Northumbria Research Link

Northumbria University
NEWCASTLE

UniversityLibrary

ORIGINAL ARTICLE

Assessing Facial Symmetry and Attractiveness using Augmented Reality

Wei Wei[1] · Edmond S. L. Ho[2] · Kevin D. McCay[2] · Robertas Damaševičius[3] · Rytis Maskeliūnas[4] · Anna Esposito[5]

Abstract

Facial symmetry is a key component in quantifying the perception of beauty. In this paper, we propose a set of facial features computed from facial landmarks which can be extracted at a low computational cost. We quantitatively evaluated the proposed features for predicting perceived attractiveness from human portraits on four benchmark datasets (SCUT-FBP, SCUT-FBP5500, FACES and Chicago Face Database). Experimental results showed that the performance of the proposed features is comparable to those extracted from a set with much denser facial landmarks. The computation of facial features was also implemented as an augmented reality (AR) app developed on Android OS. The app overlays four types of measurements and guidelines over a live video stream, while the facial measurements are computed from the tracked facial landmarks at run time. The developed app can be used to assist plastic surgeons in assessing facial symmetry when planning reconstructive facial surgeries.

Keywords Facial symmetry · Facial analysis · Augmented reality · Mobile app

1 Introduction

The human face is central to many aspects of social interaction [1]. It forms the basis from which humans are able to process, recognize and draw information from one another. Even from infancy, humans are able to demonstrate a preference for faces perceived as attractive [2]. Indeed, there have been several studies suggesting that individuals deemed as being attractive are more likely to achieve prestigious occupations, to have better prospects for personal fulfillment and to benefit from additional social advantages in their everyday lives [3, 4]. These observations have subsequently garnered the attention of researchers, in seeking to determine whether attractiveness can be considered objective or subjective. Several studies, both in the fields of psychology and medical science, suggest that facial attractiveness can indeed be quantified [5, 6].

Humans have evolved in such a way that they are able to perceive subtle deviations in what would be considered a normal facial structure. Facial symmetry and averageness have consequently been identified as key components in this perception, and several attempts to produce metrics for these elements have been proposed [7–9]. The development of these attractiveness metrics has led to the development of

✉ Robertas Damaševičius
robertas.damasevicius@polsl.pl

Wei Wei
weiwei@xaut.edu.cn

Edmond S. L. Ho
e.ho@northumbria.ac.uk

Kevin D. McCay
kevin.d.mccay@northumbria.ac.uk

Rytis Maskeliūnas
rytis.maskeliunas@vdu.lt

Anna Esposito
iiass.annaesp@tin.it

[1] School of Computer Science and Engineering, Xi'an University of Technology, Xi'an and Shaanxi Key Laboratory for Network Computing and Security Technology, Xi'an, China

[2] Department of Computer and Information Sciences, Northumbria University, Newcastle upon Tyne, UK

[3] Faculty of Applied Mathematics, Silesian University of Technology, Gliwice, Poland

[4] Department of Applied Informatics, Vytautas Magnus University, Kaunas, Lithuania

[5] Department of Psychology, Università della Campania "Luigi Vanvitelli" and IIASS, Caserta, Italy

several tools that can theoretically determine facial attractiveness based upon the proposed empirical data.

Recently, automated machine learning methods of assessing facial attractiveness using beauty metrics have been proposed [10–12]. These proposed frameworks focus on developing systems which automatically assess facial attractiveness based upon the facial proportions and specified landmarks typically associated with facial beauty. It is thought that automated technology capable of the quantifiable analysis and measurement of facial attractiveness could have many applications including anxiety recognition [13], entertainment, virtual media, cosmetics, orthodontics [14] and plastic surgery [12].

Given that facial symmetry has the potential to contribute greatly to the perception of facial attractiveness, the impact of facial asymmetry can lead to significant emotional and psychological distress [15]. The UK Equality Act 2010 [16] states that a severe facial disfigurement should be treated as a disability. In these cases, reconstructive plastic surgery is often considered necessary. In undertaking surgery to resolve facial asymmetry, surgeons will often manually determine the differences between the two sides of the face, simply by examining the patient subjectively. Working in close collaboration with the patient, surgeons tend to use the contralateral normal side as a guide rather than using the well-studied metrics discussed above. While the implementation of technology has not been widely acknowledged in this area, there is certainly the potential for it to be very helpful.

Recent studies [17–20] suggest that the implementation of computer-based assessment systems can provide an aid to surgeons in preparing, measuring and analyzing facial reconstruction procedures. These studies typically demonstrate methodologies for the objective and quantifiable measurement of facial imperfections and provide a means of tracking treatment outcomes. These approaches utilize still images for comparison, which are generated either from photographs or 3D scans of the patient. These images are then analyzed and overlaid with relevant information pertinent to the surgical processes, such as predefined landmarks or volumetric comparisons. While encouraging results using computer vision and computer-based assessment are evident from the literature, it is worth considering that the financial costs of implementing full clinical 3D systems, such as 3dMD [21], can be significant. Additionally, each of the considered related works deals only with still images, potentially neglecting vital information about the transitions between facial expressions.

To this end, we propose a set of new features which focuses on facial symmetry for use in an augmented reality (AR)-based prototype for objective assessment of facial deformation and subsequently perceived facial beauty. The facial features are computed from facial landmarks extracted from the 2D color video stream of the patient. In order to reduce the computation cost to improve the interactivity and responsiveness on AR applications, the features are computed from a compact set of facial landmarks extracted using the Google Face API. We further developed a smartphone application (with Android OS) which captures a 2D color video stream of the patient, extracts important facial features, such as facial landmarks, and analyzes the data in real time. The results are then overlaid directly onto the live video stream, to assist both the surgeon and patient in determining the most appropriate surgical options interactively.

We quantitatively evaluated the proposed features for attractiveness prediction on four benchmark datasets [22–25], and the results show that the proposed features can improve the performance of the existing geometric features. On the AR app, the on-screen visual feedback provided by the overlaid visualizations is based upon several of the quantifiable metrics and guidelines discussed, allowing for objective and fully informed decisions to be made. The proposed face assessment method is quick, cost-effective, and non-invasive.

The remainder of the paper is structured as follows; Sect. 2 discusses the related works, Sect. 3 describes our methodology, Sect. 4 presents our experimental results, Sect. 5 discusses the limitations and applicability of the proposed method, and finally, Sect. 6 presents our conclusions and discusses intended future work.

2 Related works

Assessing facial symmetry and attractiveness using computer vision and pattern recognition techniques has recently become an active area of research. This section discusses the most prominent examples in this field.

Hong et al. [12] proposed an automated framework which extracts geometric facial features from images to predict the "beauty score" of the subject. In particular, 4 types of features (*neoclassical features* [26], *golden ratio* [27], *symmetry* [27], and *8-ratio vectors* [28]) are extracted to train multiple regressors for the prediction task. The predicted scores are then fused to predict a final score in order to boost the performance. Ulrich et al. [29] proposed the use of facial proportions as a means of predicting facial attractiveness in females. In particular, 29 features are extracted from facial landmarks. Gan et al. [30] proposed a deep learning framework for multi-task transfer learning, which focuses on beauty score prediction as the primary task, treating gender recognition as an auxiliary task. This multi-task framework improves the performance of both tasks and alleviates the over-fitting problem in training the network. Xu et al. [31] also proposed a multi-task deep learning framework for facial attractiveness, gender, and race prediction, achieving

state-of-the-art performance. Zhou et al. [32] presented a system for analyzing trends in perceived attractiveness of Chinese males at different times. A large image database of Chinese male faces was constructed. The Inception v3 network was then retrained using the new data for facial shape classification. The correlation between the shape of facial landmarks and the attractiveness at different times was then compared. Liu et al. [33] proposed a method for facial attractiveness computation from 2.5D data; however, their facial landmarks (82 frontal keypoints and 40 profile keypoints) were computed from 3D faces.

In addition to general attractiveness prediction, automated facial feature assessment is also used in health and medical applications. Sajid et al. [34] proposed an image classification framework to automatically assess whether the subject is suffering from facial palsy. Image features are automatically extracted from a pre-trained convolutional neural network (CNN). To further improve the robustness of the classification framework and avoid over-fitting, a generative adversarial network (GAN) is proposed for data augmentation. To make the automated facial palsy assessment more accessible, Kim et al. [35] proposed a computationally efficient framework, enabling the image classification to be done on a smartphone. The facial landmarks are used to compute an asymmetry index as features for light-weight classification, using classifiers such as support vector machine (SVM), and linear discriminant analysis (LDA).

The metrics incorporated into our prototype system are primarily based around the perceived notion that facial symmetry is a desirable trait. Based upon the literature there is also evidence to suggest that the golden ratio may contribute to the overall aesthetic of facial construction. The implementation of metrics such as these are a commonly found approach in many related works. In a recent study, Gunes and Piccardi [36] attempted to evaluate human facial attractiveness using an automated classifier. A decision tree extracted features from the images based upon the golden ratio and, through supervised classification, calculated what the average human judgment would be regarding the facial beauty portrayed in the image. Schmid et al. [27] proposed a regression-based approach to analyze the significance of symmetry, neoclassical canon, and golden ratio in the attractiveness of a face. Their study focused on the geometry of the face by using a specific set of 29 landmarks. Their results suggest that while symmetry plays an important part in the perceived attractiveness, its role is secondary to those defined by the neoclassical canons and golden ratios.

Each of these studies suggests that symmetry and golden ratio have the potential to play a significant part in the subjective perception of facial attractiveness. It is this search for a perceptually ideal facial structure that has led to a correlation between symmetry, golden ratio, and attractiveness, and subsequently the development of models such as Marquardt Phi Mask [37]. And while the idea that a universal standard to classify beauty has been discredited in some areas [38], it does not diminish the potential of such models to function as indicative tools for the analysis of facial structure. The amalgamation of real-time video data and computer-generated visual feedback, such as the discussed indicative tools and overlay metrics, has the capacity to provide objective and quantifiable data. While augmented reality has been discussed in other works [39] as a potential diagnostic and rehabilitative tool, limited work has been done to apply it in this context.

3 Methodology

3.1 Framework

The outline of the proposed framework for facial symmetry evaluation is outlined in Fig. 1. The face measurements should be visualized in an intuitive way, such that users understand the readings easily. Also, interactivity is one of
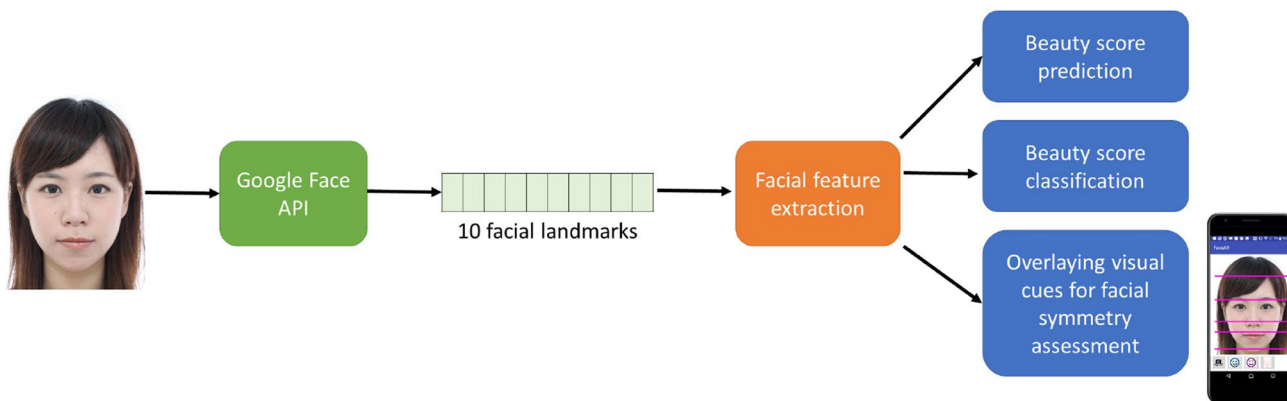

Fig. 1 The overview of the proposed framework

the most important aspects of the tool, so that the surgeon and patient can see the results immediately. Finally, the tool should be easy to use and the hardware should be portable with simple setup procedures, such that it can be used in a clinical environment (e.g., consultation room).

As a result, we decided to develop the prototype of the tool as an augmented reality (AR) app, allowing us to incorporate all the aforementioned features. The AR app is developed on the Android platform using Java in Android studio. Inspired by the FaceSpotter [40] project, which tracks faces in real time and overlays graphics on facial landmarks, the Google Face API [41] is used. The Google Face API [41] provides mobile App developers with a wide range of face-related functionalities, including face recognition and face tracking. By utilizing this, the positions of facial landmarks as well as the head orientation (in Euler angles) can be detected.

The Google Face API can track 12 facial landmarks on each face. However, we found that the tracking accuracy for the left and right ear landmarks is significantly lower than other landmarks in our experiments. As a result, only 10 facial landmarks are used in the app, including left and right eyes, left and right ear tips, left and right cheeks, left and right mouth corners, nose base, and mouth bottom. An example is shown in Fig. 2a. Having detected the facial landmarks, we then assess the facial symmetry using the facial geometric features.

3.2 Facial geometric features extraction

The developed AR app provides a wide range of quantitative measurements interactively. Since the accuracy of facial landmark tracking depends on various factors (such as the lighting conditions, camera motion, movement of the subject, head orientation), the tracked landmark locations and labels are displayed for the user to evaluate whether the landmarks are tracked correctly. In order to assist the users, we designed user interface measurement tools which can be overlaid on the live video stream (as illustrated in Fig. 3(a)) to help determine facial symmetry. Our proposed pipeline makes use of these indicative measurement tools to ensure that the user is able to extract the most accurate features for assessment. Once aligned correctly, the tracked landmark locations can then be used for further analysis.

3.2.1 Eye distance to the midline:

With the facial landmarks tracked using the Google Face API, the *nose base* and *mouth bottom* landmarks are used to define the midline. The distance between each of the eyes and the midline can then be computed to evaluate whether the eye positions are symmetrical. Directly using the distance between facial landmarks has been explored in previous works [22]. One of the limitations of this is that the distances are sensitive to scaling of the facial features, e.g., the
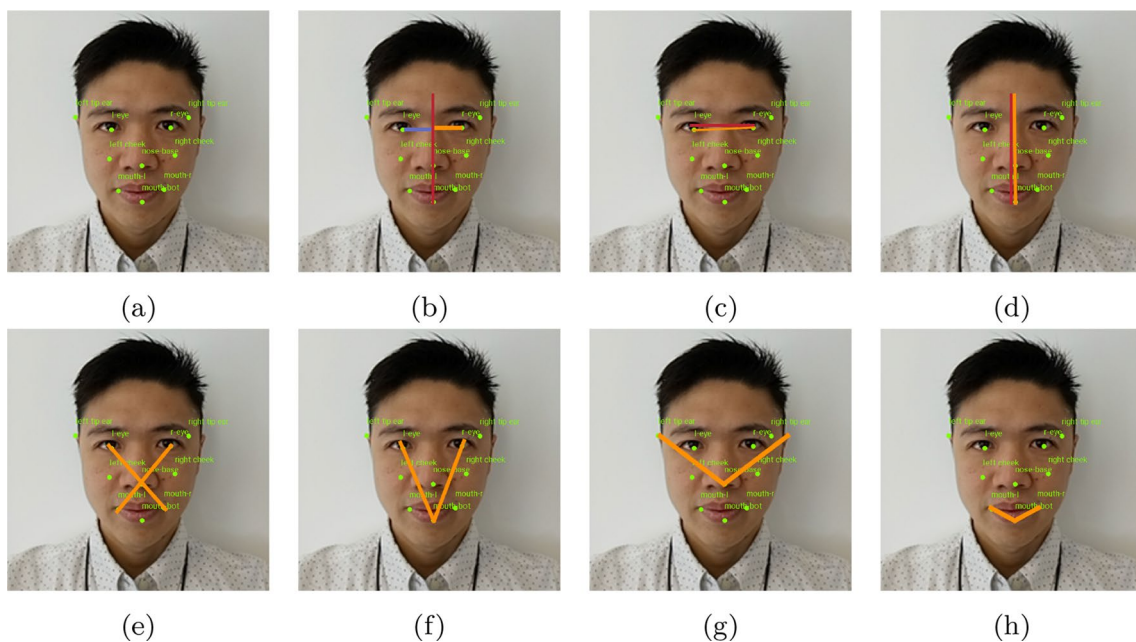
Fig. 2 The proposed facial geometric features. **a** The 10 facial landmarks detected using Google Face API. **b** The deviation of the distances between each of the eyes and the midline of the face. **c** The acute angle between the eye level and a horizontal line. **d** The acute angle between the facial midline (derived from the nose and the bottom of the mouth) and a vertical line. **e** The acute angle between the diagonal lines connecting the eyes and the mouth. **f** The acute angle between the 'v-shape' lines connecting the eyes and the bottom of the mouth. **g** The acute angle between the 'v-shape' lines connecting the ears and the nose. **h** The interior angle at the bottom of the mouth
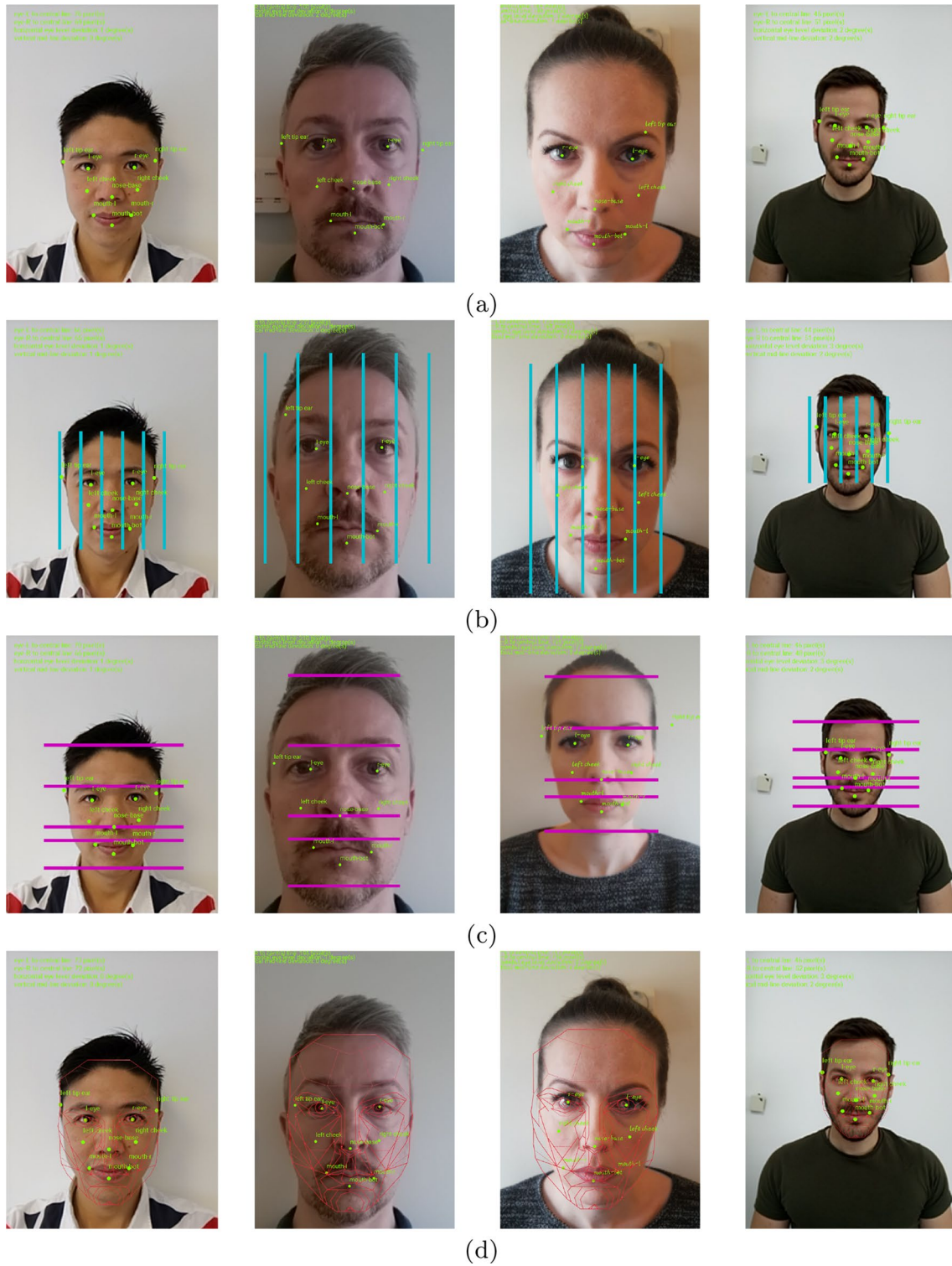
Fig. 3 The prototype AR app. Different types of information are overlaid on the live video stream. **a** Basic information. **b** The rule of fifths—uniformly dividing the face into 5 regions vertically. **c** The rule of thirds—uniformly dividing the face into 3 regions by horizontal lines. **d** Marquardt's Phi Mask [37]

distance between the subject and the camera. To alleviate the problem of scaling and normalizing the images properly, we propose a 'relative measurement' of the differences between the distances from each eye to the midline. Specifically,

$$eye_{dev} = \frac{\left||(p_{l,x} - p_{mid,x})| - |(p_{r,x} - p_{m,x})|\right|}{max(|(p_{l,x} - p_{mid,x})|, |(p_{r,x} - p_{mid,x})|)} \tag{1}$$

where $p_{l,x}$ and $p_{r,x}$ are the x coordinates of the left eye and right eyes, respectively. $p_{m,x}$ is the x coordinate of the midline derived from the *nose base* and *mouth bottom* landmarks. An example is illustrated in Fig. 2b.

3.2.2 Horizontal eye-level deviation

The deviation (in degrees) can be computed by calculating the acute angle between the line drawn between two eyes and a horizontal line:

$$\theta_{hor} = \arccos\left(\frac{|p_{l,x} - p_{r,x}|}{\sqrt{(p_{l,x} - p_{r,x})^2 + (p_{l,y} - p_{r,y})^2}}\right) \tag{2}$$

where $p_{l,x}$ and $p_{l,y}$ are the x and y coordinates of the left eye, and $p_{r,x}$ and $p_{r,y}$ are the x and y coordinates of the right eye. By this, the deviation can effectively be represented by a scale value. Since we are calculating the acute angles between two lines derived from the facial landmarks, the feature is scale-invariant and robust to image resizing. An example is illustrated in Fig. 2c.

3.2.3 Vertical midline deviation:

Similarly, the deviation (in degrees) can be computed by acute angle between the midline computed from the tracked facial landmarks and a vertical line:

$$\theta_{ver} = \arccos\left(\frac{|p_{m,y} - p_{n,y}|}{\sqrt{(p_{m,x} - p_{n,x})^2 + (p_{m,y} - p_{n,y})^2}}\right) \tag{3}$$

where $p_{m,x}$ and $p_{m,y}$ are the x and y coordinates of the *mouth bottom* landmark, and $p_{n,x}$ and $p_{n,y}$ are the x and y coordinates of the *nose base* landmark. An example is illustrated in Fig. 2d.

3.2.4 Eye-mouth diagonal

Since the symmetry among multiple facial landmarks is also important for humans to perceive facial symmetry, we further propose extracting geometric features from different combinations of facial landmarks. Here, we focus on

the central area of the face, focusing upon the eyes and the mouth. There are four landmarks in this combination: *right eye*, *left eye*, *right mouth*, and *left mouth*, forming a rough rectangular shape. We derive the diagonal lines from the 4 landmarks and compute the acute angle between these diagonal lines:

$$\theta_{eye-mouth-diag} = \arccos\left(\frac{(p_l - p_{rm}) \cdot (p_r - p_{lm})}{\|p_l - p_{rm}\|\|p_r - p_{lm}\|}\right) \tag{4}$$

where p_l and p_r are the x and y coordinates of the *left eye* and *right eye*, respectively. p_{lm} and p_{rm} are the x and y coordinates of the *left mouth* and *right mouth*. An example is illustrated in Fig. 2e.

3.2.5 Eye-mouth angle

Facial symmetry can also be evaluated using a combination of eyes and *mouth bottom*. The three landmarks form a roughly triangular shape. As the triangle should be close to an Isosceles triangle on a typical face, we focus on the interior angle at the *mouth bottom* instead of the other interior angles at the eyes. As a result, the feature is calculated by:

$$\theta_{eye-mouth-V} = \arccos\left(\frac{(p_l - p_{mb}) \cdot (p_r - p_{mb})}{\|p_l - p_{mb}\|\|p_r - p_{mb}\|}\right) \tag{5}$$

where p_{mb} contains the x and y coordinates of the *mouth bottom* landmark. An example is illustrated in Fig. 2f.

3.2.6 Ear–nose angle

Similar to $\theta_{eye-mouth-V}$, we extract another geometric feature using a combination of the ears and *nose base*. We focus on the interior angle at the nose:

$$\theta_{ear-nose} = \arccos\left(\frac{(p_{le} - p_{nb}) \cdot (p_{re} - p_{nb})}{\|p_{le} - p_{nb}\|\|p_{re} - p_{nb}\|}\right) \tag{6}$$

where p_{le} and p_{re} are the x and y coordinates of the *left ear* and p_{nb} contains the x and y coordinates of the *nose base* landmark. An example is illustrated in Fig. 2g.

3.2.7 Mouth angle

This feature focuses on the landmarks for the mouth. Similar to $\theta_{eye-mouth-V}$, we extract the feature from a triangle formed by three landmarks: *right mouth*, *left mouth*, and *mouth bottom*. We compute the interior angle at the *mouth bottom*:

$$\theta_{mouth-V} = \arccos\left(\frac{(p_{lm} - p_{mb}) \cdot (p_{rm} - p_{mb})}{\|p_{lm} - p_{mb}\|\|p_{lm} - p_{mb}\|}\right) \tag{7}$$

An example is illustrated in Fig. 2h.

3.2.8 Laplacian coordinates of facial landmarks

Within the computer graphics community, differential coordinates have been widely used to encode the local details of 3D shapes, such as preserving details in 3D mesh deformation [42], maintaining the spatial relation between close character interactions [43], and robot environment interactions [44]. Here we propose using the Laplacian coordinates of the selected facial landmarks as additional facial features to assess facial symmetry and attractiveness. For the sake of generality, we first explain how to calculate the Laplacian coordinates of a facial landmark. Given $P = p_1, ..., p_n$ where P is the vector that contains the 2D coordinates of all n facial landmarks; the Laplacian coordinates θ_{Lap-i} of the landmark p_i are computed by:

$$\theta_{Lap-i} = p_i - \frac{1}{n-1} \sum_{\substack{j=1 \\ j \neq i}}^{n} p_j \tag{8}$$

The Laplacian coordinates of a facial landmark essentially indicate the degree by which the landmark deviates from the average position computed for the rest of the landmarks. As a result, the Laplacian coordinates features should target the landmarks which are closer to the central area of the face. In particular, we selected *left eye*, *right eye*, and *nose* to calculate the Laplacian coordinates as $\theta_{Lap-eyeL}$, $\theta_{Lap-eyeR}$ and $\theta_{Lap-nose}$, respectively.

3.3 Visualizing the rule of Fifths—Vertical

Having tracked the landmarks for the ears, the face can be divided into 5 equal-width segments vertically [45]. The width of each segment should be close to the width of one of the eyes. In addition, the width of the nose base should be close to the distance between the inner corners of the eyes. An example of the overlaid information is shown in Fig. 3(b).

3.4 Visualizing the rule of Thirds—Horizontal

A face can be uniformly divided into 3 segments (i.e., the top of the head to the eye, the eye to the nose base, and the nose base to the chin) horizontally using the rule of thirds [45]. However, tracking the outline of the face on smartphones can be challenging due to illumination variations. On the other hand, having tracked the landmarks for the eyes (i.e., eye level) and the nose, the height of the middle segment (eye to nose base) can be computed. In doing so, the other two segments can be estimated. An example of this estimation is illustrated in Fig. 3(c). The bottom segment can also be further divided, whereby the distance between the nose base and the middle of the mouth should be 1/3 of the segment height.

3.5 Visualizing the Marquardt Phi Mask

Marquardt [37] proposed the Marquardt Phi Mask as a means of describing the ideal facial proportions for perceptually beautiful faces. As such, the mask can be used as a guideline for make-up or even plastic surgery, such that the facial landmarks appear to be closer to the corresponding parts of the mask. The mask is derived from mathematics and mostly related to the *Golden Ratio*. An example of the mask overlaid on a male subject is illustrated in Fig. 3(d). Positive feedback on measuring facial attractiveness using this mask has previously been documented [7, 46], and since publication, common variations of the mask have been made available for different age groups, genders, and ethnicities [47]. In our implementation, the mask can be used to show facial symmetry, and if the deviations of the tracked facial landmarks adhere to the template. In the AR app, the mask is overlaid onto the live video stream. Having detected the facial landmarks from the live video stream, the distance between two eyes is calculated and the mask is scaled accordingly, as the eye distance of the mask is known. Next, the location of the mask is updated by using the mid-point between the two eyes as a reference point.

4 Experimental results

In this section, we evaluate the proposed facial features and the developed AR tool. Qualitative results are illustrated in Fig. 3, which highlights the tracking accuracy of the facial landmarks. It can be seen that that AR tool is intuitive to use and provides visualized quantitative measurements.

4.1 Benchmark datasets

To further evaluate the proposed features quantitatively, we conducted a series of experiments on the following four benchmark datasets: SCUT-FBP [22], SCUT-FBP5500 [23], FACES [24], and Chicago Face Database (CFD) [25].

4.1.1 SCUT-FBP

The SCUT-FBP [22] dataset contains 500 portraits of different females of Asian ethnicity. Each of the portraits is rated by 70 different assessors, and for each image the average is reported as the final score. The rating process focuses entirely on the *beauty* of the subject, asking the rater to confirm to which extent they think the subject is beautiful. The scoring system ranges from 1 (strongly disagree) to 5

(strongly agree). Figure 4 shows some examples from the SCUT-FBP [22] dataset.

4.1.2 SCUT-FBP5500

The SCUT-FBP5500 [23] dataset is an extension of the SCUT-FBP [22] dataset and has 5500 facial portraits. In addition to the increased size of the dataset, the SCUT-FBP5500 provides a wider variety by including male and female subjects, Asian and Caucasian ethnicities, as well as a wider range of ages (from 15–60). This makes the SCUT-FBP5500 more challenging due to the increased diversification of facial images. Each of the portraits is rated by 60 different assessors, and for each image the average of these is reported as the final score. Again, a *beauty* rating is associated with each subject, with the scoring system once again ranging from 1 (strongly disagree) to 5 (strongly agree).

4.1.3 FACES benchmark dataset

The FACES [24] dataset contains 2,052 facial portrait images collected from 171 men and women, categorized into different age groups (young, middle-aged, and older). Multiple images are captured from each of the subjects. In particular, images from the same subjects show different facial expressions. The original usage of this dataset focused upon emotion and perceived age analysis. Recently, Ebner [48] further enhanced the dataset by labeling the images with attractiveness scores to support studies related to perceived attractiveness. The dataset is annotated by 154 participants with a scoring range from 1 to 100, where the higher the score the more attractive the subject is perceived to be. In our study, as in the reported baseline method, all images in the dataset were used; however, each image is associated with a score; as such, the same subject received different scores for different expressions. Figure 5 shows some examples in the FACES [24] dataset.

4.1.4 Chicago Face Database (CFD)

The Chicago Face Dataset (CFD) [25] contains facial portrait images collected from 158 subjects of varying genders and ethnicities, including 37 Black males, 48 Black females, 36 White males, and 37 White females. The images were annotated by 1,087 participants with a scoring range from 1 to 100, where again the higher the score the more attractive the subject is perceived to be. In our work, 597 images from this dataset are used. Figure 6 shows some examples from the CFD [25] dataset.

4.2 Evaluation methods

In this work, we follow [22, 23] in conducting experiments on *beauty score prediction* to quantitatively evaluate different features. While other facial features, such as color and texture, can also be used to assess facial beauty, these
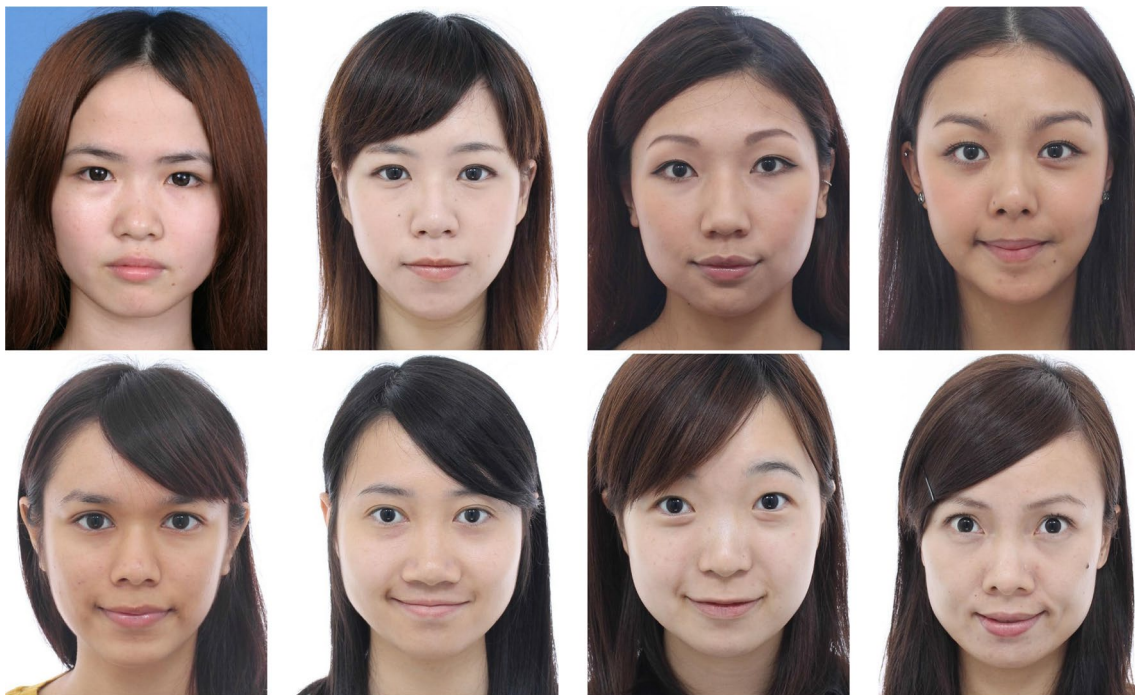

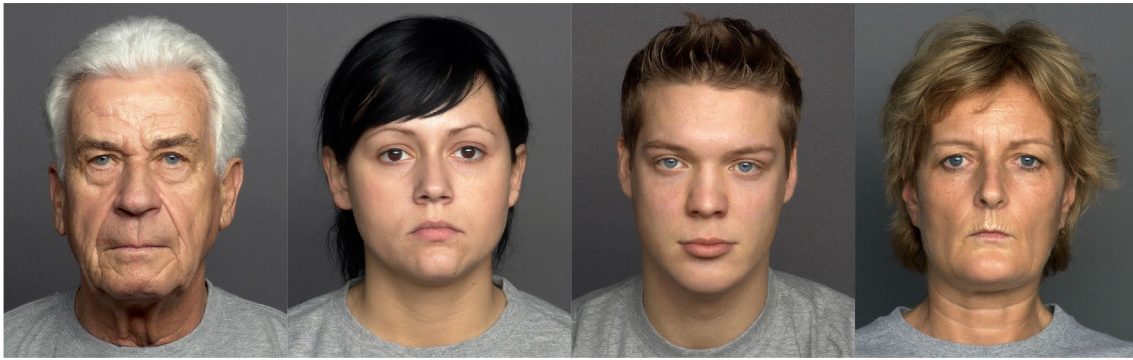**Fig. 4** Examples of portraits in the SCUT-FBP dataset [22]

Fig. 5 Examples of portraits in the FACES [24] dataset

Fig. 6 Examples of portraits in the (CFD) [25] dataset

features can, however, be easily affected by other factors such as make-up and illumination. As a result, we focus on analyzing the performance of geometric features. Specifically, given the geometric facial features extracted from the images, a beauty score prediction can be formulated as a regression problem, which takes the features as the input and treats the annotated beauty score as the output. As in [22], SVM regression (SVR) and linear regression are used in our experiments.

We also include other previous approaches for comparison. In particular, Hong et al. [12] summarized and evaluated 4 types of geometric facial features. The *neoclassical features* are presented in [26] which focused on the ratios between different facial landmarks such as the eyes and mouth. The *golden ratio* features presented in [27] are based upon the ratio between the sizes of different facial regions and landmarks. The *symmetry* features in [27] purely focus on facial symmetry using distances between facial landmarks. The *8-ratio vectors* in [28] is similar to the *neoclassical features* which are computed based on the ratios of the distances between landmarks. We follow the summary in [12] to extract all 4 different feature sets in this work.

We further compare our system with the geometric facial features produced using the method proposed in [49]. This method produces a detailed set of 18 distances [22] between

different pairs of landmarks (see Fig. 7b). However, extracting the aforementioned geometric features requires a denser set of facial landmarks, such as the 68 landmarks detected using dlib (see Fig. 7a). As a result, a subset of the features proposed in [22] is selected and illustrated in Fig. 7c. We also computed the complete set of 18 features [22] by detecting the 68 landmarks using dlib as a comparison.

Our evaluation implemented fivefold cross-validation to test the regression performance. For each dataset, all images are divided into 5 groups. In each trial, 1 of these groups is used as the testing set and the remaining 4 groups are used as the training set. In order to ensure a fair comparison is obtained, images taken from the same subject do not appear in both the training and testing sets in each trial. We further follow [22, 23] to measure the regression performance using mean absolute error (MAE) and root-mean-squared error (RMSE).

4.3 Results of beauty score prediction

With a fivefold cross-validation, we observe that the proposed features consistently outperformed the subset of features, proposed in [22], on all 4 datasets in RMSE (see Tables 3 and 4), and in 3 of the 4 datasets (SCUT-FBP [22], SCUT-FBP5500 [23], and FACES [24]) in MAE (see Tables 1 and 2).

The proposed features also outperformed the *neoclassical features* [26], *golden ratio* features [27], *symmetry* features [27], and *8-ratio vectors* [28], in 3 datasets (SCUT-FBP [22], SCUT-FBP5500 [23] and FACES [24]) and produced a comparable performance on CFD [25] in both MAE and RMSE. This highlights the effectiveness of the newly proposed features, particularly given that the proposed features are extracted from a compact set of 10 facial landmarks, rather than from a dense set of facial landmarks as per the baseline methods.

To further boost the performance, we combined the proposed features with the subset of the features proposed in

(a) (b) (c)
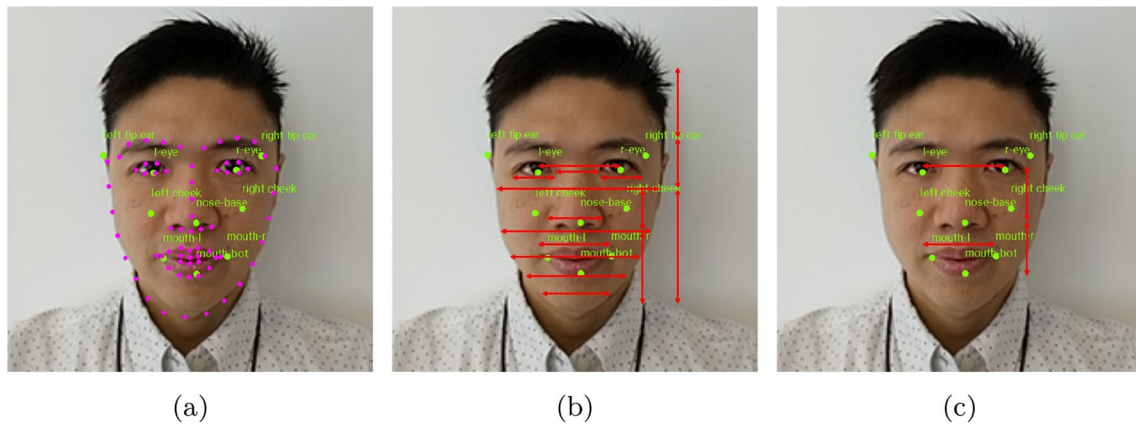
Fig. 7 a The 68 facial landmarks (magenta) detected using the dlib facial landmark detector are much denser than the 10 landmarks (green) detected using Google Face API. **b** Geometric features proposed in [22]. **c** A simplified set of features detected using Google Face API

Table 1 Beauty score prediction error (MAE) on SCUT-FBP [22] and SCUT-FBP5500 [23] benchmark datasets with a fivefold cross-validation

	SCUT-FBP [22]		SCUT-FBP5500 [23]	
	SVM Regr	Linear Regr	SVM Regr	Linear Regr
Xie et al. subset [22]	0.0948	0.0945	0.1003	0.1002
Proposed features	0.0826	0.0829	0.0889	0.0889
Proposed features with landmark distance	0.0838	0.0838	0.0878	0.0878
Neoclassical features [26]	0.0830	0.0830	0.0925	0.0925
Golden ratio [27]	0.0877	0.0875	0.0914	0.0914
Symmetry [27]	0.1004	0.1001	0.1102	0.1102
8-Ratio vectors [28]	0.0886	0.0886	0.0988	0.0988
Xie et al. [22]	0.0840	0.0841	0.0792	0.0792
Proposed features + Xie et al. [22]	**0.0785**	**0.0781**	**0.0760**	**0.0760**

The best values are shown in bold

Table 2 Beauty score prediction error (MAE) on FACES [24] and CFD [25] benchmark datasets with a fivefold cross-validation

	FACES [24]		CFD [25]	
	SVM Regr	Linear Regr	SVM Regr	Linear Regr
Xie et al. subset [22]	0.0883	0.0884	0.0840	0.0840
Proposed features	0.0874	0.0875	0.0841	0.0844
Proposed features with landmark distance	0.0801	0.0801	0.0821	0.0822
Neoclassical features [26]	0.1038	0.1044	0.0811	0.0811
Golden ratio [27]	0.0983	0.0983	0.0833	0.0832
Symmetry [27]	0.1040	0.1040	0.0900	0.0901
8-Ratio vectors [28]	0.1021	0.1022	0.0824	0.0825
Xie et al. [22]	0.0784	0.0782	0.0807	0.0807
Proposed features + Xie et al. [22]	**0.0743**	**0.0745**	**0.0771**	**0.0770**

The best values are shown in bold

[22] (namely, 'Proposed features with landmark distance' in Tables 1, 2, 3, 4 and 5). The results show that the proposed facial symmetry features can be successfully combined with widely used features, based on the distances of facial landmarks, to boost the beauty score prediction performance.

We further compare the proposed set of facial symmetry features with the more complex 18 geometric features proposed in [22]. Note that the features proposed in [22] have to be extracted from a much denser set of facial landmarks (see 68 landmarks in Fig. 7a). Although our method extracts robust features from a sparse set of only 10 facial landmarks, combining the proposed features with landmark distance obtained comparable performance when compared with the 18 features from [22]. When combining the

Table 3 Beauty score prediction error (RMSE) on SCUT-FBP [22] and SCUT-FBP5500 [23] benchmark datasets with a fivefold cross-validation

	SCUT-FBP [22]		SCUT-FBP5500 [23]	
	SVM Regr	Linear Regr	SVM Regr	Linear Regr
Xie et al. subset [22]	0.1325	0.1324	0.1277	0.1274
Proposed features	0.1150	0.1152	0.1155	0.1154
Proposed features with landmark distance	0.1165	0.1165	0.1144	0.1144
Neoclassical features [26]	0.1160	0.1158	0.1193	0.1193
Golden ratio [27]	0.1195	0.1194	0.1183	0.1182
Symmetry [27]	0.1405	0.1401	0.1410	0.1406
8-Ratio vectors [28]	0.1256	0.1256	0.1281	0.1279
Xie et al. [22]	0.1229	0.1231	0.1024	0.1024
Proposed features + Xie et al. [22]	**0.1102**	**0.1097**	**0.0990**	**0.0990**

The best values are shown in bold

Table 4 Beauty score prediction error (RMSE) on FACES [24] and CFD [25] benchmark datasets with a fivefold cross-validation

	FACES [24]		CFD [25]	
	SVM Regr	Linear Regr	SVM Regr	Linear Regr
Xie et al. subset [22]	0.1135	0.1137	0.1071	0.1070
Proposed features	0.1094	0.1096	0.1060	0.1061
Proposed features with landmark distance	0.1015	0.1014	0.1041	0.1043
Neoclassical features [26]	0.1288	0.1296	0.1027	0.1028
Golden ratio [27]	0.1460	0.1438	0.1044	0.1043
Symmetry [27]	0.1300	0.1299	0.1139	0.1139
8-Ratio vectors [28]	0.1273	0.1272	0.1041	0.1043
Xie et al. [22]	0.0979	0.0976	0.1018	0.1019
Proposed features + Xie et al. [22]	**0.0930**	**0.0932**	**0.0986**	**0.0985**

The best values are shown in bold

Table 5 Beauty score prediction error (MAE and RMSE) on SCUT-FBP5500 [23] with a fivefold cross-validation. The prediction errors are normalized and compared to the geometric-based features as in [23]

	MAE		RMSE	
	SVM Regr	Linear Regr	SVM Regr	Linear Regr
Xie et al. subset [22]	0.5017	0.5011	0.6383	0.6370
Proposed features	0.4447	0.4443	0.5775	0.5770
Proposed features with landmark distance	0.4392	0.4391	0.5722	0.5721
Neoclassical features [26]	0.4624	0.4626	0.5965	0.5965
Golden ratio [27]	0.4572	0.4572	0.5913	0.5908
Symmetry [27]	0.5512	0.5509	0.7049	0.7032
8-Ratio vectors [28]	0.4942	0.4941	0.6405	0.6396
Xie et al. [22]	0.3959	0.3959	0.5119	0.5118
Liang et al. [23]	0.3898	0.4289	0.5132	0.5531
Proposed features + Xie et al. [22]	**0.3798**	**0.3799**	**0.4952**	**0.4952**

The best values are shown in bold

proposed features with the 18 features from [22], the best performance is consistently achieved across all tests. This highlights how the proposed facial symmetry features can further improve the performance of landmark distance-based features.

We also compared the proposed features with another 18-dimensional ratio feature vector [23], which is extracted from 86 facial landmarks points. The beauty score prediction performance is evaluated on SCUT-FBP5500, which is the largest of the benchmark datasets used in this study. The evaluation results are presented in Table 5. Solely using the proposed features achieved comparable performance with [23] for linear regression on both MAE and RMSE. We observe that by combining the proposed features with Xie et al. [22], we achieved the best performance across all datasets in both MAE and RMSE.

4.4 Computational Cost

To further evaluate and compare the performance of different features in beauty score prediction, the computation time for training and predicting the scores are presented in Table 6 and 7. All the tests were executed in MATLAB R2020a on a desktop PC with an Intel Core-i7 7700K CPU and 8GB of 2133MHz DDR4 RAM. Since a fivefold cross-validation test was used, the computation times (in seconds) reported are the averaged total times for training and prediction in each fold.

Recall that, in general, our method outperformed all of the other approaches except Xie et al. [22], as presented in Tables 1 ,2, 3 and 4. For the CFD database, our method performed comparable to the best result. The computation time required by our method is comparable with other methods in general. While a smaller beauty score prediction error is obtained by using the method proposed by Xie et al. [22], their method has a significantly higher computational cost when compared with all other methods. This highlights how our method is able to provide a balance between accuracy and efficiency. Depending upon the computational power available to the device, it may be optimal to fuse our proposed features with others to further boost the performance, as demonstrated in the *Proposed features + Xie et al.* [22] example in Tables 6 and 7.

Table 6 Averaged training and beauty score prediction time (in seconds) with SVM regression on benchmark datasets

	SCUT-FBP [22]		SCUT-FBP5500 [23]		FACES [24]		CFD [25]	
	Train	Predict	Train	Predict	Train	Predict	Train	Predict
Xie et al. subset [22]	0.0138	0.0004	1.9533	0.0032	0.2811	0.0011	0.0261	0.0005
Proposed features	0.0340	0.0005	5.2256	0.0092	0.4574	0.0012	0.0495	0.0005
Proposed features with landmark distance	0.0482	0.0005	6.6754	0.0104	0.8335	0.0038	0.0631	0.0005
Neoclassical features [26]	0.0294	0.0006	4.1425	0.0113	0.3664	0.0011	0.0340	0.0005
Golden ratio [27]	0.0348	0.0005	4.9752	0.0104	0.5699	0.0033	0.0487	0.0005
Symmetry [27]	0.0358	0.0006	3.9144	0.0121	0.3805	0.0011	0.0428	0.0006
8-Ratio vectors [28]	0.0223	0.0005	3.3229	0.0078	0.2295	0.0011	0.0291	0.0005
Xie et al. [22]	0.0613	0.0005	9.8003	0.0172	1.2449	0.0044	0.0743	0.0005
Proposed features + Xie et al. [22]	0.0942	0.0005	14.0903	0.0258	1.8689	0.0057	0.1287	0.0006

Table 7 Averaged training and beauty score prediction time (in seconds) with linear regression on benchmark datasets

	SCUT-FBP [22]		SCUT-FBP5500 [23]		FACES [24]		CFD [25]	
	Train	Predict	Train	Predict	Train	Predict	Train	Predict
Xie et al. subset [22]	0.0149	0.0003	0.0128	0.0002	0.0123	0.0003	0.0108	0.0002
Proposed features	0.0114	0.0003	0.0160	0.0003	0.0152	0.0003	0.0126	0.0003
Proposed features with landmark distance	0.0199	0.0003	0.0430	0.0005	0.0211	0.0003	0.0201	0.0003
Neoclassical features [26]	0.0123	0.0003	0.0184	0.0005	0.0147	0.0003	0.0126	0.0002
Golden ratio [27]	0.0223	0.0002	0.0339	0.0002	0.0374	0.0004	0.0196	0.0003
Symmetry [27]	0.0158	0.0003	0.0253	0.0003	0.0189	0.0003	0.0158	0.0002
8-Ratio vectors [28]	0.0109	0.0003	0.0115	0.0003	0.0099	0.0003	0.0102	0.0002
Xie et al. [22]	0.0270	0.0003	0.0399	0.0004	0.0266	0.0003	0.0250	0.0003
Proposed features + Xie et al. [22]	0.0282	0.0002	0.0721	0.0003	0.0420	0.0002	0.0288	0.0002

4.5 Beauty score classification

Similar to the beauty score prediction experiments, we compared the performance of different types of features on the task of classifying the beauty scores into 2 classes: *more beautiful* and *less beautiful*. The class label for each image is determined by calculating whether its beauty score is greater than the dataset's average beauty score or not. After extracting the facial features from the images, a support vector machine (SVM) is used as the classification model. The results are presented in Table 8. The general trend of the results is similar to those obtained in the beauty score prediction tests. Specifically, the proposed features outperformed the subset of landmark distance features in [22], *neoclassical features* [26], *golden ratio* features [27], *symmetry* features [27], and *8-ratio vectors* [28] in 3 of the datasets (SCUT-FBP [22], SCUT-FBP5500 [22] and FACES [24]), and comparable results were obtained on CFD [25]. Combining the proposed features with landmark distance features boosted the classification performance in 3 of the datasets (SCUT-FBP5500 [22], FACES [24], and CFD [25]), ranging from 1.38% to a significant 5.6%. When combining the proposed features with the 18-dimensional features in [22], the classification accuracy is improved in all 4 datasets with a range of 0.16% to 2.49%.

4.6 Feature analysis

To further evaluate the performance of the proposed features, we carried out the analysis of the regression models trained for beauty score prediction. This provided us with a quantitative analysis of the significance of each proposed feature on beauty score prediction. The results (i.e., p values) are presented in Table 9. Although the *p* value of each feature varies across different datasets, 6 of the proposed features (eye_{dev}, $\theta_{eye-mouth-diag}$, $\theta_{mouth-V}$, $\theta_{Lap-eyeL}$, $\theta_{Lap-eyeR}$, and $\theta_{Lap-nose}$) showed a significant impact (i.e., $p < 0.05$)

Table 9 p values of the regression analysis on the proposed features on benchmark datasets

	SCUT-FBP [22]	SCUT-FBP5500 [23]	FACES [24]	CFD [25]
eye_{dev}	0.5917	**0.0000**	**0.0000**	**0.0003**
θ_{hor}	0.5799	0.0965	**0.0046**	**0.0092**
θ_{ver}	0.4597	**0.0000**	**0.0001**	0.3980
$\theta_{eye-mouth-diag}$	0.3852	**0.0104**	**0.0216**	**0.0053**
$\theta_{eye-mouth-V}$	0.1128	**0.0000**	0.0705	0.2728
$\theta_{eye-nose}$	**0.0005**	0.8350	**0.0194**	**0.0012**
$\theta_{mouth-V}$	**0.0273**	**0.0027**	0.3932	**0.0494**
$\theta_{Lap-eyeL}$	**0.0027**	**0.0045**	**0.0000**	0.3079
$\theta_{Lap-eyeR}$	**0.0030**	**0.0133**	**0.0005**	0.5675
$\theta_{Lap-nose}$	**0.0100**	**0.0000**	**0.0000**	0.0942

Bold values indicate a significant impact ($p < 0.05$)

on 3 out of 4 benchmark datasets we tested. Another 3 features (θ_{hor}, θ_{ver}, and $\theta_{ear-nose}$) showed a significant impact on 2 of the benchmark datasets. Additionally, 8 proposed features were identified as having significant impact on the beauty score prediction results for the largest dataset (SCUT-FBP5500 [23]), and 8 features were shown to have significant impact on the second large dataset (FACES [24]). This evaluation highlights the robustness of the proposed features.

We further justify the selection of features by conducting statistical analysis of the regression models on all landmarks, the results of which are presented in Table 10. We observe that the proposed features (i.e., $\theta_{Lap-eyeL}$, $\theta_{Lap-eyeR}$, and $\theta_{Lap-nose}$) have a significant impact on 3 out of 4 benchmark datasets in the beauty score prediction experiment. Conversely, the Laplacian coordinates computed from other landmarks only have a significant impact on 1 to 2 datasets. This further supports the facial landmark selection for the proposed features.

Table 8 Beauty score classification on benchmark datasets with a fivefold cross-validation

	SCUT-FBP [22]	SCUT-FBP5500 [23]	FACES [24]	CFD [25]
Xie et al. subset [22]	63.20%	71.06%	62.87%	59.46%
Proposed features	71.00%	76.81%	66.13%	59.30%
Proposed features with landmark distance	70.80%	78.19%	71.73%	62.31%
Neoclassical features [26]	71.00%	74.14%	67.35%	59.46%
Golden ratio [27]	70.40%	74.22%	72.42%	**63.82%**
Symmetry [27]	61.80%	63.31%	52.97%	50.59%
8-Ratio vectors [28]	66.40%	72.05%	56.04%	63.32%
Xie et al. [22]	62.80%	82.53%	73.68%	63.32%
Proposed features + Xie et al. [22]	**72.20%**	**83.72%**	**76.17%**	63.48%

The best values are shown in bold

Table 10 p values of the regression analysis of the Laplacian coordinates features on beauty score prediction in benchmark datasets

	SCUT-FBP [22]	SCUT-FBP5500 [23]	FACES [24]	CFD [25]
$\theta_{Lap-eyeL}$	**0.0027**	**0.0045**	**0.0000**	0.3079
$\theta_{Lap-eyeR}$	**0.0030**	**0.0133**	**0.0005**	0.5675
$\theta_{Lap-earL}$	0.1114	**0.0431**	**0.0006**	0.8713
$\theta_{Lap-earR}$	0.1474	**0.0273**	**0.0000**	0.6628
$\theta_{Lap-nose}$	**0.0100**	**0.0000**	**0.0000**	0.0942
$\theta_{Lap-mouthL}$	**0.0418**	0.7259	**0.0023**	0.1655
$\theta_{Lap-mouthR}$	0.1142	0.8743	**0.0016**	0.4191
$\theta_{Lap-mouthB}$	0.5341	**0.0001**	0.0530	0.0727

Bold values indicate a significant impact ($p < 0.05$)

Finally, we demonstrate the relationship between the significance of each feature and the associated contribution to the beauty score prediction performance. In particular, we compare the beauty score prediction performance of the proposed features with a larger feature set, namely *All features*, which includes all of the features listed in Tables 9 and 10. We tested the importance of *All features* on all 4 datasets. The results of this are presented in Tables 11 , 12, 13 and 14. From the results, it can be seen that the proposed set of features, determined by the regression model analysis, consistently outperforms the *All features* set across all datasets and different settings. These empirical results suggest that there is a significant relationship between the impact and the overall performance of the features.

Table 11 Beauty score prediction error (MAE) on SCUT-FBP [22] and SCUT-FBP5500 [23] benchmark datasets with a fivefold cross-validation

	SCUT-FBP [22]		SCUT-FBP5500 [23]	
	SVM Regr	Linear Regr	SVM Regr	Linear Regr
Proposed features	0.0826	0.0829	0.0889	0.0889
Proposed features with landmark distance	0.0838	0.0838	0.0878	0.0878
Proposed features + Xie et al. [22]	**0.0785**	**0.0781**	**0.0760**	**0.0760**
All features	0.0847	0.0846	0.0895	0.0895
All features with landmark distance	0.0849	0.0853	0.0887	0.0887
All features + Xie et al. [22]	0.0800	0.0798	0.0791	0.0792

The best values are shown in bold

Table 12 Beauty score prediction error (MAE) on FACES [24] and CFD [25] benchmark datasets with a fivefold cross-validation

	FACES [24]		CFD [25]	
	SVM Regr	Linear Regr	SVM Regr	Linear Regr
Proposed features	0.0874	0.0875	0.0841	0.0844
Proposed features with landmark distance	0.0801	0.0801	0.0821	0.0822
Proposed features + Xie et al. [22]	**0.0743**	**0.0745**	**0.0771**	**0.0770**
All features	0.0892	0.0895	0.0847	0.0846
All features with landmark distance	0.0808	0.0806	0.0838	0.0836
All features + Xie et al. [22]	0.0746	0.0746	0.0779	0.0780

The best values are shown in bold

Table 13 Beauty score prediction error (RMSE) on SCUT-FBP [22] and SCUT-FBP5500 [23] benchmark datasets with a fivefold cross-validation

	SCUT-FBP [22]		SCUT-FBP5500 [23]	
	SVM Regr	Linear Regr	SVM Regr	Linear Regr
Proposed features	0.1150	0.1152	0.1155	0.1154
Proposed features with landmark distance	0.1165	0.1165	0.1144	0.1144
Proposed features + Xie et al. [22]	**0.1102**	**0.1097**	**0.0990**	**0.0990**
All features	0.1172	0.1173	0.1155	0.1154
All features with landmark distance	0.1177	0.1182	0.1146	0.1145
All features + Xie et al. [22]	0.1122	0.1119	0.0993	0.0993

The best values are shown in bold

Table 14 Beauty score prediction error (RMSE) on FACES [24] and CFD [25] benchmark datasets with a fivefold cross-validation

	FACES [24]		CFD [25]	
	SVM Regr	Linear Regr	SVM Regr	Linear Regr
Proposed features	0.1094	0.1096	0.1060	0.1061
Proposed features with landmark distance	0.1015	0.1014	0.1041	0.1043
Proposed features + Xie et al. [22]	**0.0930**	**0.0932**	**0.0986**	**0.0985**
All features	0.1114	0.1118	0.1068	0.1066
All features with landmark distance	0.1011	0.1008	0.1057	0.1055
All features + Xie et al. [22]	0.0933	0.0933	0.0990	0.0991

The best values are shown in bold

5 Discussion

The evaluation of the proposed features and comparison against other baseline methods suggests that we are able to achieve excellent performance across multiple datasets, while retaining comparable computation time to class-leading methods. Our method outperforms the comparable methods on beauty score prediction and beauty score classification, including the method of Xu et al. [31], which achieved 0.2501 MAE and 0.3263 RSME on SCUT-FBP5500 using Hierarchical Multi-task Network (HMT-Net). Other deep learning-based methods (such as presented in [50]) additionally use color and texture-based features, i.e., they cannot be directly compared to our method, which is based on geometric features of a face.

Our method is only outperformed by the much more computationally expensive method proposed by Xie et al. [22]. We did, however, note that our method also compliments [22], by improving performance across the board, when combined. Our feature significance evaluation also confirms the impact that the proposed features have on overall performance.

There are of course limitations associated with the proposed features. One such limitation is in the generation of the mid-line used for the eye_{dev} calculation (illustrated in Fig. 2b). Since the mid-line is based upon a line passing through both the nose base and mouth bottom, it is possible that the mid-line may not be perfectly vertically aligned, subsequently affecting the eye_{dev} calculation. However, these two landmarks are the only available landmarks which can be used to define this vertical mid-line without also including the eye landmarks. This is due to the Google Face API, as it extracts a limited number of landmarks, which while helping to reduce the complexity of the model, can impact these calculations. Furthermore, from our evaluation we discovered that the eye_{dev} feature is particularly discriminative, and as such it was important that we include this in our proposed feature set. It is for this reason that the proposed user interface measurement overlay tools are particularly important and form a key part of the feature extraction pipeline.

The horizontal eye-level deviation (illustrated in Fig. 2c) is similarly affected by the generation of a mid-line. This once again highlights the importance of the proposed measurement tools as a means of generating a suitably accurate input image.

Given the evident links between the proposed features and the beauty score metrics used in our evaluation, it is clear that a system, such as our proposed framework, could aid in several key areas relating to plastic surgery and other rehabilitative health-related tasks. With facial cosmetic surgery having a proven positive effect upon self-esteem and self-efficacy, the ability to aid with this process has the potential to provide an impactful contribution in the healthcare sector. This is particularly true given that significant impairment of self-esteem and self-efficacy may also require additional psychological intervention. As such, the proposed framework has the capacity to serve as a quick and objective assessment for patients trying to 'improve' the perceived attractiveness of their faces, and for surgeons hoping to quantitatively evaluate facial structures and deformations.

6 Conclusion and future work

In this paper, a new set of facial features and an augmented reality (AR) tool (smartphone app) to assist the user in evaluating facial symmetry interactively is proposed. The features are computed from a compact set of facial landmarks which can be extracted at a low computational cost. We quantitatively evaluated the features proposed in this paper for predicting the attractiveness of faces from human portraits from four benchmark datasets. Experimental results showed that the performance of the proposed features is comparable to those extracted from a set with much denser facial landmarks. By further combining the proposed features with existing geometric facial features, the beauty score prediction performance can also be improved. In addition, a prototype of the AR app was developed on the Android platform. Important facial landmarks can be tracked in the live video stream and 4 different types of commonly used facial symmetry measurements are provided.

While encouraging results are obtained, there are some limitations. In particular, the proposed features are extracted from the facial landmarks. As a result, the accuracy of the landmarks extracted from the Google Face API has significant impact on the quality of the proposed facial features. In addition to the proposed geometric features, color and texture information could also potentially provide additional useful features for beauty score prediction. However, a set of identified significant landmarks depend upon a particular dataset used, which is a limitation of our approach. Moreover, the larger database of faces allows to elicit more significant features, which underscores the need to have larger and more diverse face datasets for facial beauty research.

In the future, we will evaluate the performance of the system by combining both geometric and image appearance features. Another intended area of future work is to improve the robustness of the performance in a real-world setting. Given that facial images captured from smartphone cameras in the real-world produce variable image quality (due to illumination, resolution and sharpness), and since most of the publicly available datasets were captured in an indoor and controlled environment, this could be a useful direction. In order to improve the robustness of the feature extraction and beauty score prediction, image editing-based data augmentation, such as adjusting the color, contrast, global and local illumination and sharpness [51], could subsequently be used.

Similarly, since facial expression and the underlying emotional state of the subject can also affect measurement accuracy, we are interested in normalizing the facial expression. By analyzing 2D [52] and 3D [53] facial information and the associated emotional states, we may also be able to further improve the robustness of the proposed method in the future.

Additionally, motivated by a recent study [54] which suggested that the Microsoft Kinect depth sensor can be used in a wide range of healthcare imaging applications, we would like to introduce a face assessment tool that can analyze live video and 3D information. By using AR devices which are equipped with depth cameras, such as Microsoft Hololens, our prototype could potentially be improved by providing 3D facial information for further analysis.

Finally, we would like to conduct a large-scale user study to evaluate the effectiveness of the AR tool in practical use.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Esposito Anna, Esposito Antonietta M, Vogel Carl (2015) Needs and challenges in human computer interaction for processing social emotional information. Pattern Recognit Lett 66:41–51
2. Little Anthony C (2014) Facial attractiveness. Wiley Interdiscip Rev: Cognitive Sci 5(6):621–634
3. Little Anthony C, Jones Benedict C, DeBruine Lisa M (2011) Facial attractiveness: evolutionary based research. Philosophical Transactions Royal Soc B: Biol Sci 366(1571):1638–1659
4. Talamas Sean N, Mavor Kenneth I, Perrett David I (2016) Blinded by beauty: attractiveness bias and accurate perceptions of academic performance. PLOS ONE 11(2):148284
5. Hui Shen, Desmond K. P. Chau, Jianpo Su, Ling-Li Zeng, Weixiong Jiang, Jufang He, Jintu Fan, and Dewen Hu (2016). Brain responses to facial attractiveness induced by facial proportions: evidence from an fMRI study. *Scientific Reports*, 6(1), October 2016
6. Ibáñez-Berganza Miguel, Amico Ambra, Loreto Vittorio (2019) Subjectivity and complexity of facial attractiveness. Sci Rep 9(1):1–12
7. Jefferson Yosh (2004) Facial beautyestablishing a universal standard. Int J orthod 15:9–22
8. Dantcheva A, Dugelay J-L (2015) Assessment of female facial beauty based on anthropometric, non-permanent and acquisition characteristics. Multimedia Tools Appl 74(24):11331–11355
9. Soler C, Keklinen J, Nez M, Sancho M, Nez J, Yaber I, Gutirrez R (2012) Male facial anthropometry and attractiveness. Perception 41(10):1234–1245
10. Hsiu-Hsia Lin, Lun-Jou Lo, Wen-Chung Chiang, and Chin-Fa Chen (2018). An automatic assessment of facial symmetry before and after orthognathic surgery based on three-dimensional contour features using deep learning system. In *151st The IIER International Conference*, pages 126–129, 2018
11. Chiang Wen-Chung, Lin Hsiu-Hsia, Huang Chiung-Shing, Lo Lun-Jou, Wan Shu-Yen (2014) The cluster assessment of facial attractiveness using fuzzy neural network classifier based on 3d moir features. Pattern Recognit 47(3):1249–1260
12. Hong Yu-Jin, Nam Gi Pyo, Choi Heeseung, Cho Junghyun, Kim Ig-Jae (2017) A novel framework for assessing facial attractiveness based on facial proportions. Symmetry 9(12):294
13. Šalkevicius Justas, Damaševičius Robertas, Maskeliunas Rytis, Laukienė Ilona (2019) Anxiety level recognition for virtual reality therapy system using physiological signals. Electronics 8(9):1039
14. Milutinovic Jovana, Zelic Ksenija, Nedeljkovic Nenad (2014) Evaluation of facial beauty using anthropometric proportions. Sci World J 1–8:2014
15. Millar Keith, Bell Aileen, Bowman Adrian, Brown Denise, Lo Tsz-Wai, Siebert Paul, Simmons David, Ayoub Ashraf (2013) Psychological status as a function of residual scarring and facial asymmetry after surgical repair of cleft lip and palate. Cleft Palate-Craniofacial J 50(2):150–157
16. Legislation.gov.uk. Equality act 2010, 2010
17. Mckearney Richard M, Williams John V, Mercer Nigel S (2013) Quantitative computer-based assessment of lip symmetry following cleft lip repair. Cleft Palate-Craniofacial J 50(2):138–143

18. Patel Arti, Islam Syed, Murray Kevin, Goonewardene Mithran (2015) Facial asymmetry assessment in adults using three-dimensional surface imaging. Prog Orthod 16:36. https://doi.org/10.1186/s40510-015-0106-9

19. Jia Wu, Heike Carrie, Birgfeld Craig, Evans Kelly, Maga Murat, Morrison Clinton, Saltzman Babette, Shapiro Linda, Tse Raymond (2016) Measuring symmetry in children with unrepaired cleft lip: Defining a standard for the three-dimensional midfacial reference plane. Cleft Palate-Craniofacial J 53(6):695–704

20. Nkenke Emeka, Lehner Bernhard, Kramer Manuel, Haeusler Gerd, Benz Stefanie, Schuster Maria, Neukam Friedrich W, Vairaktaris Eleftherios G, Wurm Jochen (2006) Determination of facial symmetry in unilateral cleft lip and palate patients from three-dimensional data: Technical report and assessment of measurement errors. Cleft Palate-Craniofacial J 43(2):129–137

21. van Loon B, Maal TJ, Plooij JM, Ingels KJ, Borstlap WA, Kuijpers-Jagtman AM, Spauwen PH, Berg SJ (2010) 3d stereophotogrammetric assessment of pre- and postoperative volumetric changes in the cleft lip and palate nose. Int J Oral Maxillofac Surg 39(6):534–540

22. Xie D, Liang L, Jin L, Xu J, and Li M (2015). Scut-fbp: a benchmark dataset for facial beauty perception. In *2015 IEEE International Conference on Systems, Man, and Cybernetics*, pages 1821–1826, Oct 2015

23. Liang L, Lin L, Jin L, Xie D, and Li M (2018). Scut-fbp5500: a diverse benchmark dataset for multi-paradigm facial beauty prediction. In *2018 24th International Conference on Pattern Recognition (ICPR)*, pp 1598–1603

24. Ebner Natalie C, Riediger Michaela, Lindenberger Ulman (2010) FACES - A database of facial expressions in young, middle-aged, and older women and men: Development and validation. Beh Res Methods 42(1):351–362

25. Ma Debbie S, Correll Joshua, Wittenbrink Bernd (2015) The chicago face database: a free stimulus set of faces and norming data. Behav Res Methods 47:1122–1135

26. Farkas Leslie, Hreczko Tania A, Kolar John C, Munro Ian R (1985) Vertical and horizontal proportions of the face in young adult north american caucasians. J Oral Maxillofac Surg 43(11):917

27. Schmid Kendra, Marx David, Samal Ashok (2008) Computation of a face attractiveness index based on neoclassical canons, symmetry, and golden ratios. Pattern Recognit 41(8):2710–2717

28. Aarabi P , Hughes D, Mohajer K, and Emami M (2001). The automatic measurement of facial beauty. In *2001 IEEE International Conference on Systems, Man and Cybernetics. e-Systems and e-Man for Cybernetics in Cyberspace*. 4: 2644–2647

29. Ulrich Luca, Dugelay Jean-Luc, Vezzetti Enrico, Moos Sandro, Marcolin Federica (2020) Perspective morphometric criteria for facial beauty and proportion assessment. Appl Sci 10(1):8

30. Gan J, Xiang L, Zhai Y, Mai C, He G, Zeng J, Bai Z, Donida Labati R, Piuri V, Scotti F (2020) Facial beauty prediction based on multi-task transfer learning. 2m beautynet. IEEE Access 8:20245–20256

31. Xu L, Fan H and Xiang J (2019). Hierarchical multi-task network for race, gender and facial attractiveness recognition. In *2019 IEEE International conference on image processing (ICIP)*, pp 3861–3865

32. Zhao J, Cao M, Xie X, Zhang M, Wang L (2019) Data-driven facial attractiveness of chinese male with epoch characteristics. IEEE Access 7:10956–10966

33. Shu L, Yang-Yu F, Zhe G, Ashok S, Afan A (2017) A landmark-based data-driven approach on facial attractiveness computation. Neurocomputing 238:168–178

34. Sajid M, Shafique T, Jabbar A B M, Riaz I, Amin S, Manzoor S (2018) Automatic grading of palsy using asymmetrical facial features: a study complemented by new solutions. Symmetry 10(7):242

35. Seok Kim H, Young Kim S, Ho Kim Y, Park Kwang Suk (2015) A smartphone-based automatic diagnosis system for facial nerve palsy. Sensors 15(10):26756–26768

36. Gunes H, Piccardi M (2006) Assessing facial beauty through proportion analysis by image processing and supervised learning. Int J Human-Comput Stud 64(12):1184–1199

37. Stephen R M. Dr. stephen r. marquardt on the golden decagon and human facial beauty. interview by dr. gottlieb. 36:339–47, 07 2002

38. Holland E (2008) Marquardt's phi mask: pitfalls of relying on fashion models and the golden ratio to describe a beautiful face. Aesthet Plast Surg 32(2):200–208

39. Douglas DB, Wilke CA, Gibson JD, Boone JM, Wintermark M (2017) Augmented reality: advances in diagnostic imaging. Multimodal Technol Interact 1(4):29

40. Joey deVilla (2017). Augmented reality in android with google's face api, 2017

41. Google Inc. Face detection concepts overview - mobile vision, 2016

42. Alexa M (2003) Differential coordinates for local mesh morphing and deformation. V Comput 19(2):105–114

43. Ho ESL, Taku K, Chiew-Lan T (2010) Spatial relationship preserving character motion adaptation. ACM Transactions Gr 29(4):1–8

44. Ho ESL and Shum HPH (2013). Motion adaptation for humanoid robots in constrained environments. In *Robotics and Automation (ICRA), 2013 IEEE International conference on*, pp 1–6, may 2013

45. Naini FB, Gill DS (2008) Facial aesthetics: 1. concepts and canons. Dental Update 35(2):102–107

46. Kim Y-H (2007) Easy facial analysis using the facial golden mask. J Craniofacial Surg 18(3):643–649

47. Marquardt Beauty Analysis, Inc., 2018

48. Ebner Natalie C, Luedicke J, Voelkle Manuel C, Riediger M, Lin T, Lindenberger U (2018) An adult developmental approach to perceived facial attractiveness and distinctiveness. Front Psychology 9:561

49. Mao H, Jin L, and Du M (2009). Automatic classification of chinese female facial beauty using support vector machine. In *2009 IEEE International Conference on Systems, Man and Cybernetics*, pp 4842–4846, Oct 2009

50. Zhai Y, Huang Y, Xu Y, Gan J, Cao H, Deng W, Donida Labati R, Piuri V, Scotti F (2020) Asian female facial beauty prediction using deep neural networks via transfer learning and multi-channel feature fusion. IEEE Access 8:56892–56907

51. Sakkos D, Ho ESL, Shum H, Elvin G (2020) Image editing based data augmentation for illumination-insensitive background subtraction. J Enterp Information Manag. https://doi.org/10.1108/JEIM-02-2020-0042

52. Chan JCP, Shum HPH, Wang H, Yi L, Wei W, Ho ESL (2019) A generic framework for editing and synthesizing multimodal data with relative emotion strength. Comput Anim Virtual Worlds 30(6):1871

53. Stef A, Perera K, Shum HPH, and Ho ESL (2018). Synthesizing expressive facial and speech animation by text-to-ipa translation with emotion control. In *2018 12th International Conference on Software, Knowledge, Information Management Applications (SKIMA)*, pages 1–8, Dec 2018

54. Pöhlmann STL, Harkness EF, Taylor CJ, Astley SM (2016) Evaluation of kinect 3d sensor for healthcare imaging. J Med Biological Eng 36(6):857–870

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.