Darrell, T.: Decaf: A deep convolutional activation feature for generic visual recognition. In: International Conference on Machine Learning. pp. 647–655 (2014)
8. Finsterer, J.: Management of peripheral facial nerve palsy. European Archives of Oto-Rhino-Laryngology **265**(7), 743–752 (2008)
9. Gaber, A., Faher, M.F., Wahed, M.A.: Automated grading of facial paralysis using the kinect v2: A proof of concept study. In: International Conference on Virtual Rehabilitation. pp. 258–264. IEEE (2015)
10. Gaber, A., Taher, M.F., Wahed, M.A.: Quantifying facial paralysis using the kinect v2. In: Annual International Conference of the IEEE Engineering in Medicine and Biology Society. pp. 2497–2501. IEEE (2015)
11. Guarin, D.L., Dusseldorp, J., Hadlock, T.A., Jowett, N.: A machine learning approach for automated facial measurements in facial palsy. JAMA Facial Plastic Surgery **20**(4), 335–337 (Jul 2018)
12. Guarin, D.L., Yunusova, Y., Taati, B., Dusseldorp, J.R., Mohan, S., Tavares, J., van Veen, M.M., Fortier, E., Hadlock, T.A., Jowett, N.: Toward an automatic system for computer-aided assessment in facial palsy. arXiv preprint arXiv:1910.11497 (2019)
13. Haase, D., Kemmler, M., Guntinas-Lichius, O., Denzler, J.: Efficient measuring of facial action unit activation intensities using active appearance models. In: Machine Vision Applications. pp. 141–144 (2013)

14. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 770–778. IEEE (2016)
15. House, J.W., Brackmann, D.E.: Facial nerve grading system. Otolaryngology–Head and Neck Surgery **93**(2), 146–147 (1985)
16. Howard, A.G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., Adam, H.: Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861 (2017)
17. Kazemi, V., Sullivan, J.: One millisecond face alignment with an ensemble of regression trees. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 1867–1874 (2014)
18. King, D.E.: Dlib-ml: A machine learning toolkit. Journal of Machine Learning Research **10**, 1755–1758 (2009)
19. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems. vol. 25, pp. 1097–1105 (2012)
20. Lankton, S., Tannenbaum, A.: Localizing region-based active contours. IEEE Transactions on Image Processing **17**(11), 2029–2039 (2008)
21. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proceedings of the IEEE **86**(11), 2278–2324 (1998)
22. Lin, T.Y., RoyChowdhury, A., Maji, S.: Bilinear cnn models for fine-grained visual recognition. In: Proceedings of the IEEE international conference on computer vision. pp. 1449–1457 (2015)
23. Modersohn, L., Denzler, J.: Facial paresis index prediction by exploiting active appearance models for compact discriminative features. In: International Conference on Computer Vision Theory and Applications. pp. 271–278 (2016)
24. Morales, D.R., Donnan, P.T., Daly, F., Staa, T.V., Sullivan, F.M.: Impact of clinical trial findings on bell's palsy management in general practice in the uk 2001–2012: Interrupted time series regression analysis. BMJ Open **3**(7), e003121 (2013)
25. Parkhi, O.M., Vedaldi, A., Zisserman, A.: Deep face recognition. In: British Machine Vision Conference. p. 6 (2015)
26. Plumbaum, K., Volk, G., Boeger, D., Buentzel, J., Esser, D., Steinbrecher, A., Hoffmann, K., Jecker, P., Mueller, A., Radtke, G., Witte, O., Guntinas-Lichius, O.: Inpatient treatment of patients with acute idiopathic peripheral facial palsy: A population-based healthcare research study. Clinical Otolaryngology **42**(6), 1267–1274 (2017)
27. Ross, B.G., Fradet, G., Nedzelski, J.M.: Development of a sensitive clinical facial grading system. Otolaryngology—Head and Neck Surgery **114**(3), 380–386 (1996)
28. Sagonas, C., Tzimiropoulos, G., Zafeiriou, S., Pantic, M.: 300 faces in-the-wild challenge: The first facial landmark localization challenge. In: IEEE International Conference on Computer Vision - Workshops. pp. 397–403 (2013)
29. Schaede, R.A., Volk, G.F., Modersohn, L., Barth, J.M., Denzler, J., Guntinas-Lichius, O.: Video instruction for synchronous video recording of mimic movement of patients with facial palsy. Laryngo-Rhino-Otologie (2017)
30. Simon, M., Rodner, E.: Neural activation constellations: Unsupervised part model discovery with convolutional networks. In: IEEE International Conference on Computer Vision. pp. 1143–1151. IEEE (2015)
31. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)

32. Stennert, E., Limberg, C., Frentrup, K.: An index for paresis and defective healing–an easily applied method for objectively determining therapeutic results in facial paresis (author's transl). Hno **25**(7), 238–245 (1977)
33. Tan, M., Le, Q.: EfficientNet: Rethinking model scaling for convolutional neural networks. In: Chaudhuri, K., Salakhutdinov, R. (eds.) Proceedings of the 36th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 97, pp. 6105–6114. PMLR, Long Beach, California, USA (09–15 Jun 2019)
34. Thielker, J., Geißler, K., Granitzka, T., Klingner, C., Volk, G., Guntinas-Lichius, O.: Acute management of bell's palsy. Current Otorhinolaryngology Reports **6**(2), 161–170 (2018)
35. Vapnik, V.N.: The Nature of Statistical Learning Theory. Springer-Verlag New York, Inc. (1995)
36. Volk, G.F., Geitner, M., Geißler, K., Thielker, J., Raslan, A., Mothes, O., Dobel, C., Guntinas-Lichius, O.: Functional outcome and quality of life after hypoglossal-facial jump nerve suture. Frontiers in surgery **7**, 11 (2020)
37. Volk, G.F., Schaede, R.A., Thielker, J., Modersohn, L., Mothes, O., Nduka, C.C., Barth, J.M., Denzler, J., Guntinas-Lichius, O.: Reliability of grading of facial palsy using a video tutorial with synchronous video recording. The Laryngoscope **129**(10), 2274–2279 (2019)