OPEN

PRS GLOBAL OPEN

ORIGINAL ARTICLE
Technology

Toward a Universal Measure of Facial Difference Using Two Novel Machine Learning Models

Abdulrahman Takiddin, MSc*
Mohammad Shaqfeh, PhD†
Osman Boyaci, MSc*
Erchin Serpedin, PhD*
Mitchell A. Stotland, MD, MS, FRCSC‡§

Background: A sensitive, objective, and universally accepted method of measuring facial deformity does not currently exist. Two distinct machine learning methods are described here that produce numerical scores reflecting the level of deformity of a wide variety of facial conditions.

Methods: The first proposed technique utilizes an object detector based on a cascade function of Haar features. The model was trained using a dataset of 200,000 normal faces, as well as a collection of images devoid of faces. With the model trained to detect normal faces, the face detector confidence score was shown to function as a reliable gauge of facial abnormality. The second technique developed is based on a deep learning architecture of a convolutional autoencoder trained with the same rich dataset of normal faces. Because the convolutional autoencoder regenerates images disposed toward their training dataset (ie, normal faces), we utilized its reconstruction error as an indicator of facial abnormality. Scores generated by both methods were compared with human ratings obtained using a survey of 80 subjects evaluating 60 images depicting a range of facial deformities [rating from 1 (abnormal) to 7 (normal)].

Results: The machine scores were highly correlated to the average human score, with overall Pearson's correlation coefficient exceeding 0.96 ($P < 0.00001$). Both methods were computationally efficient, reporting results within 3 seconds.

Conclusions: These models show promise for adaptation into a clinically accessible handheld tool. It is anticipated that ongoing development of this technology will facilitate multicenter collaboration and comparison of outcomes between conditions, techniques, operators, and institutions. *(Plast Reconstr Surg Glob Open 2022;10:e4034; doi: 10.1097/GOX.0000000000004034; Published online 18 January 2022.)*

INTRODUCTION

Although individuals with congenital or acquired facial conditions may present with abnormalities across a spectrum of severity, even relatively subtle differences can result in considerable psychosocial impact.[1] However, because a sensitive, objective, and universally accepted method of measuring facial deformity does not currently exist, there is also a lack of reliable means to assess the benefits of reconstructive facial surgery. Most medical practitioners are able to plan and evaluate their treatments based on some combination of laboratory values, functional measures, or radiologic and pathologic findings. Facial reconstructive surgeons, however, are resigned to working almost exclusively with subjective assessments (ie, examining "before and after" photographs) or anthropometric measurements that may not faithfully reflect the complexity of human perception of facial appearance.

For the purposes of clinical evaluation and comparison of outcomes, it would be useful to create a scale of deformity across broad populations against which any face—and any facial disorder—could objectively be measured. It is challenging, however, for human raters to establish a gradient of facial form. Sorting through large numbers of faces requires the recognition and active retention of vast amounts of perceptual information, something better suited to a machine system than to the

*From the *Electrical and Computer Engineering Department, Texas A&M University, College Station, Tex.; †Electrical and Computer Engineering Department, Texas A&M University, Doha, Qatar; ‡Division of Plastic, Craniofacial and Hand Surgery, Sidra Medicine, Doha, Qatar; and §Weill Cornell Medical College, Doha, Qatar.*

Received for publication October 7, 2021; accepted November 9, 2021.

DOI: 10.1097/GOX.0000000000004034

Disclosure: *The authors have no financial interest to declare in relation to the content of this article. This study was supported by the Qatar National Research Fund, NPRP Cycle-13 (NPRP13S-0119-200108).*

human mind. In addition, human appraisal is influenced by personal or cultural preferences, as well as cognitive biases based on factors such as age, gender, race, and professional background. Here, we describe two computer models that integrate data from an extensive and diverse population of normal faces and are able to then score any newly encountered facial image in an impartial and predictable manner. Being able to determine where a particular face falls within the spectrum of normality is an essential task for the facial reconstructive surgeon, and one that until now has relied almost exclusively on intuition.

Previously, our team used a generative adversarial network facial generator[2] to produce authentic, normalized analogues of raw facial images exhibiting deformity. The model we developed was also able to calculate the perceptual distance between the normalized face and its raw abnormal counterpart, yielding scores that correlated closely with the human ratings of deformity.[3] However, because that method turns out to be computationally inefficient, it does not lend itself easily to adaptation into a portable application for clinical use. In the current report, we propose two alternative design approaches that avoid the steps of image normalization and perceptual distance measurement, and thus require less processing power: (1) an object detector model based on an ensemble of boosted Haar Cascade classifiers, relying on the confidence level of the system to discriminate gradations of abnormality, and (2) a convolutional autoencoder model, relying on the reconstruction error of the model as an indicator of deviation from the norm. Similar to our earlier method, both the object detector and the convolutional autoencoder models are shown here to generate facial scoring of a wide variety of facial conditions that correlates closely with human scoring. Moreover, we demonstrate that the object detector approach can be tuned to not only holistically evaluate a face, but to judge discrete aesthetic units within a face, thereby potentially enhancing sensitivity to subtle differences in the orbital, nasal, and oral regions. Placing this type of technology into the hands of the clinician in the future could usher in a paradigm shift in the way patients, surgeons, researchers, and third-party payers interpret the clinical problem of facial deformity and the potential benefits of corrective surgical intervention.

Takeaways

Question: The field of facial reconstruction is hampered by the absence of an objective and clinically practical means of measuring disease severity and clinical outcome.

Findings: We designed two machine learning models that generate a numerical score of normality for any face. The models were trained using images of 200,000 normal faces, and tested on 30 images reflecting a range of facial deformity. The machine-generated data closely correlate with scores obtained from a cohort of human raters.

Meaning: Development of these systems will allow for a universal and objective means of assessing quality of clinical outcome by facilitating a meaningful comparison between techniques, surgeons, and institutions.

METHODS

Data Preparation

An estimated 200,000 images of normal faces were used to train both the object detector and the convolutional autoencoder. We produced these images using the StyleGAN facial image generator, which fabricates highly realistic facial images that are demographically well distributed and reflect a range of lighting, pose, and expression (Fig. 1).[2] All images used in this study were in RGB mode and scaled to a common size of 224 × 224 pixels to align with our lower resolution testing dataset.

For the training of the object detector, a second group of negative (ie, nonface) images was required. For this, we used the Canadian Institute for Advanced Research 100 dataset, which consists of 100 classes of images (600 images per class).[4] We excluded all facial classes from the dataset, leaving only objects such as nature, fruits, vegetables, electrical devices, and buildings. In total, this negative training group consisted of 47,500 images.

Both measurement techniques that we developed were tested using 30 open-source images of facial deformity (licensed for re-use under the Creative Commons, Mountain View, Calif.), as well as 30 normal faces fabricated with the StyleGAN (unique from the 200,000 normal faces used in the training phase). The 30 images depicting deformity included 12 women, 12 men, and six infants of indeterminate gender seven adults, 23 children; and a diversity of ethnic backgrounds as outlined in the Results section. Eighty volunteers aged 18–65 rated the 60 images on a 1–7 Likert scale (1: most deformed, 7: most normal).

Object Detector Method

We implemented the OpenCV version of the Haar Cascade object detector, which has been shown to work efficiently and reliably with resource-constrained devices.[5] The Haar Cascade object detector scans images using a sliding window approach, summating pixel data in adjacent rectangular areas as it progresses sequentially across an image. A given classifier is defined by the measured difference in pixel sums between adjacent areas, relative to a determined threshold value. The Viola-Jones method[6] was applied, computing edge, line, and diagonal (four-rectangle) image features to target known facial properties (ie, orbital region darker than upper-cheeks, nasal bridge brighter than nasal sidewalls, etc.) (Fig. 2). As per the original description, our approach involved the use of an adaptive boosting algorithm (AdaBoost)[7] that takes weak classifiers and uses them to incrementally build a much better, stronger classifier by optimizing the weights for, and adding, one weak classifier at a time. To enhance the model's accuracy and efficiency, we tuned the hyperparameters of image scaling and k-nearest neighbors using a sequential grid-search hyperparameter optimization approach.[8] As the Haar process cascaded forward through stages, the optimal hyperparameters were obtained for each

Fig. 1. Representative sample of normal faces fabricated by the facial generator (StyleGAN) that provided the 200,000-image training database for both study models.

cascade for each image, allowing background regions to be discarded more quickly and more computational attention to be placed on promising areas of the image.

Following the training phase, we derived a confidence score for our test images by taking into account the following considerations: (1) each weak classifier is essentially a one-level decision tree; (2) each decision is associated with a threshold value in pixel sums between the adjacent rectangular areas of its Haar-feature; (3) the confidence score of a single weak classifier is the difference between its value and its threshold; and (4) during the AdaBoost training stage, each selected weak classifier gets associated with a relative weight. Therefore, all weak classifiers in a single stage of the cascade can be combined in a weighted manner, and the confidence score for each stage can be calculated (Fig. 3). Because final strong classifiers are determined by weighted majority "voting" of all weak classifiers, a higher percentage of weak classifiers in favor of the presence of a face equates to a higher confidence of facial detection [expressed here on a 0 (abnormal) to 10 (normal) scale)].[9]

Convolutional Autoencoder Method

We designed an unsupervised anomaly detector based on a connected convolutional neural network and autoencoder [ie, a convolutional autoencoder (CAE)] that was

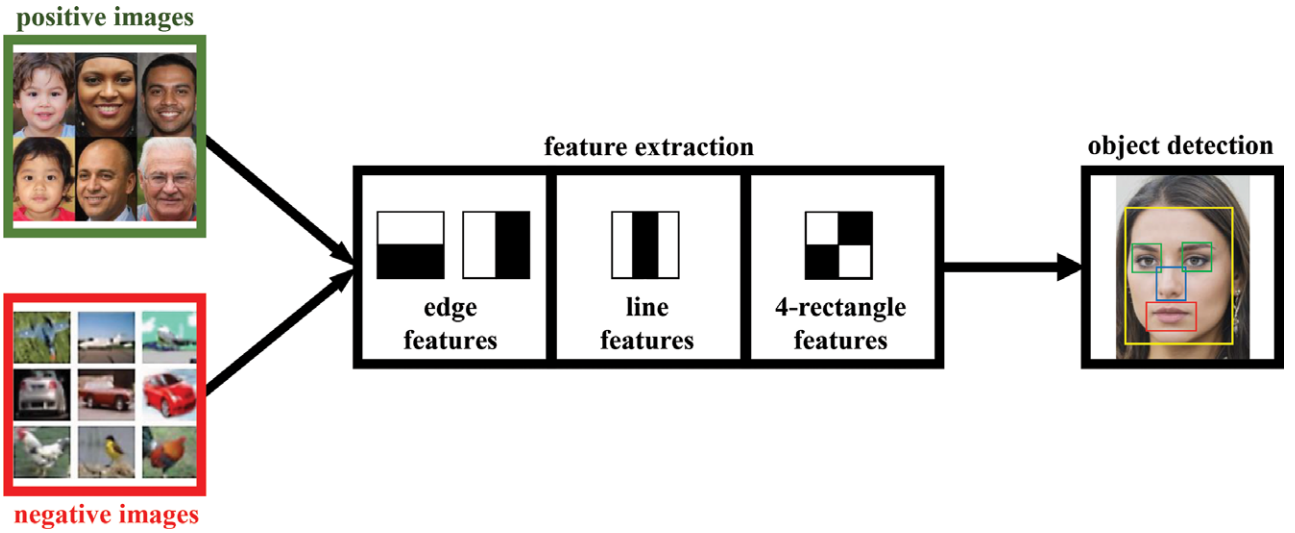

Fig. 2. Training input and example of Haar feature extraction for object detector method. The whole face, as well as defined orbital, nasal, and oral aesthetic subunits were considered.

trained on images of normal faces and tested on images of normal and abnormal faces. The architecture of our CAE, which is similar to the VGG16 architecture,[10] is depicted in Figure 4.

As depicted, the convolutional encoder takes the input image and processes it through multiple convolutional encoder layers that reduce the image dimensions from [224 × 224 × 3] to [14 × 14 × 512]. The height and width of the volumes (image input: [224 × 224] pixels) progressively decrease throughout the convolutional layers, while the depth, which represents the number of feature maps, increases as image features are extracted. Following a pooling layer, a one-dimensional vector is reached at the fully connected layer. Within the second half of the construct, the convolutional decoder receives the encoder output and reconstructs it through multiple convolutional layers. The convolutional decoder reconstructs the image by increasing the volume from [14 × 14 × 512] to the original dimensions of [224 × 224 × 3]. Similar to VGG16, all convolutional layers in our model have 3 × 3 filters. During the training process, the CAE learns the forming features of the input normal images, as the autoencoder model learns the parameters required to minimize the reconstruction error of output versus input. The score of the reconstruction error is derived from a calculated cost function (Fig. 5) expressed on a 0 (normal) to 1 (abnormal) scale, where X_{TR} denotes the training set. During testing, the reconstruction error should be small for normal facial images having similar forming representations as the training dataset, while the score should be progressively higher for those images displaying more dissimilar facial features.

Input: An input image x, the total number of cascade stages T; the number of classifiers at stage t, N_t; the function to calculate the feature value of each classifier $h_{t.n}(x)$; the weight of each classifier $\alpha_{t.n}$; the threshold value of boosted classifier at stage t, θ_t; initialize the confidence score at stage t to be zero $s_t(x) = 0$; initialize the overall confidence score to be zero $s(x) = 0$.

Output: The final confidence score $s(x)$

for $t = 1$ to T **do**
  $h_t(x) = \sum_{n=1}^{N_t} \alpha_{t.n} h_{t.n}(x)$
  $s_t(x) = h_t(x) - \theta_t$
  if $s_t(x) \geq 0$ **then**
    $s(x) = s(x) + s_t(x)$
  else
    $s(x) = 0$
    Break and exit the loop
  end if
end for

Fig. 3. Algorithm used to derive confidence score of Haar Cascade object detector.

As for the Haar object detector method, we conducted a sequential grid-search hyperparameter optimization algorithm for the CAE to select the best combination of hyperparameters. After running the algorithm, we confirmed that the activation function of the internal layers is ReLU. The training process was done using 500 epochs with stochastic gradient descent, a learning rate of 0.01, and a momentum of 0.9.

Experiments were repeated in triplicate and the results were reported in terms of the average testing set (X_{TS}) for all experiments. For all 60 images of facial deformity that we tested, a Pearson correlation was calculated between human rating and object detector confidence score, and between human rating and CAE reconstruction error. Significance was set at a P value less than 0.05.

Both machine learning models described in this study were trained and tested using the same machine setup using an NVIDIA GeForce RTX 2080 hardware accelerator (NVIDIA, Santa Clara, Calif.) and Python 3.7 (Python Software Foundation, www.python.org).

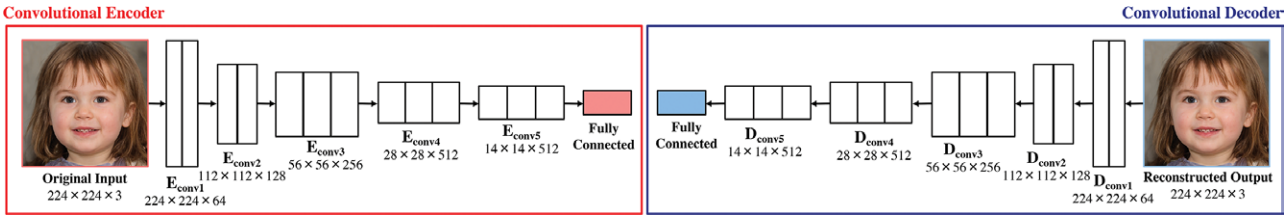

Fig. 4. Schematic illustration of implemented convolutional autoencoder architecture.

RESULTS

Diagnoses, reference numbers, human ratings, Haar Cascade object detector confidence scores, and CAE reconstruction errors for eight representative abnormal testing images (open-source and licensed for re-use through the Creative Commons) are listed in Figure 6.[11–18] Note that for the object detector method, we were also able to obtain individual confidence scores for aesthetic subunits of the face.

For all 60 images of facial deformity and normality that we tested, a close correlation was found between human rating and object detector confidence score (r = 0.96, $P < 0.00001$, Fig. 7), and between human rating and CAE reconstruction error (r = 0.98, $P < 0.00001$, Fig. 8). The diagnoses and all human and machine scoring for the 30 training images are listed in Table 1.

For the Haar Cascade, the average time to report the facial rating score for all four facial segments (orbital, nasal, oral, and full face) for one image was 3.2 seconds, whereas the CAE required 1.2 seconds to report the abnormality score for the full face.

DISCUSSION

With recent advancements in artificial intelligence, a rich set of computational methods and platforms are readily available for use and development.[19–21] This provides a great opportunity for end-users in industries such as healthcare who are seeking to match new solutions to longstanding problems.[22–24] Because the sensitive detection and recognition of faces has become almost commonplace today, it is inevitable that modern computer methods be adapted for use in the evaluation of faces within the clinical setting. Our intention in this work was to construct a universal facial rating system that would align with human perception, and offer itself as an objective and clinically accessible modality for gauging any type of facial impairment and assessing surgical outcomes. However, there are some important issues to bear in mind when considering the automated judgment of facial appearance.

Let $\boldsymbol{E} = f_\Theta(\boldsymbol{x})$ and $\boldsymbol{D} = g_\Theta(\boldsymbol{x})$ stand for the encoder and decoder, respectively, where $\boldsymbol{x}$ denotes the input image that is part of $\boldsymbol{X}_{\text{TR}}$ and Θ represents the autoencoder parameters, which are determined as minimizers of:

$$\min_{\Theta} C(\boldsymbol{x}, g_\Theta(f_\Theta(\boldsymbol{x}))), \qquad \boldsymbol{x} \in \boldsymbol{X}_{\text{TR}}. \tag{1}$$

The cost function $C(\boldsymbol{x}, g_\Theta(f_\Theta(\boldsymbol{x})))$ denotes the mean squared error (MSE),

$$MSE = \frac{1}{m}\sum_{i=1}^{m}(\boldsymbol{x} - g_\Theta(f_\Theta(\boldsymbol{x})))^2, \tag{2}$$

and it is used to penalize $g_\Theta(f_\Theta(\boldsymbol{x}))$ for being dissimilar (deviation) from $\boldsymbol{x}$. Parameter m defines the image dimension (e.g., $224 \times 224 \times 3$).

Fig. 5. Algorithm used to derive reconstruction error of convolutional autoencoder model.

First, no ground truth exists that can be applied as a reference standard. The detection and appraisal of perceptual difference within a face is inherently an idiosyncratic task. Yet, although human ratings may be influenced by bias or self-censorship, they also represent highly evolved judgments that integrate vast amounts of perceptual information within the context of cultural predilections. Existing measurement methods that target anatomic landmarks and focus discretely on factors such as symmetry, proportion, or gender averageness are unable to capture the holistic information instinctively being processed by a human rater.[25–32]

Another issue that should be considered when measuring faces is the concept referred to as *lookism*, a form of social discrimination that has been widely discussed elsewhere.[33] Despite an extensive body of literature outlining the biological and evolutionary foundations of our attraction to beauty,[34] it is appreciated that a range of injustice is unwittingly committed upon those whose faces are perceived as relatively less appealing.[35,36] The introduction of a mechanized rating of appearance could therefore be seen as posing a hazard if used to expose and disparage individuals who rate poorly. However, precisely because evidence shows that the perception of facial appearance is so fundamentally hardwired into our cognitive processes,[37] it is unreasonable to expect humans to refrain from appraising faces. Ultimately, the ethical implementation of a facial measurement tool should be seen as no different than the expectation for right-minded behavior in response to any of the various and readily perceived differences between people (age, gender, race, habitus, etc.).

In terms of functioning as a meaningful arbiter of the human face, a computer system ideally should demonstrate (1) alignment with human judgment, (2) order-preservation, (3) indifference to extraneous variations, and (4) sensitivity to subtle structural discrepancies. The high correlation between the output of our two machine learning models and the human ratings of the test images plainly confirms the models' alignment with human judgment. The order-preservation of the systems is reflected

Image # ref	Diagnosis	Human Rating (1-7)	Object Detector Confidence Score (0-10)				CAE Reconstruction Error (1-0)
		Total Face	Total Face	Oral	Nasal	Orbital	Total Face
1[11]	NF	1.0	0.45	0.45	1.11	0.32	0.990
2[12]	TC	1.13	1.75	2.70	1.11	1.67	0.998
3[13]	UC/L	2.02	2.31	0.33	1.32	5.01	0.982
4[14]	MFH	2.30	3.09	3.02	4.50	2.30	0.843
5[15]	FP	3.44	3.25	2.32	2.31	5.43	0.600
6[16]	VL	3.51	4.25	3.91	4.91	5.15	0.556
7[17]	FP	3.53	4.90	2.54	3.26	6.02	0.516
8[18]	R BC/L	3.92	4.96	3.51	3.14	6.12	0.536

Fig. 6. Subset of test images (with diagnoses and source reference number[11–18]) used to evaluate the object detector and convolutional autoencoder machine learning models. Scoring ranges from abnormal to normal: human rating (1–7); object detector (0–10); CAE (1–0). NF: Neurofibromatosis; TC: Treacher Collins; UC/L: Unilateral cleft lip; MFH: Midface hypoplasia; FP: Facial palsy; VL: Vascular lesion; R BC/L: Repaired bilateral cleft lip.

by the gradient of scoring across the training set that matched closely between machine and human (eg, consistently superior scores for repaired versus unrepaired cleft lip deformities), and by the capacity of the Haar object detector to reliably distinguish the abnormal from the normal subunits of the face (Fig. 6, Table 1). Extraneous variations in the diversity of age, gender, and race of our test dataset did not seem to affect the reliability of the computer ratings, although we were not able to specifically test that with our limited sample size of test images. With regard to feature sensitivity, our machine learning models appeared to discern subtle changes in facial appearance as reflected in the gradient of scores ranging from an individual with extensive neurofibromatosis to one with modest jaw asymmetry.

A growing body of work involving the use of artificial intelligence in plastic surgery is now emerging, the vast majority of which has employed supervised learning and

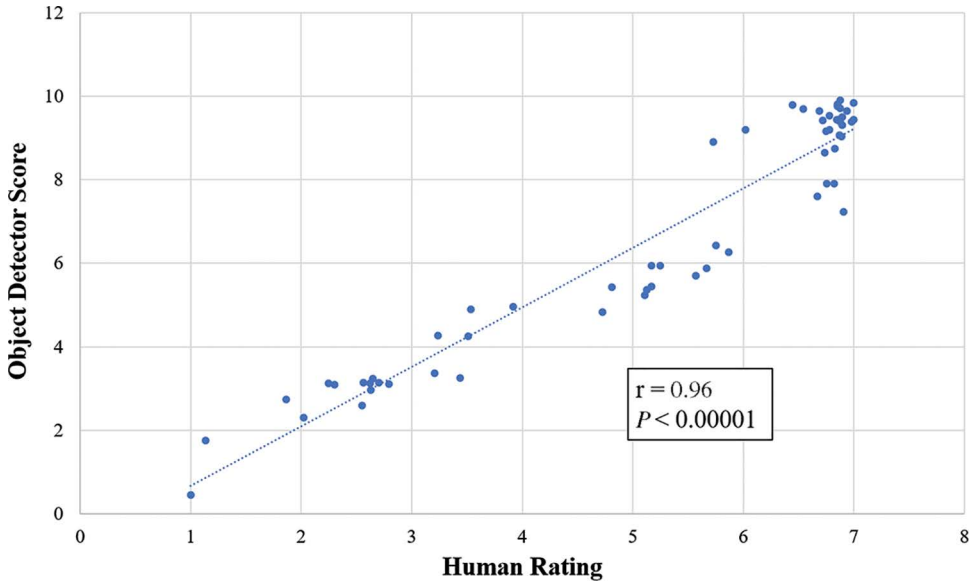

Fig. 7. Scatter plot of object detector confidence score relative to human ratings for all 60 test images (30 abnormal and 30 normal). For object detector the rating scale is 0 (abnormal) to 10 (normal), and for human rating it is 1 (abnormal) to 7 (normal).

classification models to focus on diagnostic and outcome prediction.[38] A recent systematic review references a host of studies predicting burn wound depth and clinical outcome using machine learning models primed with various combinations of patient images, demography, and laboratory parameters. Several groups have described automated systems used to diagnose and assess the severity of craniosynostosis based on shape analysis, whereas only a handful of studies have considered the assessment of facial parameters.[39–43] The current study can be clearly distinguished from any of the preceding work in that our models combine a rich training dataset, a lack of requirement for anatomic landmark recognition or the use of arbitrary human scales, and a holistic consideration of faces that can theoretically be applied to any category of deformity. We are currently compiling a new database of high-resolution clinical images that are approved for analysis and publication. For the purpose of this project, however, we relied on testing open-source images of lower quality, which compelled us to downgrade the resolution of the training set. This could theoretically limit sensitivity to facial feature details. Another avenue

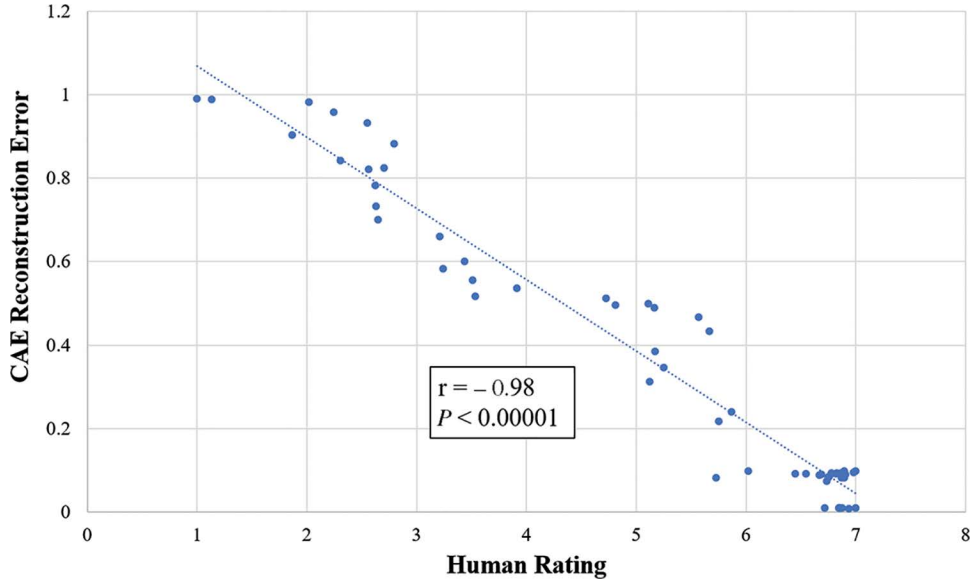

Fig. 8. Scatter plot of convolutional autoencoder reconstruction error relative to human ratings for all 60 test images (30 abnormal and 30 normal). For convolutional autoencoder the rating scale is 0 (normal) to 1 (normal), and for human rating it is 1 (abnormal) to 7 (normal).

Table 1. Diagnosis, Human Rating, Haar Cascade Object Detector Confidence Score, and Convolutional Autoencoder Reconstruction Error for All 30 Test Images

	Human Rating (1–7)	Object Detector Confidence Score (0–10)				Convolutional Autoencoder Reconstruction Error (1–0)
Diagnosis	Total Face	Total Face	Oral	Nasal	Orbital	Total Face
Neurofibromatosis	1.00	0.45	0.45	0.45	0.32	0.99
Treacher Collins	1.13	1.75	2.70	1.11	1.67	1.00
Complete R C/L	1.87	2.73	1.00	2.22	3.60	0.90
Complete L C/L	2.02	2.31	0.33	1.32	5.01	0.98
Complete R C/L	2.24	3.12	1.35	2.17	5.05	0.96
Midface hypoplasia	2.30	3.09	3.02	4.50	2.30	0.84
Complete L C/L	2.55	2.59	0.91	2.33	3.21	0.93
Incomplete R C/L	2.57	3.14	1.77	2.90	5.00	0.82
Incomplete L C/L	2.63	3.13	1.70	2.35	5.02	0.78
Incomplete L C/L	2.63	2.96	1.33	3.10	3.76	0.73
Complete R C/L	2.65	3.24	1.81	2.89	5.31	0.67
Incomplete R C/L	2.70	3.14	1.95	2.43	5.22	0.82
Incomplete L C/L	2.80	3.10	1.51	2.91	3.19	0.88
Repaired R C/L	3.21	3.37	3.00	3.21	4.22	0.66
Down syndrome	3.24	4.27	3.74	5.00	3.91	0.58
Facial palsy preoperative*	3.44	3.25	2.32	2.31	5.43	0.60
Vascular lesion	3.51	4.25	3.91	4.91	5.15	0.56
Facial palsy	3.53	4.90	2.54	3.26	6.02	0.52
Repaired B C/L	3.92	4.96	3.51	3.14	6.12	0.54
Down syndrome	4.72	4.83	4.44	5.41	4.21	0.51
Repaired B C/L	4.81	5.42	4.31	5.10	5.81	0.50
Facial palsy	5.11	5.23	4.21	4.94	5.94	0.50
Facial asymmetry	5.13	5.35	3.91	5.36	6.25	0.31
Facial asymmetry	5.17	5.94	6.13	4.98	6.66	0.49
Macrostomia	5.17	5.45	5.15	6.01	5.42	0.38
Facial palsy postoperative*	5.25	5.95	5.91	5.01	6.23	0.35
Orbital asymmetry	5.57	5.69	6.21	6.31	5.32	0.47
Down syndrome	5.67	5.88	5.81	5.89	5.81	0.43
Facial asymmetry	5.75	6.42	5.99	6.00	6.55	0.22
Moebius syndrome	5.87	6.25	6.05	5.98	6.21	0.24

*Pre- and postoperative images are of the same individual.
Scoring ranges from abnormal to normal: human rating (1–7); object detector (0–10); CAE (1–0). R: right; L: left; B: bilateral; C/L: cleft lip.

we are exploring to enhance discernment is to re-train our models separately based on gender and adult/child status of the faces, as well as on isolated facial aesthetic subunits. However, acknowledging any possible limitation in the sensitivity of our method, we believe that the initial results we report here are impressive and notable—particularly in comparison with testing we performed with off-the-shelf, state-of-the-art detection tools, including YOLOv3[44] and fast.ai.[45] The main issue we found with these latter tools is that they are designed to classify objects as either "human face or not human face"; therefore, even images depicting clinical deformities were rated with extremely high confidence. Our custom tailored models provided far better granularity of measurement in a computationally efficient manner. We anticipate that development of this new technology into a clinically accessible, portable platform will usher in new opportunities for multicenter collaboration and objective comparison of outcomes between conditions, techniques, operators, and institutions.

Mitchell A. Stotland, MD, MS, FRCSC
Department of Surgery
Division of Plastic, Craniofacial and Hand Surgery
Sidra Medicine
Weill Cornell Medical College-Qatar
Room C1-121, Sidra OPC
Qatar Foundation
PO Box 26999
Doha, Qatar
E-mail: mstotland@sidra.org

REFERENCES

1. Thompson A, Kent G. Adjusting to disfigurement: Processes involved in dealing with being visibly different. *Clin Psychol Rev.* 2001;21:663–682. .
2. Karras T, Laine S, Aila T. A Style-based generator architecture for generative adversarial networks. Paper presented at: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); June 2019; Long Beach, Calif. Published online February 2, 2020. Available at https://openaccess.thecvf.com/content_CVPR_2019/html/Karras_A_Style-Based_Generator_Architecture_for_Generative_Adversarial_Networks_CVPR_2019_paper.html. Accessed August 27, 2021.
3. Boyaci O, Serpedin E, Stotland MA. Personalized quantification of facial normality: a machine learning approach. *Sci Rep.* 2020;10:21375. .
4. Krizhevsky A. Learning multiple layers of features from tiny images. 2009. Available at https://scholar.google.com/scholar?hl=en&as_sdt=0%2C5&q=A.+Krizhevsky%2C+%E2%80%9CLearning+multiple+layers+of+features+from+tiny+images%2C%E2%80%9D+Tech.+Rep.%2C+2009.&btnG= Accessed August 26, 2021.
5. OpenCV. Cascade Classifier. Available at https://docs.opencv.org/3.4/db/d28/tutorial_cascade_classifier.html. Accessed August 26, 2021.
6. Viola P, Jones M. Rapid object detection using a boosted cascade of simple features. Paper presented at: Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition; December, 2001; Kauai, HI. Available at https://ieeexplore.ieee.org/abstract/document/990517?casa_token=oujeGr8lnCUAAAAA:5An3eB5DkMVQ0hRfaocq_Whdgt_MUsQ_L0kUoRuKKhGW_nkQ2UKcwUPpvqSSPjtG4sNH1qURj3ms. Accessed August 27, 2001.

7. Freund Y, Schapire RE. Experiments with a new boosting algorithm. 1996. Available at https://scholar.google.com/scholar?hl=en&as_sdt=0%2C5&q=Freund%2C+Y.+and+R.+Schapire.+%E2%80%9CExperiments+with+a+New+Boosting+Algorithm.%E2%80%9D+&btnG= Accessed August 26, 2021.
8. Bergstra J, Bardenet R, Bengio Y, et al. Algorithms for hyper-parameter optimization. Paper presented at: Proceedings of the 24th International Conference on Neural Information Processing Systems. 2011:2546–2554. Granada, Spain; December 2011. Available at https://www.researchgate.net/publication/216816964_Algorithms_for_Hyper-Parameter_Optimization/link/02e7e537d951197f0d000000/download. Accessed August 27, 2021.
9. Horton M, Cameron-Jones M, Williams R. Multiple classifier object detection with confidence measures. Paper presented at: Proceedings of the 20th Australian Joint Conference on Advances in Artificial Intelligence (AI). Vol. 4830. Gold Coast Australia; December 2–6, 2007.
10. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. Paper presented at: Proceedings of the 3rd International Conference on Learning Representations, ICLR 2015; Available at: http://arxiv.org/abs/1409.1556. Accessed August 26, 2021.
11. Bibbs R. Untitled [Photograph]. Available at: https://www.flickr.com/photos/reggiebibbs/6138648983/. 2011. Licensed under https://creativecommons.org/licenses/by-nc/2.0/. Accessed August 27, 2021.
12. Wikimedia Commons, The Free Media Repository. Treacher Collins syndrome. Available at: https://commons.wikimedia.org/w/index.php?title=File:Treacher Collins syndrome %28MedMedicine%29.jpg&oldid= 232695857. 2017. Licensed under https://creativecommons.org/licenses/by-sa/4.0/deed.en. Accessed August 27, 2021.
13. Trinity Care Foundation. Free cleft lip surgery India. Available at https://www.flickr.com/photos/trinitycarefoundation/15880128986/. 2014. Licensed under https://creativecommons.org/licenses/by-nc-nd/2.0/. Accessed July 14, 2021.
14. Wikimedia Commons, The Free Media Repository. Baby with Crouzon syndrome. Available at https://commons.wikimedia.org/w/index.php?title=File:Baby with Crouzon Syndrome.jpg&oldid=297658046. 2018. Licensed under https://creativecommons.org/licenses/by-sa/4.0/deed.en. Accessed July 14, 2021.
15. Wikimedia Commons, The Free Media Repository. Fazialislähmung. https://commons.wikimedia.org/w/index.php?title=File:Fazialisl%C3%A4hmung Tag 031000.jpg&oldid=222325005. 2016. Licensed under https://creativecommons.org/licenses/by-sa/4.0/. Accessed July 14, 2021.
16. Trinity Care Foundation. Free cleft surgery India. Available at https://www.flickr.com/photos/trinitycarefoundation/21142841765/in/photostream/. 2015. Licensed under https://creativecommons.org/licenses/by-nc-nd/2.0/. Accessed July 14, 2021.
17. Wikimedia Commons, the Free Media Repository. Bell's palsy. Available at https://commons.wikimedia.org/w/index.php?title=File:Bellspalsy.JPG&oldid=368535279. 2019. Licensed under https://creativecommons.org/licenses/by-sa/3.0/deed.en. Accessed July 14, 2021.
18. ReSurge International. Luis after his cleft palate repair. Available at https://www.flickr.com/photos/interplast/309022750/. 2006. Licensed under https://creativecommons.org/licenses/by-nc-nd/2.0/. Accessed July 14, 2021.
19. OpenCV library. Available at https://opencv.org/. Accessed August 27, 2021.
20. TensorFlow library. Available at https://www.tensorflow.org/. Accessed August 27, 2021.
21. Keras library. Available at https://keras.io/. Accessed August 27, 2021.
22. Briganti G, Le Moine O. Artificial intelligence in medicine: Today and tomorrow. *Front Med (Lausanne)*. 2020;7:27.
23. Jarvis T, Thornburg D, Rebecca AM, et al. Artificial intelligence in plastic surgery: Current applications, future directions, and ethical implications. *Plast Reconstr Surg Glob Open*. 2020;8:e3200.
24. Miller MQ, Hadlock TA, Fortier E, et al. The auto-eFACE: Machine learning-enhanced program yields automated facial palsy assessment tool. *Plast Reconstr Surg*. 2021;147:467–474.
25. Boonipat T, Brazile TL, Darwish OA, et al. Measuring visual attention to faces with cleft deformity. *J Plast Reconstr Aesthet Surg*. 2019;72:982–989.
26. Sinko K, Jagsch R, Prechtl V, et al. Evaluation of esthetic, functional, and quality-of-life outcome in adult cleft lip and palate patients. *Cleft Palate Craniofac J*. 2005;42:355–361.
27. Carruthers J, Flynn TC, Geister TL, et al. Validated assessment scales for the mid face. *Dermatol Surg*. 2012;38(2 Spec No.):320–332.
28. Edler R, Rahim MA, Wertheim D, et al. The use of facial anthropometrics in aesthetic assessment. *Cleft Palate Craniofac J*. 2010;47:48–57.
29. Mercan E, Oestreich M, Fisher DM, et al. Objective assessment of the unilateral cleft lip nasal deformity using three-dimensional stereophotogrammetry: Severity and outcome. *Plast Reconstr Surg*. 2018;141:547e–558e.
30. Tse RW, Oh E, Gruss JS, et al. Crowdsourcing as a novel method to evaluate aesthetic outcomes of treatment for unilateral cleft lip. *Plast Reconstr Surg*. 2016;138:864–874.
31. Rhee JS, McMullin BT. Outcome measures in facial plastic surgery: Patient-reported and clinical efficacy measures. *Arch Facial Plast Surg*. 2008;10:194–207.
32. Campbell A, Restrepo C, Deshpande G, et al. Validation of the unilateral cleft lip severity index for surgeons and laypersons. *Plast Reconstr Surg Glob Open*. 2017;5:e1479.
33. Tietje L, Cresap S. Is lookism unjust?: The ethics of aesthetics and public policy implications. *J Libert Stud*. 2005;19:31–50.
34. Cui X, Cheng Q, Lin W, et al. Different influences of facial attractiveness on judgments of moral beauty and moral goodness. *Sci Rep*. 2019;9:1–12.
35. Warhurst C, Van den Broek D, Hall R, et al. Great expectations: Gender, looks and lookism at work. *Int J Work Organ Emot*. 2012;5:72–90.
36. Waring P. Keeping up appearances: Aesthetic labour and discrimination law. *J Ind Relat*. 2011;53:193–207.
37. Chatterjee A, Thomas A, Smith SE, et al. The neural response to facial attractiveness. *Neuropsychology*. 2009;23:135–143.
38. Mantelakis A, Assael Y, Sorooshian P, et al. Machine learning demonstrates high accuracy for disease diagnosis and prognosis in plastic surgery. *Plast Reconstr Surg Glob Open*. 2021;9:e3638.
39. Dusseldorp JR, Guarin DL, van Veen MM, et al. In the eye of the beholder: Changes in perceived emotion expression after smile reanimation. *Plast Reconstr Surg*. 2019;144:457–471.
40. Boonipat T, Asaad M, Lin J, et al. Using artificial intelligence to measure facial expression following facial reanimation surgery. *Plast Reconstr Surg*. 2020;146:1147–1150.
41. Patcas R, Bernini DAJ, Volokitin A, et al. Applying artificial intelligence to assess the impact of orthognathic treatment on facial attractiveness and estimated age. *Int J Oral Maxillofac Surg*. 2019;48:77–83.
42. Chen K, Lu SM, Cheng R, et al. Facial recognition neural networks confirm success of facial feminization surgery. *Plast Reconstr Surg*. 2020;145:203–209.
43. McCullough M, Ly S, Auslander A, et al. Convolutional neural network models for automatic preoperative severity assessment in unilateral cleft lip. *Plast Reconstr Surg*. 2021;148:162–169.
44. Redmon J, Farhadi A. Yolov3: An incremental improvement. *arXiv preprint*. 2018. Available at https://arxiv.org/pdf/1804.02767.pdf.
45. Howard J, Gugger S. Fastai: A layered API for deep learning. *Information*. 2020;11:108. Available at https://arxiv.org/abs/2002.04688.