https://doi.org/10.2352/ISSN.2470-1173.2018.14.HVEI-513

Assessing gloss perception of human facial skin across subject

Jing Wang, Thrasyvoulos N. Pappas; Northwestern University; Evanston, IL/US
Carla Kuesten, Gopa Majmudar, Jim Mayne; Amway; Ada, MI/US

Abstract

We propose novel techniques for the evaluation of perceived facial gloss across subjects with varying surface reflections. Given a database of facial skin images from multiple subjects, ordered according to perceived gloss within each subject, we propose a head-tail (least and most glossy image of each subject) selective comparison approach for ordering the entire database. We conducted a two-alternative forced-choice empirical study to compare the facial gloss across subjects within each group. Using the gloss scores of selected candidates and the gloss range of a reference subject, we fit each within-subject gloss range to a global gloss range and quantized the scores into distinct gloss levels. We then conducted another empirical study to validate the quantized gloss levels. The results show that in 90% of the cases, the levels are consistent with human judgments. Based on the database with quantized gloss levels, we develop a max-margin learning model for facial skin gloss estimation. The model relies on gloss related statistics extracted from surface and subsurface reflection images obtained using multimodal photography. The predicted gloss level is decided by the nearest neighbors using the learned scoring function. Performance tests demonstrate that the best performance, with 82% accuracy, is obtained when we combine local statistics from both surface and subsurface reflections.

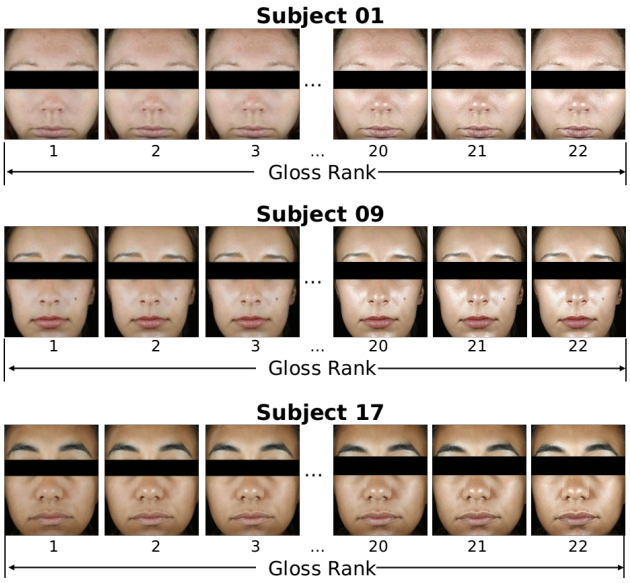


Figure 1: Examples of images with within-subject gloss rankings (Eye regions are masked to protect personal identities). Note that the same ranking of different subjects does not guarantee the same gloss level.

Introduction

Human facial skin gloss is important in multiple applications like image rendering, skin condition estimation, and skin-care product evaluation [1]. Previous gloss studies on synthetic materials found that perceived gloss is affected by the lighting conditions [2–4], object geometry [5–7], and object color [8]. However, the findings relying on simplified synthetic stimuli cannot be directly applied to facial gloss perception as facial appearance involves complicated contextual information. On the other hand, previous studies on facial skin gloss perception have mainly focused on the visual difference between two conditions (before and after makeup, before and after cleansing) within each subject [9, 10]. However, such within-subject perception is not adequate for developing a global visual gloss range across subjects. Facial appearance across subjects involves larger variances in the skin intrinsic condition, such as complexion, surface texture, face shape, and the production of sebum.

The focus of this paper is on analyzing facial gloss perception across subjects and building an efficient model for quantitative estimation of skin gloss level. Our empirical study and model development utilize a facial skin gloss image dataset labeled with existing within-subject gloss ranking scores [10]. The dataset contains facial images of 25 subjects with wide variations in surface reflection (specular reflection). Figure 1 shows example images for three subjects. The images of each subject have been ranked in increasing order of perceived gloss from 1 (least glossy) to 22 (most glossy) [10]. Note that the same ranking for different subjects does not guarantee the same perceived gloss level. A straightforward procedure to link all within-subject gloss ranges is to collect human judgments on all possible pairwise combinations of stimuli across all subjects. However, such exhaustive evaluation is not feasible, because it would require a massive amount of trials, during which human attention is prone to drifting.

Instead of exhaustive comparisons, we propose a head-tail selective comparison method to fit all appearances of different subjects together by scaling each within-subject gloss range using selected candidates. We collect the least glossy images of each subject as the *head* set of stimuli and the most glossy images of each subject as the *tail* set of stimuli. With comparative studies within each set, we are able to get the relative gloss positions among all head/tail stimuli. The new gloss score of each subject is scaled based on the relative positions of the stimuli to the gloss scores of a reference subject whose gloss range is kept unchanged. We finally quantize all gloss scores into distinct gloss levels with agglomerative clustering. To check if the global gloss levels agree with human perception, we select random triplets with non-decreasing gloss levels from the entire dataset, and ask

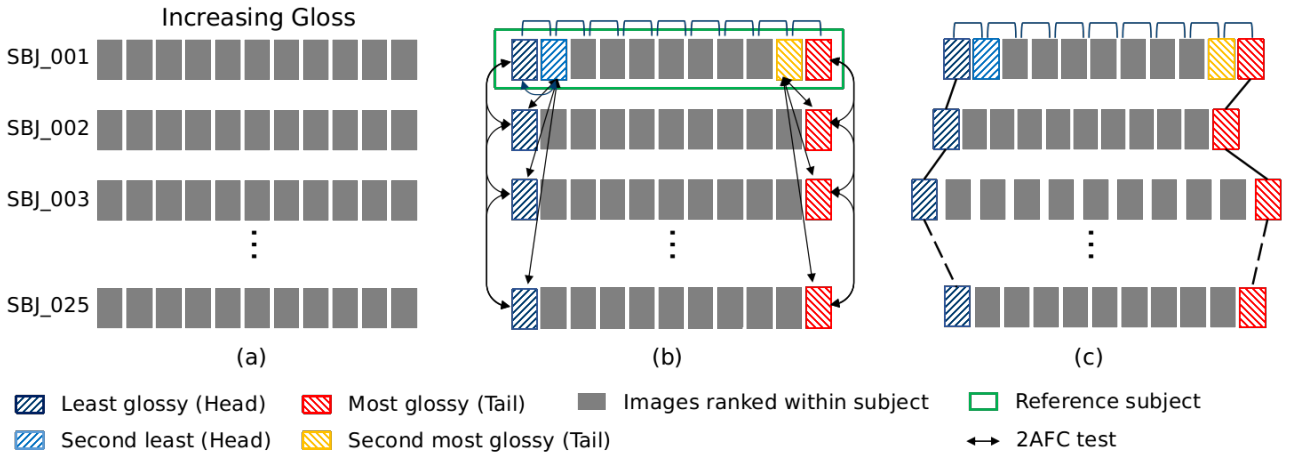


Figure 2: Head-tail selective comparison framework (Best viewed in color print).

the participants to indicate whether they agree with the triplet ordering. The participant responses show 90% agreement with the triplet ordering.

With the updated gloss levels as ground truth, we propose a model for perceived gloss level prediction. We extract facial gloss features by separating surface reflection and subsurface reflection using multimodal photography. We then extract local image statistics of each reflection image from five face sub-regions (left and right cheeks, forehead, nose, and lips) based on facial landmarks. Concatenating all statistics together, we learn a ranking function constrained by the similarity of image pairs. The gloss level is estimated based on nearest neighbors with closest ranking scores.

Gloss Perception Across Subjects

Head-tail Selective Method

As mentioned in the introduction, our study of across-subject gloss perception is based on a facial skin dataset with existing perceived gloss scores within each subject. The dataset contains frontal view face images from $K = 25$ different subjects. For each subject k there are $M = 22$ images with varying surface reflection. The images were generated from two original images, before and after cleansing, using image manipulations to vary the surface reflection [10]. All images within subject k were ranked by increasing perceived gloss. Each image $I_{k,m}$ of subject k at condition m was assigned a unique gloss ranking score $y_{k,m}$, where $m \in [1, M]$. $I_{k,1}$ denotes the least glossy image of subject k and $I_{k,M}$ denotes the most glossy image of subject k.

The head-tail selective method across subjects is shown in Figure 2 and detailed as follows:

1. Generate Head set H and Tail set T as candidate stimuli.
   - For each subject k, add $I_{k,1}$ to *head* candidate set H and $I_{k,M}$ to *tail* candidate set T;
   - Randomly select one subject r as reference subject;
   - Add the second least glossy image $I_{r,2}$ to H and add the second most glossy image $I_{r,M-1}$ to T.
2. Gloss evaluation on selected candidates
   - Conduct a two-alternative forced choice (2AFC) test on on all possible pairs in H and all possible pairs in T [1];
   - Compute relative gloss strength $g_{k,m}$ of each image in H and T using the Bradley-Terry model [11].
3. For each subject k, update the gloss score of each image
   - Get the new gloss score $\hat{y}_{k,1}$ of head image and $\hat{y}_{k,M}$ of tail image using Equation (1a) and Equation (1b);
   - Calculate the scaling factor α_k as the ratio between the new gloss range and the old gloss range using Equation (1c);
   - For each image $I_{k,m}$, update the gloss score by scaling the original score by α_k using Equation (2).

We first elaborate the test described in Step 1 and Step 2.

Step 1-2: Gloss Evaluation with 2AFC Tests

Stimuli and Procedure: As mentioned above, the head stimuli include the least glossy images of all subjects plus the second least glossy image of the reference subject. The tail stimuli include the most glossy images of all subjects plus the second most glossy image of the reference subject. The images of all subjects are frontal viewed, with the eye region masked, and 590×500 pixel resolution. In total we generated 325 head pairs and 325 tail pairs. All head pairs and all tail pairs were mixed and randomly shuffled. In each trial, the participants were shown one pair of facial images at a time and asked to choose the one that looks glossier.

Apparatus and Participants: The tests were conducted using a calibrated LCD screen with linear gamma and 1920×1080 resolution. The viewing distance was approximately 600 mm such that a 256 pixel image subtended an angle of 9.39 degrees. We collected evaluations from 10 observers in total with normal or correct-to-normal vision. Before the test, all observers were asked to read and sign consent forms.

Results: Once we collected results from all pairwise comparisons, we applied the Bradley-Terry model to estimate the perceived gloss strength $g_{k,m}$ of each image in H and T, as shown in Figure 3. Note that the gloss strength values are relative values

[1] For fair comparison, all head pairs and tail pairs are shuffled for random occurrence in each trial.

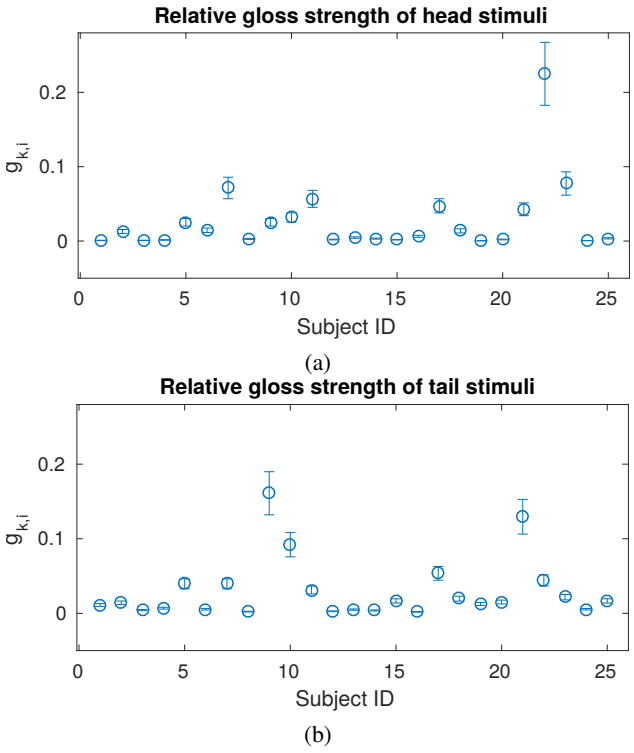


Figure 3: Relative gloss strength on (a) head stimuli and (b) tail stimuli from 2AFC tests.

within the comparison stimuli.

$$\hat{y}_{k,1} = \frac{y_{r,2} - y_{r,1}}{g_{r,2} - g_{r,1}} \cdot (g_{k,1} - g_{r,1}) + y_{r,1}, \tag{1a}$$

$$\hat{y}_{k,M} = \frac{y_{r,M-1} - y_{r,M}}{g_{r,M-1} - g_{r,M}} \cdot (g_{k,M} - g_{r,M}) + y_{r,M}, \tag{1b}$$

$$\alpha_k = \frac{\hat{y}_{k,M} - \hat{y}_{k,1}}{y_{k,M} - y_{k,1}} \tag{1c}$$

$$\hat{y}_{k,m} = \alpha_k \cdot (y_{k,m} - y_{k,1}) + \hat{y}_{k,1} \tag{2}$$

Step 3: Update by Linear Scaling

To link the estimated relative strength $\{g_{k,m}\}$ for each within-subject gloss score $\{y_{k,1}\}$, we refer to the relative gloss difference in the reference subject r using Equations (1a) - (1c). As shown in Figure 2(b), the gloss scores of all images in the reference subject r are kept unchanged. For head candidates in Equation (1a), $g_{r,2} - g_{r,1}$ across subjects and $y_{r,2} - y_{r,1}$ within subject both refer to the gloss difference between images $I_{r,2}$ and $I_{r,1}$. Therefore, the ratio $(y_{r,2} - y_{r,1})/(g_{r,2} - g_{r,1})$ connects the within-subject perception and across-subject perception as a constant scaling factor. The new gloss score of head image $\hat{y}_{k,1}$ is then updated by scaling its gloss difference with the reference head image $g_{k,1} - g_{r,1}$ by $(y_{r,2} - y_{r,1})/(g_{r,2} - g_{r,1})$. Similarly, Equation (1b) gets the new gloss score of each tail image $\hat{y}_{k,M}$ by scaling its gloss difference with the reference tail image $g_{k,M} - g_{r,M}$ by $(y_{r,M-1} - y_{r,M})/(g_{r,M-1} - g_{r,M})$. With the new gloss scores of the head images and tail images, we are able to calculate the gloss range of each subject for the entire dataset. The ratio between the new gloss range and the old gloss range is α_k. The remaining images of each subject between the head and the tail images are updated by scaling the gloss difference with α_k, as shown in Equation (2) and Figure 2(c).

The updated continuous across-subject scores $\{\hat{y}_{k,m}\}$ are then quantized into discrete gloss levels with bottom-up agglomerative clustering. The quantization guarantees that images from different gloss levels have obvious gloss differences. Table 1 summarizes the distribution of subjects and images on five distinct gloss levels. Note that the least glossy (head) images across subjects lie from gloss level 1 to gloss level 3, and the most glossy (tail) images across subjects lie from gloss level 3 to gloss level 5.

Table 1: Facial gloss dataset across subjects

Gloss level	Lvl 1	Lvl 2	Lvl 3	Lvl 4	Lvl 5
Subjects per level	18	25	25	25	15
Images per level	126	124	103	101	96
Head images per level	13	7	5	0	0
Tail images per level	0	0	5	10	10

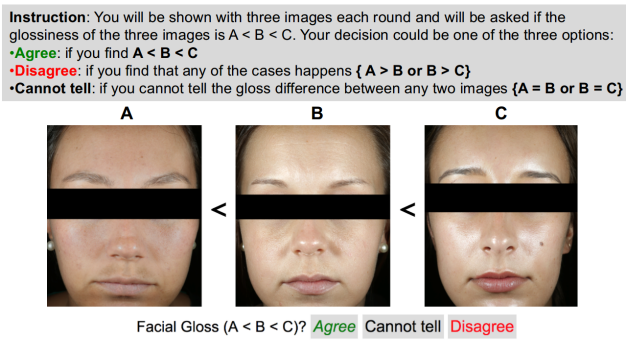


Figure 4: Interface of subjective validation on triplets.

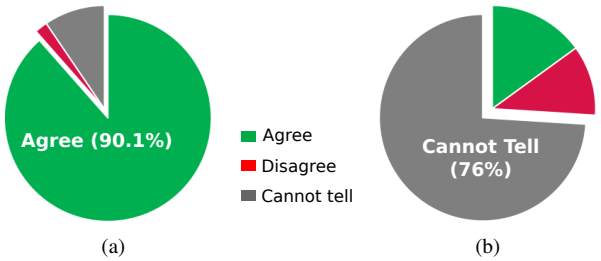
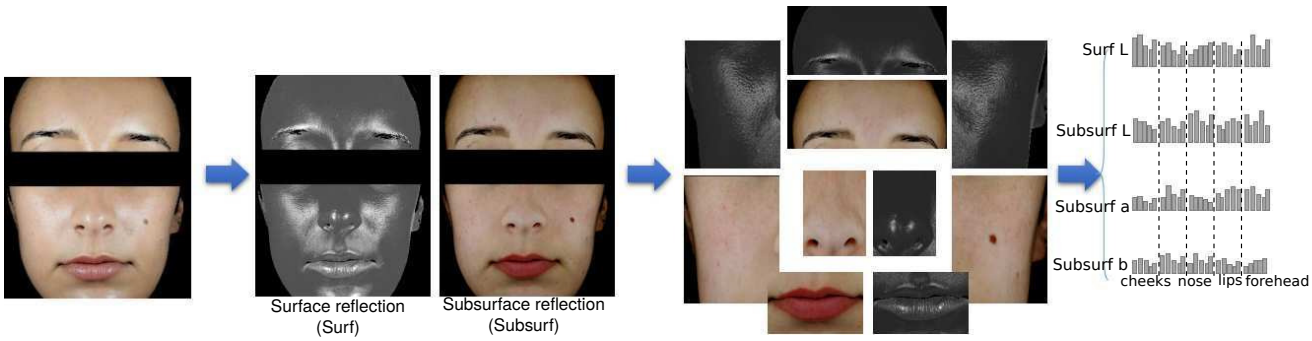


Figure 5: Empirical validation on quantized gloss levels. (a) Evaluation on triplets from different gloss levels and (b) evaluation on triplets containing same gloss level images.

Empirical Validation

Given the five distinct levels as labels to the facial gloss dataset, we conducted an empirical study to test whether the proposed method agrees with human judgments.

Stimuli and procedure: Given the 550 images in the dataset, we randomly selected 20% of images from each gloss level and formed 600 triplets. In each trial, observers were shown one triplet of images (A, B, and C) sorted in non-decreasing order of



(a) (b) (c) (d)

Figure 6: Facial feature extraction. (a) Original input image (b) Separation of (a) into surface reflection (grayscale) and subsurface reflection (color) (c) Separate whole face into five sub-regions (left cheek, right cheeks, forehead, nose, and lips) (d) Statistical features extracted from each sub-region and channel.

gloss level as shown in Figure 4. The observers were asked to indicate whether they agree with the image order. They were given three options: *Agree* means the observer finds that the triplet is ordered according to increasing gloss level ($A < B < C$). *Disagree* means the observer finds that either pair of adjacent images in the triplet is in reverse gloss order ($A > B$ or $B > C$). *Cannot tell* means the observer cannot find a gloss difference in either pair of adjacent images in the triplet ($A = B$ or $B = C$).

Apparatus and participants: The apparatus was the same as the one we used in the 2AFC tests. We collected evaluations from 3 observers in total with normal or correct-to-normal vision. Before the test, all observers were asked to read and sign consent forms.

Results: Figure 5 displays a summary of the observer opinions. For images with distinct gloss levels, we found that for 90.1% of the triplets the gloss orders are consistent with human judgments, while for about 1% of the triplets the ranking orders conflict with human perception, as shown in Figure 5(a). When triplets contain more than one image from the same gloss level (Figure 5(b)), the participants agreed and indicated that they "cannot tell" for about 76% of the images. The empirical validation tests demonstrate that images from distinct gloss levels have obvious perceived gloss differences, while images from the same gloss level have no perceptual differences, both with high probability.

Gloss Ranking model

With the selective head-tail comparison method, we successfully generated a global gloss range with five distinct levels for all images across subjects. In this section, we focus on developing a learning method for predicting the gloss level for a given facial image.

Image based gloss feature

As a material composed of multiple layers, the appearance of human skin is composed of surface reflection from outer air-oil layer and subsurface reflection from inner epidermis and dermis layers. We rely on multimodal photography and a polarization approach to separate the surface reflection from the subsurface reflection [10]. As shown in Figure 6(b), the surface reflection is a single lightness map that contains specular reflection. The subsurface reflection depends on intrinsic skin properties such color and skin condition.

With an off-the-shelf facial landmark detection method, we can automatically segment the whole face region into five separate sub-regions: left cheek, right cheek, forehead, nose, and lips as illustrated in Figure 6(c). Inspired by the relationship between statistics and gloss perception [10], we extract statistical features (mean, standard deviation, skewness, kurtosis, and entropy) from each sub-region and each color channel (surface lightness and subsurface color, in *CIE* $L^*a^*b^*$ coordinates). By concatenating all statistics together, we generate a gloss related feature vector with 200 parameters.

Max-margin rank-based model

We are given a set of training images $I_{train} = \{I_i\}$ represented in $\mathbb{R}^n$ by feature vector $\{x_i\}$ and the corresponding gloss level $\{y_i\}$. The data is formatted into a set of ordered pairs $T_O = \{(x_i, x_j) | y_i > y_j\}$, *i.e.,* image I_i has a stronger gloss level than image I_j, and a set of similar pairs $T_S = \{(x_i, x_j) | y_i = y_j\}$, *i.e.,* image I_i and image I_j have the same gloss level. The problem can then be described as learning a ranking function $R(\cdot)$.

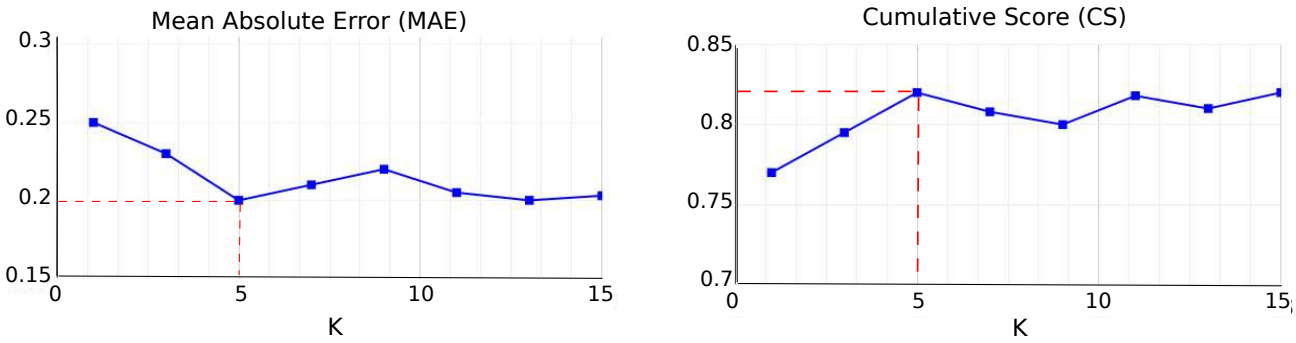
$$R(x_i) = \omega^T x_i, \tag{3}$$

which satisfies the following constraints:

$$\forall (x_i, x_j) \in T_O : \omega^T x_i > \omega^T x_j, \tag{4a}$$
$$\forall (x_i, x_j) \in T_S : \omega^T x_i = \omega^T x_j \tag{4b}$$

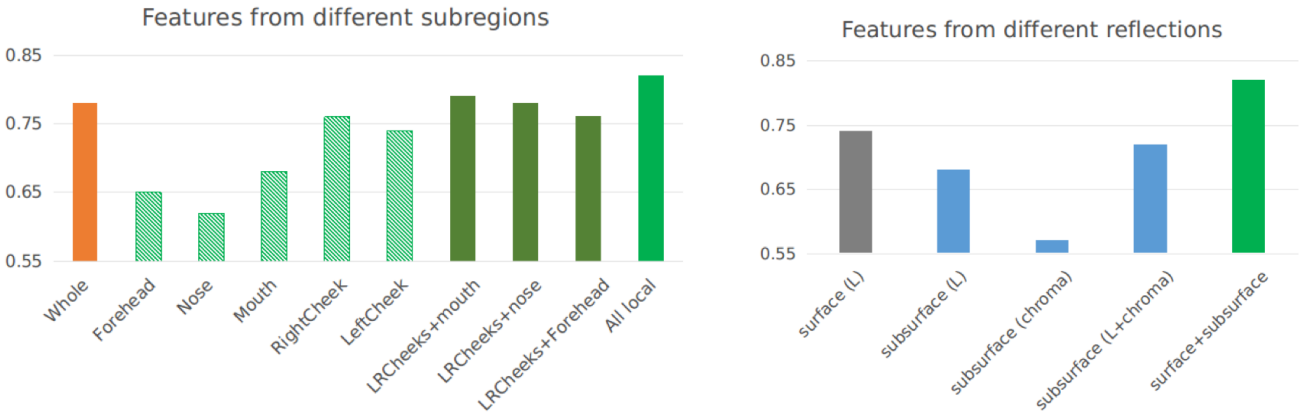
Inspired by [12], we approximate the solution with two slack variables added to ranking SVM as shown in Equation (5).

$$\text{minimize} \quad (\frac{1}{2}\|\omega^T\|^2 + C(\sum \xi_{i,j}^2 + \sum \gamma_{i,j}^2)) \tag{5a}$$
$$\text{s.t.} \quad \omega^T x_i \geq \omega^T x_j + 1 - \xi_{i,j}; \forall (x_i, x_j) \in T_O \tag{5b}$$
$$|\omega^T x_i - \omega^T x_j| \leq \gamma_{i,j}; \forall (x_i, x_j) \in T_S \tag{5c}$$
$$\xi_{i,j} \geq 0; \gamma_{i,j} \geq 0, \tag{5d}$$



(a) (b)

Figure 7: Performance evaluation with different number of neighbors



(a) (b)

Figure 8: Performance comparison with (a) different sub-regions (b) different reflections.

where $\xi_{i,j}$ and $\gamma_{i,j}$ are the slack variables and C is the trade-off constant between maximizing the margin and satisfying the pairwise relative constraints. The above primal problem is solved using Newton's method.

Experiments

To test the efficiency of the ranking model, we randomly split the facial gloss dataset of 25 subjects into a training set (80%) and a test set (20%). We then learn the ranking score function $R(\cdot)$ using Equation (5) on the training set. For an arbitrary test image I_t, we first calculate the ranking score $R(x_t)$ and then obtain the estimated gloss level $\tilde{y}_t$ based on majority votes of the k-nearest neighbors in the training set.

We use two performance indices, mean absolute error (MAE) and cumulative score (CS), in Equation (6) to evaluate the performance:

$$MAE = \frac{\sum_{t=1}^{N} |\tilde{y}_t - \hat{y}_t|}{N}, \tag{6a}$$

$$CS = \frac{\sum_{t=1}^{N} \mathbb{1}(|\tilde{y}_t - \hat{y}_t| < e)}{N}, \tag{6b}$$

where N is the total number of test images and $\hat{y}_t$ is the ground-truth gloss level and $\tilde{y}_t$ is the estimated gloss level. $\mathbb{1}(\cdot)$ is an indicator function. Note that CS is equivalent to *accuracy* with error level $e = 1$.

Figure 7 shows the performance as a function of the number of nearest neighbors for gloss level estimation. The best performance is obtained for $k = 5$ neighbors with $MAE = 0.2$ and $CS = 0.82$ with error level $e = 1$. We further use CS to test different combinations of features. Figure 8(a) shows the results when using different sub-regions. Using single sub-region features cannot compete with the features from the whole face. Features from the cheek regions result in higher CS than other regions, while features from the nose result in the lowest CS. One possible reason for this is that the cheek region occupies a larger area over the face with the largest specular reflections. Combining all subregion features results in the best performance. Figure 8(b) shows the effect of using different combinations surface and subsurface reflection and color components. Using only surface reflection lightness features results in better performance than using subsurface reflection lightness features, because surface reflection contains most of the specular reflections. The chroma (a* and b*) statistics of subsurface reflection do not contribute much to gloss estimation. However, it will boost the performance of subsurface reflection with L+chroma. The best performance is obtained by combining the features of the surface and subsurface reflections.

Conclusions and future work

We proposed a head-tail selective comparison method to fit within-subject gloss perception to a global gloss range across subjects. Without exhaustive comparisons on all possible combinations, we are able to update gloss scores of all images using just around 10% of the data in 2AFC tests. Empirical validation tests demonstrate the effectiveness of our method. We also developed a max-margin rank-based model to estimate the gloss level of a human facial image. The performance tests show that features from diverse local regions and reflection layers provide complimentary information in gloss perception. The best performance was gained by concatenating all local features together.

References

[1] T. Igarashi, K. Nishino, S. K. Nayar, *et al.*, "The appearance of human skin: A survey," *Foundations and Trends® in Computer Graphics and Vision* **3**(1), pp. 1–95, 2007.

[2] R. W. Fleming, R. O. Dror, and E. H. Adelson, "Real-world illumination and the perception of surface reflectance properties," *Journal of vision* **3**(5), pp. 3–3, 2003.

[3] M. Olkkonen and D. H. Brainard, "Perceived glossiness and lightness under real-world illumination," *Journal of vision* **10**(9), pp. 5–5, 2010.

[4] I. Motoyoshi and H. Matoba, "Variability in constancy of the perceived surface reflectance across different illumination statistics," *Vision Research* **53**(1), pp. 30–39, 2012.

[5] M. Olkkonen and D. H. Brainard, "Joint effects of illumination geometry and object shape in the perception of surface reflectance," *i-Perception* **2**(9), pp. 1014–1034, 2011.

[6] M. W. A. Wijntjes and S. C. Pont, "Illusory gloss on lambertian surfaces," *Journal of Vision* **10**(9), p. 13, 2010.

[7] Y.-X. Ho, M. S. Landy, and L. T. Maloney, "Conjoint measurement of gloss and surface texture," *Psychological Science* **19**(2), pp. 196–204, 2008.

[8] G. Wendt, F. Faul, V. Ekroll, and R. Mausfeld, "Disparity, motion, and color information improve gloss constancy performance," *Journal of vision* **10**(9), pp. 7–7, 2010.

[9] R. Ohtsuki, S. Tominaga, and R. Hikima, "Appearance analysis of human skin with cosmetic foundation," in *Color Imaging XVII: Displaying, Processing, Hardcopy, and Applications*, **8292**, p. 82920Q, International Society for Optics and Photonics, 2012.

[10] J. Wang, J. Mayne, C. Kuesten, G. Majamudar, and T. N. Pappas, "Determining the influence of image-based cues on human skin gloss perception," in *SPIE/IS&T Electronic Imaging*, pp. 195–202, International Society for Optics and Photonics, 2017.

[11] R. A. Bradley and M. E. Terry, "Rank analysis of incomplete block designs: I. the method of paired comparisons," *Biometrika* **39**(3/4), pp. 324–345, 1952.

[12] D. Parikh and K. Grauman, "Relative attributes," in *2011 International Conference on Computer Vision*, pp. 503–510, IEEE, 2011.