Importance Estimation for Scene Texts Using Visual Features

著者	OODAIRA Kota, MIYAZAKI Tomo, SUGAYA Yoshihiro, OMACHI Shinichiro
journal or publication title	Interdisciplinary information sciences
volume	28
number	1
page range	15-24
year	2022
URL	http://hdl.handle.net/10097/00137320

doi: 10.4036/iis.2022.A.06

Interdisciplinary Information Sciences Vol. 28, No. 1 (2022) 45–53
©Graduate School of Information Sciences, Tohoku University
ISSN 1340-9050 print/1347-6157 online
DOI 10.4036/iis.2022.A.02

Prediction of Image Preferences from Spontaneous Facial Expressions

Yoshiyuki SATO[1,2,*], Yuta HORAGUCHI[3], Lorraine VANEL[3] and Satoshi SHIOIRI[1,2,3]

[1] *Advanced Institute for Yotta Informatics, Tohoku University, Sendai 980-9577, Japan*
[2] *Research Institute of Electrical Communication, Tohoku University, Sendai 980-8577, Japan*
[3] *Graduate School of Information Sciences, Tohoku University, Sendai 980-8579, Japan*

With advances in digital technologies, the number of images we are subjected to every day has increased significantly. Predicting and recommending human subjective preferences for images is useful for selecting image data efficiently to avoid the unnecessary use of valuable storage space. In this study, we investigate the use of a machine learning model for estimating human preferences for images from spontaneous facial features extracted from video images of human faces while they are performing a natural preference evaluation task. We use two image categories and compare the results between categories. We also conduct an experiment to assess the performance of human raters in predicting the preferences of others from facial videos. As a standard to compare predictive performance from facial expressions, we also test prediction from high-level image features by training a deep learning model using the obtained experimental data. The results show that the spontaneous facial features produce prediction performance comparable with, and for lunch box images, marginally better than, the image features specifically trained for our dataset, and clearly outperform the human raters. We further examine which facial expression features are important for prediction and show that the important facial features differ between image categories. Our results show that facial expressions can be used to predict the preference for images, to some extent, although we need to be careful when generalizing the learned model to other image categories. Our machine learning approach also provides insights into the differences in the cognitive mechanisms used for preference evaluation for different image categories.

KEYWORDS: image preference, spontaneous facial expression, LightGBM model, importance analysis

1. Introduction

With the evolution of Internet technology and mobile devices, the number of images that we are subjected to every day is increasing rapidly. We often make subjective judgements about images in situations such as internet shopping, rating photos on social networking sites, or choosing which photos to keep when there are too many for the available storage space. As the number of images overwhelmingly increases, technologies for predicting and recommending human subjective preferences for images become more important [1, 2]. In response to this need, in recent years, there has been considerable research on predicting subjective evaluation using machine learning [3–10]. The recent advancements of researches also allow us to automatically detect text importance from images [11] and extract body postures of students related to class engagement from videos [12]. Human studies of attention also provide insights into the selection of information. There are two types of attention: top-down attention, whereby subjects intentionally control where their attention is focused, and bottom-up attention, whereby the subject's attention is attracted to the location of salient stimuli, such as a flashing light or loud sound [13, 14]. Under the assumption that attention is oriented to process important information, it is worth considering the effects of both top-down and bottom-up attention on preference judgements. We consider image features, which are related to bottom-up attention, and facial expressions while evaluating preference, which are perhaps related to top-down attention.

In predicting preference judgements, it is preferable to use implicit information that does not interfere with people's choice behavior. One important piece of implicit information is the facial expressions that can be obtained by the cameras in modern digital devices, such as PCs and mobile phones. Humans make facial expressions in response to images [15, 16], and recent advances in machine learning techniques have made it possible to automatically analyze these facial expressions in response to the presented images. In many recent studies, machine learning techniques have been used to estimate human decision making and preferences from facial expressions extracted from video recordings. Examples of these studies include predictions of voting behavior [5], music preferences [6], engagement of learners in education [10, 17], and the estimation of emotions during video advertisement viewing [7, 9].

Automatic facial expression recognition has been investigated for many years in a number of research fields. In many early studies, researchers used facial images or videos in which actors or other people attempted to show emotions

Received November 4, 2021; Accepted March 2, 2022
This work is supported by Yotta Informatics Project by MEXT, Japan.
[*] Corresponding author. E-mail: yoshi.yk.sato@gmail.com

intentionally [18, 19]. Recently, spontaneous and natural facial expressions have become a topic of great interest [17–21]. Generally, spontaneous facial expressions are weaker and more diverse than intentionally made facial expressions, which makes it challenging to evaluate spontaneous facial expressions [21–24]. Among the investigations of spontaneous facial expressions, in one study, researchers showed that human preferences for images could be predicted [25]. In their work, the task was an alternative-choice judgement, that is, choosing one of two images presented for a fixed duration. In the present study, we examine prediction performance in a more natural setting. We recorded facial videos while the participants were forming a preference between images on a social networking service, as they may do in real life. The participants were asked to "like" images to reflect their preferences, and were able to view the images for as long as they wished and to "like" as many images as they wished, rather than being forced to make a decision. Assessing predictive performance in such a natural scenario is vital for real-world applications.

There are two additional difficulties related to preference prediction in natural scenarios. One is that humans use different properties of images to determine their preference depending on the category of the images. For example, familiarity is important for the preference of faces and novelty is important for the preference of natural scene images [26]. The other difficulty is the fact that humans seem to express their preferences differently for different image categories. A study has shown that changes in facial expressions according to the preference for the image/video being watched vary among image categories [27]. Therefore, it is important to investigate the performance of preference prediction using multiple image categories and analyze which facial features contribute to the prediction.

In this study, we develop a method for estimating users' preferences for images using machine learning from spontaneous facial videos taken while the users are performing a natural preference evaluation task. To compare the results for different image categories, we conduct a preference judgement experiment using two image categories. As a standard to compare predictive performance from facial expressions, we also test prediction from high-level image features by training a deep learning model using the obtained experimental data. We also conduct an experiment to assess human performance in predicting preference judgements from face videos. Furthermore, we examine the importance of the facial expression features used by the model to predict preference judgements from face videos to investigate which facial features contribute to the prediction and how they differ between image categories.

2. Experiments

2.1 Image preference evaluation experiment

We conducted preference evaluation experiments to obtain data on human preference judgements for images and the facial expressions of the evaluators while making preference judgements. To create a natural scenario that may occur in our daily lives, we presented the participants with images on a screen resembling the interface of Instagram [28] (Fig. 1). The images were obtained from the Internet in real time using the tag search function of Instagram. We chose two tags that are commonly used on Instagram, that is, "#お弁当" (lunchbox in Japanese) and "#landscape," to obtain data for the image domains.

Each trial started when a visual stimulus was presented on a screen (Fig. 1). The participants were asked to indicate whether they "liked" the image or not by pressing a key on a keyboard. Because the images were obtained automatically using hashtags, the presented images were not always in the intended category. In that case, the participants were asked to press a button that indicated that the presented image was irrelevant, and the trial was abandoned. The participants were given as long as they needed to make a judgement in each trial. After they provided a

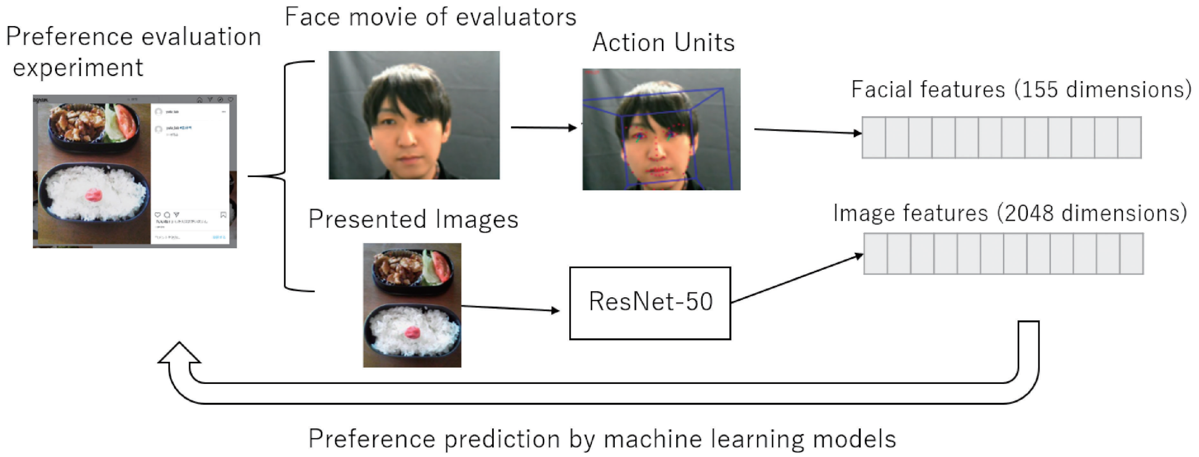

Preference evaluation experiment
Face movie of evaluators
Action Units
Facial features (155 dimensions)
Presented Images
Image features (2048 dimensions)
ResNet-50
Preference prediction by machine learning models

Fig. 1. Schematic diagram of the image preference evaluation experiment and the analysis of image preference estimation in this study. Preference estimation was performed by LightGBM using features extracted from the face image of the evaluator during the experiment, along with the presented image.

response, the next trial started after an interval of 0.5 s. During the trials, the facial videos of the participants were recorded with a camera mounted on a PC for later analysis, together with their responses. For each trial, we recorded the face video from the time of stimulus presentation until the start of the next stimulus presentation and used the video for analysis. Because some participants performed the experiment from home using their own PCs, the video resolution and frames per second of the recordings varied across participants. To record natural facial expressions, we did not instruct the participants to make any particular facial expressions. Ten participants, including two of the authors, participated in the experiment (eight males and two females), and each participant evaluated 600 images for each of the two tags.

This study was approved by the Ethics Committee of the Research Institute of Electrical Communication, Tohoku University, and was conducted in accordance with the Code of Ethics of the World Medical Association (Declaration of Helsinki).

2.2 Human performance of preference prediction

To establish a standard against which to compare the prediction performance of the machine learning model, we measured the human performance of preference prediction from face videos. We recruited six raters (five males and one female, age 22.8 ± 1.6) who were not participants in the preference evaluation experiment.

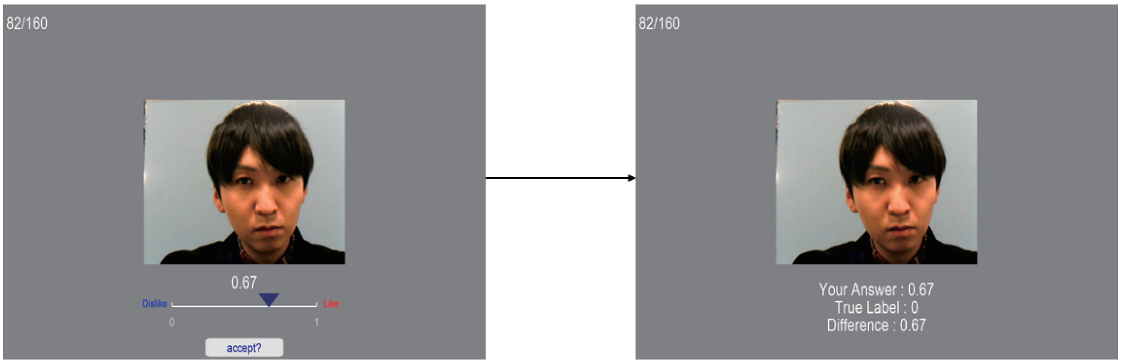
As a test face video dataset, we randomly chose ten face videos from each category and for each participant in which the participant "liked" and "not liked" an image. This procedure resulted in 400 face videos to be judged by the human raters (ten videos × two judgement categories × two image categories × ten participants). We divided each set of ten videos described above into training and test sets according to the ratio 8 : 2.

The experiment started with a training session in which the training face videos were randomly chosen and shown to the raters. A trial started with a face video from one trial in the preference evaluation experiment being presented on a gray background (Fig. 2) with a slide bar below the face video. The raters were asked to rate their confidence about whether the participant liked the image or not by moving the slider to a place between the leftmost point ("not liked" with 100% confidence) and the rightmost point ("liked" with 100% confidence). This continuous evaluation allowed us to calculate the area under the curve (AUC) score described in Sect. 3.3 and compare the performance of human raters and the machine learning model. In the training session, the raters' rating, the true label, and the distance between their rating and the true label were presented to them as feedback. After 160 trials of the training session, the raters proceeded to the test phase of 40 trials. In the test phase, the experimental settings were almost the same as in the training session, but without the feedback of the true label.

After the experiment, the raters were asked to answer two questions on a 5-point scale: "How difficult was the task?" and "How well did you learn in the training phase?"

3. Analysis

In this study, we constructed a binary classification model for predicting the image preference. We compared different sets of input features to predict the preference: facial expression features extracted from face videos during judgements, image features extracted from presented stimulus images, and a combination of both sets of features. Additionally, we performed importance analysis to investigate the facial expression features that were most important in predicting the preference ratings.



Your Answer : 0.67
True Label : 0
Difference : 0.67

Fig. 2. Example screens in the experiment to measure human performance of preference judgements. A face video from a trial in the preference evaluation experiment was presented (left panel), and the raters moved the slide bar below the face video to indicate their confidence about whether the person in the video "liked" the image they were seeing. In the training session, feedback about the true label was given to the rater (right panel).

3.1 Extraction of facial expression features

The Facial Action Coding System (FACS) [29] is the most widely used framework for analyzing facial expressions. In FACS, the movements of face parts are expressed as action units (AUs) to characterize the elements of facial expressions. We used open-source OpenFace [30] software to extract AUs from each face image frame of the video recorded during the preference evaluation. OpenFace outputs two types of AUs: quantities that represent the intensity of each AU on a continuous scale from 0 to 5 (AUr) and binary quantities that represent whether AUs exist in the image (AUc). OpenFace outputs 17 types of AUs for AUr and 18 types of AUs for AUc. For each trial, we used the face video from the time of stimulus presentation until the start of the next stimulus presentation. By applying OpenFace to the images of each frame of the face video, we obtained time series data of the changes in the AUs in each trial. In this study, to summarize the dynamic features of the time series data, we calculated seven features from the time series of AUr for each trial: the mean, minimum, maximum, standard deviation, and three quartiles. Similarly, for AUc, we calculated the mean and standard deviation for each trial. Combining these facial features, we extracted a total of 155 (119 for AUr plus 36 for AUc) dimensions of facial expression for each trial.

3.2 Image features of presented images

As a baseline to compare predictive performance from facial expressions, we also tested the predictions of the preference evaluation from stimulus image features. Deep learning models are powerful models that can capture the complex features of input images. We used a ResNet-50 model [31], one of the most widely used deep learning models, to extract high-level image features from the stimulus images presented to the participants during the trials. To obtain the image features related to image preference evaluation, we trained a ResNet-50 model to predict the preference judgement of the participants. We used the pre-trained ResNet-50 model that was trained using the ImageNet dataset, but replaced the last fully connected layer with a fully connected layer with two output units that represented the two classes of the participant's response (liked or not liked). We used the weights of the convolutional layers in the pretrained model as the initial state of our model and fine-tuned the entire network to fit the data obtained in our experiment. We divided the data into training and validation sets, as explained in the next section. We resized the input stimulus images to match the input of the ResNet-50 model. For training the stimulus images, to improve the generalization performance of the trained model, we also applied image augmentations using random horizontal flip, random rotation (up to $\pm 15°$), random translation (up to 0.1 of the image width and height), random rescaling (between 0.9 and 1.1), and random brightness change (between 0.9 and 1.1). We trained the network to minimize the cross-entropy loss in the last output layers using the Adam algorithm with a learning rate of $1e^{-5}$. We repeated training using the training data for 10 epochs, at which point the training effect was almost saturated. After training, we used the final network output of the convolutional layers right before the fully connected layer as the high-level image features (2048 dimensions) of the presented images.

3.3 Model for predicting image preference

We used the LightGBM model [32], which is a gradient boosting method, to estimate the preference for images from facial expression features and stimulus image features. We trained the model to predict a binary variable of whether the participants liked the image from the facial expression features and/or the stimulus image features extracted for each trial. We compared three sets of features as the input to the model and examined the difference in prediction performance that resulted from the type of features used for prediction: only facial features, only image features, and both types of features. We used the AUC score, which is a measure of the performance of a classifier, as the index of prediction performance. The machine learning model output a continuous value rather than a binary value, and we calculated the AUC score by setting various criteria to transform the continuous value to the final binary decision. We assessed the generalized performance of the model using a stratified 5-fold cross-validation analysis in which we divided the data into training (4/5) and validation (1/5) sets while maintaining the distribution of the target variable in the original data for each of the training and validation target distributions. For prediction based on image features, we first used the training data in each cross-validation set to train the ResNet-50 model, and then used the high-level image features as the input image features for both the training and evaluation of the LightGBM model.

3.4 Importance analysis of facial image features

To identify the facial image features that contributed to the preference prediction by the machine learning model, we conducted importance analysis using Shapley additive explanations (SHAP) values [33]. The prediction of the model can be approximated using SHAP values as

$$\text{Prediction}_i = \varphi_0 + \sum_{j=1}^{M} \varphi_j, \tag{3.1}$$

where Prediction_i is the predicted value for sample i, φ_0 is the mean of the predicted values, M is the number of features, and φ_j is the contribution of feature j to the prediction. The sign and magnitude of φ_j represent the effect of each feature on the prediction. The average importance of a feature p can be calculated by taking the average of φ_j over samples as

$$\text{FeatureImportance}_p = \frac{1}{N}\sum_{i=1}^{N} |\varphi_p^i|, \tag{3.2}$$

where $\text{FeatureImportance}_p$ is the average importance of feature p, N is the number of samples in the data, and φ_p^i is the contribution of predicted sample i to feature p.

4. Results

4.1 Distribution of participants' responses in the preference evaluation experiment

Table 1 presents the distribution of the preference judgements of all participants in the experiment. The total number of data samples is less than the expected 6,000 (600 trials × 10 participants) because we excluded data with missing feature values. For both the "lunchbox" and "landscape" tags, the "liked" and "not liked" proportions were nearly 50%, which indicates that there was no significant bias in the distribution of the preference judgement data.

Table 1. Distribution of the preference judgements of all participants.

Tag	Number of data	Proportion of "liked" (%)	Proportion of "not liked" (%)
lunchbox	5921	53	47
landscape	5276	51	49

4.2 Predictive performance of the model and human raters

Figure 3 compares the predictive performance of the model from different feature sets, that is, facial expression features, image features, and both features, for each of the image categories, together with the performance of the human raters. The facial features achieved prediction performance comparable with, and for lunch box images, marginally better than, the image features specifically trained for our dataset. The predictive performance using both features was not significantly different from when only facial features were used, which means that most of the predictive power comes from facial features. The performance of the human raters was nearly the same as, or slightly worse than, the model performance using only image features. In particular, for the landscape images, the performance of the human raters was not significantly different from the chance level (1 sample t-test, $p = 0.74$).

4.3 Importance analysis of facial expression features

Next, we conducted importance analysis to investigate the facial expression features that contributed to the preference prediction by the machine learning model using the SHAP values. We used the LightGBM model trained to

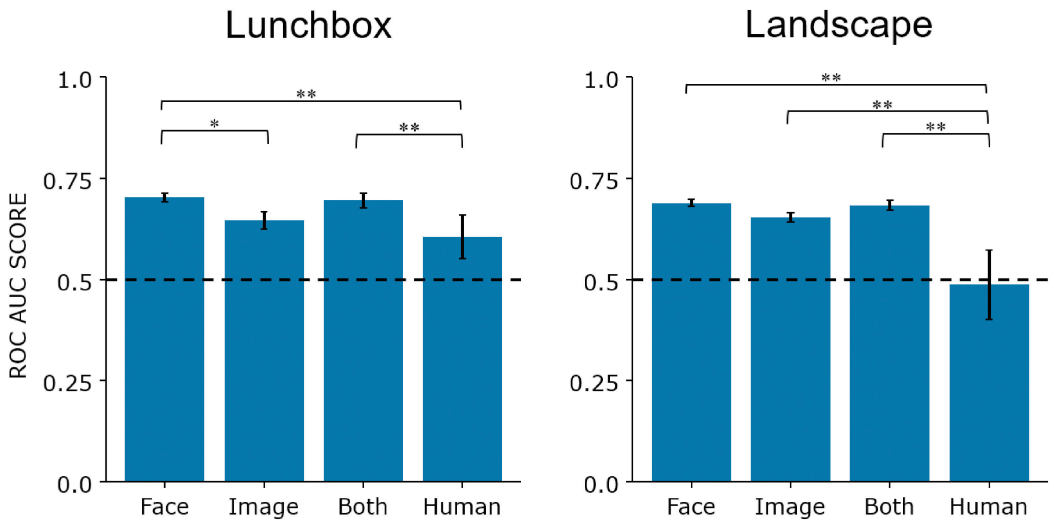


Fig. 3. Predictive performance of the model using different feature sets and human raters. The performance of the model is compared for facial expression features, image features, and both features as input. The vertical axis is the AUC score as the measure of predictive performance. The error in the model results represents the standard deviation across the cross-validation sets, and the error in human performance is the standard deviation across all raters. The dotted line represents the chance level of 0.5. Pairs that were significantly ($p < 0.05$) and marginally different ($p < 0.1$) according to Tukey-HSD multiple comparison tests are indicated by ** and *, respectively.

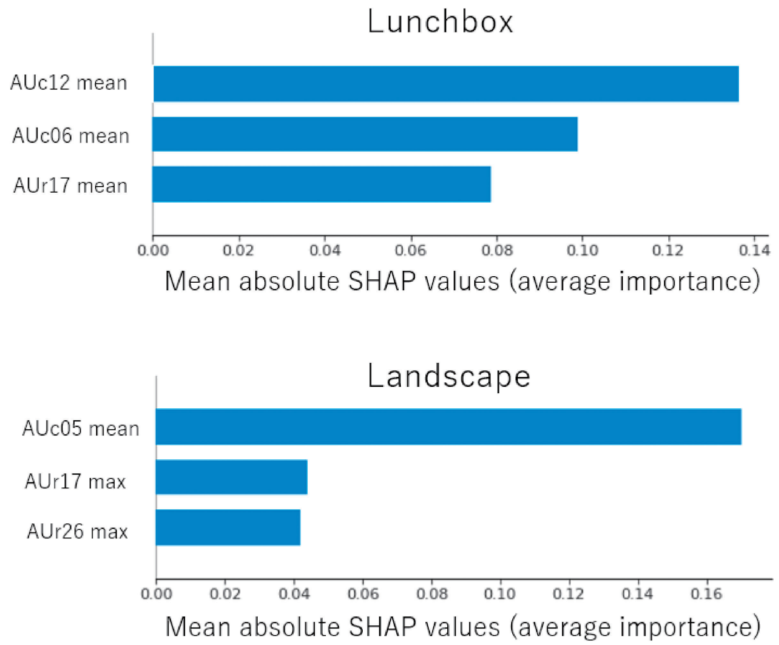


Fig. 4. Top three facial expression features in terms of their contribution to the prediction of human preference judgements for lunchbox (top) and landscape (bottom) image categories. The horizontal axis represents the average absolute value of the SHAP values across all training samples, as defined in Eq. (3.2), which is the measure of the average contribution of the feature to the prediction.

predict human preferences from only facial features for the analysis. Figure 4 shows the contribution of the top-three features out of 155 facial expression features for the lunchbox and landscape image categories. The results show that different facial features made different contributions to each image category. For the lunchbox images, AU12 (Lip corner puller) made the largest contribution, followed by AU06 (Cheek raiser) and AU17 (Chin raiser). For the landscape image, the contribution of AU05 (Upper lid raiser) was dominant, followed by minor contributions from AU17 and AU26 (Jaw drop).

5. Discussion and Conclusion

In this study, we developed a method for estimating human preferences for images using a machine learning model that examined facial videos taken while the users were evaluating the images. We conducted a preference judgement experiment using two image categories. We also conducted an experiment to assess the human performance of predicting preference judgements from face videos to compare the predictive performance of human raters with that of the proposed machine learning model.

The results showed that the facial features achieved prediction performance comparable with, and for lunch box images, marginally better than, the image features specifically trained for our dataset. We also found that our model performed better at predicting image preferences than the human raters. Because we did not ask the participants in the image preference experiment to make deliberate facial expressions, most of the participants' facial expressions could not be distinguished by the human raters. Our results showed that the preference judgements predicted by the human raters had a low success rate. For the landscape images, in particular, human performance was at the chance level, whereas the machine learning model predicted preferences as accurately as for the lunchbox images.

It should be noted that we trained the machine learning model using approximately 4,000–5,000 samples in each cross-validation set, whereas the human raters practiced with 160 trials for each image category. It could be argued that a direct comparison between the performance of the machine learning model and human raters may not be fair. In the post-experiment questionnaire, we asked the human raters two questions: "How difficult was the task?" and "How well did you learn in the training phase?" All the raters rated the task difficulty as 4 or 5, which means that the task was difficult for all of them. However, the answers to the learning performance question varied across raters. We calculated the ROC AUC score for each rater from their responses and averaged the score across two image categories, and examined the relationship between the mean AUC score and raters' self-report about learning performance (Fig. 5). There was a significant and strong correlation between these two quantities. In future work, it will be interesting to investigate whether these individual differences diminish with more intensive training or whether they reflect the diversity in the raters' cognitive ability to predict preference from facial expressions.

The participants' facial expressions changed dynamically during a trial, and the temporal change could contain some useful information about their preference judgement. In our analysis, we calculated some statistics of the dynamics of

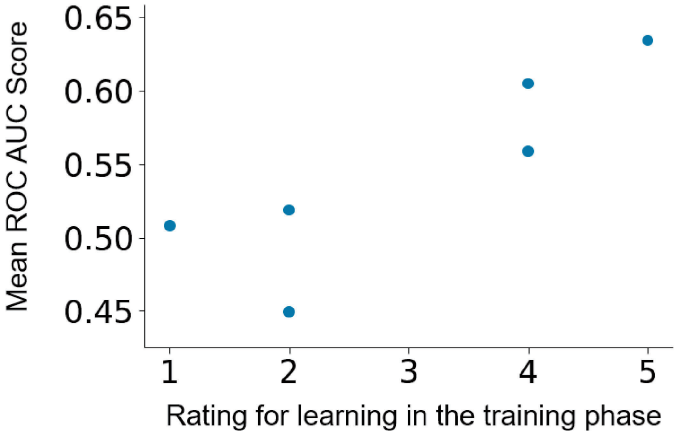


Fig. 5. Relationship between the rating for learning in the training phase (horizontal axis) and the mean ROC AUC score (vertical axis) in the human preference judgement experiment. They were significantly correlated ($r = 0.85$, $p = 0.03$).

expression from the entire duration of the face video in the trial. Humans show brief emotional facial expressions shorter than half a second, particularly when they try to conceal their true emotion [34]. There have also been recent advancements in automatically detecting micro expressions [35]. Because our current analysis method cannot capture such a short change of facial expressions, predictive performance could be further improved by incorporating techniques to detect micro expressions.

We also examined the importance of the facial expression features used by the model to predict preference judgements. We showed that the important facial features differed between the two image categories, that is, lunchbox images and landscape images. For lunchbox images, AU12 (Lip corner puller) and AU06 (Cheek raiser) were important facial expression features. These features are related to smiling and associated with positive emotions, such as happiness and joy [36, 37]. For the landscape images, AU05 (Upper lid raiser) was dominant and is associated with rather negative emotions, such as fear, anger, and surprise.

This difference in the associated emotions between the two image categories might explain the poor performance of the human raters, particularly the almost-chance level performance in the landscape category, because humans recognize happy facial emotions better and more consistently than other emotions [24, 38]. It is intuitive to assume that people exhibit happy face features when they prefer the images they are looking at, as we observed for the lunchbox image category. However, the major contribution of AU05 for the landscape images is counterintuitive. One possible explanation is that people exhibit feelings of awe, a complex emotion that is most closely related to fear among the basic emotions, when exposed to natural scenes [39, 40]. Another possibility is that novelty, which is supposedly related to surprise, is important for the preference judgement of natural scenes [26].

These results show that, although the proposed model can predict the preference judgements for images in different categories with nearly the same level of accuracy, there are significant differences in the facial expression features that are useful for making predictions. This suggests that we need to be careful when generalizing a model trained on a dataset with limited image categories to predict the preferences for images in categories that were not in the trained data. Our machine learning approach also suggests that the cognitive mechanism used for preference judgements may be different for different image categories, in line with previous research. In our experiment, we used lunchbox and landscape image categories. Although landscape seems to be a broad image category, lunchbox is a relatively narrow category that is included in broader categories such as food. Future work is required to investigate the image category broadness to which we can generalize our results and at what broadness level the difference in relevant facial features appears.

To summarize, we have attempted to predict human preference judgements for images using facial images recorded during the judgement process and the features of the images being judged. The results verify the usefulness of implicit facial expressions. These should be added to the list of important implicit processes for human cognitive and active behaviors, such as the effect of implicit learning on the actions performed during a visual search or automatic attention to objects that are near to the subject's hands [41–43].

Acknowledgments

This study was supported by Yotta Informatics Project by MEXT, Japan.

REFERENCES

[1] Shioiri, S., Sato, Y., Horaguchi, Y., Muraoka, H., and Nihei, M., "Quali-informatics in the Society with Yotta Scale Data," *53rd IEEE International Symposium on Circuits and Systems, ISCAS 2021*, Institute of Electrical and Electronics Engineers

Inc. (2021) doi: 10.1109/ISCAS51556.2021.9401161.

[2] Muraoka, H., *et al.*, "Gigantic amount information and storage technology: Challenge to Yotta-byte-scale informatics," *IEICE Tech. Rep.*, **116(441)**: 27–32 (2017).

[3] Talebi, H., and Milanfar, P., "NIMA: Neural image assessment," *IEEE Trans. Image Process.*, **27(8)**: 3998–4011 (2018) doi: 10.1109/TIP.2018.2831899.

[4] Wang, H., *et al.*, "The Evaluation of Images Based on Human Preference with Convolutional Neural Networks," *Asia-Pacific Conference on Vision 2018* (2018).

[5] McDuff, D., El Kaliouby, R., Kodra, E., and Picard, R., "Measuring Voter's Candidate Preference Based on Affective Responses to Election Debates," *Proceedings - 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction, ACII 2013*, 369–374 (2013) doi: 10.1109/ACII.2013.67.

[6] Tkalčič, M., Elahi, M., Maleki, N., Ricci, F., Pesek, M., and Marolt, M., "Prediction of Music Pairwise Preferences from Facial Expressions," *Int. Conf. Intell. User Interfaces, Proc. IUI*, vol. Part F1476, 150–159 (2019) doi: 10.1145/3301275.3302266.

[7] Lewinski, P., Fransen, M. L., and Tan, E. S. H., "Predicting advertising effectiveness by facial expressions in response to amusing persuasive stimuli," *J. Neurosci. Psychol. Econ.*, **7(1)**: 1–14 (2014) doi: 10.1037/npe0000012.

[8] Goldberg, P., *et al.*, "Attentive or not? Toward a machine learning approach to assessing students' visible engagement in classroom instruction," *Educ. Psychol. Rev.*, **33(1)**: 27–49 (2021) doi: 10.1007/s10648-019-09514-z.

[9] Pham, P., and Wang, J., "Attentive video: A multimodal approach to quantify emotional responses to mobile advertisements," *ACM Trans. Interact. Intell. Syst.*, **9(2–3)**: (2019) doi: 10.1145/3232233.

[10] Thomas, C., and Jayagopi, D. B., "Predicting Student Engagement in Classrooms Using Facial Behavioral Cues," *MIE 2017-Proceedings of the 1st ACM SIGCHI International Workshop on Multimodal Interaction for Education, Co-located with ICMI 2017* (2017) doi: 10.1145/3139513.3139514.

[11] Oodaira, K., Miyazaki, T., Sugaya, Y., and Omachi, S., "Importance estimation for scene texts using visual features," *Interdiscip. Inf. Sci.* (in press).

[12] Hatori, Y., Nakajima, T., and Watabe, S., "Body posture analysis for the classification of classroom scenes," *Interdiscip. Inf. Sci.* (in press).

[13] Shioiri, S., Honjyo, H., Kashiwase, Y., Matsumiya, K., and Kuriki, I., "Visual attention spreads broadly but selects information locally," *Sci. Rep.*, **6**: 35513 (2016) doi: 10.1038/srep35513.

[14] Carrasco, M., "Visual attention: The past 25 years," *Vision Res.*, **51(13)**: 1484–1525 (2011) doi: 10.1016/j.visres.2011.04.012.

[15] Dimberg, U., "Facial electromyography and emotional reactions," *Psychophysiology*, **27(5)**: 481–494 (1990) doi: 10.1111/j.1469-8986.1990.tb01962.x.

[16] Dimberg, U., and Karlsson, B., "Facial reactions to different emotionally relevant stimuli," *Scand. J. Psychol.*, **38(4)**: 297–303 (1997) doi: 10.1111/1467-9450.00039.

[17] Murshed, M., Dewan, M. A. A., Lin, F., and Wen, D., "Engagement Detection in e-Learning Environments Using Convolutional Neural Networks," *2019 IEEE Intl Conf on Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing, Intl Conf on Cloud and Big Data Computing, Intl Conf on Cyber Science and Technology Congress (DASC/PiCom/CBDCom/CyberSciTech)*, 80–86 (2019) doi: 10.1109/DASC/PiCom/CBDCom/CyberSciTech.2019.00028.

[18] Langner, O., Dotsch, R., Bijlstra, G., Wigboldus, D. H. J., Hawk, S. T., and van Knippenberg, A., "Presentation and validation of the Radboud faces database," *Cogn. Emot.*, **24(8)**: 1377–1388 (2010) doi: 10.1080/02699930903485076.

[19] Lucey, P., Cohn, J. F., Kanade, T., Saragih, J., Ambadar, Z., and Matthews, I., "The Extended Cohn-Kanade Dataset (CK+): A Complete Dataset for Action Unit and Emotion-specified Expression," *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*, 94–101 (2010) doi: 10.1109/CVPRW.2010.5543262.

[20] Haines, N., Southward, M. W., Cheavens, J. S., Beauchaine, T., and Ahn, W. Y., "Using computer-vision and machine learning to automate facial coding of positive and negative affect intensity," *bioRxiv* (2018) doi: 10.1101/458380.

[21] Krumhuber, E. G., Küster, D., Namba, S., Shah, D., and Calvo, M. G., "Emotion recognition from posed and spontaneous dynamic expressions: Human observers versus machine analysis," *Emotion*, **21(2)**: 447–451 (2021) doi: 10.1037/emo0000712.

[22] Höfling, T. T. A., Gerdes, A. B. M., Föhl, U., and Alpers, G. W., "Read my face: Automatic facial coding versus psychophysiological indicators of emotional valence and arousal," *Front. Psychol.*, **11(June)**: 1–15 (2020) doi: 10.3389/fpsyg.2020.01388.

[23] Höfling, T. T. A., Alpers, G. W., Gerdes, A. B. M., and Föhl, U., "Automatic facial coding versus electromyography of mimicked, passive, and inhibited facial response to emotional faces," *Cogn. Emot.*, **35(5)**: 874–889 (2021) doi: 10.1080/02699931.2021.1902786.

[24] Krumhuber, E. G., Küster, D., Namba, S., and Skora, L., "Human and machine validation of 14 databases of dynamic facial expressions," *Behav. Res. Methods*, **53(2)**: 686–701 (2021) doi: 10.3758/s13428-020-01443-y.

[25] Masip, D., North, M. S., Todorov, A., and Osherson, D. N., "Automated prediction of preferences using facial expressions," *PLoS One*, **9(2)**: 1–5 (2014) doi: 10.1371/journal.pone.0087434.

[26] Park, J., Shimojo, E., and Shimojo, S., "Roles of familiarity and novelty in visual preference judgments are segregated across object categories," *Proc. Natl. Acad. Sci. U.S.A.*, **107(33)**: 14552–14555 (2010) doi: 10.1073/pnas.1004374107.

[27] North, M. S., Todorov, A., and Osherson, D. N., "Inferring the preferences of others from spontaneous, low-emotional facial expressions," *J. Exp. Soc. Psychol.*, **46(6)**: 1109–1113 (2010) doi: 10.1016/j.jesp.2010.05.021.

[28] "Instagram," [Online], Available: https://www.instagram.com/.

[29] Ekman, P., and Friesen, W. V., Facial Action Coding System, Consulting Psychologists Press, Palo Alto, CA (1978).

[30] Baltrusaitis, T., Robinson, P., and Morency, L. P., "OpenFace: An Open Source Facial Behavior Analysis Toolkit," *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 1–10 (2016) doi: 10.1109/WACV.2016.7477553.

[31] He, K., Zhang, X., Ren, S., and Sun, J., “Deep Residual Learning for Image Recognition,” *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 770–778 (2016) doi: 10.1109/CVPR.2016.90.

[32] Ke, G., *et al.*, “LightGBM: A Highly Efficient Gradient Boosting Decision Tree,” *Advances in Neural Information Processing Systems* (2017).

[33] Lundberg, S. M., and Lee, S. I., “A Unified Approach to Interpreting Model Predictions,” *Advances in Neural Information Processing Systems* (2017).

[34] Ekman, P., and Friesen, W. V., “Nonverbal leakage and clues to deception,” *Psychiatry*, **32(1)**: 88–106 (1969) doi: 10.1080/00332747.1969.11023575.

[35] Davison, A. K., Lansley, C., Costen, N., Tan, K., and Yap, M. H., “SAMM: A spontaneous micro-facial movement dataset,” *IEEE Trans. Affect. Comput.*, **9(1)**: 116–129 (2018) doi: 10.1109/TAFFC.2016.2573832.

[36] Ekman, P., Friesen, W. V., and Hager, J. C., *The Facial Action Coding System: A Technique for the Measurement of Facial Movement*, Consulting Psychologists Press, San Francisco, CA (2002).

[37] Clark, E. A., *et al.*, “The facial action coding system for characterization of human affective response to consumer product-based stimuli: A systematic review,” *Front. Psychol.*, **11(May)**: 1–21 (2020) doi: 10.3389/fpsyg.2020.00920.

[38] Mollahosseini, A., Hasani, B., and Mahoor, M. H., “AffectNet: A database for facial expression, valence, and arousal computing in the wild,” *IEEE Trans. Affect. Comput.*, **10(1)**: 18–31 (2019) doi: 10.1109/TAFFC.2017.2740923.

[39] Keltner, D., and Haidt, J., “Approaching awe, a moral, spiritual, and aesthetic emotion,” *Cogn. Emot.*, **17(2)**: 297–314 (2003) doi: 10.1080/02699930302297.

[40] Shiota, M. N., Keltner, D., and Mossman, A., “The nature of awe: Elicitors, appraisals, and effects on self-concept,” *Cogn. Emot.*, **21(5)**: 944–963 (2007) doi: 10.1080/02699930600923668.

[41] Shioiri, S., Kobayashi, M., Matsumiya, K., and Kuriki, I., “Spatial representations of the viewer’s surroundings,” *Sci. Rep.*, **8(1)**: 7171 (2018) doi: 10.1038/s41598-018-25433-5.

[42] Reed, C. L., Stone, V. E., Grubb, J. D., and McGoldrick, J. E., “Turning configural processing upside down: Part and whole body postures,” *J. Exp. Psychol. Hum. Percept. Perform.*, **32(1)**: 73–87 (2006) doi: 10.1037/0096-1523.32.1.73.

[43] Shioiri, S., Sasada, T., and Nishikawa, R., “Visual attention around a hand location localized by proprioceptive information” (in preparation).