Received: 6 September 2022 | Accepted: 27 March 2023

DOI: 10.1111/srt.13324

ORIGINAL ARTICLE

WILEY

Comparative analysis of original face and skin-warped average face images for the scoring of skin attributes

Quentin Marin[1] | Tatsuya Honda[2] | Yuri Okano[3] | Marie Cherel[1] | Elodie Prestat-Marquis[1]

[1]Newtone Technologies, Lyon, France

[2]DRC Co., Ltd., Osaka, Japan

[3]CIEL Co., Ltd., Tokyo, Japan

Correspondence
Elodie Prestat-Marquis, Newtone Technologies, 13 bis Place Jules Ferry, 69006 Lyon, France.
Email: eprestat@newtone.fr

Funding information
Newtone Technologies

Abstract

Objectives: Representative of a panel, an average face image could be used to analyse/display skin changes while alleviating image rights constraints. Therefore, we used landmark-based deformation (warping) of individual skin images onto their panel's average face, evaluating this approach's relevance and possible limits.

Methods: An average front face image was constructed from images of 71 Japanese women (50–60 years old). After warping individual skin images onto this average face, the resulting skin-warped average faces were presented to three experts who graded: forehead wrinkles, nasolabial fold, wrinkle of the corner of the lips, pore visibility and skin pigmentation homogeneity. Two experts estimated subjects' age. Results were compared to gradings performed on original images.

Results: Inter-expert grading shows excellent to good correlation whatever image type: from 0.918 (forehead wrinkles) to 0.693 (visibility of pores). Correlations between scoring of both image types are almost always higher than inter-expert correlations (maximum: 0.939 for forehead wrinkles—minimum: 0.677 for pore visibility). Frequencies of grades/ages are similar when scoring original and skin-warped average face images. Experts scores are similar in 90.6%–99.3% of the cases. Average deviations upon scoring both image types are smaller than average inter-expert deviations on original images.

Conclusions: Scoring facial characteristics in original images and skin-warped average face images show an excellent agreement, even for perceived age, a complex feature. This opens the possibility of using this approach to grade facial skin features, monitor changes over time, and to valorise results on a face deprived of image rights.

KEYWORDS
ageing signs, average face, claim substantiation, digital image, skin grading, statistics

1 | INTRODUCTION

In the cosmetic field, claim substantiations face several challenges. One that could seem trivial is illustrating the performance of a product. With the growing awareness of subjects' image rights and the implementation of more regulations to protect personal data, presenting the effects of a product using the very image of a subject becomes increasingly complicated. Besides, the beauty standards of subjects are rarely an inclusion criterion in clinical studies, and therefore, their image may not match the communication constraints of cosmetic brands.

Another point is that claim substantiations require objective measurements of treatment outcomes. Despite the development of several instruments, the evaluation of facial features such as wrinkles, pigmentation spots or pores still largely relies on visual scoring performed by experts. For this purpose, structured scales have been developed, especially photonumeric grading scales, as they proved useful in increasing rating accuracy.[1–7] Yet, upon validation, inter-expert variability was found superior to intra-expert reproducibility,[5,7–9] indicating that the grading is subjected to experience and subjectivity.[10]

One solution to both issues could be taking advantage of average faces. Indeed, when, in 1878, Sir Francis Galton constructed the first average faces by multiple exposures of the same photographic plate,[11] he was astonished by the beauty of the resulting superimposed image. He also realised that average faces highlight common characteristics of the group it is made from.

Since this initial work, average face construction has taken advantage of progresses made in computational graphics and computer performances. The first attempt would now seem rudimentary but was a pioneering challenge considering the capacities of the 1990s desktop personal computers.[12] It consisted of blending deformed pictures that were manually adjusted to match the locations of eye pupils and the middle of the lip line. Another milestone came with image morphing, which shows that the best way to transform a face into another is to deform both using the average position of landmarks.[13] Several studies took advantage of this work, relying on the deformation of individual faces using average positions of selected characteristic points and leading to the construction of average faces with increasing quality.[14,15] Despite obvious advantages, studies taking advantage of this approach were rather restricted to psychological studies.

The use of average faces in biology or cosmetic is sparse. Yet, it was successfully used to perform a preliminary analysis of the age-related decrease in the contrast between facial features (eyes, lips and brows) and the colour/luminance of the surrounding skin in different ethnic groups.[16] It also enabled identifying morphological differences between Chinese and Caucasian faces and how an increase in body mass index differently affects morphological characteristics of each group.[17]

In both works, average face construction also relied on the deformation of individual faces based on the average position of characteristic morphological points. Yet, technical improvements led to high-resolution, sharp average face images and the interest of average faces was further demonstrated in a work analysing changes in facial skin colour upon a cosmetic treatment.[18] The authors reconstructed an average face from two subgroups of subjects: one applying a cosmetic cream with an active ingredient and the other a vehicle cream. They analysed skin colour in both groups at two time points. As the deformation approach ensured a perfect morphological match, they could extract and directly compare skin colour parameters in both subgroups, following their changes over time.

With advances in computing, additional applications can be envisioned. Once computed for a group of subjects, the individual skin pictures of each subject can be warped onto their average face. The resulting image is then a face having the morphology of the average face of the panel with the actual skin image of a subject. Applying this approach to all subjects of a panel results in a skin-warped average face image of each subject. All these images are highly standardised, being identical by their underlying morphology, but each holds a subject's unique skin image.

Accurate skin-warped average faces could have many applications. Indeed, the average face being constructed, it does not have the morphological characteristic of any subject it is made from. Still, the warping of the facial picture enables presenting skin characteristics of several subjects on a face with identical morphology. In the field of cosmetics, this could be of particular interest to score clinical signs on highly standardised faces, possibly alleviating interferences from individual morphological differences. It could also be of interest to present the actual result of a cosmetic on the face of a subject that has been 'made' more attractive and closer to the generally admitted beauty canons by the average face reconstruction process while respecting image privacy and limiting image rights constraints. Yet, applications are possible only if the average face construction and warping processes do not alter facial features. Therefore, our primary goal was to evaluate the accuracy of clinical cosmetic scoring on skin-warped average faces. To do so, we compared assessments performed on original pictures and skin-warped average faces. We did so for local facial features (wrinkles, pores, pigmentation) and, to test the limits of the approach, on a global facial characteristic: perceived age.

2 | MATERIALS AND METHODS

2.1 | Subjects

This non-invasive study was performed following the principles of the Declaration of Helsinki. All subjects were informed of the purpose of the study. They received detailed information about the procedures and gave written informed consent before enrolment.

The subjects were 71 healthy women living in the region of Osaka or Tokyo (Japan), with an age ranging between 50 and 60 years (53.7 ± 2.6 years old). This rather narrow age range has been selected as it corresponds to the period of life during which the clinical characteristics evaluated significantly evolve.[19,20] The inclusion criteria were the absence of severe skin alteration or skin diseases, aesthetic surgery or therapy on the face, excessive sun exposure during the month preceding the study. Before their pictures were taken, the subjects were asked

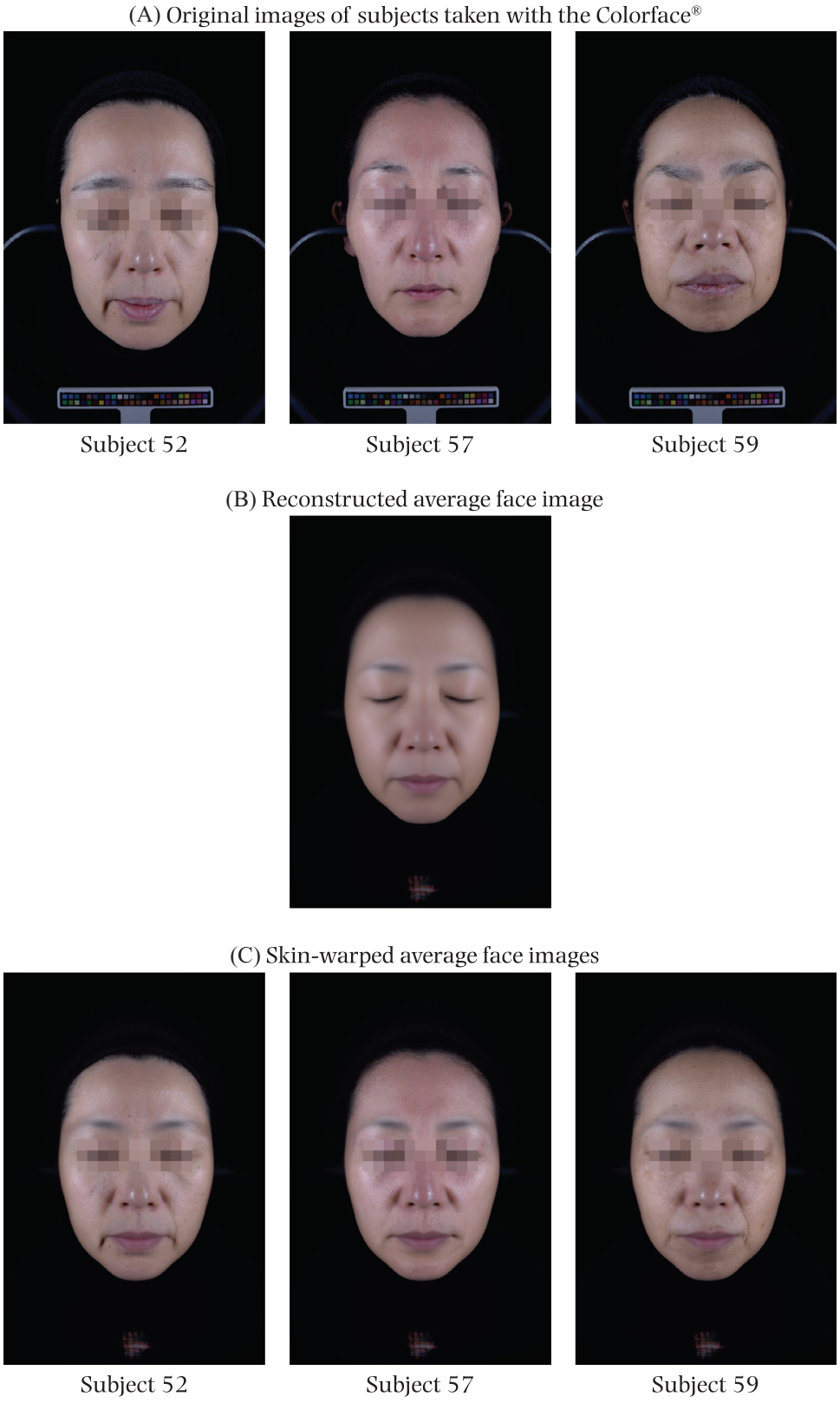

FIGURE 1 Overview of the reconstruction of skin-warped average face images. (A) Examples of original images of subjects. (B) Reconstructed average front face image of the 71 subjects. (C) Examples of skin-warped average face images.

to wash their face with a gentle face cleanser and allowed to rest for 20 min in a climatized room (21 ± 1°C and 50% ± 5% relative humidity).

2.2 | Acquisition and pre-processing of front and facial side images

Individual standardised front images of the entire face were taken under normal diffused light using the Colorface (Newtone Technologies, Lyon, France), a device equipped with a built-in single-lens reflex (focal length: 20 mm) Nikon D5300 camera (Nikon Inc., Japan). Images were saved as high-resolution (4000 × 6000 px) jpeg files, and all included a 48-colour reference chart (ColorChart, Newtone Technologies, Lyon, France).

As a preliminary step, to prevent variations in lighting that could affect luminosity, contrast and colour consistency, a colour correction was performed using the StudyManager software (Newtone Technologies (Lyon, France) and the ColorChart as a reference.

2.3 | Reconstruction of the average front face and extraction of individual images warped onto the average face

Front-face images from all 71 subjects were used to construct an average full-face image according to a procedure previously published.[17–18] Briefly, a Random Forest algorithm automatically detects 76 characteristic morphological points on each image.[21,22] For each picture, the coordinates of these points are registered in a common reference space and moved to their average positions by a deformable registration algorithm based on spline.[23–25] To ensure the best colorimetric consistency, a statistical analysis is then performed on every pixel, and the median colour value is used to generate the average face image. In a last step, the individual skin images of the 71 subjects were each warped onto the panel's average face image so that each characteristic morphological points of each individual face matched their position of the average face. This procedure uses the deformable registration algorithm based on spline that served for average face construction. As our focus was to generate faces only differing in their skin characteristics, the eyes, eyebrows, nostrils and lips of the average face were preserved in all individual skin-warped average faces. An overview of the complete process is depicted in Figure 1.

2.4 | Grading of facial features

Three skin experts graded the severity of forehead wrinkles, the nasolabial fold and the wrinkle of the corner of the lips. They also scored the visibility of the sebaceous pores and the homogeneity of skin pigmentation at the level of the cheek. These characteristics were evaluated separately using cropped front face images (Figures 2-3). Except for the forehead wrinkles that encompass the entire top part of the face, which was scored once by subject, experts scored both

FIGURE 2 Location of the cropped zones for the scoring of: 1, forehead wrinkles, 2, visibility of the sebaceous pores and homogeneity of skin pigmentation, 3, nasolabial fold and wrinkle of the corner of the lips. For clarity reasons, cropped zones are presented on one side of the face, but left and right sides were scored using, when needed, mirror images to present them according to Bazin and Doublet.[14]

facial sides—one in the actual image orientation and the other one mirrored—of each subject according to the orientation and the scale of Bazin and Flament.[4] Grades given upon scoring of wrinkles, sebaceous pores and the homogeneity of skin pigmentation on both facial sides were considered independent and treated as such during the statistical analysis.

The last characteristic, perceived age, was scored by two dermatologist experts using entire front face images. Experts were informed of the age range of subjects and graded randomly five times each of them by entering the digits of their perceived age. The perceived age considered for each subject was the median of the five repeated scorings.

All gradings were performed on personal computers. More precisely, this implies that each expert performed his/her grading on the same screen (that was colour calibrated prior to scoring sessions) and in the same environment, but that both conditions could slightly differ between experts. For all gradings, the images of subjects were presented in random order, using the Photoscale online image grading application (Newtone Technologies, Lyon, France). Evaluations were performed in two independent series: the original images taken with

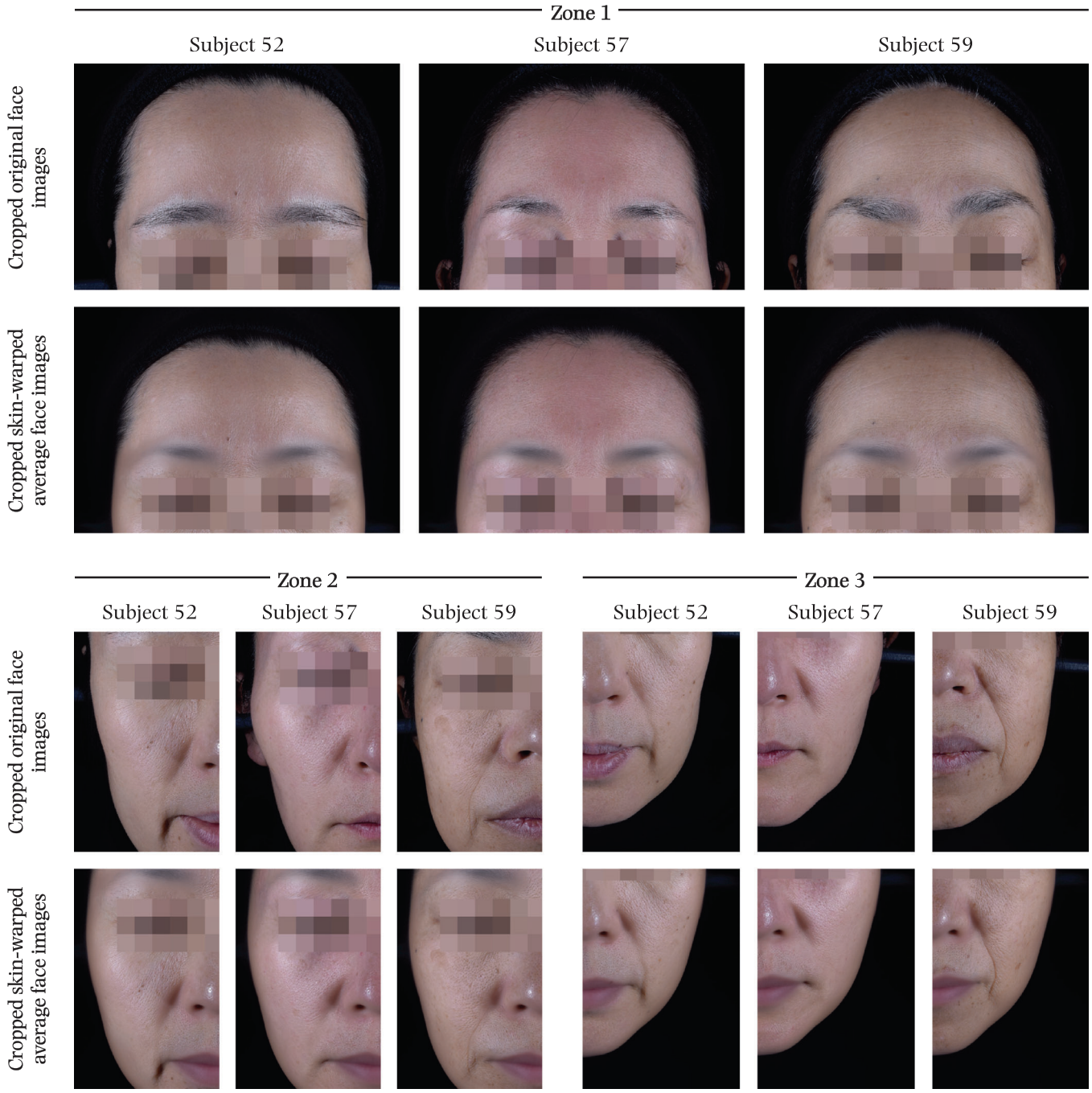

FIGURE 3 Representative original and skin-warped average face cropped images used for the scoring of the different facial features.

the Colorface and the individual skin-warped average face images. Experts could select the type of images they wanted to start grading but had to score each image set of an attribute in a row.

2.5 | Statistical analysis

The inter-expert scoring homogeneity in one image type was performed by calculating the Pearson coefficient of correlation between scores given by two experts. The Pearson correlation coefficient was also used to analyse grades given by one expert upon scoring of original images and skin-warped average face images.

When considering grades given by all experts to a subject, whether from original images or skin-warped average face images, results are presented as the mean value ± standard deviation of the grades scored by experts. For perceived age, as two experts only performed the scoring, the standard deviation was regarded as meaningless and not reported. The frequency at which grades were given on both image sets (original face and skin-warped average faces) were compared using chi-square and considering a significance level of 0.05.

TABLE 1 Inter- and intra-expert correlation for the scoring of wrinkles on original and skin-warped average face images

	Original images			Skin-warped average face images			Original versus skin-warped average face images		
	Exp 1	Exp 2	Exp 3	Exp 1	Exp 2	Exp 3	Exp 1	Exp 2	Exp 3
Forehead wrinkles									
Exp 1	–	0.904	0.918	–	0.900	0.914	0.935	–	–
Exp 2	0.904	–	0.894	0.900	–	0.895	–	0.939	–
Exp 3	0.918	0.894	–	0.914	0.895	–	–	–	0.938
Nasolabial fold									
Exp 1	–	0.853	0.829	–	0.841	0.837	0.845	–	–
Exp 2	0.853	–	0.825	0.841	–	0.828	–	0.825	–
Exp 3	0.829	0.825	–	0.837	0.828	–	–	–	0.826
Wrinkle at the corner of the lips									
Exp 1	–	0.845	0.895	–	0.785	0.883	0.880	–	–
Exp 2	0.845	–	0.828	0.785	–	0.803	–	0.901	–
Exp 3	0.895	0.828	–	0.883	0.803	–	–	–	0.877

Grade differences upon scoring original and skin-warped average face images were analysed by calculating the frequency of a defined score difference. They are presented as the mean value + standard deviation of the grade differences of the three experts, except for perceived age for which standard deviation is not reported as two experts only are involved. Mean absolute deviations and mean relative deviations were also calculated to highlight differences. For all these analyses, original face images were used as a reference; namely, differences are calculated as the grade given upon scoring original faces images minus the score given upon scoring skin-warped average face images.

3 | RESULTS

3.1 | Comparison of original images and skin-warped average face images for the scoring of facial wrinkles

To identify possible differences between skin-warped average face and original images scoring, wrinkles from different regions of the face were first analysed: forehead wrinkles, the nasolabial fold and the wrinkle at the corner of the lips.

Grades given by the three experts to these wrinkles show high correlations whatever type of image rated—original or warped (Table 1). When scoring original images, the inter-expert correlation coefficient of grades ranges from a maximum of 0.918 (between expert 1 and 3 for the forehead wrinkle) to a minimum of 0.825 (between expert 2 and 3 for the nasolabial folds). The overall coefficient of correlation is 0.866 ± 0.037. Similar results are obtained when evaluating skin-warped average face images. The overall inter-expert correlation coefficient of the grades is 0.854 ± 0.46. The maximum correlation is 0.914 (between expert 1 and 3, similarly to the scoring of original images), and the minimum of 0.785 (between expert 1 and 2 for the nasolabial folds).

For the three wrinkles studied, comparisons of results obtained on the two types of images show that the intra-expert correlation between gradings on original versus skin-warped average face images is generally higher than inter-expert correlations achieved when scoring one image type (Table 1). In addition, the overall shape of the graph presenting the mean frequency of scorings by experts is similar for original and skin-warped images (Figure 4). No statistical differences were found between the frequencies at which experts gave a grade when using original or skin-warped average face images. This is true for the mean frequency of grades given by the three experts and for each expert individually (data not shown).

The analysis of grade differences upon scoring original and skin-warped average face images shows that experts give identical grades in more than half of the cases when scoring both image types and that large grade differences are scarce (Figure 5). Identical grading happened in 60.1% of the cases for forehead wrinkles and in 98.6% of the cases with a +1 or −1 difference. The lowest value is for the nasolabial fold for which the very same grades are given in 51.2% of the cases and similar grading (± 1 grade difference) is achieved in 92.0% of the cases.

Calculation of the mean absolute deviations (Figure 5) indicates evenly distributed low differences along the forehead wrinkles scoring scale—from a minimum of 0.13 points for grade 6 to a maximum of 0.59 for grade 4. Mean relative deviations show that differences relate to a low tendency to under-score skin-warped average face images compared to original images. For the nasolabial fold (Figure 5), only the highest grade (7) presents a larger mean absolute and relative deviations (1.38 points), most probably due to the small number of subjects. Most of the scoring scale (grades 0–6) shows low deviations (from 0.45 to 0.78 points). Only the two lowest grades tend to be slightly underscored on skin-warped average face images (−0.45 and −0.20 points),

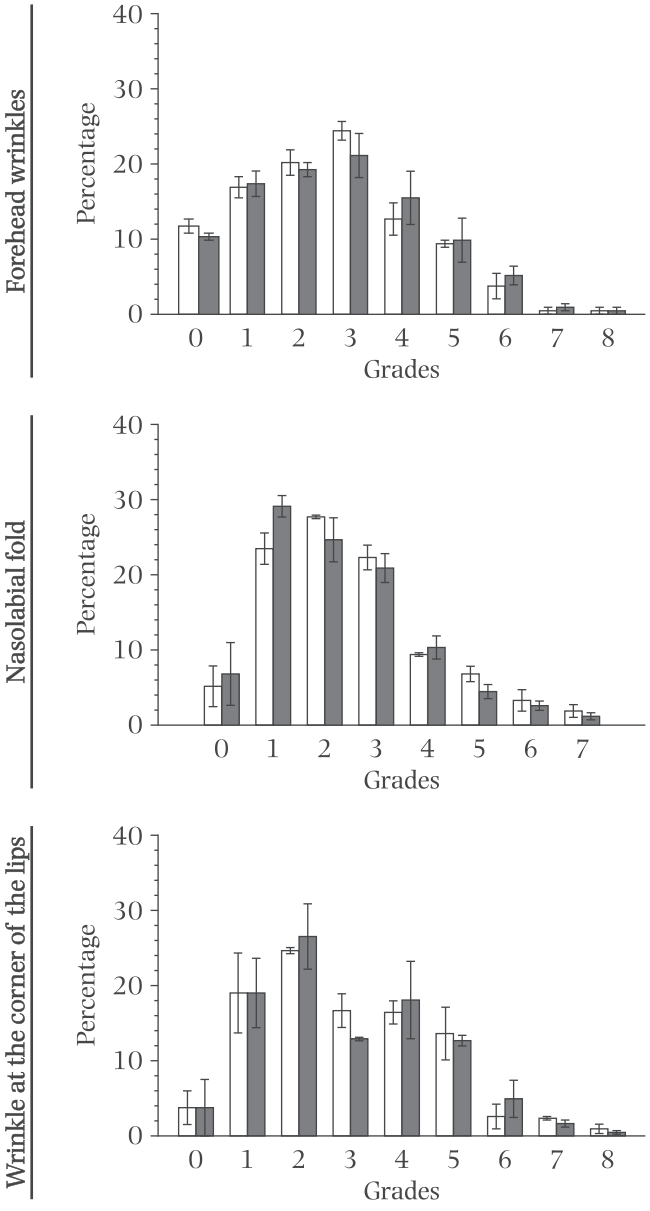

FIGURE 4 Mean grades given upon wrinkle scoring of original (in white) and skin-warped average face (in dark grey) images. Results are presented as the mean value ± standard deviation of the grades given by the three experts.

while most of the scoring scale (grades 2 to 6) tends to be slightly over-scored (from +0.22 for grade 2 to a maximum of +0.53 for grade 5). The wrinkle at the corner of the lips presents the highest mean absolute deviations, yet still of low values: from a minimum of 0.43 points for grade 3 to a maximum of 0.91 for grade 6. Besides, the mean relative deviations reveal that grades below 1 tend to be under-scored on skin-warped average face images while being over-scored for grades of 6 and over.

A last important result is that the average of all mean absolute deviations between original and skin-warped average face images is always lower than the average of deviations between experts upon the scoring of original images: 0.41 and 0.47 points respectively for the forehead wrinkles, 0.59 and 0.61 for the nasolabial fold and 0.53 and 0.88 for the wrinkle at the corner of the lips.

3.2 | Comparison of original images and skin-warped average face images for the scoring of cheek characteristics

The analysis then focused on two facial skin characteristics that are best evaluated at the level of the cheek. They are the visibility of sebaceous pores and the homogeneity of skin pigmentation.

Whatever type of images scored, the overall coefficient of correlation between the grades given by the different experts is lower than those obtained for wrinkles, yet still acceptable (Table 2). On original images, they range from 0.760 to 0.693 for the visibility of sebaceous pores and from 0.755 to 0.702 for the homogeneity of skin pigmentation. Again, similar values are obtained when scoring is performed using skin-warped average face images and inter-expert correlations range from 0.824 to 0.682 for the visibility of sebaceous pores and from 0.786 to 0.731 for the homogeneity of skin pigmentation.

For both cheek characteristics, the intra-expert correlation obtained when comparing grades given upon scoring original and skin-warped average face images is almost always better than for inter-expert correlation on one image type (Table 1). Considering grades given by an expert (data not shown) or mean grades of the three experts, the graphs presenting the frequency of grades given by the expert also have very similar shapes (Figure 6). Statistical analyses of grade frequencies reveal no difference between the evaluations of original images and scoring performed on skin-warped average face images.

The difference in the grades given upon scoring original and skin-warped average face images show no differences in 45.3% of the cases for the visibility of pores and 50.9% of the cases for the homogeneity of pigmentation (Figure 7). Similar grading (± 1 grade difference) is achieved in 93.6% of the cases for the visibility of pores and in 93.6% for skin pigmentation.

The last point is that the scoring scales of the visibility of sebaceous pores and homogeneity of skin pigmentation present homogenous low mean absolute deviations (Figure 7): from 0.46 to 0.73 for pores and 0.46 to 1.09 points for skin pigmentation. For both, mean relative deviations indicate a tendency to under-score grades of 2 and below on skin-warped average face images, while higher grades tend to be slightly over-scored. Finally, for the two cheek features, the averages of all mean absolute deviations between both image types are always lower than the average of deviations between experts upon the scoring of original images. The average of all mean absolute deviations between both image types is 0.65 for the visibility of sebaceous pores and 0.60 for the homogeneity of skin pigmentation. They are respectively 0.79 and 1.02 for the average deviations between experts when using original images.

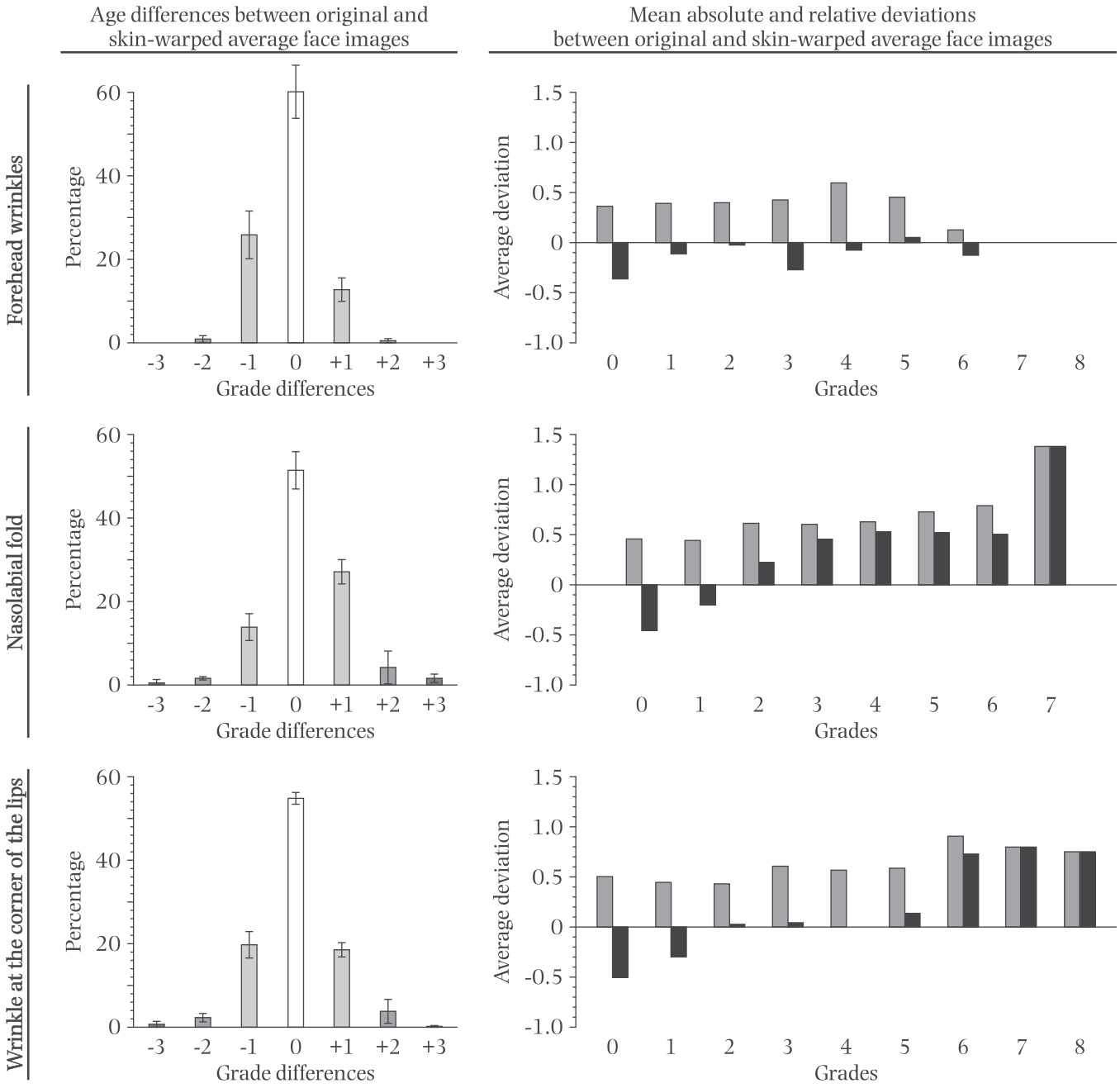

FIGURE 5 Grade differences and mean absolute (in grey) as well as mean relative (in black) deviations of the grades given upon wrinkle scoring of original and skin-warped average face images. For all analyses, original images were used as a reference. Grade differences are presented as the mean value ± standard deviation of the differences from the three experts.

3.3 | Comparison of original images and skin-warped average face images for the grading of perceived age

Lastly, the analysis focused on a complex characteristic involving several attributes of the entire face: perceived age. The median values of the age given by the two experts upon five randomly repeated scorings they performed for each subject on original images show a correlation coefficient of 0.765 (Table 3). This inter-expert correlation is 0.705 when perceived age is estimated using skin-warped average face images.

Focusing on intra-expert differences between the two image types, the coefficients of correlation of the mean age when using original and skin-warped average face images range from 0.813 to 0.842 (Table 3). These values are higher than the correlation coefficients obtained when comparing grades given by two experts on original or skin-warped average face images, correlations that are similar whatever image type considered. In addition, the overall shape of the

TABLE 2 Inter- and intra-expert correlation for the scoring of cheek characteristics on original and skin-warped average face images

	Original images			Skin-warped average face images			Original versus skin-warped average face images		
	Exp 1	Exp 2	Exp 3	Exp 1	Exp 2	Exp 3	Exp 1	Exp 2	Exp 3
Visibility of the sebaceous pores									
Exp 1	–	0.708	0.693	–	0.725	0.682	0.677	–	–
Exp 2	0.708	–	0.760	0.725	–	0.824	–	0.806	–
Exp 3	0.693	0.760	–	0.682	0.824	–	–	–	0.84,1
Homogeneity of skin pigmentation									
Exp 1	–	0.755	0.736	–	0.786	0.767	0.856	–	–
Exp 2	0.755	–	0.702	0.786	–	0.731	–	0.821	–
Exp 3	0.736	0.702	–	0.767	0.731	–	–	–	0.836

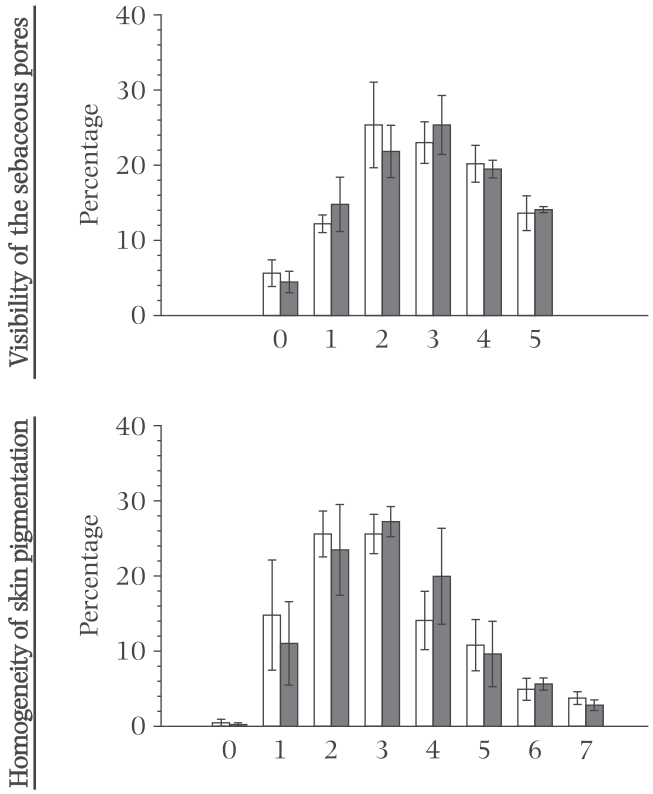

FIGURE 6 Mean grades given upon scoring of cheek characteristics on original (in white) and skin-warped average face (in dark grey) images. Results are presented as the mean value ± standard deviation of the grades given by the three experts.

graph presenting the mean frequency of perceived age groups by both experts is similar for original and skin-warped images (Figure 8) and no age group frequencies significantly differ. These results are not only true for mean perceived ages from the two experts but also when considering medians of each expert (data not shown).

In 73.2% (expert 1) to 80.3% (expert 2) of the cases (mean: 76.8%), perceived age presents a 1-year maximum difference when experts apprehended perceived age using original or skin-warped average face images (Figure 9). In 18.3% of the cases, the age differences range from more than a year to a maximum of 2 years. Larger differences are sparse despite a single instance of a 5-year difference (expert 2). Therefore, in 89.4% of the cases, experts give a similar perceived age (maximum age difference of 2 years) when grading original and skin-warped average face images.

Finally, mean absolute deviations (Figure 9) show higher values (up to 2.83 points) for the lower perceived age groups analysed (52-year-old and below). This relates to an under-scoring of skin-warped average face images compared to original images for these age groups, as shown by mean relative deviations. On the opposite, age groups of 56 and over present a tendency to be only slightly over-scored on skin-warped average face images (from 0.71 to a maximum of 1.00 points).

4 | DISCUSSION

The grades given by experts upon scoring skin wrinkles—forehead wrinkles, nasolabial fold, and wrinkle at the corner of the lip—and cheek features—pigmentation homogeneity and visibility of sebaceous pores—on original and skin-warped average face images present high correlations. Besides, the bar charts of grading scores are almost superposable, and grade distributions show no statistical differences. In more than 90% of the cases, a maximum grade difference of ±1 is observed when comparing evaluations performed on both image types, and 45% to 60% of the grades are identical. Mean differences are below 0.5 points for most grades, a low value. If mean deviations over 0.5 points are observed for the lowest or highest grades, these higher values should be considered with caution as they correspond to grades scored at a low frequency. Indeed, small differences between the grading of original images and skin-warped average face images have a large impact on mean deviations when only a few subjects are involved.

Similar results are achieved upon determining perceived age. As expected from such a complex and subjective evaluation, results show more variability than upon scoring wrinkles or cheek features. Yet, the correlation between perceived age scored on original and skin-warped average face images by each expert is similar to those of wrinkles and cheek features. Bar charts of perceived age present little difference whatever the image type used, and, in 90% of the cases, a similar age is

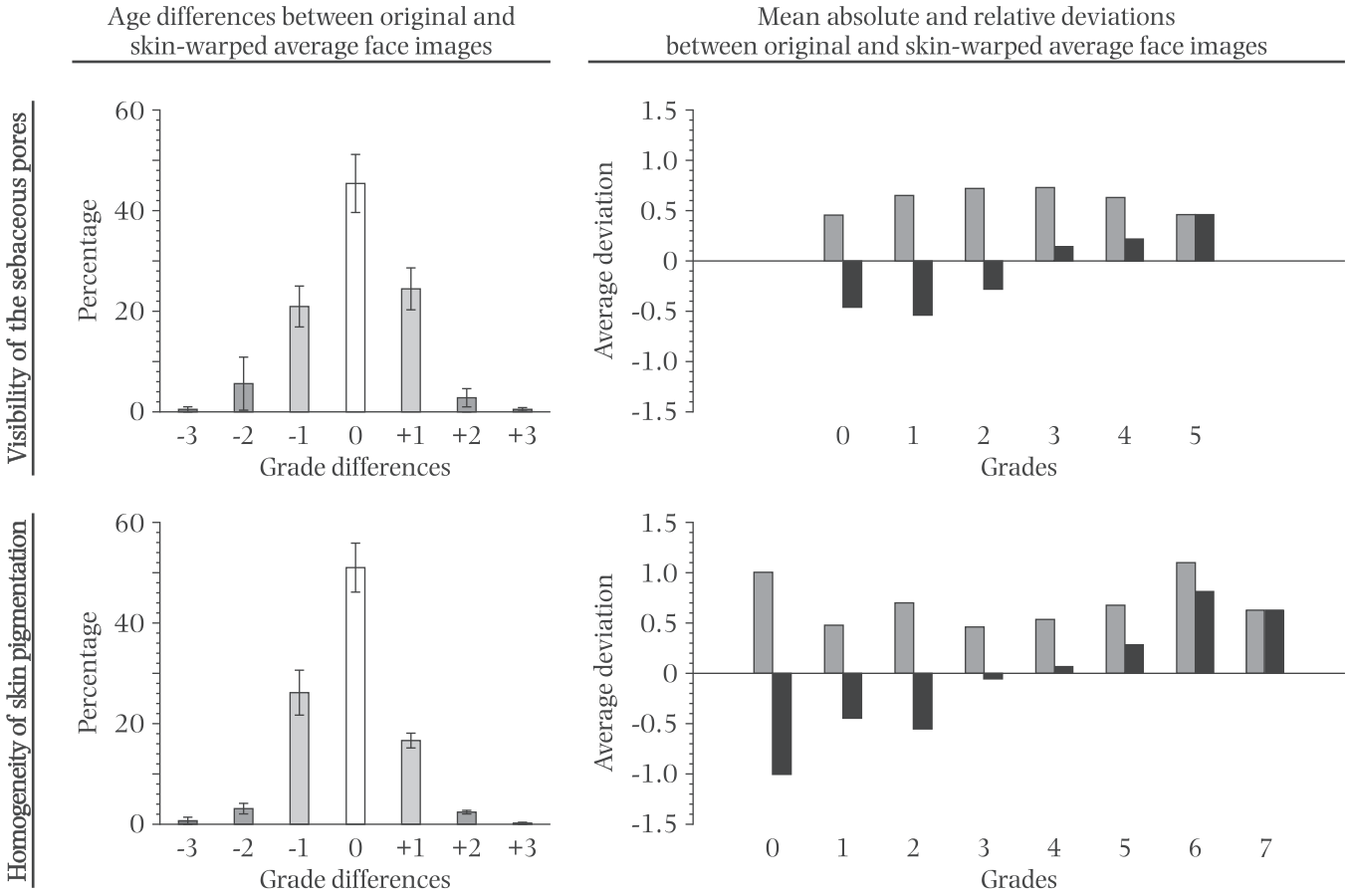

FIGURE 7 Grade differences and mean absolute (in grey) as well as mean relative (in black) deviations of the grades given upon scoring of cheek characteristics on original and skin-warped average face images. For all analyses, original images were used as a reference. Grade differences are presented as the mean value ± standard deviation of the differences from the three experts.

TABLE 3 Inter- and intra-expert correlation for the scoring of perceived age on original and skin-warped average face images

	Original images		Skin-warped average face images		Original versus skin-warped average face images	
	Exp 1	Exp 2	Exp 1	Exp 2	Exp 1	Exp 2
Perceived age						
Exp 1	–	0.765	–	0.705	0.842	–
Exp 2	0.765	–	0.705	–	–	0.813

given (±2 years difference). Furthermore, the mean absolute deviation between the scoring of original and skin-warped average face images is below 1.0 point. Yet, it can present some higher values for the younger age classes, which can be attributed to the higher variability induced by the small number of subjects estimated to be of these ages. This good agreement on perceived age between evaluations performed using original images and skin-warped average face images was not totally expected. Indeed, all skin-warped average face images present the same morphology. Thus, experts could not rely on the age-induced sagging of the lower face to perform their evaluation. Yet, ptosis is a major feature in estimating perceived age of Japanese women, for which it was shown to have a similar preponderant weight than wrinkles.[26] The 50- to 60-year-old age group selected for the study seems unlikely to have played a role in masking the relevance of ptosis since it still largely evolves at these ages.[19,20] If experts were guided in their evaluation by knowing the minimum and maximum age of subjects, this is not sufficient to fully explain the good results. Therefore, we can speculate that, as experts, they used other facial features to perform their evaluation and/or that the skin-warped average face still retains part of the morphological information as light and shadows that provided clues about the original morphological characteristics and helped them.

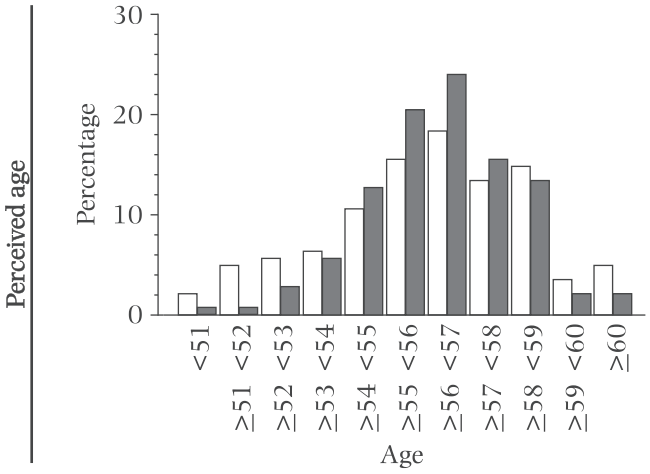

FIGURE 8 Mean perceived age groups upon scoring original (in white) and skin-warped average face (in dark grey) images. Results are presented as the mean value of the perceived age groups given by the two experts.

The individual skin images being warped onto the average face, the deformations they undergo are produced by the spline-based diffeomorphic deformation that is applied so that all characteristics morphological points match their average position. Construction of an average face using the average position of landmarks is not a new approach,[12–15,27] nor is picture warping according to landmarks.[28] Yet, the quality of the average face has benefited from technical improvements: automatic detection of morphological points, improved computer power and statistical analysis of each pixel of the average face. Furthermore, it is the first time that individual skin images are warped onto an average face to present the skin of a subject on faces having an identical morphology and made neutral by replacing the eyes, eyebrows and nostrils with those of the average face.

Collectively, results show that the distortions induced by the average face and the individual skin image warping processes have no impact on experts' perception of the analysed skin features. One such alteration is the previously mentioned alteration of individual facial contour. This is also the case for the region of the upper eyes and eyebrows. The original skin image is not applied to these regions, yet not altering the scoring of forehead wrinkles. Another region subjected to variation between original and skin-warped average face images is the nose, especially around the nostril wings. In this region, the original skin image is adjusted to the average face nose shape, which might induce some deformations. Again, these alterations did not impact the scoring of the visibility of sebaceous pores, the homogeneity of skin pigmentation or the severity of the nasolabial fold. Finally, the warping process must have deformed the wrinkles and other facial features we scored. If their shape and size were modified, their colour and the shade they induce were maintained. Furthermore, their proportions remain very similar between original and skin-warped average face. Nevertheless, having a comprehensive view of the potential of skin-warped average face images and of its limits will require additional studies. It will be essential to determine if other facial characteristics from regions close to those undergoing distortions can be faithfully rated. This should include the tear through, the upper eyelids and the eyelashes surroundings, especially for scoring dark circles. The region close to the lips might also need to be scrutinised in more detail as individual peri-oral skin is warped around the average lips.

Except for perceived age, the inter-expert correlation coefficients of all other features graded are almost always identical whether the grading is carried out on original or skin-warped average face images. What is striking is that the inter-expert variability is generally higher than the variability of the evaluation of a given expert on both image types. This is not only shown by the analysis of correlations but also by the average of mean absolute deviations that are always lower than the inter-expert deviations when grading is performed on original images. The inter-expert variability is well-documented, including upon formal validation of some structured photographic scales.[7–9] The variability between grading original and skin-warped average face images being lower than inter-expert variability shows that skin-warped average face images can be trustfully used to grade skin features.

This work is a first study of skin-warped average face images and the reliability of this approach. If skin-warped average face images are more standardised than original images of subjects, results do not show that this standardisation decreases inter-expert rating variability. Indeed, this variability is generally attributed to the difficulty experts can be confronted with when having to choose between mid-scores, forcing them to select the score below or over. Judgment subjectivity and differences in grading experience and/or training are also invoked. If, obviously, skin-warped average face images cannot change these sources of variability, our study was poorly adapted to determine if the standardisation of subjects' images could reduce inter-expert variability. Additional works involving more experts and analysing reproducibility might answer this question. It will also be essential to validate formally the skin-warped average face image approach that can be adapted to the scoring of facial profile attributes.

In conclusion, although one should be careful not to conclude too quickly, as this work only relies on a small number of experts, the comparative analysis of gradings of facial skin characteristics performed on skin-warped average face images revealed no significant differences compared to scorings carried out on original images of subjects. The difference between grading original and skin-warped average face is even lower than inter-expert variability. These results open the possibility of using this new approach to monitor skin changes, assess the effects of cosmetics and present actual results on a poorly recognisable face to support claim substantiation. Whether these changes will be recognisable by laypeople is another question that remains to be answered. Yet, it will be interesting as skin-warped average face images also open up the possibility of analysing facial images with the very same morphology but differing only by their skin characteristics, which could be of interest for other applications.

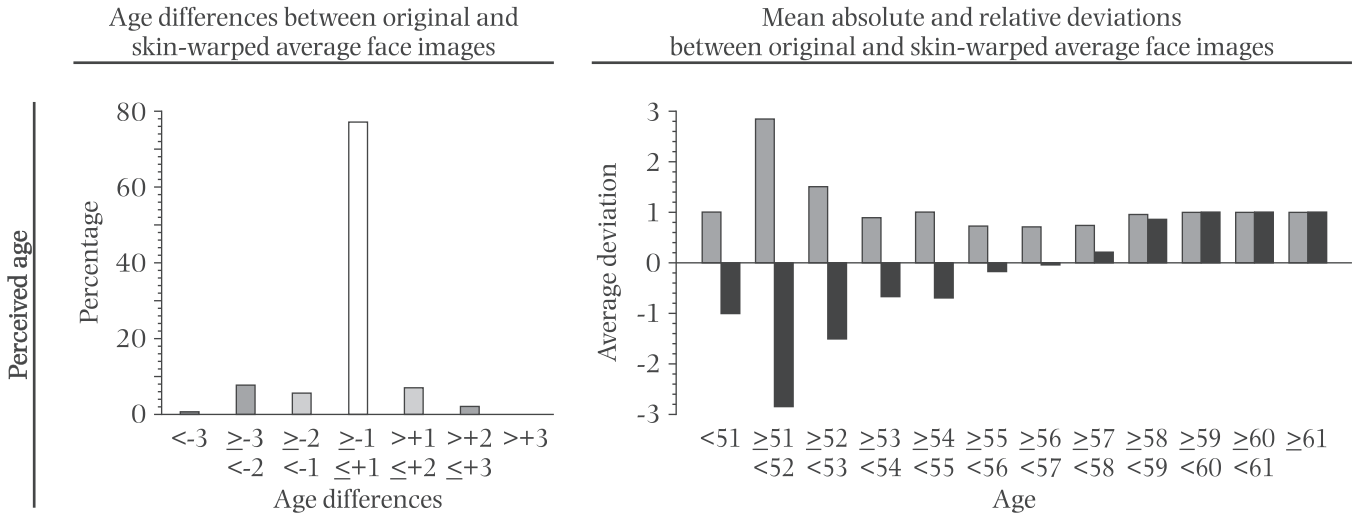

FIGURE 9 Perceived age differences and mean absolute (in grey) as well as mean relative (in black) deviations of the perceived age group upon scoring of original and skin-warped average face images. For all analyses, original images were used as a reference. Perceived age differences are presented as the mean value ± standard deviation of the differences from the three experts.

ACKNOWLEDGEMENTS

The authors wish to thank Philippe Crouzet, PhD, from Estium-Concept, for scientific writing services. This study was entirely funded by Newtone Technologies, Lyon, France.

CONFLICT OF INTEREST STATEMENT

QM, MC and EPM are full-time employees of Newtone Technologies, a company specialised in the design of innovative solutions for skin imaging and analysis. TH is a full-time employee of DRC Co., a company dedicated to the testing of cosmetic products. YO is a full-time employee of CIEL Co. Ltd., a company dedicated to the testing of cosmetic products.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request.

REFERENCES

1. Griffiths CE, Wang TS, Hamilton TA, et al. A photonumeric scale or the assessment of cutaneous photodamage. *Arch Dermatol*. 1992;128:347-351.
2. Bazin R. Skin Aging Atlas. Volume 1, Caucasian Type. *Med'com*. 2007.
3. Carruthers A, Carruthers J. A validated facial grading scale: the future of facial ageing measurement tools? *J Cosmet Laser Ther*. 2010;12:235-241.
4. Bazin R, Flament F. Skin Aging Atlas. Volume 2, Asian Type. *Med'com*. 2010.
5. Bernois A, Huber A, Derome C, et al. A photographic scale for the evaluation of facial skin aging in Indian women. *Eur J Dermatol*. 2011;21:700-704.
6. Bazin R, Flament F, Giron F. Skin Aging Atlas. Volume 3, Afro-American Type. *Med'com*. 2012.
7. Jdid R, Latreille J, Soppelsa F, Tschachler E, Morizot F. Validation of digital photographic reference scales for evaluating facial aging signs. *Skin Res Technol*. 2018;24:196-202.
8. Buchner L, Vamvakias G, Rom D. Validation of a photonumeric wrinkle assessment scale for assessing nasolabial fold wrinkles. *Plast Reconstr Surg*. 2010;126:596-601.
9. Rhoda S, Narins RS, Carruthers J, et al. Validated assessment scales for the lower face. *Dermatol Surg*. 2012;38:333-342.
10. Conkling N, Bishawi M, Phillips BT, Bui DT, Khan SU, Dagum AB. Subjective rating of cosmetic treatment with botulinum toxin type A: do existing measures demonstrate interobserver validity? *Ann Plast Surg*. 2012;69:350-355.
11. Galton F. Composite Portraits. *J Anthropolog Inst G B and Irel*. 1878;8:132-142.
12. Langlois JH, Roggman LA. Attractive faces are only average. *Psychol Sci*. 1990;1:115-121.
13. Beier T, Neely S. Feature-based image metamorphosis. Paper presented at: *SIGGRAPH'92 Proceedings*. Chicago, USA, July 26-31, 1992:35-42.
14. Rhodes G, Sumich A, Byatt G. Are average facial configuration attractive only because of their symmetry? *Psychol Sci*. 1999;10:52-58.
15. Fink B, Grammer K, Thornhill R. Human (*Homo sapiens*) facial attractiveness in relation to skin texture and color. *J Comp Psychol*. 2001;115:92-99.
16. Porcheron A, Mauger E, Soppelsa F, et al. Facial contrast is a cross-cultural cue for perceiving age. *Front Psychol*. 2017;8:1208.
17. Machard A, Jomier M, Hottelart D, Vié K. Identification of new morphological differences between Chinese and Caucasian faces and influence of BMI on these characteristics. *Skin Res Technol*. 2016;22:137-147.
18. Séroul P, Campiche R, Gougeon S, Cherel M, Rawlings AV, Voegeli R. An image-based mapping of significance and relevance of facial skin colour changes of females living in Thailand. *Int J Cosmet Sci*. 2020;42:99-107.
19. Tsukahara K, Fujimura T, Yoshida Y, et al. Comparison of age-related changes in wrinkling and sagging of the skin in Caucasian females and in Japanese females. *J Cosmet Sci*. 2004;55:351-371.
20. Tsukahara K, Sugata K, Osanai O, et al. Comparison of age-related changes in facial wrinkles and sagging in the skin of Japanese, Chinese and Thai women. *J Dermatol Sci*. 2007;47:19-28
21. Dihua X, Seong-Whan L. Face detection and facial feature extraction using support vector machines. Paper presented at: Proceedings of the

16th International Conference on Pattern Recognition. Quebec City, Canada, August 11-15, 2002;4,209-212.
22. Guo ZQ, Yang J. A new method to extract face features based on combination of mean face SVD and KFDA. Paper presented at: International Conference on Biomedical Engineering and Computer Science (ICBECS). Wuhan, China, April 23-25, 2010.
23. Joshi S, Davis B, Jomier M, Gerig G. Unbiased diffeomorphic atlas construction for computational anatomy. *Neuroimage*. 2004;23:S151-S160.
24. Li B, Christensen GE, Hoffman EA, McLennan G, Reinhardt JM. Establishing a normative atlas of the human lung: computing the average transformation and atlas construction. *Acad Radiol*. 2012;19:1368-1381.
25. Fletcher PT, Venkatasubramanian S, Joshi S. The geometric median on Riemannian manifolds with application to robust atlas estimation. *Neuroimage*. 2009;45:S143-S152.
26. Flament F, Abric A, Adam AS. Evaluating the respective weights of some facial signs on perceived ages in differently aged women of five ethnic origins. *J Cosmet Dermatol*. 2021;20:842-853.
27. Rhodes G, Yoshikawa S, Clark A, Lee K, McKay R, Akamatsu S. Attractiveness of facial averageness and symmetry in non-western cultures: in search of biologically based standards of beauty. *Perception*. 2001;30:611-625.
28. Fink B, Grammer K, Matts PJ. Visible skin color distribution plays a role in the perception of age, attractiveness, and health in female faces. *Evol Hum Behav*. 2006;27:433-442.

How to cite this article: Marin Q, Honda T, Okano Y, Cherel M, Prestat-Marquis E. Comparative analysis of original face and skin-warped average face images for the scoring of skin attributes. *Skin Res Technol*. 2023;29:1–13.
https://doi.org/10.1111/srt.13324