Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 00 (2023) 000–000

Procedia Computer Science

www.elsevier.com/locate/procedia

27th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems (KES 2023)

Assessing the attractiveness of human face based on machine learning

Adriana Żejmo[a], Maciej Gielert[a], Marcin Grabski[a] and Bożena Kostek[b*]

[a]*Gdańsk University of Technology, ETI Faculty, Narutowicza 11/12, 80-232 Gdańsk, Poland*
[b]*Gdańsk University of Technology, ETI Faculty, Audio Acoustics Laboratory, Narutowicza 11/12, 80-232 Gdańsk, Poland*

Abstract

The attractiveness of the face plays an important role in everyday life, especially in the modern world where social media and the Internet surround us. In this study, an attempt to assess the attractiveness of a face by machine learning is shown. Attractiveness is determined by three deep models whose sum of predictions is the final score. Two annotated datasets available in the literature are employed for training and testing the algorithms, i.e., a dataset named SCUT-FBP5500 to train the deep learning models to predict facial attractiveness and Face Research Lab London Set designated for the test. The first model pays attention to the dominant background colors in the photo; the second model is based on a pre-trained deep neural network. Finally, for facial proportion assessment, distances between key points on the face are linked with attractiveness ratings, so the last dataset considers face proportions. Several algorithms are trained and tested, including baseline machine learning algorithms, i.e., LinearSVR, SDGRegressor, Lasso, RandomForestRegressor, and deep models, such as Xception VGG19 ResNet50v2, and MobileNetv2. A discussion of the results, as well as some concluding remarks, are also provided. The results from the trained models based on SCUT-FBP5500 show a systematic error for the Face Research Lab London Set database. This is probably caused by a different type of image evaluation in both databases. Although the results obtained show no visible winner among the algorithms employed, the best results are seen for five clusters and five colors fed onto the regressor.

Keywords: Facial attractiveness, Deep learning, Photogenicity

* Corresponding author. Tel.: 48-58-3472717.
E-mail address: bokostek@audioakustyka.org

1. Introduction

Physical appearance and especially facial attractiveness play an important role across cultures and beauty stereotypes, as well as in everyday life, especially in the modern world surrounded by social media and the Internet. This topic attracts a lot of attention in many science fields, such as medical sciences [1,2], psychology and social sciences [3,4], and computer science leading to an increased interest in facial attractiveness [5-7].

Moreover, the aspect of 'photogenicity' is also interesting to pursue. There are some answers to what makes a face desirable to the camera [8,9]. Most cues regarding whether a face is photogenic or not refer to the facial structure proportions, known as the 'golden ratio.' Obviously, a person may be perceived as attractive from the viewer's perspective without being photogenic. Hence, there should be some additional factors influencing 'photogenicity.' A photographer artist would say that light plays an important role when defining a face structure. Another mentioned factor refers to eyes, magnification, or smiling effects [9]. It was shown that the eye region has a greater impact on face attractiveness examination and judgment than other parts of the face [10-13].

In recent years, the means of predicting facial attractiveness or the so-called Facial beauty prediction (FBP) much advanced, both traditional approaches and machine learning methods [14,15], notably deep neural networks [7,16]. However, due to the subjective nature of the issue, assessing facial attractiveness remains difficult. Of importance is the fact that facial attractiveness judgment depends not only on personal preferences, which differ across time and cultures but also on the cognitive representation of the face category in one's mind [6]. To this end, Tong et al. [17] indicated that a deep model could be trained to learn facial ratios considered ideal in the sense of 'golden' proportions based only on categorical annotation when no annotated facial features for attractiveness are explicitly given. The schema of their experiments seems very intuitive as one of the deep models learns female/male and high/low attractiveness, and the second one generates face-like images by reversing the DNN model for facial attractiveness (FADNN). Finally, the third experiment simulates human-like judgments on facial images with varying ratios of features that reveal changes in the activity of the category-specific neurons that are remarkably similar to those observed in the 2010 study by Pallet et al. [4].

Often, machine learning models are trained and tested on a chosen dataset [18]. So, the motivation for this study is to use datasets available in the literature but created by different authors. In our contribution to the research on face attractiveness, we propose an approach that aims to process data of varying origins. Two databases retrieved from the Internet are employed for this purpose, i.e., a dataset named SCUT-FBP5500 [14,15] to train the deep learning models to predict facial attractiveness and Face Research Lab London Set designated for the test [19,20]. Moreover, distances between key points on the face are linked with attractiveness ratings for facial proportion assessment, so the last dataset is created for face proportion annotation.

The first dataset contains 5500 front-on-face portraits with a diverse proportion of Asian females, Asian males, Caucasian females, and Caucasian male subjects with neutral expressions. They were subjectively rated on a 1-5 scale.

Face Research Lab London Set contains images of 102 adult faces in full color with a size of 1350x1350 pixels. Moreover, the template files show 189 coordinates delineating face shapes for use with Psychomorph or WebMorph.org. Some additional information, such as self-reported age, gender, and ethnicity, is also included in the file london_faces_info.csv. In contrast to SCUT-FBP5500, attractiveness was rated by the annotators on a 1-7 scale, starting from a description "much less attractiveness than average" to "much more attractive than average" for the neutral front faces from 2513 people (ages 17-90) [20].

It should be pointed out that there are several noteworthy differences between the datasets retrieved from the Internet. The SCUT-FBP5500 dataset comprises 5500 faces exhibiting diverse characteristics, including sex, race (Caucasian and Asian), and age. Additionally, it provides various labels, such as facial landmarks, beauty scores on a scale of 5, and distribution of beauty scores. The images within this dataset feature are either homogeneous solid-colored backgrounds or heterogeneous backgrounds with color gradients or non-homogeneous patterns. Many faces display slight rotation and are captured from varying distances, although most are taken from a frontal position. These images exhibit diverse qualities and colors, as most are in color, while some are in black and white. It is worth noting that certain pictures may have undergone retouching.

In contrast, the Face Research Lab London dataset comprises 102 adult face images measuring 1350x1350 pixels. These images include self-reported information regarding age, gender, ethnicity (not only Caucasian and Asian), and attractiveness ratings on a 7-point scale, ranging from "much less attractive than average" to "much more attractive

than average." Notably, this dataset offers less variation, as all pictures are captured against a consistent gray gradient background, featuring identical face positions, camera distances, and quality.

The facial attractiveness in these databases was rated and labeled subjectively. We try to check whether it is possible to create a machine learning model that may help label attractiveness automatically based on parametrizing facial features. Also, an aspect worth exploring was judging the entire photo's attractiveness, not just the face.

Moreover, to measure facial proportion, distances between key points on the face are linked with attractiveness ratings, so this aspect is also researched in our study.

In the study, several algorithms are trained and tested, including baseline machine learning algorithms, i.e., LinearSVR, SDGRegressor, Lasso, RandomForestRegressor, and deep models, such as Xception VGG19 ResNet50v2, and MobileNetv2. Training and test parameters were chosen experimentally. The results obtained are discussed further on. Conclusions are also provided. As already said, the main contribution of this paper is to use several baseline algorithms and deep models to train and test them on different data.

2. Methods

The starting point in this study was creating an application that enables to send a selected photo and return its rating on a scale of 1 to 5. To obtain the photo rating, the combined prediction results from the three methods described below were used.

2.1. Evaluation of the dominant background colors

Studies by Minami et al. [21] and Nakajima et al. [22] have shown that both background and facial color inflect judgments of facial expressions. Even though facial color influences the perceived images more significantly than background color effects [22], it was decided to check to what extent the background in the photo impacts the subjective assessment of the attractiveness of a given person's face. The solution proposed is an algorithm that extracts the dominant colors from a photo, omitting the part of the image where the assessed face is located. The algorithm consists of the following steps:

- a face is found in the photo, and then a mask is applied in place of the face,
- clustering is performed on the photo without a face,
- the most common colors are taken, and a feature vector is created on their basis,
- the feature vector is trained in the regressor.

The number of clusters in the clustering method and the number of colors taken for training were tested. Values from 0 to 9 were assigned for both variables. The best results were obtained for five clusters and five colors fed further onto the regressor. In this case, the following regressors were used: LinearSVR [23], SGDRegressor [24], Lasso [25], and RandomForestRegressor [26]. In the case of RandomForestRegressor the hyperparameters tuning was induced. A maximum depth of 50 was returned as the best metric. The remaining parameters for all models were left as default. The results obtained are shown in Section 3.

2.2. Learning with neural networks

Then, MobileNetV2, VGG19, ResNet50V2, Xception were chosen as deep models for face attractiveness processing. These models allow for extracting features from face images automatically, and they appear in the literature sources related to this subject. The training was carried out on these four selected trained deep networks employing the ImageNet set. Photos were resized to the size of 160 by 160. The sequential model, in addition to the trained model, includes layers such as batch normalization, global average pooling 2D, Dropout, and Dense with a linear activation function. The outcome of this approach is shown in Section 3.

2.3. Assessment of facial proportions

Finally, for facial proportion assessment, distances between key points on the face were linked with attractiveness ratings. In each photo, first, a face was detected, and then the points were marked to have the distances calculated.

Face proportions considered were as follows (see Fig. 1):

- eye width – the distance between the eyes,
- distance from the center of the eyes to the tip of the nose – from the tip of the nose to the chin,
- face width – face length,
- face width – mouth width,
- face width – width between eyes,
- nose width – nose length,
- eyebrow width – eye width,
- nose width – eye width,
- nose width – mouth width,
- from the edge of the face to the eye - the distance between the eyes.

Based on these 10 calculated values, a regressor indicating the attractiveness of a given face was trained for each photo from the dataset. The following algorithms were used: RandomForestRegressor, LinearSVR, and SDGRegressor. For the first of them, hyperparameter tuning was induced. A maximum depth of 51 was determined to be the best result. All other parameters were left at their default values. The results are summarized further on in Section 3.

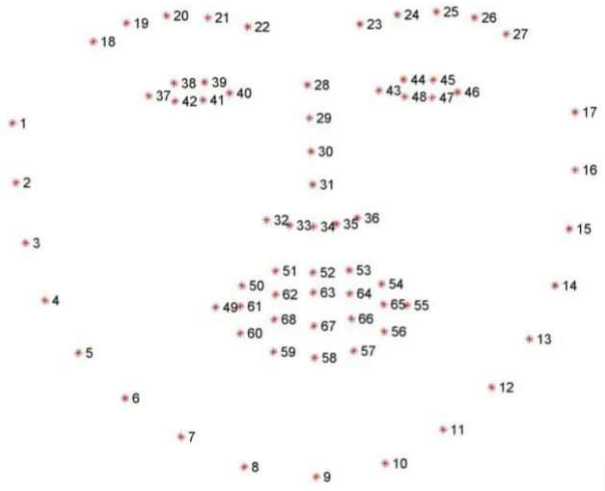

Fig. 1. Distances between key points on the face and attractiveness ratings.

3. Results

3.1. Dominant color results

Models were trained and tested on the SCUT-FBP5500 basis with averaged scores on a scale of 1-5. Since cross-validation provides a statistically more reliable estimation of the performance than a single training/test set split, that is why such a method was chosen. Moreover, we decided on 5-fold cross validation as it enables to train five different models.

With 5-fold validation, the following averaged results were obtained for the models based on dominant background colors (see Table 1). As mentioned, Linear SVR, SDGRegressor, Lasso, and RandomForestRegressor were employed.

RandomForestRegressor was chosen for further analysis as it received the best results in metrics. As seen above, the model evaluating the dominant background colors strongly averages the results. This may be due to insufficient training data. There is an under-prediction of major grades and an over-prediction of low rates.

Table 1. Averaged results obtained for the models based on dominant background colors.

Model	r2	MAE	RMSE
LinearSVR	0.083	2.208	0.66
SDGRegressor	0.08	2.07	0.66
Lasso	0.100	2.11	0.65
RandomForestRegressor	0.2	2.26	0.61

3.2. Learning with neural networks

The results for 5-fold cross-validation average values for neural networks models trained with batch size 32 for 100 epochs, the Adam optimizer, and the loss function set to MAE (Mean Absolute Error), as it is more robust to outliers, are shown in Table 2. Also, R2 (a relative measure of fit) and RMSE (an absolute measure of fit) loss functions were checked. The model using ResNet50v2 performed relatively well compared to the other three models.

Table 2. Results for 5-fold cross-validation average values for neural networks models.

Model	R2	MAE	RMSE
Xception	0.49	2.01	0.49
VGG19	0.58	1.87	0.44
ResNet50v2	0.60	1.70	0.43
MobileNetv2	0.57	1.78	0.45

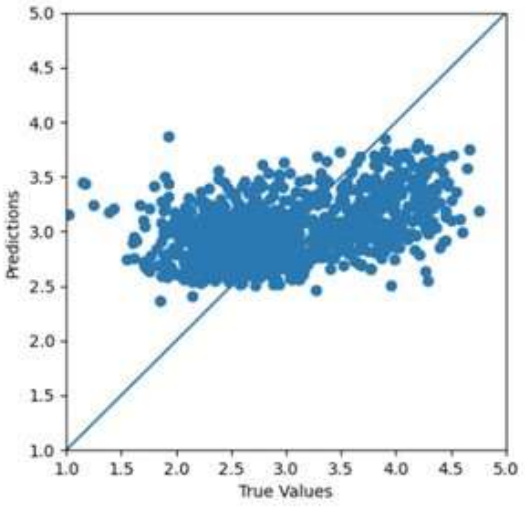

Fig. 2. Results obtained with RandomForestRegressor learned on dominant background colors.

3.3. Assessment of facial proportions−regressor results for method no. 3

The regressors were trained on 4,500 examples, and the effectiveness was checked on 1,000. As shown above (see Fig. 2), the model evaluating facial proportions averages the results, but the prediction is more reliable than based on the dominant background color model. This may be due to the lack of gender division in the test cases. There is an under-prediction of higher grades and an over-prediction of low rates, similar to the model evaluating dominant background colors.

The three above models were launched in parallel in the prepared application. The models received the proportions of importance in the final evaluation. The first model got 11% of the rating, the second 67%, and the third 22%. To test the obtained models, the Face Research Lab London Set was used (see Table 3). Only frontal smiling photos were selected. The original ratings on a scale of 1-7 contained in this database needed to be rescaled to a scale of 1-5.

The results obtained are rather unexpected; it is evident that the application significantly overestimated them. At this point, the minimum, maximum and average values in both used datasets were checked. The results are presented in Table 4.

Table 3. Results for training on SCUT-FBP5500 and test on Face Research Lab London dataset.

Model	R2	MAE	RMSE
RandomForestRegressor	0.233	2.344	0.613
LinearSVR	0.146	2.433	0.647
SDGRegressor	0.182	2.509	0.634

Table 4. Minimum, maximum, and average values in both used datasets.

Dataset	min	max	avg
SCUT-FBP5500	1.02	4.75	2.99
Face Research Lab London	1.11	4.05	2.15

Predicted values obtained with the pre-trained ResNetV50 backbone and resized images as input are shown in Fig. 3. Further, prediction values while using a facial ratio-trained RandomForestRegressor model are contained in Fig. 4. Finally, predicted values obtained with three models, described above but concatenated, are presented in Fig. 5.

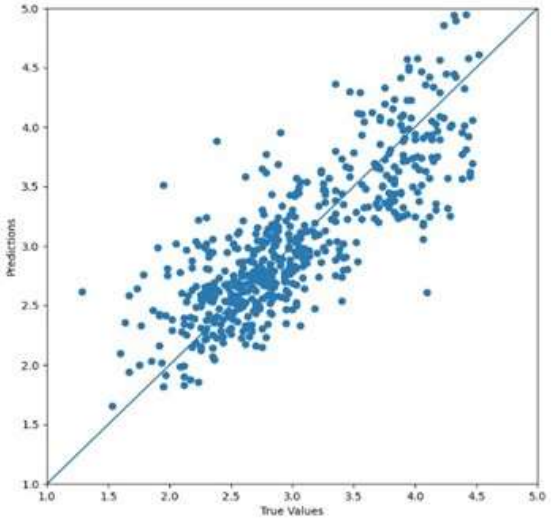

Fig. 3. Predicted values obtained with the pre-trained ResNetV50 backbone and resized images as input.

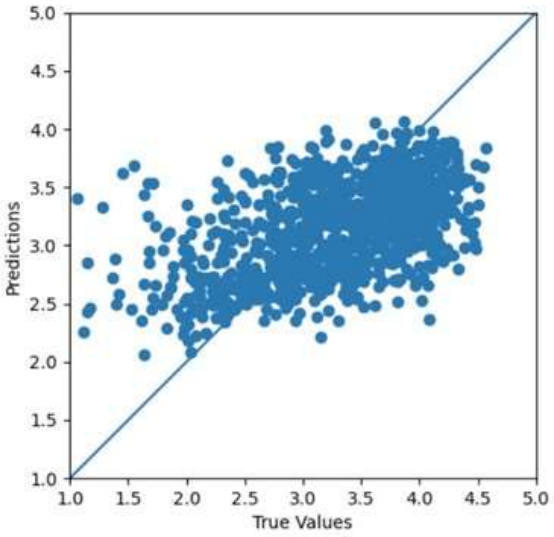

Fig. 4. Prediction values obtained using a facial ratio-trained RandomForestRegressor model.

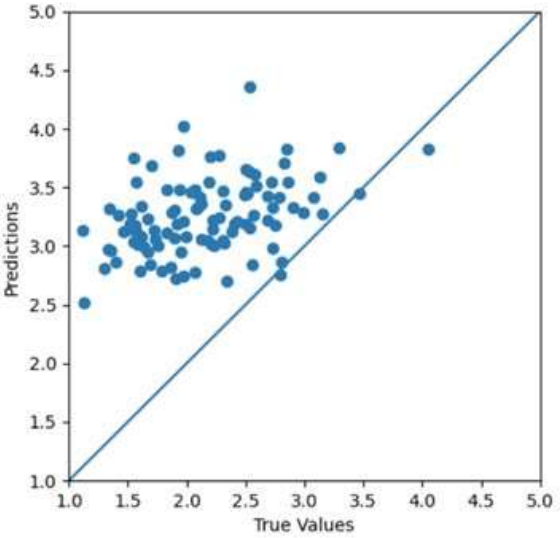

Fig. 5. Predicted values obtained with three models described above but concatenated.

The added value of the study performed is to employ different datasets. The difficulty of a different attractiveness annotation scale within the two datasets was overcome by rescaling them. The stability of the results trained and tested on different data may be assessed as high. This was confirmed by changing model hyperparameters. However, the results from the models trained on SCUT-FBP5500 show a systematic error for the Face Research Lab London Set database.

With the solution proposed, we wanted to focus as much as possible on the background in the image. Using the Haar cascade object detection algorithm, frontal faces in photos were found. Based on the boxes created by the cascade classifier, a mask on a face was built. The pixel values were then converted to a vector, excluding the masked pixels. The K-means algorithm was run on the obtained vector, and then the most frequently occurring values were selected. Relatively small or negative values of the measured r2 and RMS metrics were found in subsequent trials during

subsequent algorithm runs on the database. The background of the photograph of a human face has a small but significant impact on its attractiveness, but not sufficient to be negligible. There is a high degree of instability in this solution due to the algorithm's sensitivity to the number of clusters set in the K-means algorithm.

Also, there are some general remarks that may be drawn for the experiments performed. Even if some authors report a high correlation between facial metrics generated from manually and automatically placed image landmarks [5], the attractiveness assessment – as seen from the results obtained – is still subjective. Therefore, creating a general model remains a big challenge. Moreover, aesthetics is highly correlated with emotion [27], so this aspect should also be considered along with the emotion model chosen [28] when judging facial attractiveness.

4. Conclusion

Several conclusions may be drawn from the study performed. It occurred that models should be trained separately for photos of women and men due to different facial proportions and other features considered desirable. The following improvement of the model that calculates face proportions may be proposed – the model should consider all possible combinations of key points on the face. Then a selection of features should be performed to check which proportions affect the attractiveness of the face the most. As already said, the attractiveness assessment remains a subjective process; that is why it is challenging to create a general model of 'attractiveness.' Moreover, aesthetics is highly correlated both with the emotion expressed on the face evaluated and the emotion evoked in an annotator, so this aspect should also be considered along with the emotion model chosen when judging facial attractiveness.

Due to the subjectivity of image ratings, obtaining an accurate model for different datasets is not trivial. A factor with a significant impact is the scale on which the respondents rated the photos. The results from the trained models based on SCUT-FBP5500 show a systematic error for the Face Research Lab London Set database. This was probably caused by a different type of image evaluation in both databases.

Also, we pointed out that employing different datasets often requires some rescaling or other types of processing as the origin of the dataset may differ much. This may be an example of dealing with data from different backgrounds. Moreover, this may show guidelines for creating own data concerning this subject.

References

[1] Adamson, Peter A., and Matthew B. Zavod. (2006) "Changing perceptions of beauty: a surgeon's perspective." *Facial Plastic Surgery* **22 (3)**: 188–193. doi: 10.1055/s-2006-950176.

[2] Bashour, Mounir. (2006) "An objective system for measuring facial attractiveness". *Plastic and reconstructive surgery* **118 (3)**: 757–74; discussion 775–6 doi: 10.1097/01.prs.0000207382.60636.1c.

[3] Little, Anthony C., Benedict C., Jones and Lisa M., DeBruine. (2011) "Facial attractiveness: evolutionary based research". *Philos Trans R Soc Lond B Biol Sci*. **366(1571)**: 1638-59 doi: 10.1098/rstb.2010.0404.

[4] Pallett, Pamela M., Sthephen Link, and Kang Lee. (2010) "New "golden" ratios for facial beauty". *Vision Research* **50(2)**: 149–154. doi: 0.1016/j.visres.2009.11.003.

[5] Jones, Alex, Lee, Christoph Schild, and Benedict C. Jones. (2020) "Facial metrics generated from manually and automatically placed image landmarks are highly correlated". *Evolution and Human Behavior* **42 (3)**: 186-193. doi: 10.1016/j.evolhumbehav.2020.09.002.

[6] Kagian, Amit, Gideon Dror, Tommer Leyvand, Isaac Meilijson, Daniel Cohen-Or, and Eytan Ruppin. (2008) "A machine learning predictor of facial attractiveness revealing human-like psychophysical biases", *Vision Research* **48 (2)**: 235-243. https://doi.org/10.1016/j.visres.2007.11.007.

[7] Bougourzi, Fares, Fadi Dornaika, and Abdelmalik Taleb-Ahmed. (2022) "Deep learning based face beauty prediction via dynamic robust losses and ensemble regression". *Knowledge-Based Systems* **242:** 108246. doi:10.1016/j.knosys.2022.108246.

[8] Experts reveal what makes people more photogenic than others, https://hnmagazine.co.uk/beauty/experts-reveal-what-makes-people-more-photogenic-than-others/ (accessed on March 32, 2023).

[9] Wen, Wen, and Hideaki Kawabata. (2014) "Why am I not photogenic? Differences in face memory for the self and others". *Iperception* **5 (3)**: 176-87. doi: 10.1068/i0634.

[10] Etcoff, Nancy (2011) "Survival of the prettiest: The science of beauty". Anchor Books, A Division of Random House, Inc., New York.

[11] Kwart, Dylan G., Tom Foulsham, and Alan Kingstone. (2012) "Age and beauty are in the eye of the beholder". *Perception* **41**: 925–938. doi: 10.1068/p7136.

[12] Langlois, Judith H., Lisa Kalakanis, Adam J. Rubenstein, Andrea Larson, Monica Hallam, M., and Monica Smoot. (2000) "Maxims or myths of beauty? A meta-analytic and theoretical review". *Psychological Bulletin* **126**: 390–423. doi: 10.1037/0033-2909.126.3.390.

[13] Thornhill, Randy, and Steven W. Gangestad. (1999) "Facial attractiveness". *Trends in Cognitive Sciences* **3**: 452–460. doi: 10.1016/S1364-6613(99)01403-5.

[14] Liang, Lingyu, Luojun Lin, Lianwen Jin, Duorui Xie, and Mengru Li. (2018) "SCUT-FBP5500: A diverse benchmark dataset for multi-paradigm facial beauty prediction". In: Proceedings of the 24th International Conference on Pattern Recognition, ICPR .

[15] Xie, Duorui, Liang Lingyu, Lianwen Jin, Jie Xu, and Mengru Li. (2015) "SCUT-FBP: A benchmark dataset for facial beauty perception". IN: Proceedings of the 2015 IEEE International Conference on Systems, Man, and Cybernetics, IEEE, 1821-1826. doi: https://doi.org/10.1109/SMC.2015.319.

[16] Shi, Shengjie, Fei Gao, Xuantong Meng, Xingxin Xu, Jingjie Zhu. (2019) "Improving Facial Attractiveness Prediction via Co-attention Learning". In: Proceedings of the ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 4045-4049. doi: 10.1109/ICASSP.2019.8683112.

[17] Tong, Song, Xuefeng Liang, Takatsune Kumada, and Sunao Iwaki. (2021) "Putative ratios of facial attractiveness in a deep neural network". *Vision Research* **178**: 86-99. doi: 10.1016/j.visres.2020.10.001.

[18] Chen, Fan, Hui Li, and Yinglin Zheng. (2021) Attribute-induced Attractiveness Regression of Facial Images with Multi-task Convolution Neural Network. In: Proceedings of the 16th International Conference on Computer Science Education (ICCSE), 816-821. doi: 10.1109/ICCSE51940.2021.9569262.

[19] DeBruine, Lisa M., and Benedict C. Jones. (2017) "Face Research Lab London Set. Psychology" (Data set). doi:10.6084/M9.FIGSHARE.5047666.V2

[20] Ratings of the London dataset.london_faces_ratings.csv; https://figshare.com/articles/dataset/Face_Research_Lab_London_Set/5047666/2 (accessed on March 32, 2023).

[21] Minami, Tetsuto, Kae Nakajima, and Shigeki Nakauchi. (2018) "Effects of Face and Background Color on Facial Expression Perception". *Frontiers in Psychology* **9**. doi:10.3389/fpsyg.2018.01012.

[22] Nakajima, Kae, Tetsuto Minami, and Shigeki Nakauchi. (2017) "Interaction between facial expression and color". *Scientific Reports* **7**. 41019. doi: 10.1038/srep41019.

[23] sklearn.svm.LinearSVR, https://scikit-learn.org/stable/modules/generated/sklearn.svm.LinearSVR.html (accessed on March 32, 2023).

[24] sklearn.linear_model.SGDRegressor, https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.SGDRegressor.html (accessed on March 32, 2023).

[25] sklearn.linear_model.Lasso, https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.Lasso.html

[26] sklearn.ensemble.RandomForestRegressor, https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html (accessed on March 32, 2023).

[27] Yu, Jun., Chaoran, Cui, C., LeiLei Geng, Yuling Ma, and Yilong Yin. (2019) Towards Unified Aesthetics and Emotion Prediction in Images. In: Proceedings of the 2019 IEEE International Conference on Image Processing (ICIP), Taipei, Taiwan. doi: 10.1109/ICIP.2019.8803388.

[28] Plewa, Magdalena, and Bozena Kostek, (2015) "Music Mood Visualization Using Self-Organizing Maps". *Archives of Acoustics* **40 (4)**: 513-525. https://doi.org/10.1515/aoa-2015-0051.