Shape Analysis Approach towards Assessment of Cleft Lip Repair Outcome*

Paul Bakaki[1,4][0000−0001−8277−2554], Bruce Richard[2][0000−0002−4712−2860], Ella Pereira[1][0000−0002−6013−3935], Aristides Tagalakis[1][0000−0002−4610−0803], Andy Ness[3][0000−0003−3548−9523], and Yonghuai Liu[1][0000−0002−3774−2134]

[1] Faculty of Arts and Science, Edge Hill University, Lancashire L39 4QP, UK
{bakakip, pereirae, Aristides.Tagalakis, yonghuai.liu}@edgehill.ac.uk
[2] Birmingham Children's Hospital, Steelhouse Lane Birmingham B4 6NH, UK
brucerichard@blueyonder.co.uk
[3] British Dental School, University of Bristol, Bristol BS1 2LY, UK
Andy.Ness@bristol.ac.uk
[4] Department of Computer Science, Makerere University,
P.O. Box 7062, Kampala, Uganda

Abstract. Current methods of assessing the quality of a surgically repaired cleft lip rely on humans scoring photographs. This is only practical for research purposes due to the
resources necessary and is not used in routine audit. It has poor validity due to human subjectivity and thus low inter-rater reliability. An automatic method for aesthetic outcome assessment of cleft lip repair is required. The appearance and shape of the lips constitute the region of interest for analysis. The mouth borderline and corner points are detected using a bilateral semantic network for real-time segmentation. The bisector of the line linking the mouth corners is estimated as the vertical symmetric axis. By splitting the mouth blob into two parts, they are analyzed for similarity and a numeric score ranging from 1 to 5 is then generated. Pearson correlation coefficient between automatically generated scores and human-assigned ones serves as a validation metric. A correlation of about 40% indicates a good agreement between human and computer-based assessments. However, better automatic scoring correlation of 95.9% exists between the automatically detected mouth regions and those manually drawn by human experts, the third ground truth set in scenario two. Our method has the potential to automate an outcome estimation of the aesthetics of cleft lip repair with human bias reduced, easy implementation and computational efficiency.

Keywords: Cleft Lip, Aesthetic Assessment, Segmentation, Symmetry, Structural Similarity, Correlation coefficient.

1 Introduction

Cleft lip (CL) is one of the most common maxillofacial congenital deformities affecting about 1 in 500 Asians, 1 in 1000 Caucasians and 1 in 2500 Africans [1].

* Supported by Graduate Teaching Assistantship, Edge Hill University.

Children with this condition face socio-economic challenges, including high costs for treatment and care (specialized feeding bottles and multiple surgeries), social integration with speech, hearing, and dental problems and potential rejection due to poor facial appearance [2].

Treatment of cleft lip and nose deformity is by surgical repair, usually when the child is 3 to 6 months old and again in early childhood and adolescence to revise or improve the facial appearance [3][4]. An attractive, symmetric and normal appearance of the lip repair is a primary purpose, since people with symmetric faces are more socially acceptable, more confident and have better educational and employment opportunities in life [5]. Current audit practice is to take standard 2D colour photographs to allow an evaluation of the aesthetic appearance of the lip when the child is five years old. But in practice, they are rarely evaluated unless the child is in a research project. Whilst other outcome measures such as mid facial growth, speech, hearing, dental and psychological well-being, which all have internationally accepted and validated outcome measures which are used for audit and research. If there was a validated, efficient outcome measure for the appearance of the lip, it would allow an effective evaluation of the surgical result, be a tool for comparisons of the techniques and protocols of care, and patient/parent satisfaction.

Predominantly, outcome assessment following CL repair is done through qualitative analysis of facial images of the patient. Whilst lip closure is necessary for normal eating, drinking and speaking, the facial beauty aspect is also a primary outcome of the procedure, and is referred to as facial aesthetics (facial appearance). Aesthetic outcome assessment is a research field, that has attracted attention because it has few objective measures. The different approaches for aesthetic outcome assessment are largely indirect in nature, although direct clinical assessment through physical expert observation of the patients is also possible.

Experts create a score of the facial aesthetics based on visualization of images presented to them, either as hard copies, projected on a screen or increasingly through a digital platform. This results in a descriptive qualitative assessment. The Asher-McDade method uses a five-point Likert scale [6] and has been widely used internationally. The image is described as either "Excellent", "Very Good" "Good", "Fair" or "Poor" as each individual expert or lay person may decide. A semi-automatic method, SymNose, was developed to improve objective scoring in [7]. Analyse It Doc (A.I.D.) [8] is an analysis software with modules for subjective and objective assessment/evaluation of aesthetic outcomes. These approaches are still subjective and rely on an emotive interpretation of what is "good" by different human subjects [9].

Given the advancement of computer vision and deep learning technology, this study advances the notion that minimizes human involvement in aesthetic outcome assessment, it will increase the objectivity and validity of any score derived [10]. This study leverages on the fact that digital aesthetic images contain a lot of useful information that can be used in aesthetic assessment research. Such information can be extracted and analyzed to support automatic aesthetic outcome assessment. This study proposes an automatic approach for aesthetic

outcome assessment following CL surgery, based on low level features of the lips and mouth region. Our approach uses lip aesthetic assessment method based on the mouth boundary following successful lips segmentation, proven by ground truth.

2 Method

The method has the following main components/steps in the pipeline: mouth detection, symmetrical axis determination, similarity measurement, and numerical score estimation. Mouth detection is vital for clear determination of the visual features of lips, vermilion lines and mouth corners within a given image.

2.1 Dataset and tools

The data set has 4 classes of 25 facial images, which have been anonymized for ethical reasons to reveal only the nose and mouth/lips. In addition, it was also intended that human assessors are not biased by any other facial features. The first class of 25 images constitute the raw data for aesthetic assessment. The other 3 classes (dubbed as GT1, GT2, and GT3) are ground truth (GT) whose mouth/lip region boundary was already manually drawn by three different human experts respectively using the open source ImageJ software. The 3 ground truths serve as validation for the segmentation approach and the assessment prediction mechanism discussed in this paper. Human numeric scores (HNS) were generated through a subjective aesthetic assessment process aided by statistical coding of assessor's description of the individual images in the raw dataset.

Using our method, all the images of the 4 classes are automatically assessed and a numeric score is then generated. These scores are named AENS, short form for automatically estimated numeric score, with a name prefix of the respective data set, for example, GT1-AENS, and so forth.

The implementation programming language is Python 3.7. The supporting libraries are OpenCV, Matplotlib, PyTorch and Keras.

2.2 Feature Description and Detection

All the images have been anonymized for ethical and other reasons stated previously, implying that some facial features are not available for detection, and thus only limited features can be identified. Our focus is on the features of the mouth region, starting with segmentation. The anatomy of the human mouth region consists of the following key parts: the vermillion border (upper and lower), oral commissures (left and right) and the philtra ridges (left and right, separated by philtrum) [11].

Ideals of facial beauty indicate that the mouth region should be in the lower third of a given facial image [4]. Because the skin color and the lips may be indistinguishable, contrast enhancement and selection of suitable color transform is inevitable. To mitigate this, the segmentation method we used considers the

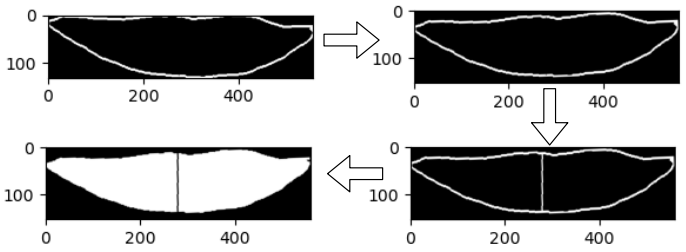

Fig. 1. An example for boundary extraction, rotation, and symmetry axis detection of a cropped mouth lip image. Top row - left: mouth corners are at different elevations from the horizontal axis. Top row - right: After anticlockwise rotation mouth corners are at the same elevation. Bottom row shows the symmetric axis (black and white).

semantics of individual pixels, first discussed in 1987 [12]. While traditional techniques which perform segmentation as a binarization task usually underperform at medical imagery analysis tasks [13], the deep learning based semantic segmentation method [14] has been employed in this study. Even so, residues such as scars, open mouth and runny nose still influence the segmentation outcome.

Semantic segmentation enhances edge detection by creating a sharper contrast between the surrounding skin and the mouth region, hence facilitating shape identification and feature extraction. The ideal mouth region mainly consists of soft tissue features defined below.

- PR_L and PR_R are philtra ridges identified as one of the upper most extreme pixels on the left-hand and right-hand sides of the philtrum, found along the mouth boundary.
- OC_L and OC_R are the left-hand and right-hand side mouth corners identified as the most extreme pixels on the left-hand and right-hand sides, located along the mouth boundary.
- VB_U and VB_B are a list of pixels constituting the upper and lower mouth region boundaries, stretching between OC_L and OC_R.

$$VB_U = \{u_1, u_2, ..., u_n\} \tag{1}$$
$$VB_B = \{b_1, b_2, ..., b_m\} \tag{2}$$

  where u_i and b_j are pixels in a given $2D$ grayscale image I.
- The mouth boundary B is a combined list of VB_B and VB_U. Collectively, it is also known as the longest non-nested detected contour in the face, represented in Eq. 3 as

$$B = VB_U \cup VB_B \tag{3}$$

  where $VB_U \cap VB_B = \{OC_L, OC_R\}$, $PR_L, PR_R \subset VB_U$ and $OC_L, OC_R \in B$.
- The linking line of OC_L and OC_R is not always parallel to the horizontal axis of the image. Its orientation angle θ to the horizontal axis dictates the magnitude of the rotational transformation (Fig. 1). if $\theta < 0$, rotate *anticlockwise*; otherwise, rotate *clockwise*. Such orientation may influence

how human subjects visualise and assign the numerical score to a given facial image, to be investigated in Section 3 below.

2.3 Symmetrical axis detection and measurement

Several approaches have been previously used in general detection of symmetry. Related methods are discussed in [10]. However, those techniques utilized many more local and invariant object features with higher contrasts. This study utilizes basic lip and mouth features instead, similar to the perception of human assessors. The midpoint D given in Eq. 4 is a position where the vertical symmetric axis is plotted through in the image plane.

$$D = (OC_L + OC_R)/2. \tag{4}$$

A vertical straight line plotted through D and crossing the lower and upper mouth boundaries ensures slicing the mouth region into two shapes, left-side shape, sh_l and right-side shape, sh_r. The evenness or variance is computed and categorized using the structural similarity measure, S [15]. S is an aggregated rational number ranging between -1 and 1 for color images or 0 and 1 for binary images. We consider sh_l and sh_r as independent and unique shapes over which to compute S. S is an aggregate of luminance l, contrast c, and structure s, adapted from [15] and expressed in Eq. 5 below as:

$$S(sh_l, sh_r) = [l(sh_l, sh_r)^{\alpha} \cdot c(sh_l, sh_r)^{\beta} \cdot s(sh_l, sh_r)^{\gamma}] \tag{5}$$

where $\alpha = 1$, $\beta = 1$ and $\gamma = 1$ for easy implementation. Since the dimensions of sh_l and sh_r should be similar, sh_r is vertically flipped along the vertically plotted symmetric axis. Setting the different statistical parameters of l, c and s as described in [15] gives the usable form of the parameter S in Eq. 6 below as:

$$S(sh_l, sh_r) = \frac{(2\mu_{sh_l}\mu_{sh_r} + C_1)(2\sigma_{sh_l sh_r} + C_2)}{(\mu_{sh_l}^2 + \mu_{sh_r}^2 + C_1)(\sigma_{sh_l}^2 + \sigma_{sh_r}^2 + C_2)} \tag{6}$$

where μ_{sh_l}, σ_{sh_l}, μ_{sh_r}, and σ_{sh_r} are the mean and standard deviation of pixels in shapes sh_l and sh_r respectively, $\sigma_{sh_l sh_r}$ is the standard deviation of the pixels in sh_l and sh_r, $C_1 = (k_1 L)^2$, $C_2 = (k_2 L)^2$, $k_1 = 0.01$, $k_2 = 0.03$, $L = 2^p - 1$ and p is the number of bits per pixel.

2.4 Conversion of similarity measure to a numeric score

The structural similarity S is computed and converted to a numeric score in the range of 1 and 5, where 1 = "excellent", 2 = "Very good", 3 = "Good", 4 = "Fair" and 5 "Poor". The transformation $f(S)$ should fulfill the following boundary and monotonicity conditions: $f(0) = 5$, $f(1) = 1$, and $f(S)$ is monotonically decreasing. Therefore, $f(S)$ is thus the finally $AENS$. The following three models

(Eqs. 7, 8 and 9) are designed and selected for a comparative study about what relationship is between S and $AENS$.

$$f(S) = 5 - 4S \tag{7}$$
$$f(S) = 5 - 4S^3 \tag{8}$$
$$f(S) = 1/(0.2 + 0.8S^2) \tag{9}$$

Three scenarios are also considered in Fig. 2 for the generation of sh_l and sh_r for further comparative studies how the two shapes should be defined:

- Scenario 1: Parameters calculated over the entire mouth blob.
- Scenario 2: Parameters calculated over the entire mouth boundary only.
- Scenario 3: Parameters calculated over the upper lip blob only.

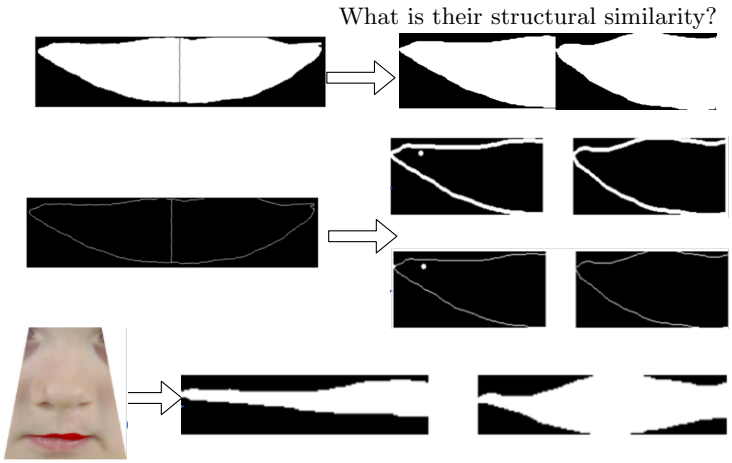

Fig. 2. Different scenarios for parameter calculation. Top: entire mouth region blob (upper and lower lips) has been split into right and left blobs, sh_l and sh_r. sh_r has been flipped. Middle: Scenario 2 with the boundaries defined with different thicknesses of 1 and 3 pixels respectively; Bottom: Scenario 3.

3 Experimental results

In this section, we present both the qualitative and quantitative experimental results of the proposed automated programmed rating (PR) method compared with the others when applicable.

3.1 Image segmentation

Facial images were segmented using the bilateral real-time semantic network (SN) [14]. Traditional approaches such as threshold-based segmentation (such

as Otsu and moment preservation) and clustering method (K-means (KM) and mean shift (MS)) usually produce unsatisfactory results. A comparative study between the MS (spatial bandwidth=20, color bandwidth=7), KM (k=3) and SN is presented in Fig 3 where the performance is measured in F1-score in percentage: the higher the better. Clustering-based approaches yielded worse

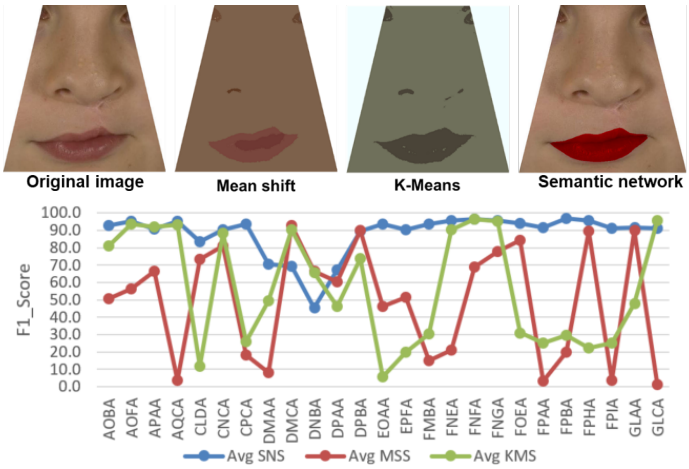

Fig. 3. Segmentation results of images using different techniques.

outputs with discontinuous areas and boundaries. Gaussian blurring, morphing and dilation were usually used to mitigate such issues. The segmented mouth region (our RoI) is found in the bottom third of the facial image. Standardization with a bounding box was also used to reduce the background from the image as seen in Fig. 1.

3.2 Evaluation of Aesthetic Assessment

After computing S based on Scenarios 1, 2, and 3 and converting it to an aesthetic numeric score, our method was evaluated using Pearson correlation coefficient against HNS: the higher the better. Table 1 shows that the orientation standardization is helpful to improve the $AENS$ using the bounding box. This shows that the mouth orientation may affect how the human subjects perceive and thus assign numerical scores to given images.

The performance metrics of our method are presented in Fig. 4 over 3 scenarios, 3 models, and 2 options of the symmetric axis crossing position, D and D_2:

$$D_2 = d(OC_L + OC_R)/2 \tag{10}$$

where shift factor d by 5% inward being most effective, considering that the mouth corners may not be accurately detected due to imaging noise and shadows.

Table 1. AENS before and after standardization of mouth orientation in Scenario 1.

Category	Range		Average	
	before	after	before	after
PR	$0.24 < S < 0.82$	$0.55 < S < 0.89$	0.60	0.79
GT1	$0.30 < S < 0.84$	$0.49 < S < 0.83$	0.60	0.72
GT2	$0.28 < S < 0.84$	$0.39 < S < 0.86$	0.60	0.69
GT3	$0.35 < S < 0.82$	$0.51 < S < 0.88$	0.64	0.72

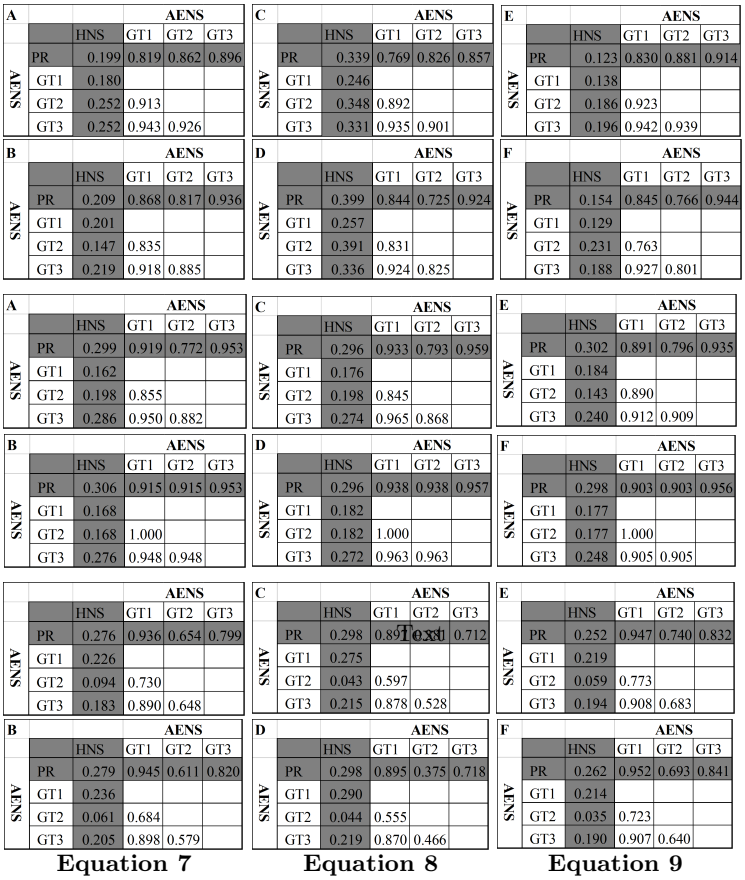

Fig. 4. Correlation coefficient results for different scenarios over different transformation models. Odd row: symmetric axis plotted at D; Even row: symmetric axis plotted at D_2. Top two rows: Scenario 1; Middle two rows: Scenario 2; Bottom two rows: Scenario 3.

It can be seen that the correlation between HNS and $PR - AENS$ is considered most significant because it is a test directly made between human and fully automated computer-based assessment. The highest score is about 40% in Fig 4 *Table D* in Scenario 1 and the lowest is about 15% in *Table F*, due to

inconsistency for human subjects to assign scores from one image to another. Overall, shifting the mouth corners inward improves the most significant correlation across the three models. However, the model in Eq. 8 is the most robust, implying that the mapping from shape similarity measurements to their aesthetic scores is non-linear. In sharp contrast, correlation between $PR-AENS$ and either GT_1-AENS, GT_2-AENS or GT_3-AENS is significantly higher, as high as 94% in $Table\ F$ in Scenario 1. This implies that the automatic segmentation of the mouth regions is accurate, compared to human manually drawn ones.

In Scenario 2, the most significant correlation is about 31%, $Table\ B$. There is little difference in the various correlations over different setups, indicating that the mouth boundaries may not be as predictive as expected. This is somewhat contradictory to the practice that focuses on the vermilion lines and thus requires further investigation. Scenario 3 has produced the lowest correlation value in the category of $PR-AENS$ and either GT_1-AENS, GT_2-AENS or GT_3-AENS on record of as low as 38% in $Table\ D$. This indicates that the determination of the RoI is still challenging.

However, determining the symmetric axis using fewer features is a potential limitation of the proposed method, future research studies utilizing deep learning techniques such as transfer learning will target improving results. Additionally, the benchmark for the validity of our approach is based on a single method, spearheaded by human experts.

4 Conclusion

This paper proposed an automatic assessment approach that utilizes lips and mouth features. These features are considered appealing to humans and can be distinguishable to support aesthetics judgement. These include oral commissures and the vermillion border. Once the mouth region has been detected using the bilateral network segmentation method and split through the midpoint of the horizontal line linking the mouth corners, the two ensued blobs are analyzed for evenness or difference. To this end, the widely used structural similarity measure [15] was employed. The measure is a rational number, that was then converted non-linearly to a numeric score in the range of 1 and 5, like the Asher-McDade five-point Likert Scale used by human experts. A numerical similarity computation following a symmetric axis computation is a better objective aesthetics assessment of the repaired lips compared to the qualitative measures proposed in [7], [8] and [9]. The experimental results show that the automatically estimated scores have relatively low correlation coefficients with human assigned ones but have high correlation coefficients with those estimated from the human manually drawn mouth regions.

It is also noted that inward shift of the mouth corners by 5% improves the accuracy of the midpoint D_2 and offers an alternative for a symmetric axis position to combat the challenging nature in identifying the mouth corners with improved aesthetics assessment scores. Further research will investigate more

accurate estimation of the symmetrical axis and difference measurement between the two sides of the mouth regions.

Acknowledgments

The facial images are the cropped and anonymised anteroposterior (A/P) photos of 5-year-old children from the Cleft Care UK (CCUK). This publication presents data derived from the Cleft Care UK Resource (an independent study funded by the National Institute for Health Research (NIHR) under its Programme Grants for Applied Research scheme RP-PG-0707-10034).

References

1. Zhang, Q., et al: A Bibliometric Analysis of Cleft Lip and Palate-Related Publication Trends From 2000 to 2017. Cleft Palate-Craniofacial J. 56, 658–669 (2019).
2. Shkoukani, M.A., Chen, M., Vong, A.: Cleft lip - A comprehensive review. Front. Pediatr. 1, 1–10 (2013).
3. Mosmuller, D.G.M., et al: Scoring systems of cleft-related facial deformities: A review of literature. Cleft Palate-Craniofacial J. 50, 286–296 (2013).
4. Kar, M., Muluk, N.B., Bafaqeeh, S.A., Cingi, C.: È Possibile Definire Le Labbra Ideali? Acta Otorhinolaryngol. Ital. 38, 67–72 (2018).
5. Little, A.C., Jones, B.C., Debruine, L.M.: Facial attractiveness: Evolutionary based research. Philos. Trans. R. Soc. B Biol. Sci. 366, 1638–1659 (2011).
6. Asher-McDade, C., Roberts, C., Shaw, W.C., Gallager, C.: Development of a method for rating nasolabial appearance in patients with clefts of the lip and palate, Cleft Palate Craniofac J 28(4), 385-390 (1991).
7. Piggot, R.W. and Piggot, B.B.: Quantitative measurement of symmetry from photographs following surgery for unilateral cleft lip and palate. Cleft Palate-Craniofacial Journal. 47 (4), 363–367 (2010).
8. Pietrski, P., Majak, M., and Antoszewski, B.: Clinically Oriented Software for Facial Symmetry, Morphology, and Aesthetic Analysis. Aesthetic surgery journal. 38 (1) NP19–NP22 (2017).
9. Deall, C.E., et al: Facial Aesthetic Outcomes of Cleft Surgery: Assessment of Discrete Lip and Nose Images Compared with Digital Symmetry Analysis. Plast. Reconstr. Surg. 138, 855–862 (2016).
10. Deng, Y., Loy, C.C., Tang, X.: Image Aesthetic Assessment: An experimental survey. IEEE Signal Process. Mag. 34, 80–106 (2017).
11. Berlin, N.F., et al: Quantification of facial asymmetry by 2D analysis - A comparison of recent approaches. J. Cranio-Maxillofacial Surg. 42, 265–271 (2014).
12. Chen, S., Leung H.: Chaotic spread spectrum watermarking for remote sensing images. J. Electron. Imaging. 13(1) (2004).
13. Kuruvilla, J., et al: A review on image processing and image segmentation. Proc. 2016 Int. Conf. Data Min. Adv. Comput. SAPIENCE 2016. 198–203 (2016).
14. Yu, C., et al: BiSeNet: Bilateral segmentation network for real-time semantic segmentation. Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics). 11217 LNCS, 334–349 (2018).
15. Wang, Z., et al: Image quality assessment: From error visibility to structural similarity. IEEE Trans. Image Process. 13, 600–612 (2004).