Estimation of Semantic Impressions from Portraits*

Mari MIYATA\(^{(a)}\), Nonmember and Kiyoharu AIZAWA\(^{(b)}\), Fellow

SUMMARY In this paper, we present a novel portrait impression estimation method using nine pairs of semantic impression words: bitter-majestic, clear-pure, elegant-mysterious, gorgeous-nature, modern-intellectual, natural-mild, sporty-agile, sweet-sunny, and vivid-dynamic. In the first part of the study, we analyzed the relationship between the facial features in deformed portraits and the nine semantic impression word pairs over a large dataset, which we collected by a crowdsourcing process. In the second part, we leveraged the knowledge from the results of the analysis to develop a ranking network trained on the collected data and designed to estimate the semantic impression associated with a portrait. Our network demonstrated superior performance in impression estimation compared with current state-of-the-art methods.

key words: portrait, semantic impression, deformation, learning to rank, relative attribute

1. Introduction

In the last seven years, the number of social networking service (SNS) users has increased from 0.97 billion to 2.62 billion. Given this very large number of SNS users, portraits play an important role in creating impressions. In fact, it is observed that it only takes a marginal change to a portrait (facial angle, lighting, etc.) to change the impression made on the observer\(^{[1], [2]}\) even for the same subject. In particular, a change in the facial appearance of the subject of a portrait has a major effect on the impression associated with the portrait\(^{[3]-[5]}\). For this reason, people tend to manipulate their portraits before uploading them to an SNS.

There are numerous types of portrait manipulation techniques, such as color filters, application of digital makeup, modification of the facial expression, and deformation of facial features. These manipulation techniques help control the color of the portrait, shape of the face, and so on. However, as there are no obvious rules for making such a manipulation, it is difficult to know whether the result obtained is ideal. For instance, people might ask themselves questions such as “How can I make my face prettier?” or “Will this manipulation look unnatural to my friends?” To address this problem, many studies have proposed a recommendation model for digital makeup style\(^{[6], [7]}\) or portrait color manipulation\(^{[8]}\) and applications for automatically enhancing the attractiveness or beauty score of a portrait\(^{[9]-[12]}\). As the goal of these studies is primarily to generate and recommend new portraits based on the recommendation model or application, they contain little mention of the details of changes in the facial appearance in the portrait due to the manipulation. Meanwhile, in the field of psychology, there have been studies attempting to unravel the relationship between the impression made by a person and the facial features of that person\(^{[4], [13]}\). These studies focus on the original facial appearance of the person and the impression it makes. Thus, as it is not their goal to provide techniques for manipulating portraits, these conventional studies do not provide explicit evidence regarding the specific facial feature manipulation(s) that played a crucial role in changing the impression made by the portrait after the manipulation. Moreover, to the best of our knowledge, there is no existing method for manipulating portraits based on semantic impression words, which are words that represent impressions made by people, such as “sunny,” “gorgeous,” or “warm.”

Toward developing an intuitive portrait manipulation method based on semantic impression words, we present a novel technique for estimating different types of semantic impressions associated with manipulated human portraits. Here, portrait manipulation is limited to the deformation of facial features in order to more easily identify the diverse changes in portraits that are associated with different impressions. Our method can be used to identify the change in the strength of semantic impression before and after the deformation during the manipulation process. For example, if a person tries to make his/her portraits “pure”, he/she can check how much each deformation makes their own por-
traits “pure” compared with before.

There are two key tasks in developing the impression estimation method: (1) collecting human evaluations for semantic impressions associated with deformed portraits and (2) constructing an impression estimation network. Even though these tasks were already carried out in [14], the dataset created therein can be improved, particularly with regard to the number of impression words. Moreover, even though the landmark-based impression estimation network in [14] achieved the highest accuracy among existing methods, it does not consider texture, which is also important in impression evaluation [15], [16]. Thus, in this study, we devise new settings for the crowdsourcing task and collect a new dataset consisting of semantic impression words and portraits with human evaluations. Furthermore, to focus on the change in the impression due to manipulation, we propose a new architecture for the impression estimation network that uses both facial landmarks and portrait images as input.

The main contributions of this study are as follows.

- We constructed a novel dataset consisting of 806 portraits along with their deformed versions and human evaluations for nine pairs of semantic impression words: bitter-majestic, clear-pure, elegant-mysterious, gorgeous-mature, modern-intellectual, natural-mild, sporty-agile, sweet-sunny, and vivid-dynamic.
- We analyzed the human evaluations and demonstrated that the deformation of facial features results in impression changes and that these changes differ among impressions.
- We constructed a novel impression estimation network that uses as input the portraits themselves, the coordinates of facial landmarks, and the labels indicating the relative relationship between portraits. The proposed network achieved impression estimation results that are comparable to or better than those of existing methods.

2. Related Work

In this section, we review methods that infer the attributes of an image and generate an image based on its attributes.

2.1 Estimation of Portrait Attributes

The estimation of an image’s attributes has become a common task in several areas of research. There are two types of attributes: absolute attributes, which show how well the image fits the attribute, and relative attributes, which show how well the image fits the attribute compared with other images. As our study deals with differences in impressions associated with portraits, here we focus on reviewing methods that handle relative attributes.

Relative attributes are used in various tasks that attempt to infer the strength of attributes: skill determination [17], prediction of a person’s beauty [18], [19], prediction of an image’s aesthetics [20], [21], and prediction of an image’s attributes [22]–[25]. For handling relative attributes, learning-to-rank algorithms are generally used and work effectively. Parikh and Grauman [25] first developed an algorithm for learning the ranking function of the relative visual attributes of images. While considering the difference of the attributes between images in a dataset, their algorithm can evaluate the extent to which the individual image has the attribute strength, based on the concept of the ranking support vector machine (Rank SVM) [26]. After their work was published, an SVM-based method was developed that focuses more on the local similarity of images [27]. Beyond these successes, network-based methods [23], [24], [28], [29] have performed better for learning relative attributes, especially Siamese-structured networks, which use a pair of image as inputs. Souri et al. [23] developed a new Siamese-structured network that consists of convolutional feature extraction layers and ranking layers. Because of the simplicity of the structure of the network and its high accuracy based on the RankNet algorithm [30], their method is frequently referenced as a baseline. Yu and Grauman [28], [29] attempted to increase the data by generating synthetic images based on attributes and to learn the discrimination using local similarity the same as the method [27]. In another approach, Meng et al. [22] performed multi-task learning using a graph-based neural network. Miyata et al. [14] developed a ranking network to learn the difference of the impression between a pair of deformed portraits. Though conventional methods only handle images as input, their network uses a pair of facial landmarks as input to specialize in the deformation of facial features. They achieved the highest accuracy among existing methods regardless of their simpler architecture without convolutional layers.

Following the success of the conventional method [14], we propose a novel ranking network whose input is both facial landmarks and portraits to more specialize in the deformation of facial features.

2.2 Portrait Manipulation Techniques

In recent years, it has become possible for users to manipulate portraits relatively easily by using web service smartphone applications such as BeautyPlus††, SNOW†††, and VSCO†††, even if they have little experience in editing portraits. Meanwhile, researchers have developed a system that can automatically estimate a person’s attributes and generate new portraits based on those attributes [31].

Rule-based and/or sample-based methods are frequently used to generate new portraits for purposes such as makeup style transfer [6], [32], [33] and facial image editing based on attributes [12], [34]–[36]. Conditional generative models, in particular, have been well studied for the generation of new portraits based on conditions of the sub-

†††web.beautyplus.com
†††snow.me
†††vsco.co
ject’s appearance such as facial expression, social appearance, and choice of accessories; the conditional generative adversarial network is a typical model of this type. M. Liu et al. [37] proposed a generative network that can edit subjects’ attributes such as hairstyle, skin color, and presence or absence of mustache while maintaining subjects’ identities and the resolutions of images.

In the context of our study, however, changes in impression caused by facial feature deformation are thought to be inconsistent among subjects (unlike changes in color, emotion, or age), and therefore it was expected to be difficult for a network to learn such ambiguous attributes. The method could be expected to fail to generate new portraits based on impression words. Thus, our method of portrait generation based on attributes considers manipulation of the portrait based on the behavior of the impression estimation network and does not use the architecture of the conditional generative adversarial network.

3. Data Collection

There are many datasets with facial images for psychology and computer vision research [38], [39]. Most recent studies have collected images of people with a certain “attribute” such as golden hair or a double chin and also include accessories such as glasses and hats [40]. Other studies have introduced portrait images for impression evaluation based on visual appearance [12], [41], facial expression [42], age, and gender. However, in the most of these datasets, the conditions of the portraits (resolution, subject conditions, etc.) are heterogeneous, and it is difficult to review them. Thus, for this study, we needed to build our own dataset.

In [14], the authors collected portraits and their deformed versions along with scores for five impression words. However, the number of portraits was relatively small, and thus the dataset is not well suited for use in practical applications. Accordingly, in this study, we selected a new set of words to describe the impressions associated with deformed portraits; moreover, we enriched the dataset with new portraits.

In this section, we describe the collection of human evaluations regarding the impressions associated with deformed portraits. First, we present the process used for collecting portraits and impression words; next, we describe how these portraits can be deformed. Then, we describe the process used for collecting human evaluations, and we provide a detailed analysis of these evaluations.

3.1 Semantic Impression Words

To express the impression of a person, many social studies use words that are related to human emotion or personality, such as the Big Five personality traits [43]. To develop an intuitive method for manipulating portraits, however, we need impression words that can express more abstract impressions from a subject’s visual appearance, which need not be related to the subject’s emotion or personality. For our study, we decided to select semantic impression words from Kobayashi’s list of keywords [44]. Kobayashi proposed 180 keywords that describe semantic impressions of shapes, patterns, materials, and industrial products, including, for example, fashion styles, images, and fonts [45], [46]. Examples of Kobayashi’s keywords are shown in Fig. 2. As shown in the figure, each keyword has unique coordinate values in the two-dimensional semantic space with the axes warm–cool and soft–hard. From these keywords, we carefully selected those words that can describe a human’s visual appearance and that are not related to facial expression, emotion, or personality. Then, we created nine pairs using these selected impression words as follows: bitter-majestic, clear-pure, elegant-mysterious, gorgeous-mature, modern-intellectual, natural-mild, sporty-agile, sweet-sunny, and vivid-dynamic. Each pair consists of conceptually similar words, and they have different coordinates in the semantic space.

3.2 Portraits

After selecting the semantic impression words, we collected the portraits for the study. We manually selected high-quality portraits that do not exhibit compression degradation or in which the face of the subject is not occluded. Moreover, to avoid bias caused by factors irrelevant to understanding the impression associated with a portrait, we restricted the profiles of subjects and their portraits as follows: 1) the ethnicity indicates the same country, 2) sex is female, 3) age group is from 20 to 40 years, 4) facial expression is not intense, and 5) head position is almost frontal. Following these conditions, we collected portraits of 806 Japanese females in the age group of 20 to 40 years. Among them, 32 portraits were taken in a fixed environment; the others were obtained from the Internet. Examples of the portraits are shown in Fig. 3.

3.3 Deformation of Portraits

Among the various types of portrait manipulation techniques, only the deformation of facial features is considered in this study. This restriction was applied to make it
Fig. 3 Examples of portraits in our dataset (aligned). All subjects are Japanese females in the age group of 20 to 40 years.

Fig. 4 Eight deformations of a portrait. In the portraits, we deformed four parts of the face—facial contour, eyes, nose, or mouth—based on facial landmarks.

easier to understand the changes in impressions associated with portraits. Furthermore, we restricted the area of deformation to four facial features—eyes, nose, mouth, and facial contour—because these play a major role in creating impressions.

Examples of portrait deformations are shown in Fig. 4. Before performing a deformation, we cropped the portraits to 500 × 500 pixels and aligned them by the position of the eyes. To generate deformed portraits, we first used the Face++ API† to detect 106-point facial landmarks from each portrait as the initial facial landmarks. We then calculated the coordinates of the facial landmarks for the deformed portrait by adding the deformation vector—obtained from an existing portrait beautification product—to the initial facial landmarks. To avoid the unnatural deformation of facial features, the degree of deformation was set heuristically so that the difference between portraits can be visually perceived. In the case of the eyes, the coordinates of the landmarks were approximated by similarity transform to avoid unnaturalness, as discussed in [12]. We deformed the portraits based on the calculated facial landmarks by utilizing the moving least-squares method (MLS) [47] using affine lines deformation and the Open Graphics Library. In particular, we mapped the texture of the original portrait from a mesh with a 100 × 100 grid to a new mesh based on the deformation function derived from MLS. In this way, we deformed one facial feature in the original portrait bigger or smaller as shown in Fig. 4. As a result, we generated 6448 deformed portraits from 806 collected portraits with deformed facial contour, eyes, nose, or mouth. Other examples of deformed portraits are shown in the experimental result part.

3.4 Collection of Human Evaluations

To collect human evaluations of the impressions associated with a deformed portrait, we adopted a pairwise comparison approach [48]. Under pairwise comparison, participants can observe the difference between portraits with partial deformation and can more easily determine the impression associated with each portrait. An example of this task is shown in Fig. 5. In each task, we displayed two portraits and the nine pairs of semantic impression words. We asked the participants to evaluate the difference in the impression associated with the portraits using a five-point scale. In each task, we showed thirteen pairs of portraits, each of which was randomly selected and displayed. Moreover, the position of the portraits (left or right) in a pair was randomly assigned. For each task, nine participants were asked to compare the pairs of portraits.

We used Amazon Mechanical Turk to collect the human evaluations. For each Human Intelligent Task (HIT), we paid $0.30.

To facilitate review of the evaluations, we inserted three control questions within each task to help determine the quality of a participant. In each control question, we showed a pair of portraits of the same person, whose eyes were obviously of different sizes in the two portraits, and asked the participant to select the portrait in which the subject had the larger eyes.

3.5 Data Validation and Analysis

After rejecting tasks in which the participant answered two or more of the control questions incorrectly, we collected the results of 2232 HITs and analyzed the results from 107 different participants, all of whom live in the U.S. To analyze the evaluations, we tested the significance of the difference between the impressions of a pair of portraits by an unpaired t-test. For this test, we divided the evaluations into two groups: those indicating that the impression changed (1, 2, 4, or 5 on the five-point scale) and

†www.faceplusplus.com
those indicating that the impression did not change (3 on the scale). A significance level $p < 0.01$ was obtained for all impression words and all deformations. This suggests that the deformation of facial features causes a change in the facial impression. Figure 6 shows the average proportion of human evaluations that indicated that the impression was changed by each deformed facial feature on each impression word pair.

Even though a significance level $p < 0.01$ was obtained for all impression words in the unpaired t-test, it is clear from the figure that the proportion differs among impression word pairs and deformation areas. Thus, it can be said that deforming different facial features results in different amounts and different types of impression changes.

4. Impression Estimation Network

To effectively utilize our dataset, which contains data on the relative relationship of the deformed portraits, we consider a ranking-based method. Ranking-based methods are known to be suitable for learning to regress attribute strength using the relative relationship of objects as input [26], [30]. In [14], the authors demonstrated that a landmark-based method exhibits superior performance in estimating impressions associated with deformed portraits. In this paper, we propose a novel impression estimator with a multiple-ranking network architecture. The proposed model uses both the portrait image and the $x$-$y$ coordinates of facial landmarks as inputs and generates the absolute rank of the impression associated with the portrait as an output.

In this section, we introduce a new ranking network to estimate the “rank,” defined as how well a given portrait fits an impression word.

4.1 Learning Algorithm

In order to estimate the absolute rank of the impression for each portrait, we rank each portrait according to the pairwise comparison data (Sect. 3.5). For learning the ranking using the relative relationship between pairs, it is known that it is better to rank by solving the problem as a classification task rather than directly regressing the strength of the attribute [30]. To train the network, we follow the ranking algorithm given in [30]. We map the ranks $r_i$, $r_j$ of the portraits $I_i$, $I_j$ to the probability $p_{i,j}$ by using a logistic function as follows: $p_{i,j} = 1/(1 + e^{-\alpha(r_i-r_j)})$. For the loss function, we use the binary cross-entropy loss with the probability $p_{i,j}$ and the target relative attribute label $t_{i,j}$ as follows:

$$L_{i,j} = -t_{i,j} \log(p_{i,j}) - (1-t_{i,j}) \log(1-p_{i,j}).$$

(1)

The loss function indicates the similarity of the target label $t_{i,j}$ and the relative relationship of portraits in the rank. The three values of target label $t_{i,j} \in \{0, 0.5, 1\}$ are defined in Sect. 5.1.

4.2 Network Architecture

The architecture of the proposed network is shown in Fig. 7. As shown in the figure, the network has two Siamese structures, which consist of the feature extractors and the ranking layers. There are two types of feature extractors: one for facial landmarks and one for portrait images. The former comprises three fully connected layers for converting each $n$-point facial landmark $v_i \in \mathbb{R}^{2n}$, $v_j \in \mathbb{R}^{2n}$ into a feature vector $f_{Li} \in \mathbb{R}^d$, $f_{Lj} \in \mathbb{R}^d$. The image feature extractor comprises the VGGFace-SENt-50 pre-trained network [49] and three fully connected layers for converting the output vectors of the network into feature vectors $f_{Li} \in \mathbb{R}^d$, $f_{Lj} \in \mathbb{R}^d$. Through the ranking layer, the latent feature vectors $f_{Li}$, $f_{Lj}$, $f_{Li}$, $f_{Lj}$ are mapped onto the estimated ranks $r_{Li}$, $r_{Lj}$, $r_{Li}$, $r_{Lj} \in \mathbb{R}$. We use a fully connected layer with a linear activation function as the ranking layer. To calculate the ranks $r_i$ and $r_j$ of portrait $I_i$ and $I_j$, we merge the ranks for facial landmarks and the portrait image as follows:

$$r = \alpha r_L + (1-\alpha)r_j,$$

(2)

where $\alpha$ denotes the weighting coefficient for the ranks. In the output of the network, we calculate $p_{i,j}$, which is an estimate of the probability of the relationship between $r_i$ and $r_j$. The network parameters are updated according to the loss function (Eq. (1)).
5. Experiments

5.1 Setup

In our experiments, we performed ten-fold cross-validation. Specifically, we used one-tenth of the pairs as test data and divided the remainder into ten sets for training and validation. To obtain reasonable results, we split the data pairs into the training, validation, and test sets in such a way that no subject was included in more than one set. We used 2900 pairs as training data and 324 pairs as test data.

To define the label $t_{i,j}$ for the paired portraits $I_i$ and $I_j$, we divided the five-point-scale evaluations (Fig. 5) into two groups: group $G_s$ that gave both portraits the same impression strength, and group $G_d$ that gave each of the two portraits different impression strengths. First, we assigned to $G_s$ those pairs for which more than 60% of the participants selected 3 on the five-point scale. From the remaining pairs, we identified those for which more than 60% of the participants selected 1, 2, 4, or 5 and added these to group $G_d$, and we added the rest to $G_s$. After grouping, we defined label $t_{i,j}$ for the paired portraits $I_i$ and $I_j$ to be 0.5 for the pairs in $G_s$, and 1 or 0 for the pairs in $G_d$.

In the implementation of the proposed network, all feature extractors consisted of three fully connected layers whose dimensionalities were 128, 64, and 32. To avoid overfitting, we used dropout layers with a dropout rate of 0.1. For training, the learning rate was set to $10^{-3}$ initially and was changed dynamically by stochastic gradient descent [50]. The weighting coefficient $\alpha$ was set to 0.5. We trained the network separately for each semantic impression word pair.

5.2 Baselines

We compared the proposed method with the following methods:

**RankNet [23]** This network uses pairs of portrait images as input. The network comprises a Convolutional-Network-based architecture for feature generation, and a fully connected layer for ranking. We used a VGGFace-SENet-50 pre-trained network [49] for the feature generation layers.

**Landmark-based network [14]** This network specializes in deformed portraits and uses facial landmarks as input. As in the case of RankNet, it has feature generation layers, and a ranking layer consisting of fully connected layers.

5.3 Results

5.3.1 Quantitative Results

To show the behavior of the proposed network, we first report the ranges of the estimated ranks for the entire dataset in Table 1. As seen in the table, the ranges of the rank values vary according to the semantic impression word pair as there are no constraints on the ranges of the values during training. Therefore, we note here that it is impossible to compare the rank strengths of different semantic impression word pairs.

Following the previous studies [23, 25, 27], we also report the percentage of correctly ordered pairs to indicate the ranking accuracy (Table 2). It can be seen that the proposed network achieved the best performance on average.
Specifically, in seven of the nine pairs of semantic impression words, the proposed network, which uses portrait images and facial landmarks as input, outperformed the state-of-the-art method [14], which uses only facial landmarks as input. Moreover, across all impression word pairs, the landmark-based method [14] and our proposed network outperformed RankNet [23], which uses only portrait images as input. These results indicate that using facial landmarks as input works better to describe changes in facial appearance and impression than portrait images, and using portrait images can additionally improve the performance. Thus, we conclude that the proposed network effectively uses portrait images and facial landmarks for impression estimation.

Furthermore, it is observed that the performance of each method varied considerably according to the semantic impression word pair. For instance, “natural-mild” was the most difficult (56.1% on average), and “gorgeous-mature” was the easiest (74.1% on average) for all three methods. As different identity has different changes in impression due to deformation of facial features, the difficulty of estimating impression can vary according to the impression words and deformations.

5.3.2 Qualitative Results

We show the qualitative results of our proposed network in Fig. 8. As shown in the figure, the proposed network can facilitate discussion of the behavior of our network, we calculated the average face for each semantic impression as representative of a typical face for the impression (Fig. 9). Specifically, we averaged the 30 portraits having the highest impression values among all of the deformed portraits. In the results, each average face has distinct facial features. For example, the “bitter-majestic” face and the “sweet-sunny” face look quite different: The “bitter-majestic” face has a small mouth, narrow eyes, and an angular facial contour, and the “sweet-sunny” face, in contrast, has thick lips, round eyes, and a round facial contour.

We also measured the correlations between impressions and facial feature deformations and between impressions. First, we analyzed the correlation between the impression ranks and the facial feature properties. Specifically, we calculated Spearman’s rank correlation coefficients

![Figure 9 Average faces.](image)

| Words            | 1st                  | 2nd                  | 3rd                  | 4th                  | 5th                  |
|------------------|----------------------|----------------------|----------------------|----------------------|----------------------|
| bitter-majestic  | mouth width          | mouth area           | head height          | jaw deviation        | lip saturation       |
| clear-pure       | eye area             | eyebrow saturation   | head height          | eye height           | head area            |
| elegant-mysterious | right-head-to-right-eye | chin curve          | right jaw grad      | mouth position       | head width           |
| gorgeous-mature  | eye width            | head area            | eyebrow width        | head height          | eye separation       |
| modern-intellectual | mouth-to-nose          | nose width          | head width           | mouth position       | eyebrow height       |
| natural-mild     | right jaw gradient   | right-head-to-right-eye | head front-profile  | left-head-to-left-eye | chin curve           |
| sporty-agile     | eye separation       | eye width            | head area            | eye area             | head width           |
| sweet-sunny      | mouth area           | bottom lip height    | mouth height         | eye width            | eye area             |
| vivid-dynamic    | eye width            | eye area             | right-head-to-right-eye | right jaw gradient  | skin value           |

![Table 3 List of 49 selected attributes.](image)

Table 3: List of 49 selected attributes.

| Attribute type | Attributes |
|----------------|------------|
| Head           | Area       |
|                | Front-profile | Height       |
|                | Up-down     | Width        |
|                |              | Skin color (H/S/V)* |
| Eyebrows       | Area       |
|                | Color (H/S/V)* | Height       |
|                | Position    | Width        |
|                | Gradient    |
| Eyes           | Area       |
|                | Height     |
|                | Position   |
|                | Eye separation |
| Nose           | Area       |
|                | Curve       |
|                | Height     |
|                | Position   |
|                | Width      |
|                | Separation |
| Jawline        | Height     |
|                | Chin curve  |
|                | Gradient (Left/Right) |
|                | Deviation  |
| Mouth          | Area       |
|                | Mouth gap   |
|                | Bottom lip curve |
|                | Lip color (H/S/V)* |
|                | Height     |
|                | Position   |
|                | Width      |
|                | Top lip curve |
| Distance       | eyes-to-mouse |
|                | Mouth-to-nose |
|                | Left-head-to-left-eye |
|                | Right-head-to-right-eye |

* H/S/V: Hue, Saturation, Value
between the ranks of each impression and 49 facial attributes selected from [4] (see the list in Table 3) as the facial feature properties. We obtained the attributes of the portraits from their 106-point facial landmarks and color information. We show the correlation coefficients in Table 4. As shown in the table, there is no strong correlation between the ranks of semantic impression words and facial attributes, and only “sporty-agile” has a slightly stronger correlation between them. The result suggests that the estimated impression of portraits does not depend on a single facial property such as color, position and size of the facial feature. Therefore, it can be said that our impression estimation network estimates the rank by considering information from both the facial landmarks and the portrait images, not simply from one of them alone. In fact, one study found that humans tend to give importance to facial textures [15] rather than solely to positions of facial features; thus, we can say that our network has achieved a human-like reaction to portraits. Second, we analyzed the correlation between impression ranks. In this case, we calculated Pearson’s rank correlation coefficients between the ranks of the impressions. No notable correlation was found between the ranks (Table 5). In other words, no underlying relationship was found between impressions. This result suggests that each impression captures a different impression associated with portraits and that we successfully selected different types of impression words.

From the results above, we conclude that our network learned human-like behavior from the collected human evaluations.

6. Conclusion

In this study, we proposed a novel portrait impression estimation method. First, we constructed a novel dataset consisting of portraits with their deformed versions, nine-types of semantic impression word pairs, and human evaluations. Then, we analyzed the human evaluations and confirmed that the deformation of a facial feature results in impression changes on the portrait. Utilizing the collected dataset, we constructed a novel impression estimation network for deformed portraits that uses portrait images and facial landmarks as inputs. The results demonstrated that the highest accuracy was achieved by our network.

In future work, we will implement an automatic portrait manipulation technique based on two tasks: 1) Estimating the strength of an unknown semantic impression word by combining our ranking networks of nine impression word pairs, 2) Constructing a portrait manipulation network which can manipulate the facial landmarks based on the strength of an semantic impression.

References

[1] A. Todorov and J.M. Porter, “Misleading first impressions,” Psychological Science, vol.25, no.7, pp.1404–1417, 2014.
[2] R. Jenkins, D. White, X. Van Montfort, and A.M. Burton, “Variability in photos of the same face,” Cognition, vol.121, no.3, pp.313–323, 2011.
[3] J.C. Silveira Jacques Junior, Y. Güçlütürk, M. Perez, U. Güçlü, C. Andujar, X. Baró, H.J. Escalante, I. Guyon, M.A.J. Van Gerven, R. Van Lier, and S. Escalera, “First impressions: a survey on vision-based apparent personality trait analysis,” IEEE Transactions on Affective Computing, p.1, 2019.
[4] R.J.W. Vernon, C.A.M. Sutherland, A.W. Young, and T. Hartley, “Modeling first impressions from highly variable facial images,” Proc. National Academy of Sciences, vol.111, no.32, pp.E3353–E3361, 2014.
[5] A. Khosla, W.A. Bainbridge, A. Torralba, and A. Oliva, “Modifying the memorability of face photographs,” Proc. International Conference on Computer Vision, pp.3200–3207, 2013.
[6] T. Alashkar, S. Jiang, S. Wang, and Y. Fu, “Examples-rules guided deep neural network for makeup recommendation,” Proc. AAAI Conference on Artificial Intelligence, pp.941–947, 2017.
[7] L. Liu, J. Xing, S. Liu, H. Xu, X. Zhou, and S. Yan, “Wow! you are so beautiful today!”, ACM Transactions on Multimedia Computing, Communications, and Applications, vol.11, no.1s, pp.1–22, 2014.
[8] Y. Koyama, I. Sato, D. Sakamoto, and T. Igarashi, “Sequential line search for efficient visual design optimization by crowds,” ACM Transactions on Graphics, vol.36, no.4, 2017.
[9] F. Chen, X. Xiao, and D. Zhang, “Data-driven facial beauty analysis: prediction, retrieval and manipulation,” IEEE Transactions on Affective Computing, vol.9, no.2, pp.205–216, 2018.
[10] M. Sun, D. Zhang, and J. Yang, “Face attractiveness improvement using beauty prototypes and decision,” Proc. Asian Conference on Pattern Recognition, pp.283–287, 2011.
[11] S. Melacci, L. Sarti, M. Maggini, and M. Gori, “A template-based approach to automatic face enhancement,” Pattern Analysis and Applications, vol.13, no.3, pp.289–300, 2010.
[12] T. Leyvand, D. Cohen-Or, G. Dror, and D. Lischinski, “Data-driven enhancement of facial attractiveness,” ACM Transactions on Graphics, vol.27, no.3, p.1, 2008.
[13] R.M. Stolier, E. Behman, M.D. Keller, M. Walker, and J.B. Freeman, “The conceptual structure of face impressions,” Proc. National Academy of Sciences, vol.115, no.37, pp.9210–9215, 2018.
[14] M. Miyata and K. Aizawa, “Impression estimation for deformed portraits with a landmark-based ranking network,” Proc. IEEE International Conference on Image Processing, pp.1950–1954, 2019.
[15] S. Suwajanakorn, S.M. Seitz, and I. Kemelmacher-Shlizerman,
“What makes Tom Hanks look like Tom Hanks,” Proc. IEEE International Conference on Computer Vision, pp.3952–3960, 2015.

[16] B. Fink, K. Grammer, and R. Thornhill, “Human (Homo sapiens) facial attractiveness in relation to skin texture and color,” Journal of Comparative Psychology, vol.115, no.1, pp.92–99, 2001.

[17] H. Doughty, D. Damen, and W. Mayol-Cuevas, “Who’s better? who’s best? pairwise deep ranking for skill determination,” Proc. Conference on Computer Vision and Pattern Recognition, pp.6057–6066, 2018.

[18] C. Cao, I.S. Kwak, S. Belongie, D. Kriegman, and H. Ai, “Adaptive ranking of facial attractiveness,” Proc. IEEE International Conference on Multimedia and Expo, pp.1–6, July 2014.

[19] H. Altwaijry and S. Belongie, “Relative ranking of facial attractiveness,” Proc. IEEE Workshop on Applications of Computer Vision, pp.117–124, 2013.

[20] J. Ren, X. Shen, and Z. Lin, “Personalized image aesthetics,” Proc. IEEE International Conference on Computer Vision, pp.638–647, 2017.

[21] A. Deza, S. Barbara, D. Parikh, and V. Tech, “Understanding Image Virality,” Proc. IEEE International Conference on Computer Vision and Pattern Recognition, pp.1818–1826, 2015.

[22] Z. Meng, N. Adhur, H.J. Kim, G. Fung, and V. Singh, “Efficient relative attribute learning using graph neural networks,” Proc. European Conference on Computer Vision, pp.1–16, 2018.

[23] Y. Souri, E. Noury, and E. Adeli, “Deep relative attributes,” IEEE Trans. Multimedia, vol.18, no.9, pp.118–133, 2017.

[24] K.K. Singh and Y.J. Lee, “End-to-end localization and ranking for relative attributes,” Proc. European Conference on Computer Vision, pp.753–769, 2016.

[25] D. Parikh and K. Grauman, “Relative attributes,” Proc. International Conference on Computer Vision, pp.503–510, 2011.

[26] T. Joachims, “Optimizing search engines using clickthrough data,” Proc. International Conference on Knowledge Discovery and Data mining, pp.133–142, 2002.

[27] A. Yu and K. Grauman, “Fine-grained visual comparisons with local learning,” Proc. Conference on Computer Vision and Pattern Recognition, pp.192–199, 2014.

[28] A. Yu and K. Grauman, “Thinking outside the pool: active training image creation for relative attributes,” Proc. IEEE International Conference on Computer Vision and Pattern Recognition, pp.708–718, 2019.

[29] A. Yu and K. Grauman, “Semantic jitter: dense supervision for visual comparisons via synthetic images,” Proc. IEEE International Conference on Computer Vision, vol.1, pp.5571–5580, 2017.

[30] C. Burges, T. Shaked, E. Renshaw, A. Lazier, M. Deeds, N. Hamilton, and G. Hullender, “Learning to rank using gradient descent,” Proc. International Conference on Machine learning, pp.89–96, 2005.

[31] X. Zheng, Y. Guo, H. Huang, Y. Li, and R. He, “A survey of deep facial attribute analysis,” International Journal of Computer Vision, vol.128, pp.2002–2034, 2020.

[32] S. Liu, X. Ou, R. Qian, W. Wang, and X. Cao, “Makeup like a superstar: deep localized makeup transfer network,” Proc. International Joint Conference on Artificial Intelligence, pp.2568–2575, 2016.

[33] D. Guo and T. Sim, “Digital face makeup by example,” Proc. IEEE International Conference on Computer Vision and Pattern Recognition, pp.73–79, 2009.

[34] R. Yeh, Z. Liu, D.B. Goldman, and A. Agarwala, “Semantic facial expression editing using autoencoded flow,” arXiv preprint arXiv:1611.09961, 2016.

[35] Z. Shu, “Eyecopen: editing eyes in the wild,” ACM Transactions on Graphics, vol.36, no.1, 2016.

[36] F. Yang, D. Metaxas, J. Wang, E. Shechtman, and L. Bourdev, “Expression flow for 3d-aware face component transfer,” ACM Transactions on Graphics, vol.30, no.4, pp.1–10, 2011.

[37] M. Liu, Y. Ding, M. Xia, X. Liu, E. Ding, W. Zuo, and S. Wen, “STGAN: a unified selective transfer network for arbitrary image attribute editing,” Proc. IEEE International Conference on Computer Vision and Pattern Recognition, pp.3668–3677, 2019.

[38] G.B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, “Labeled faces in the wild: a database for studying face recognition in unconstrained environments,” Tech. Rep. 07-49, University of Massachusetts, Amherst, Oct. 2007.

[39] W.A. Bainbridge, P. Isola, and A. Oliva, “The intrinsic memorability of face photographs,” Journal of Experimental Psychology: General, vol.142, no.4, pp.1523–1534, 2013.

[40] Z. Liu, P. Luo, X. Wang, and X. Tang, “Deep learning face attributes in the wild,” Proc. International Conference on Computer Vision, 2015.

[41] N.E. Etoff, S. Stock, L.E. Haley, S.A. Vickery, and D.M. House, “Cosmetics as a feature of the extended human phenotype: modulation of the perception of biologically important facial signals,” PLoS ONE, vol.6, no.10, e25656, 2011.

[42] J.-Y. Zhu, A. Agarwala, A.A. Efros, E. Shechtman, and J. Wang, “Mirror mirror: crowdsourcing better portraits,” ACM Transactions on Graphics, vol.33, no.6, pp.1–12, 2014.

[43] M.R. Barrick and M.K. Mount, “The big five personality dimensions and job performance: a meta-analysis,” Personnel Psychology, vol.44, no.1, pp.1–26, 1991.

[44] S. Kobayashi, “The aim and method of the color image scale,” Color research and application, vol.6, no.2, pp.93–107, 1981.

[45] J. Jia, J. Huang, G. Shen, T. He, Z. Liu, H. Luan, and C. Yan, “Learning to appreciate the aesthetic effects of clothing,” Proc. AAAI Conference on Artificial Intelligence, pp.1216–1222, 2016.

[46] A. Jahanian, S. Keshvari, S.V.N. Vishwanathan, and J.P. Allebach, “Colors – messengers of concepts,” ACM Transactions on Computer-Human Interaction, vol.24, no.1, pp.1–39, 2017.

[47] S. Schaefer, T. McPhail, and J. Warren, “Image deformation using moving least squares,” ACM Transactions on Graphics, vol.25, no.3, pp.533–540, 2006.

[48] K. Tsukida and M.R. Gupta, “How to analyze paired comparison data,” tech. rep., Dept. of Electrical Engineering, University of Washington, 2011.

[49] J. Hu, L. Shen, and G. Sun, “Squeeze-and-excitation networks,” Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp.7312–7314, 2018.

[50] I. Loshchilov and F. Hutter, “SGDR: stochastic gradient descent with warm restarts,” ACM Transactions on Graphics, vol.36, no.1, 2017.

Mari Miyata received the B.E. and M.E. degrees in Electrical Engineering from Tokyo university of Science in 2014 and 2016, respectively. During 2016–2020, she stayed in Graduate School of Information Science and Technology, the University of Tokyo as a Ph.D. student.
Kiyoharu Aizawa received the B.E., the M.E., and the Dr.Eng, degrees in Electrical Engineering all from the University of Tokyo, in 1983, 1985, 1988, respectively. He is currently a Professor at Department of Information and Communication Engineering of the University of Tokyo. He was a Visiting Assistant Professor at University of Illinois from 1990 to 1992. His research interest is in image processing and multimedia applications. He received the 1987 Young Engineer Award and the 1990, 1998 Best Paper Awards, the 1991 Achievement Award, 1999 Electronics Society Award from IEICE Japan, and the 1998 Fujio Frontier Award, the 2002 and 2009 Best Paper Award, and 2013 Achievement award from ITE Japan. He received the IBM Japan Science Prize in 2002. He is currently a Senior Associate Editor of IEEE Trans. Image Processing, and on Editorial Board of ACM TOMM, APSIPA Transactions on Signal and Information Processing, and International Journal of Multimedia Information Retrieval. He served as the Editor in Chief of Journal of ITE Japan, an Associate Editor of IEEE Trans. Image Processing, IEEE Trans. CSVT and IEEE Trans. Multimedia. He has served a number of international and domestic conferences; he was a General co-Chair of ACM Multimedia 2012. He is a council member of Science Council of Japan.