ARTIFICIAL INTELLIGENCE RATERS: NEURAL NETWORKS FOR RATING PICTORIAL EXPRESSION

THOMAS GENGENBACH
Karlsruhe Institute of Technology
thomas.gengenbach@gmx.net

KERSTIN SCHOCH
University of Applied Sciences and Arts Ottersberg,
The Pop-up Institute
kerstin.schoch@hks-ottersberg.de

https://doi.org/10.34632/jsta.2022.10196

Journal of Science and Technology of the Arts, vol. 14, n. 1 (2022): pp. 49-71

ABSTRACT

Previous studies on classification of fine art show that features of paintings can be captured and categorized using machine learning approaches. This progress can also benefit art psychology by facilitating data collection on artworks without the need to recruit experts as raters. In this study a machine learning approach is used to predict the ratings of RizbA, a Ratinstrument for two-dimensional pictorial works. Based on a pre-trained model, the algorithm was fine-tuned via transfer learning on 886 pictorial works by contemporary professional artists and non-professionals. As quality criterion, artificial intelligence raters (ART) are compared with generic raters (GR) created from the real human expert raters, using error rate and Mean Squared Error. ART ratings have been found to have the same error range as randomly chosen human ratings. Therefore, they can be seen as equivalent to real human expert raters for almost all items in RizbA. Further training with more data will close the gap to the human raters on all items.

Keywords: Artificial intelligence raters; Machine learning; Neural nets; Pictorial expression; Visual art.

INTRODUCTION

The last decade has been marked by advances in digitization, artificial intelligence, and big data. Progress in these fields stems mainly from the huge amount of computational capacity that is available in big computing clusters. These technologies are now finding their way into new fields. Many online collections of artworks have been created in the past decades with detailed annotations from art experts comprising genre, subject, style, technique, artist etc. They can be used to train machine learning algorithms to automate these categorization tasks and annotate further image catalogues (Cetinic et al., 2018; Santos et al., 2021).

The breakthrough in computer vision happened in 2012 (Krizhevsky et al., 2012). It opened up new possibilities to use the algorithms for a range of classification tasks, thanks to the Open Source availability of the pre-trained deep learning neural nets. They are applicable for comparatively small amounts of annotated training data and require relatively little computational power using the transfer learning approach (Razavian et al., 2014; Rodriguez et al., 2018; Sarkar et al., 2018). To date, this progress has been mainly used for digitization in art history and art science (e.g., Banerji & Sinha, 2016; Cetinic et al., 2018, 2020; Hua et al., 2020; Madhu et al., 2019), but can also benefit and advance art psychological research and documentation (Amirshahi et al., 2014; Castro et al., 2014).

In this work a machine learning approach is used to predict the ratings of the Rating instrument for two-dimensional pictorial works (RizbA; Schoch, 2020, April 24; Schoch et al., 2017), an art psychological questionnaire measuring pictorial expression. Its goal is to model and train different neural nets for each item of the rating scale. To compare the different artificial intelligence raters (ART), generic raters (GR) are built up from real human expert raters of the original training data stemming from previous validation studies on RizbA (Schoch & Ostermann, 2020, 2021, July 13, 2022). To evaluate the results, we compare the different ARTs with different GRs with respect to error rate and Mean Squared Error (MSE).

The resulting application can be used for image analysis in art-related research and documentation, such as art psychology. It facilitates various possibilities for inquiries related to visual art that involve large amounts of numeric measurements of images. Its application allows a relatively easy rating experience without expert raters having to go through the tedious process of manually rating 26 items per picture. It makes research and documentation more efficient while garnering large sample sizes and retaining statistical power.

PREVIOUS RESEARCH

The Rating instrument for two-dimensional pictorial works (RizbA)

RizbA is a 26-item questionnaire referring to formal picture analysis (Streb, 1984; Stuhler-Bauer & Elbing, 2003). The scale assesses pictorial

https://doi.org/10.34632/jsta.2022.10196

expression, which is defined as artistic creation in the form of a picture, whereby aiming for a maximization of objectivity. The instrument focuses on the concept of a formal picture analysis (Bauer, 1996) analyzing formal aspects like representation, color, shaping, spatiality, motion, composition, and expression which can be found across art literature (e.g., Arnheim, 2000; Bauer, 1996; Kandinsky, 1955; Meyer, 2011; Vollmar, 2008). This approach is rooted in the tradition of phenomenological picture analysis (Betensky, 1991; Stuhler-Bauer & Elbing, 2003, 2004), seeking to overcome accidental judgement, preconception, and association (Streb, 1984).

The scale (see Table 1) uses a six-point Likert scale, which is discretely scaled and verbally anchored in shades of agreement (0 = *strongly disagree*, 5 = *strongly agree*). Raters using the questionnaire receive a brief instruction to rate the presented image using the RizbA questionnaire. They are asked to focus on the predominant overall expression of a picture and no single details, while endorsing there is no right or wrong, but only their evaluation.

Table 1

RizbA items: English translation and original German version.		
No.	**English translation**	**Original version**
1	The picture includes graphic elements	Das Bild enthält zeichnerische Elemente
2	The picture includes pictorial elements	Das Bild enthält malerische Elemente
3	The manner of representation is concrete	Die Darstellungsweise ist gegenständlich
4	The manner of representation is abstract	Die Darstellungsweise ist abstrakt
5	The color application is pastose	Der Farbauftrag ist pastos
6	The predominant coloring is vibrant	Die vorherrschende Farbgebung ist leuchtend
7	In the picture primary colors are prevalent	Im Bild befinden sich vorwiegend reine Farben

https://doi.org/10.34632/jsta.2022.10196

RizbA items: English translation and original German version.		
8	In the picture mixed colors (secondary colors) are prevalent	Im Bild befinden sich vorwiegend Mischfarben (Sekundärfarben)
9	In the picture there are complementary contrasts	Im Bild sind Komplementärkontraste vorhanden
10	In the picture organic shapes are prevalent	Im Bild enthaltene Formen sind vorwiegend organisch
11	In the picture geometric shapes are prevalent	Im Bild enthaltene Formen sind vorwiegend geometrisch
12	The layout of the line is predominantly curved	Die Linienführung verläuft vorwiegend gebogen
13	The layout of the line is predominantly angled	Die Linienführung verläuft vorwiegend gebogen
14	The picture includes unworked areas	Das Bild enthält unbearbeitete Flächen
15	The picture appears to be deep	Das Bild wirkt tief
16	The picture is perspectival	Das Bild ist perspektivisch
17	The picture is without perspective (aperspectival)	Das Bild ist frei von Perspektive (aperspektivisch)
18	The picture is restless	Das Bild ist unruhig
19	The picture is wild	Das Bild ist wild

https://doi.org/10.34632/jsta.2022.10196

RizbA items: English translation and original German version.		
20	The global composition is laid out vertically	Die Gesamtkomposition ist senkrecht angelegt
21	The global composition is laid out horizontally	Die Gesamtkomposition ist waagrecht angelegt
22	The global composition is laid out diagonally	Die Gesamtkomposition ist diagonal angelegt
23	The global composition is laid out area-wide without a main subject (All-Over-Structure)	Die Gesamtkomposition ist flächendeckend ohne Hauptmotiv (All-Over-Structure)
24	The picture appears to be diffuse	Das Bild wirkt diffus
25	The picture appears to be precise, accurate	Das Bild wirkt präzise, exakt
26	The picture appears to be harmonic	Das Bild wirkt harmonisch

RizbA enables a reliable quantitative capture of artworks and is validated for use on two-dimensional pictorial works such as sketches, drawings, paintings, collages, and mixed techniques. It allows a more detailed examination of psychological correlates of art, such as aesthetic appreciation, personality, or clinical diagnoses. As thoroughly discussed in previous literature (Schoch & Ostermann, 2020) it can be used for a variety of art psychological applications related to art reception, aesthetic appreciation, artmaking, and creativity, in both practice and research.

The test was developed in a pilot study (Schoch et al., 2017; Schoch & Ostermann, 2022) and successfully validated in three studies with samples of pictorial works by non-professionals (Schoch & Ostermann, 2022), contemporary artworks by professional artists (Schoch & Ostermann, 2020) and a mixed sample of both (Schoch & Ostermann, 2021, July 13). The data collected in these studies – image material and RizbA ratings for each image – serve as training data for the neural net.

https://doi.org/10.34632/jsta.2022.10196

Machine learning algorithms for 2D image recognition

Previous studies on classification of visual art show that certain features of fine art paintings can be captured and categorized (e.g., Banerji & Sinha, 2016; Cetinic et al., 2018, 2020; Hua et al., 2020; Madhu et al., 2019; Tan et al., 2016). These works focus on artist, genre, style, and time period and were trained on publicly available datasets like the WikiArt (2021) dataset. Other studies focus on the difference between art and non-art (Brachmann et al., 2017) and the aesthetics of artworks (Cetinic et al., 2018; Zhao et al., 2020). Most approaches use transfer learning (Razavian et al., 2014; Rodriguez et al., 2018; Sarkar et al., 2018) meaning that they take neural nets like VGG16 (Simonyan & Zisserman, 2014) as base network and fine-tune the last layers to learn new object classes.

The pre-trained neural nets that serve as a base are trained on the ImageNet (Deng et al., 2009) dataset with around 1.2 million tagged images. The dataset originates from the ImageNet Large Scale Visual Recognition Challenge (ILSVRC; ImageNet, 2012, 2016), a yearly competition in which a set of 100,000 test pictures have to be classified into 1,000 categories. This object recognition task for 2D images had its breakthrough with a Convolutional Neural Network (CNN) called AlexNet (Krizhevsky et al., 2012). The error rate was enhanced from 25.8% to 16.4% (ImageNet, 2012) and since 2016 it is less than 3% (ImageNet, 2016). These base networks are Open Sourced and available for download and usage.

METHOD

Experimental setup and CNN architecture

In the current approach the VGG16 (Simonyan & Zisserman, 2014) net is used. The architecture of this model is as follows: The first layer (or input layer) uses the RGB color values of each pixel in the image as input. This is followed by 13 convolutional layers and five pooling layers. These layers comprise the base of the neural network. The header comprises three fully connected layers at the end with dropout layers in between for regularization. For the base network, pre-trained weights are used that result from the training on the ImageNet dataset. Using these, the CNN can be considered as already having some knowledge about images, for instance detecting structures.

DATA SELECTION

Image material

The neural nets were trained on the image material collected during three previous validation studies on RizbA (Studies A, B, and C; see Table 2; availability of Open Data see Appendix). The image sample involved a total of 886 pictures. Of these, 461 are artworks by contemporary professional artists and 426 are pictorial works by non-professionals.

https://doi.org/10.34632/jsta.2022.10196

Non-professional artworks were defined as pictures created by a person without professional education in visual art or related disciplines. This means that contributors never received a professional art training, neither in terms of an academic degree nor having ever studied a subject related to visual arts. Some of these images were collected in an attendance study and digitized. Others were gathered via an online call to submit photographed, digitized pictures.

Professional artworks were defined as pictures created by international contemporary professional artists from different geographic regions, including Europe, Asia, Africa, Northern, and Southern America, dating from 2005 to 2020. The digital images were systematically retrieved from prometheus (prometheus, 2021; Simon & Verstegen, 2004), a distributed digital image archive which connects databases from various institutes, research facilities, museums and from WikiArt (2021), an open visual art encyclopedia, using the WikiArt API (WikiArt, 2020).

Real human expert raters

As an initial dataset, the according image ratings gathered in the three previous studies (Studies A, B, and C; see Table 2) were used, in which real human experts rated the image material using the RizbA scale. Inclusion criterion for raters was to have expertise in visual arts. Thus, only experts with an academic degree related to visual arts, or students who have been studying an art-related subject for at least one year, could participate. Studies A and B were conducted in a test-retest design. Since the retest data only served as a measure for test-retest reliability, only the first ratings were used.

Table 2

Collection of image material and RizbA ratings: Previous studies on RizbA

		Study A (Schoch & Ostermann, 2022)	**Study B** (Schoch & Ostermann, 2020)	**Study C** (Schoch & Ostermann, 2021, July 13)	
Images	*N*	294	318	131	143
	created by	non-professionals	professional artists	non-professionals	professional artists
	gender (coded)	107 female 39 male one diverse	59 female 233 male 1 diverse 25 unknown	104 female 25 male 2 diverse	no metadata available
	age	17 - 71 years (M = 33.86, SD = 14.02)	no metadata available	21 - 79 years (M = 45.80, SD = 15.12	no metadata available
	data collection	attendance study	prometheus	online submission	WikiArt

https://doi.org/10.34632/jsta.2022.10196

Collection of image material and RizbA ratings: Previous studies on RizbA				
Real human expert raters	*N*	880	506	179
	gender (self-identified)	750 female 122 male 8 diverse	432 female 66 male 8 diverse	159 female 16 male 4 diverse
	age	18 - 74 years (*M* = 31.59, *SD* = 10.18)	19 - 77 years (*M* = 32.92, *SD* = 11.68)	22 - 72 years (*M* = 37.59, *SD* = 12.51).
	expertise	academic degree 65.5 % studying 34.5 %	academic degree 63.6 % studying 36.4 %	academic degree 75.42 % studying 24.58 %
	disciplines	art history 30.6 % art therapy 16.8 % art pedagogy 13.0 % visual art 11.5 % graphic design 5.3 % design 5.3 % art history and aesthetics 4.8 % restoration 1.3 % other 11.4 %	art therapy 24.9% art history 19.2 % visual art 14.6 % art pedagogy 12.8 % graphic design 4.7 % design 4.7 % art sciences 3.4% restoration 2.4 % image sciences 0.4 % other 12.8 %	art therapy 31.84 % art education 15.64 % art history 13.41 % fine arts 8.94 % art science 5.59 % graphic design 5.59 % restoration 5.03 % design 3.35 % other 10.61 %
Ratings per image		16 - 31 (M = 20.35, SD = 5.23)	1 - 19 (M = 9.54, SD = 2.50)	0 - 7 (M = 3.91, SD = 1.39)
M = mean, N = sample size, SD = standard deviation				

Generic raters (GR)

Since no individual human rater had rated all 886 images, different human ratings were combined to have full sets of ratings for all images. Assuming a normal distribution of the ratings, it is canonical to consider the mean value over the different human ratings of each item for each image to be the true value. Consequently, the mean was used to train the CNN. Additionally, five generic raters were constructed to evaluate the results:

- Random rater (RR): takes a random value between 0 and 5
- Extreme rater (ER): takes always the minimum or maximum of the ratings from previous studies, whichever value is further away from 2.5, the mean of the used Likert scale
- Over rater (OR): takes either the 25% or 75% percent quartile, whichever is further away from the mean
- Under rater (UR): takes the 25% or 75% value closer to the mean
- Mean rater (MR): takes the mean value.

Journal of Science and Technology of the Arts, vol. 14, n. 1 (2022): pp. 49-71 https://doi.org/10.34632/jsta.2022.10196

To further evaluate the results to human ratings, four random raters (RR1 to RR4) were created. They were constructed by randomly picking one of the given values from the real human expert ratings for each picture and each item.

TRAINING

For the training of the neural network the data was split up into three sets: a training set with 690 images, a validation set with 173 images, and a test set with 23 images. The training and validation set were used during the training period, while the test set was used for the performance measurement. The test set was randomly chosen prior to the data split. The code is written in Python (2021) using the Keras framework (Chollet, 2018) and is available as Open Source (Gengenbach & Schoch, 2021). Each value on the six-point Likert scale $N = 0, .., 5$, was defined as a separate class and the sparse categorical crossentropy (Goodfellow et al., 2016, p. 62f.) was used. Consequently, one neural net per item had to be trained, i.e., a total of 26 different neural nets. For the training a GeForce RTX 2080 Ti (rev a1) with 11019 MiB memory graphics card from NVIDIA was used.

Parameter setup

The parameter setup of the neural net was as follows: For the net architecture a base net *VGG16* with *ImageNet weights* and a dense net with *128 / 32 / 6 with 0.5 dropout layers* (Goodfellow et al., 2016, p. 258–268) in between to reduce overfitting was used. As activation function, *relu* (Goodfellow et al., 2016, p. 193-195); and on the last layer *softmax* (Goodfellow et al., 2016, p. 184-187) to obtain probabilities for each of the classes was used. *Adam* with learning rate 10^-4 (Goodfellow et al., 2016, p. 303–305) was used as optimizer and *sparse categorical crossentropy* as loss function. The *accuracy* interpreted as *1-error_rate* was used as metric.

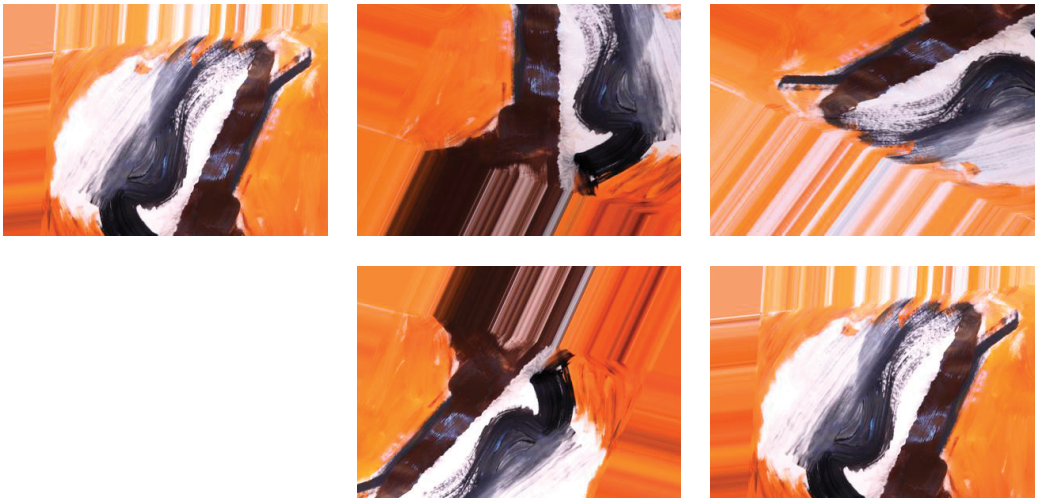
Augmentation

Image augmentation is used to achieve a better performance and prevent overfitting of the neural networks when only a small image set is available for training. Augmentation artificially creates more training images by creating randomized transformations of the existing images (Chollet, 2018). The image data was augmented as illustrated in Figure 1 using the following settings: *Rotation* up to 30°, *pixel shift* width and height range up to 0.25 (up to 25% of the image up or down, left or right), *shear range* up to 20°, *zoom range* between 0.8 and 1.2, *channel shift* range up to 0.2, *horizontal* and *vertical flipping*, *brightness range* between 0.8 and 1, *scaling* of the images to 224x224 pixels. *Color values* were normalized by 1/255 to be in the range [0,1]. The *fill in mode* ‘nearest’ was used for the new pixels that come in, due to the used augmentation.

https://doi.org/10.34632/jsta.2022.10196

Figure 1

Augmentation example: top left: original image from the sample next to four augmentations of the same image



Steps

The neural networks, one separate neural network for each of the 26 items, were trained and validated in two steps: The first step with 250 epochs had all VGG16 layers frozen, i.e., the pre-trained weights stay unchanged. The training therefore modified only the head, the dense (fully connected) part of the neural net. Only the best model according to the validation accuracy was stored throughout all epochs. The second step consisted of 350 epochs with the last block of VGG16 unfrozen, i.e., the weights can be adapted, if necessary. As a starting point for the second step, the best model from the first step was chosen. Again here, only the best model according to the validation accuracy was stored throughout all epochs.

Performance values

The test set, comprising 23 randomly chosen images, was used to compare the results of the trained artificial neural network to the GR and real human expert ratings. Their performance was reported using the following key performance indicators:

- Error rate in %: The percentage of errors, i.e., the prediction of the Likert scale value not matching the assumed true value
- MSE: The average squared difference between the predicted Likert scale value and the assumed true value
- Rank: The ranking of the error rates and the MSE, individually, from lowest to highest values
- Combined rank: The mean value of the error rate rank and the MSE rank from lowest to highest value; in case of a tie, the best individual rank was prioritized

https://doi.org/10.34632/jsta.2022.10196

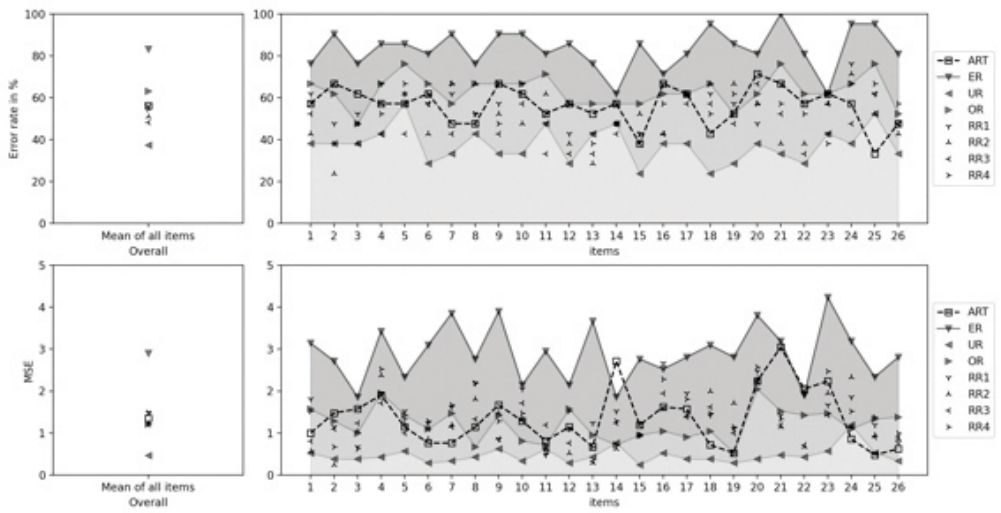
RESULTS

On the test set, ART achieved a combined rank of 6 out of 9. The error rate is 56.23%, which is rank 6 out of 9; the MSE is 1.35, which is rank 5 out of 9. The performance in relation to the other raters with exact values is reported in Table 3. A more detailed view on each item‘s neural network with error rate and MSE is provided in Table 3. The confusion matrices (see Figure 3) show the aggregated number of hits and misses of the ART for each Likert scale value and each item‘s neural network. Values on the diagonal are hits, values next to the diagonal are misses. The further away from the diagonal, the higher the deviation from the assumed true value. The confusion matrix for all the different rater types over all items can be found in Figure 4.

Table 3

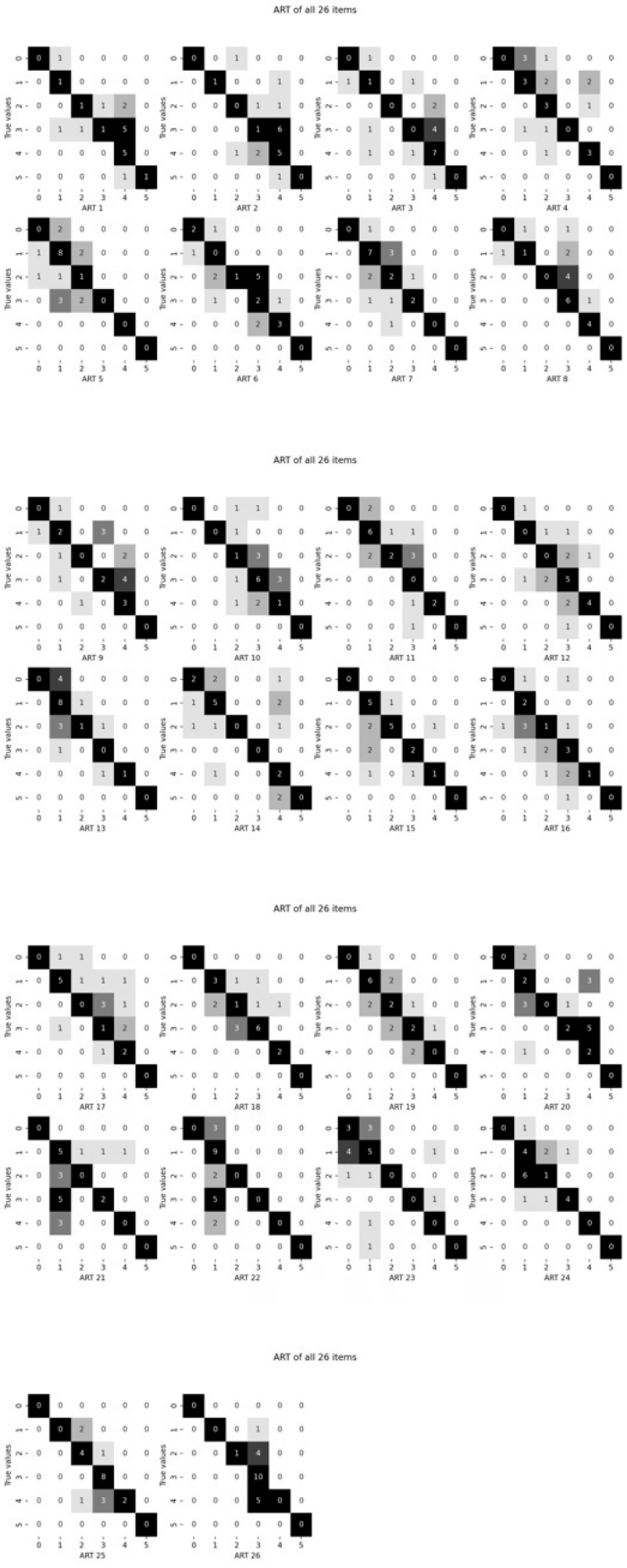
Raters' performance: Error rate, MSE and combined rank			
Rater	**Error rate (Rank)**	**MSE (Rank)**	**Combined rank**
ART	56.23 % (6)	1.35 (5)	6
RR	84.07 % (9)	4.91 (9)	9
ER	83.15 % (8)	2.90 (8)	8
OR	63.19 % (7)	1.21 (3)	4
UR	37.36 % (1)	0.47 (1)	1
RR1	56.04 % (5)	1.49 (7)	7
RR2	50.92 % (3)	1.23 (4)	3
RR3	48.17 % (2)	1.20 (2)	2
RR4	54.58 % (4)	1.48 (6)	5

Figure 2



https://doi.org/10.34632/jsta.2022.10196

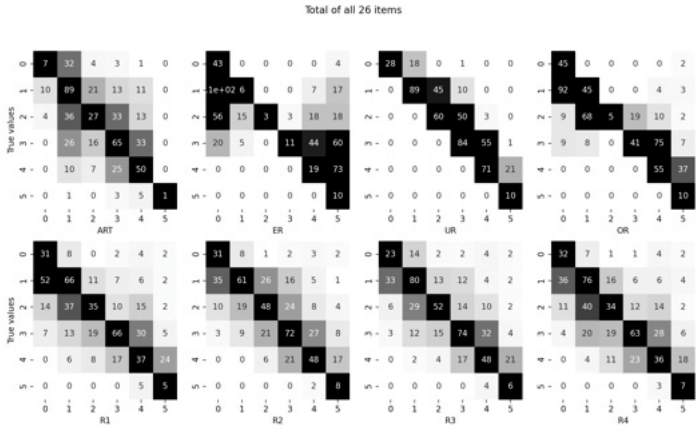
Figure 3



Confusion matrices of the ART for each item's neural network

Journal of Science and Technology of the Arts, vol. 14, n. 1 (2022): pp. 49-71 https://doi.org/10.34632/jsta.2022.10196

Figure 4



Aggregated confusion matrices for all raters

DISCUSSION

Rating with only a pre-trained model could be interpreted as equivalent to non-expert human raters rating pictures. These raters know what a picture is and what it usually looks like but lack knowledge in art history or RizbA. The fine-tuning in the transfer learning approach adds this expert knowledge to the ART.

The results of ART in the middle grey area in Figure 2 can be interpreted as 'good' in terms of error rate and MSE, since they reflect the area between the ORs and URs and are therefore in the range of the results of the human raters, i.e., RR1 to RR4. The error rate and MSE of the UR yields by construction the best result, while the OR can be considered as a lower bound for a model to be good. Only item 14 is worse in the MSE value than the ER's MSE value. Considering the error rate for item 14 being in the expected range, this is probably an artefact in the randomly selected sample of pictures taken for testing. Also, the validation error rate during training was in the same range as the validation error of the other items. Furthermore, the assumption of this artefact in the sample can be confirmed based on the confusion matrix: The assumed true values are located on both ends of the Likert scale with only a few distributed in the center. More data – especially on items with a higher MSE and error value than the average performance – will most likely close the gap to the ratings of real human experts.

In the overall confusion matrix (see Figure 3) for the ART, a tendency towards the middle is apparent, which is stronger than for RR1 to RR4. One explanation for this might be that the training objective prioritized accuracy rather than MSE. Taking only the accuracy as loss function removes the interval scale level property of the Likert scale used. Using the MSE as a term in the loss function as well, would add this property to the ART. Another explanation is the imbalanced number of images per class in the training set. However, since the data is normally distributed, a tendency towards the middle is to be expected.

https://doi.org/10.34632/jsta.2022.10196

From a machine learning point of view, the performance could still be considerably improved, (i.e., closer in terms of error rate and MSE to the assumed true value) by using either more data or tuning the neural network parameters. However, the ART is in the range of UR and OR and in between RR1 to RR4 (see Figure 2) and hence indistinguishable from the real human expert raters. Finally, with a critical view on objectivism there is no absolute truth in RizbA ratings, only an assumed true value given by the mean of the real human expert ratings. However, comparing the neural network's performance values, MSE and error rate, the results are promising and allow the conclusion that the ARTs are just as reliable as the real human expert ratings.

LIMITATIONS

Sample size

The major limitations of this work are the size and diversity of the sample. In general, a sample of 886 images is a good start for an art psychological study. For the training of neural networks using the transfer learning approach, samples of 100 images per class are supposed to be sufficient. However, the topic is highly debated and depends on the task at hand (Balki et al., 2019). Since the Likert scale ratings of RizbA are normally distributed for each item and the current sample has a minimum of 70 images per class, about 1200 images would make this neural network approach more reliable.

Image augmentation was used to mitigate the small sample size. However, augmentation itself introduces variations to the images which might in turn affect the ratings. Rotating an image for example influences the horizontal and vertical layout of the image. Not introducing augmentation would however have led to an overfitted neural net. Since the augmentations were applied randomly on images and different in each epoch, we can expect that this effect cancels itself out.

Image material

The non-professional pictorial works in study A were collected in an attendance study in Germany and therefore, by definition, they cannot be globally representative. In study C an online call in German and English language was launched and internationally distributed, which resulted in an international sample. Still, using only two languages is an exclusion criterion for participants speaking only other languages.

The images by professional artists were collected via two digital databases: prometheus and WikiArt (data sets see Appendix: Open Data). Prometheus (Simon & Verstegen, 2004) includes works by international artists, but, being based in Germany, it is still biased towards European art. Also 73.4 % of the artists names are coded male, only 18.4 % female, and 0.3 % diverse. For the remaining 7.9 % coded gender is not assignable, based on the available metadata. WikiArt (2021) is a shared knowledge database, open for use by anyone without GLAM institutions (galleries, libraries, archives, museums) serving as gatekeepers.

https://doi.org/10.34632/jsta.2022.10196

Theoretically speaking, this should enable more data diversity. Practically speaking, however, the metadata lacks the artists' self-identified gender and other diversity indicators (Nishikawa-Pacher et al., 2021). Unfortunately, the gender and diversity bias in the data reflects the current art world, which is dominated by *white* cis male Eurocentric and Northern American perspectives (Fraiberger, Sinatra, Resch, Riedl, & Barabási, 2018). This bias in the image material is inherent to the ART. Only images from artists with the same background as the training data set can be rated reliably.

Real human rater sample

The majority of raters self-identify as female, some male, and some diverse. All experts needed to be German speakers in order to participate, which resulted in most raters coming from Germany, Austria, and Switzerland. Consequently, the rater sample and as a result also the ARTs are afflicted with a Eurocentric bias.

IMPLICATIONS

The resulting application can be used for a more efficient analysis in art psychology and various other art-related fields of research, documentation, and art databases.

Firstly, it can be used to facilitate a variety of research related to visual art which involves the need for large numbers of numeric measurements of images. The application allows for a relatively easy rating experience without having to conduct the tedious process of rating 26 items per picture by many real human expert raters. Possible fields of research are for example, art psychology, in particular empirical aesthetics, but also art pedagogy, art history, art therapy.

Secondly, there are various possible fields of application. For example, it has potential as an economic tool for documenting and analyzing artworks in different application areas such as artistic processes, art education, and other art-related practices. For example, the application can be used in art therapy to consistently model the progress of therapeutic sessions on a pictorial level. In clinical practice there is little time for extensive documentation. By using ARTs this can be done with relatively little effort, providing a delta analysis for art therapy sessions. As a safety check, the recorded changes can be analyzed against the therapist's impressions. A third field of application is the systematization of art-related digital databases. By rating whole image inventories, the systematic search in databases can be enhanced. Instead of searching artworks only by keywords (e.g., epochs, artist, subject) it could also be browsed by genuinely formal characteristic features such as representation, color, or composition. Other fields of application might be computer vision, marketing, or advertisement.

https://doi.org/10.34632/jsta.2022.10196

FURTHER RESEARCH

The current training approach uses *accuracy* as metric with the *sparse categorical crossentropy* as the loss function. A different loss function using the interval scale property of the data might give a lower MSE. Furthermore, hyperparameter-optimization of the model might lead to better validation results in terms of accuracy, i.e., hitting the assumed true value. However, this will not necessarily result in a better model compared to human raters.

Further attention should be paid to the fact that the items' neural networks differ in the quality of their performance. A variance in the statistical quality of items is also reflected in the data of previous studies with human raters only. For example, items such as *The manner of representation is abstract* (item 4) provide more reliable results than items such as *The picture appears to be diffuse* (item 24; Schoch et al., 2017; Schoch & Ostermann, 2020, 2022). This is probably due to some tasks being easier to rate since they refer to more objective criteria than others. Further studies should empirically compare and theoretically discuss possible similarities and differences between human raters and neural networks when it comes to statistical quality criteria.

Since art and several psychological variables differ across geographical regions (Cattaneo, 1994), it can be assumed that there are global differences in creating and perceiving art as well. Future studies should take gender, geographical region, and other relevant diversity variables of the image material more into account with a critical view on Eurocentrism (Mosquera, 1992) and paternalistic structures. Also the real human rater sample needs to be extended to experts from all geographical regions, aiming for a decolonial coding (Ali, 2014). Since the ARTs reflect biases in image material and rater data, they must be trained on diverse samples of both. Only if neural nets are fed with diverse data, their results can be generalized.

Due to colonial continuities, European and Northern American art science still lacks truly diverse and universal perspectives, claiming universality for something that is not universal (Grant & Price, 2020). Thinking this further, not only the samples are biased but also the art theory behind the approach of formal picture analysis. In its way of thinking, RizbA refers to European theorists, such as Husserl (Uzelac, 1998), Panofsky (2006), or Mersch (2019) and is biased towards a *white* male gaze on art, which needs to be mitigated in future research. For this purpose, various global perspectives on art reception (e.g., indigenous art practices) should be considered and, if compatible, implemented in machine learning.

CONCLUSION

Although the image and human rater samples should prospectively become larger and more diverse, the results of the neural networks are

https://doi.org/10.34632/jsta.2022.10196

promising. ART ratings can be seen as equivalent to human expert raters for almost all items in RizbA. Future training with more data will probably close the gap to the human expert raters on all items and offers broad implications for art psychological research and practice.

REFERENCES

Ali, M. (2014). Towards a decolonial computing. *Ambiguous Technologies: Philosophical Issues, Practical Solutions, Human Nature, International Society of Ethics and Information Technology*, 28-35.

Amirshahi, S. A., Hayn-Leichsenring, G. U., Denzler, J., & Redies, C. (2014). *Jenaesthetics subjective dataset: analyzing paintings by subjective scores.* European Conference on Computer Vision.

Gengenbach, T., & Schoch, K. (2021, August 23). ARTificial Intelligence Raters: Neural Networks for Rating Pictorial Expression. *Open Science Framework*.
https://doi.org/10.17605/OSF.IO/QUDHX

Arnheim, R. (2000). *Kunst und Sehen: Eine Psychologie des schöpferischen Auges [Art and seeing: A psychology of the creative eye]* (3 ed.). Walter de Gruyter.

Balki, I., Amirabadi, A., Levman, J., Martel, A. L., Emersic, Z., Meden, B., Garcia-Pedrero, A., Ramirez, S. C., Kong, D., & Moody, A. R. (2019). Sample-size determination methodologies for machine learning in medical imaging research: a systematic review. *Canadian Association of Radiologists Journal*, *70*(4), 344-353.
https://doi.org/10.1016/j.carj.2019.06.002

Banerji, S., & Sinha, A. (2016). Painting classification using a pre-trained convolutional neural network. International conference on computer vision, graphics, and image processing.

Bauer, H. (1996). Form, Struktur, Stil: Die formanalytischen und formgeschichtlichen Methoden [Form, structure, style: The methods of form analysis and form history]. In H. Belting, H. Dilly, W. Kemp, W. Sauerländer, & M. Warnke (Eds.), *Kunstgeschichte: Eine Einführung [Art history: An introduction]* (Vol. 6, pp. 157-173). Reimer.

Betensky, M. G. (1991). Kunsttherapie und künstlerische Äußerung aus phänomenologischer Sicht [Art therapy and artistic expression from a phenomenological perspective]. In J. A. Rubin (Ed.), *Richtungen und Ansätze der Kunsttherapie: Theorie und Praxis [Approaches to art therapy: Theory and technique]* (pp. 176-183). Gerardi.

https://doi.org/10.34632/jsta.2022.10196

Journal of Science and Technology of the Arts, vol. 14, n. 1 (2022): pp. 49-71

Brachmann, A., Barth, E., & Redies, C. (2017). Using CNN features to better understand what makes visual artworks special. *Frontiers in Psychology*, *8*, 830.
https://doi.org/10.3389/fpsyg.2017.00830

Castro, L., Perez, R., Santos, A., & Carballal, A. (2014). Authorship and aesthetics experiments: comparison of results between human and computational systems. International Conference on Evolutionary and Biologically Inspired Music and Art.

Cattaneo, M. (1994). Addressing culture and values in the training of art therapists. *Art Therapy*, *11*(3), 184-186.
https://doi.org/10.1080/07421656.1994.10759081

Cetinic, E., Lipic, T., & Grgic, S. (2018). Fine-tuning convolutional neural networks for fine art classification. *Expert Systems with Applications*, *114*, 107-118.
https://doi.org/10.1016/j.eswa.2018.07.026

Cetinic, E., Lipic, T., & Grgic, S. (2020). Learning the principles of art history with convolutional neural networks. *Pattern Recognition Letters*, *129*, 56-62.
https://doi.org/10.1016/j.patrec.2019.11.008

Chollet, F. (2018). *Deep Learning mit Python und Keras: Das Praxis-Handbuch vom Entwickler der Keras-Bibliothek [Deep Learning with Python and Keras: The Practice Manual from the Developer of the Keras Library]*. Manning.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. 2009 IEEE conference on computer vision and pattern recognition.
Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. MIT press.
https://doi.org/10.4258/hir.2016.22.4.351

Grant, C., & Price, D. (2020). Decolonizing art history. *Art History*, *43*(1), 8-66.
https://doi.org/10.1111/1467-8365.12490

Hua, K.-L., Ho, T.-T., Jangtjik, K.-A., Chen, Y.-J., & Yeh, M.-C. (2020). Artist-based painting classification using Markov random fields with convolution neural network. *Multimedia Tools and Applications*, *79*(17), 12635-12658.
https://doi.org/10.1007/s11042-019-08547-4

ImageNet. (2012). *Large Scale Visual Recognition Challenge (ILSVRC)*.
http://image-net.org/challenges/LSVRC/2012/index

https://doi.org/10.34632/jsta.2022.10196

ImageNet. (2016). *Large Scale Visual Recognition Challenge (ILSVRC)*. *http://image-net.org/challenges/LSVRC/2016*

Kandinsky, W. (1955). *Punkt und Linie zu Fläche: Beitrag zur Analyse der malerischen Elemente [Point and Line to Plane: Contribution to the analysis of pictorial elements]*. Benteli.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, *25*, 1097-1105.

Madhu, P., Kosti, R., Mührenberg, L., Bell, P., Maier, A., & Christlein, V. (2019). Recognizing characters in art history using deep learning. Proceedings of the 1st Workshop on Structuring and Understanding of Multimedia heritAge Contents (SUMAC '19).

Mersch, D. (2019). Ästhetisches Denken: Kunst als Theoria [Aesthetic thinking: Art as theoria]. In D. Mersch, S. Sasse, & S. Zanetti (Eds.), Ästhetische Theorie [Aesthetic theory] (pp. 241-259). Diaphanes.

Meyer, G. (2011). *Sprache der Bilder: Kunst verstehen: Form, Farbe, Komposition [Language of the pictures: Understanding art: Shape, color, composition]*. Seemann.

Mosquera, G. (1992). The Marco Polo syndrome: Some problems around art and Eurocentrism. *Third Text*, *6*(21), 35-41. *https://doi.org/10.1080/09528829208576382*

Nishikawa-Pacher, A., Heck, T., & Schoch, K. (2021). Open editors: A dataset of scholarly journals' editorial board positions. *SocArXiv*. *https://doi.org/https://doi.org/10.31235/osf.io/jvzq7*

Panofsky, E. (2006). *Ikonographie und Ikonologie: Bildinterpretation nach dem Dreistufenmodell [Iconography and iconology: Image interpretation according to the three-stage model]*. DuMont.

prometheus. (2021). *prometheus: Das verteilte digitale Bildarchiv für Forschung & Lehre [prometheus: the distributed digital image archive for research & education]* *https://www.prometheus-bildarchiv.de*

Python. (2021). *Python*. *https://www.python.org*

Razavian, A. S., Azizpour, H., Sullivan, J., & Carlsson, S. (2014). CNN features off-the-shelf: An astounding baseline for recognition. Proceedings of the IEEE conference on computer vision and pattern recognition workshops.

https://doi.org/10.34632/jsta.2022.10196

Rodriguez, C. S., Lech, M., & Pirogova, E. (2018). Classification of style in fine-art paintings using transfer learning and weighted image patches. 2018 12th International Conference on Signal Processing and Communication Systems (ICSPCS).

Santos, I., Castro, L., Rodriguez-Fernandez, N., Torrente-Patino, A., & Carballal, A. (2021). Artificial neural networks and deep learning in the visual arts: A review. *Neural Computing and Applications*, *33*(1), 121-157. *https://doi-org/10.1007/s00521-020-05565-4*

Sarkar, D., Bali, R., & Ghosh, T. (2018). *Hands-On Transfer Learning with Python: Implement advanced deep learning and neural network models using TensorFlow and Keras*. Packt Publishing Ltd.

Schoch, K. (2020, April 24). Ratinginstrument für zweidimensionale bildnerische Arbeiten (RizbA): Fragebogen mit Erläuterungen in deutscher Sprache [Rating instrument for two-dimensional pictorial works (RizbA): Questionnaire with explanatory notes in German language]. *Zenodo*. *https://doi.org/10.5281/zenodo.3765221*

Schoch, K., Gruber, H., & Ostermann, T. (2017). Measuring art: Methodical development of a quantitative rating instrument measuring pictorial expression (RizbA). *The Arts in Psychotherapy*, *55*, 73-79. *https://doi.org/10.1016/j.aip.2017.04.014*

Schoch, K., & Ostermann, T. (2020). Giving the art greater weight in art psychology: RizbA, a quantitative questionnaire for formal picture analysis. *Creativity. Theories – Research – Applications*, *7*(2), 373-410. *https://doi.org/10.2478/ctra-2020-0019*

Schoch, K., & Ostermann, T. (2021, July 13). Empirics vs. art theory: Exploring a factor structure of pictorial expression based on contemporary artworks. *SocArXiv*. *https://doi.org/https://osf.io/preprints/socarxiv/xyrf5*

Schoch, K., & Ostermann, T. (2022). Psychometrics of art: Validation of RizbA, a quantitative rating instrument for pictorial expression. *Creativity research journal*. *https://doi.org/10.1080/10400419.2021.1987734*

Simon, H., & Verstegen, U. (2004). prometheus: Das verteilte digitale Bildarchiv für Forschung und Lehre. Neuartige Werkzeuge zur Bereitstellung von verteiltem Content für Wissenschaft und Forschung [prometheus: Digital image database for research and education. New tools for providing shared content for science and research]. *Historical Social Research/Historische Sozialforschung*, 247-257.

https://doi.org/10.34632/jsta.2022.10196

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *ArXiv*.
https://doi.org/arXiv:1409.1556

Streb, J. H. (1984). Thoughts on phenomenology, education, and art. *Studies in Art Education*, *25*(3), 159-166.
https://doi.org/10.1080/00393541.1984.11650370

Stuhler-Bauer, A., & Elbing, U. (2003). Die phänomenologische Bilderfassung: Ein kunsttherapeutisches Instrument [The phenomenological picture survey: An art therapeutic instrument]. *Musik-, Tanz- und Kunsttherapie*, *14*, 32-46.
https://doi.org/10.1026//0933-6885.14.1.32

Stuhler-Bauer, A., & Elbing, U. (2004). Kritik standardisierter empirischer Bilderfassung aus kunsttherapeutisch-phänomenologischer Sicht: Die Nürtinger Beurteilungsskala und die Diagnostic Assessment of Psychiatric Art, deutsche Version [Review on standardized empirical picture analysis from an art therapeutic phenomenological view: The Nürtingen Rating Scale and the Diagnostic Assessment of Psychiatric Art, German version]. *Musik-, Tanz- und Kunsttherapie*, *15*(1), 5-15.
https://doi.org/10.1026/0933-6885.15.1.5

Tan, W. R., Chan, C. S., Aguirre, H. E., & Tanaka, K. (2016). Ceci n'est pas une pipe: A deep convolutional network for fine-art paintings classification. 2016 IEEE international conference on image processing (ICIP).

Uzelac, M. (1998). Art and phenomenology in Edmund Husserl. *Axiomathes*, *9*(1), 7-26.

Vollmar, K. (2008). *Die faszinierende Welt der Farben: ein Glossar von A-Z [The fascinating world of colors: A glossary of A-Z]*. ars momentum.

WikiArt. (2020). *WikiArt API*.
https://www.wikiart.org/en/App/GetApi

WikiArt. (2021). *WikiArt: Visual art encyclopedia*.
https://www.wikiart.org

Zhao, L., Shang, M., Gao, F., Li, R., Huang, F., & Yu, J. (2020). Representation learning of image composition for aesthetic prediction. *Computer Vision and Image Understanding*, *199*, 103024.
https://doi.org/10.1016/j.cviu.2020.103024

https://doi.org/10.34632/jsta.2022.10196

ACKNOWLEDGEMENTS

The authors would like to thank Rebecca Kahn for the valuable feedback on the manuscript. Also, thanks to Thomas Ostermann for discussing initial ideas with the authors.

INDICES

ICC	*Intra-Class Correlations*
M	Mean
MSE	Mean Squared Equrror
N	Sample size
η^2	Eta-squared
r	Test-retest reliability
SD	Standard deviation

ABBREVIATIONS

ART	Artificial Intelligence Rater
API	Application programming interface
CNN	Convolutional Neural Networks
ER	Extreme rater
GR	Generic rater
ILSVRC	ImageNet Large Scale Visual Recognition Challenge
MR	Mean rater
OR	Over rater
RGB	Red, green, blue
RizbA	Rating instrument for two-dimensional pictorial works
RR	Random rater
RR1	Randomly picked generic rater 1
RR2	Randomly picked generic rater 2
RR3	Randomly picked generic rater 3
RR4	Randomly picked generic rater 4
UR	Under rater

ADDITIONAL MATERIAL

This study is conducted under terms of Open Science. The application's source code as Open Source, Open Data, and Open Methodology are freely available under a CC-BY 4.0 license as follows.

https://doi.org/10.34632/jsta.2022.10196

Open Source

Gengenbach, T., & Schoch, K. (2021, August 23). ARTificial Intelligence Raters: Neural Networks for Rating Pictorial Expression. *Open Science Framework*.
https://doi.org/10.17605/OSF.IO/QUDHX

Open Data

Study A On request to the according authors for research purposes only

Study B Schoch, K. (2021, July 13). Giving the art greater weight in art psychology: RizbA, a quantitative questionnaire for formal picture analysis. *Open Science Framework*.
https://doi.org/10.17605/OSF.IO/P84XW

Study C Schoch, K. (2021, July 21). Empirics vs. art theory: Exploring a factor structure of pictorial expression based on contemporary artworks. *Open Science Framework*.
https://doi.org/10.17605/OSF.IO/JNZ67

Open Methodology

Schoch, K. (2020, April 24). *Ratinginstrument für zweidimensionale bildnerische Arbeiten (RizbA): Fragebogen mit Erläuterungen in deutscher Sprache* [Rating instrument for two-dimensional pictorial works (RizbA): Questionnaire with explanatory notes in German]. *Zenodo*.
https://doi.org/10.5281/zenodo.3765221

Article received on 10/09/2021 and accepted on 27/04/2022.

https://doi.org/10.34632/jsta.2022.10196