EVA: An Explainable Visual Aesthetics Dataset

Chen Kang, Giuseppe Valenzise, Frédéric Dufaux

EVA: An Explainable Visual Aesthetics Dataset

Chen Kang
Université Paris-Saclay, CNRS, CentraleSupélec, L2S
Gif-sur-Yvette, France
chen.kang@centralesupelec.fr

Giuseppe Valenzise
Université Paris-Saclay, CNRS, CentraleSupélec, L2S
Gif-sur-Yvette, France
giuseppe.valenzise@centralesupelec.fr

Frédéric Dufaux
Université Paris-Saclay, CNRS, CentraleSupélec, L2S
Gif-sur-Yvette, France
frederic.dufaux@centralesupelec.fr

ABSTRACT

Assessing visual aesthetics has important applications in several domains, from image retrieval and recommendation to enhancement. Modern aesthetic quality predictors are data driven, and leverage the availability of large annotated datasets to train accurate models. However, labels in existing datasets are often noisy, incomplete or they do not allow more sophisticated tasks such as understanding *why* an image looks beautiful or not to a human observer. In this paper, we propose an Explainable Visual Aesthetics (EVA) dataset, which contains 4070 images with at least 30 votes per image. Compared to previous datasets, EVA has been crowdsourced using a more disciplined approach inspired by quality assessment best practices. It also offers additional features, such as the degree of difficulty in assessing the aesthetic score, rating for 4 complementary aesthetic attributes, as well as the relative importance of each attribute to form aesthetic opinions. A statistical analysis on EVA demonstrates that the collected attributes and relative importance can be linearly combined to explain effectively the overall aesthetic mean opinion scores. The dataset, made publicly available[1], is expected to contribute to future research on understanding and predicting visual quality aesthetics.

KEYWORDS

dataset; aesthetic quality assessment

1 INTRODUCTION

The goal of image aesthetic quality assessment is to predict how beautiful an image looks to a human observer. It has been used in different applications, such as ranking [10], recommendation [14, 15], enhancement [5], and memorability prediction [3]. Many models involve different attributes to predict the aesthetic score [1], including color and contrast, photographic attributes such as layout, and semantic influence. In this context, one of the objectives of this paper is to address the following key question: which attributes are important for an image to be perceived as beautiful and with a high aesthetic score?

Previous research has studied the relationship between aesthetic attributes and aesthetic scores. A multi-column network is proposed in [23] to learn this relationship automatically, but the features rely on manual design, which may introduce a bias. The approaches in [2, 24, 26] aim at finding the relative importance of aesthetic attributes based on text comments. However, the attributes definitions are ambiguous and subjective, which limits the reliability of the methods. The work in [10] tried to identify which attributes influence the aesthetic of an image by directly asking the subjects, but the pool of subjects is small and with little variety. One of the recurring issues that hinder the progress of research in aesthetics is the lack of properly labeled data.

In order to overcome the lack of explicit aesthetic labels, previous work has sometimes employed different, but supposedly related annotations, such as the “faves” and "views" in social networks and media platforms [19]. However, aesthetic labels obtained with this indirect approach may be influenced by external factors, e.g., the level of interest, popularity, content and even personal social relationship. Since these aspects are very complex, it is difficult to disentangle and quantify the factors producing the aesthetic appraisal. Thus, collecting aesthetic annotations directly from users remains the most reliable way to provide accurate ground truth labels to predict aesthetic quality.

AVA [13] is the most popular image dataset with aesthetic annotations. 255k images with 78-549 general scores for each image have been directly collected from thematic challenges. However, data collection lacks a precise and properly defined methodology. In particular, as each challenge has a pre-defined theme, when crawling data from online photographic challenges, the aesthetic scores have different interpretations across challenges, and within the competition, voters may have followed different evaluation criteria. Other works like [7, 24] crawl additional information on the same website with the goal to compensate the labels of AVA. However, it is very difficult to separate the contribution of purely aesthetic factors to AVA scores from any other contextual factor, without knowing the behavior of the original voters. Different from other quality assessment and computer vision tasks, such as object recognition and detection, the ground-truth labels given by human observers in image aesthetic quality remain very subjective. In particular, the user training, or lack thereof, can affect the reliability of the final score.

In summary, to know which attributes are important in aesthetics, it is important to collect data directly and properly. In order to overcome the disadvantages of previous datasets, the main objective of this paper is to build a better-defined *Explainable Visual Aesthetic (EVA) quality assessment dataset*, with the goal to investigate which attributes influence the perceived aesthetics on an image. More specifically, our contributions are:

- We propose the first dataset that simultaneously contains subjective labels for aesthetic attributes ranging from low-level visual factors (e.g., light, color, exposure) to higher-level semantic preference. In addition, we also record the importance of each attribute while collecting votes, by explicitly asking observers to indicate which factors influenced their aesthetic judgment for a given image.
- We ask for the uncertainty of subjects' votes while voting. This is meaningful to evaluate the average aesthetic score's reliability.

[1]The dataset is publicly available at https://github.com/kang-gnak/eva-dataset

- We combine crowd-sourcing with the best practices from quality assessment recommendations such as subject training and a clear definition of the attributes to select test stimuli and guarantee the quality of the collected data. Personal background and demographic information is also collected. Finally, user voting time is recorded in order to identify outliers and to clean the data.
- We analyze the collected attributes and difficulty, showing that light, color, composition and depth are generally the most important in the overall aesthetic quality of images. Different content categories display a different relative importance of the attributes. We observe that the personal difficulty in judging aesthetic is also very subjective and is only very loosely correlated with the standard deviation of mean aesthetic score, which is a measure of group aesthetic subjectivity [8].

2 RELATED WORK

A traditional way to judge aesthetics is based on well-established photographic rules [4]. However, it is challenging to draw conclusions on how each aspect affects the aesthetic score. Another typical approach is to define features involving different attributes, which are then integrated into classification and regression models to match the scores given by observers [4]. For example, attributes such as colorfulness, tone, clarity, depth, and sharpness are computed in [1]. With the growing interest for deep learning methods, several works have tried to add high-level attributes which cannot be well explained by hand-crafted features. In [12], the authors have verified how high-level attributes, like style and semantic, affect aesthetic scores. In [23], authors use a multi-column neural network to first train different visual factors and semantic features, then combine them together with a column for unknown attributes, to imitate the general aesthetic values. As an intrinsically subjective task, the collection of reliable data labels is a significant challenge, which greatly impacts the development of effective models. There are two main trends in collecting the aesthetic labels. The first one is by inferring aesthetic information indirectly from human raters. In [19], the authors used the probability of "faves" in "views" as a Flickr image's measurement of the photos' aesthetic appeal. However, in these cases, aesthetics is difficult to distinguish from other subjective values, such as the level of interest, humour, or popularity.

The second one is by eliciting aesthetic quality from subjects directly. However, existing datasets are often limited in terms of reliability, variety or quantity. The popular benchmark dataset AVA [13] collected voting scores "in the wild" and related image data from DPChallenge, where photography amateur competitions are held online. Subsequently, the dataset has been expanded [12], to meet the needs of deep learning models, but AVA remains the most popular dataset in aesthetic studies. Several datasets have continued to crawl data based on AVA to augment information from the users and photographers. As each subject voted for images in the competitions, some of them left text comments, which have been collected in [24]. Conversely, photographers' demographic information are collected by [7]. However, these data are mostly collected in unconstrained conditions, exhibiting noise and making them unreliable. For instance, simple challenges like "animal" and challenges with complex themes like "humour" are mixed, which may impact the assessment of aesthetics from one challenge to another. Contrary to the crawling, authors in [11] collected aesthetic scores in a laboratory environment and controlled conditions. More precisely, they collected users' information and trained them with many images. However, overall, they only have 33 subjects and 1000 images from *photo.net*. It is obviously not easy to acquire significantly more data under these conditions. Guidelines for gathering reliable and repeatable aesthetic ground-truth scores are discussed in [20]. It is found there that using a discrete ACR (Absolute Category Rating) scale generally gives a better consistency across voters, and a good repeatability of the scores among lab-based and crowdsourcing experiments. In our work, we also adopt an 11-point (from 0 to 10) discrete ACR scale for general aesthetic score, which has been found to have a lower standard deviation around MOS than a 5-point scale in [25]. This range is also in line with the scores in popular aesthetic datasets such as AVA.

Exploiting these existing datasets, researchers have aimed at explaining visual aesthetics. In [11, 23], attribute labels are computed from images instead of asking the subjects directly. As a consequence, the relative importance among attributes depend on the feature extraction models. In [4], the authors have used a set of attributes to assess their aesthetic importance, but the approach ignores high level features such as the semantic content. The authors in [24, 26] extracted the aesthetic attributes importance based on text comments. However, the attributes definitions are ambiguous and may differ for each user, which limits the reliability of the method. In AADB [10], the subjects have been asked to choose the level of various attributes directly. Still, the study was only conducted with 5 or 6 observers per image, recruited using Amazon Mechanical Turk. No personal background information of the observers are included, which makes it difficult to make further studies such as impact of demographics on aesthetics.

In summary, the lack of well-labeled data is negatively impacting progress towards understanding the importance of attributes in image aesthetics. Based on the above considerations, we conclude that we need a new dataset with a reasonable variety and quantity of annotated images in order to better explain visual aesthetics.

3 METHOD

In this section, we present in details the methodology for collecting data in the proposed EVA dataset. First, we describe the whole process and the survey questions. Then, we discuss the selection of test images. After collecting data, we finally investigate how to identify outliers in order to clean the subjective data.

3.1 Experiment Settings and Work Flow

The whole process is anonymous. Observers first need to provide background information including year of birth, region, gender, and whether they are color blind or wearing glasses. Then, they have to indicate, by their self-assessment, their experience in photography, as either beginner (without any specific knowledge about photography); intermediate (a casual photographer without specific training); or advanced (having followed some specific training in photography). After registering this background information,

observers have to undergo a training phase in order to better understand the test. More precisely, each survey question is explained and sample images (not present in the test stimuli) are shown for illustration. This step is especially important to ensure that different naive observers receive the same instructions and understand the meaning of each attribute they will have to assess. To verify that they have carefully read the instructions, observers have to click several check boxes intertwined with the text. After the registration and training phases, observers can start voting, as detailed in the next section. Images are randomly selected for display, and each image is only displayed once for an observer. The images were resized automatically to fit the display width of the device. In order to stabilize judgements, the first two images are dummy stimuli, i.e., their scores are discarded. For the sake of flexibility, subjects are allowed to leave voting at any time and come back later, while being identified with the same cookie account. While voting, user behaviour is recorded. More precisely, we record the time when a subject submits each vote. In particular, this information is useful to identify when a subject is voting very fast. It could also be potentially related to the difficulty of assessing a given image.

The experiments reported in this paper were carried out online from February 2020 to July 2020. To get more data as well as diversity, two web hosts have been used, the first one in English hosted outside of mainland China, and the second one in simplified Chinese is hosted in mainland China. Due to the size of the collected dataset and the limited available budget, we did not resort to any recruitment platform (such as MTurk) to enroll participants in this study. Instead, the study was advertised through authors' social networks, targeting mainly acquaintances, colleagues, scientists and students in vision-related topics. The volunteers were invited to vote for at least one session (30 images) or more. The users that voted for a large number of images were rewarded with some small gifts. Most of the votes come from France and China, reflecting the geographical location of the authors.

3.2 Survey Design

Considering the attributes in previous work [1, 8, 10, 11, 13] and inspired by methods for subjective quality assessment experiments in laboratory conditions [11], we design the survey considering four main attributes and one measure of difficulty to judge image aesthetic quality. More specifically, the survey is composed of several questions detailed hereafter:

Question 1: "What is the overall aesthetic quality of this picture?" We employ an 11-point discrete ACR scale. However, instead of the usual categories (excellent, good, etc.), we label the extremes of the scale as "least beautiful" (corresponding to 0) and "most beautiful" (corresponding to 10), and let subjects rate through a slider bar. When a new image is displayed, the slider default position is always set to 5.

Question 2: "How difficult is it for you to judge this image's aesthetic quality?" Subjects have to select an option over a four-level Likert scale: very difficult, difficult, easy, and very easy. Indeed, aesthetic quality is very subjective, and sometimes it is difficult to assign a score to an image. We set the number of options in the Likert scale to be even to avoid the possible tendency of voters to select effortlessly the middle, neutral option. While it is obvious that the consensus on the overall aesthetic score varies significantly across images, predicting the subjectivity is a difficult task [8]. The purpose of this question is to directly ask the subjects about the difficulty to score a given image, with the objective to support further studies on aesthetic subjectivity.

Question 3: "How do you like this attribute?", where we consider four attributes: *light and color*, *composition and depth*, *quality*, and *semantics* of the image. For each attribute, subjects have to vote on a four-level Likert scale with the following options: very bad, bad, good, and very good. We choose these four attributes, as they have been previously studied [1, 9, 10, 23] and they are relatively easy to understand by naive subjects. The attributes have been defined as follows in the user training phase prior to the test: Light and color relates to visual perception, including brightness, contrast, and color saturation. Composition and depth relates to the position and spatial relationship between objects in the scene. Quality can be impacted by different types of distortions, including blur, compression, noise, and other artifacts. Semantics is related to how much the subject likes the content of the image. Notice that these attributes span different levels of factors affecting image aesthetics, from perceptual (light/color and visual quality), photographic technique (composition, depth) to higher level features of the scene. We purposely keep the number of attributes to 4, without further detailing them (in particular for composition and semantics), to avoid complex categorization which might require more advanced photographic knowledge as well as longer training/test time.

Question 4: "Which factor(s) do you think is (are) important in your judgement of this photo? (choose at least one factor)" where people can choose multiple options among the four attributes mentioned above. The subjects are required to vote for at least one option. We set this question as binary check boxes, in order to avoid making the voting time for an image being too long.

3.3 Image Selection

Creating a comprehensive and representative image dataset is a challenging task. AVA is well-known for its very large size and variety. Moreover, many research works have built upon AVA, for example, [7] augments AVA data with photographer information (AVA-PD). In this paper, we select images which are present also in the AVA-PD dataset. Our goal is to select more than 5000 images in total, with the procedure highlighted below.

Since the semantic content can influence aesthetics [22], we choose images from different content classes, so that the users are shown different photography categories with a similar probability. Inspired by the 5 content types mentioned in [11], we divide images into 6 categories: animals; architectures and city scenes; human; natural and rural scenes; still life; other. In order to get a rough categorization of the test images, we use Yolo V3 [17] to detect and classify the objects in images. We manually group the object labels given by Yolo V3 model (which has 80 pre-trained classes) into our first five categories. Then, we assume that the category of an image is defined by its main objects. For this purpose, we compute the cumulative surface of the detection frames corresponding to each category. If this surface is over 50% of the entire image or larger than the total surface of the other objects, then this image is classified in the associated category. Otherwise, it is considered in

the "other" category. The latter case may therefore correspond to images with several significant objects or no significant object.

In the AVA dataset, the images with very low quality scores typically have poor technical quality. Compared to the time when the AVA dataset images have been collected, nowadays the technical quality of photo sensors and imaging system is greatly improved, and even low-end smartphone cameras are capable of capturing pictures with little noise, blur or compression artifacts. Therefore, in EVA we choose to consider only images with reasonable technical quality, as those with very little quality can be easily detected nowadays with existing methods [21]. Moreover, in this way, one could learn a more precise predictor of aesthetics over a smaller range of aesthetic qualities, which might be more useful in an image recommendation or enhancement scenario. Therefore, we selected images from AVA with associated scores within the range [4,9]. More precisely, we divided the scores in four intervals: [4,5), [5,6), [6,7), [7,9), and selected a similar number of images in each group. Three persons, including two photography amateurs, have manually checked the images in each photography category to ensure the validity of the classification.

To take into consideration potentially harmful or uninteresting content, images with specific characteristics are also removed. Grounds for removal include sexual content, religion sarcasm, drugs, horror, artwork, images comprising a lot of text, and commercial advertisements. Based on the above procedure, we have selected 5101 images, nearly evenly distributed in terms of Mean Opinion Score (MOS).

4 RESULTS AND DISCUSSION

4.1 Data Cleaning

After collecting the data, we get 4734 voting sessions. Of these, 1251 voting sessions contain at least one image after the two dummy stimuli. Notice that the same individual person might have voted in different voting sessions, if the latter are far apart in time, as the cookies expired after a few weeks of inactivity. In designing the experiment we did not include online quality controls or trap questions. On one side, given the subjective nature of aesthetic scores, it is difficult to detect whether a vote deviating from the average is due to a malicious behavior or simply to a personal judgment. On the other hand, other kinds of controls such as content questions [20] might be used; however, these can be easily circumvented as users learn to anticipate them when voting many images. Instead, we relied on the personal engagement of the voters, most of which volunteered the task. We carefully checked the votes of potentially suspect participants, by monitoring constantly the evolution of the votes and eliminating those we deemed to be unreliable. This was indeed a time-consuming activity during the dataset collection.

In total, 172934 raw votes (before data cleaning) have been collected. We apply statistical a posteriori analyses to filter out the collected votes. However, while some inter-rater agreement indicators [6] such as Cronbach's alpha or Intra-Class Correlation (ICC) have been proposed for aesthetic subjective analysis [20], the high number of votes collected in EVA required that each image is evaluated in general by a different combination of users. As a result, inter-rater variability is difficult to evaluate and interpret. On the other hand, post-filtering approaches such as CrowdMOS [18], which compares the consistency of individual votes with the population, assume Gaussianity of the image score distribution, which is generally not the case for aesthetics. Therefore, we consider alternative approaches to post-filtering the votes (see below), but we release also the raw data of EVA to allow further analyses and data cleaning methodologies to be employed in future research. In our analysis, we employ the recorded voting time associated to each individual vote as an indicator of possible under-commitment of users to the task: voting times that are too short might imply that a voter assigned scores randomly [20]. Specifically, we obtain the voting time as the time interval of two consecutive votes. In order to collect reliable statistics about the minimum voting time, we identify a group of 13 trusted voters, including users that have previously participated to other lab-based user studies organized by the authors. By inspecting the distribution of voting times for this pool of users, we observed that the minimum voting time is 7 seconds, which we select as a threshold on the minimum voting time to consider a vote valid. About 3% of the votes in the dataset correspond to a voting time smaller than 7 seconds, and are then discarded as possible outliers.

As a second criterion for outlier detection and removal, we consider the standard deviation of votes for a user. More specifically, we observe that voters with very small standard deviation in their judgment of global aesthetic score (e.g., those who gave the same vote to all images they voted) might be unreliable. We empirically set a threshold of 0.1 on the standard deviation of an individual user's votes, in order to decide whether he/she is an outlier. This leads to removing 1094 voting sessions from the dataset. Notice that most of these voting sessions actually consists of very few images, so the impact over the whole dataset is rather limited.

To have a robust estimation of aesthetic scores and similar number of votes in each image,

we remove from the dataset images having less than 30 votes. After this data cleaning process, 4070 images have been retained, with 30 to 40 valid votes each.

4.2 Data Summary

In this section, a brief summary of EVA dataset is given. The cleaned dataset includes 4070 images, with a total of 136943 votes from 1094 voting sessions.

Figure 1 reports statistics about the participants of the study. Around 30% of users use computer or laptop, and the rest uses mobile devices (including smartphones and tablets). Most people know little about photography, a part of subjects are amateurs, and a small amount of voters have professional skills. This likely reflects a realistic distribution of photographic skills across the population and is an intended feature of the EVA dataset, which targets aesthetic perception at large.

The average aesthetic score distribution is illustrated in Figure 2 and 3. The highest peak for MOS is around 6, rather than the medium score 5, probably because the images have relatively high quality. The peak for standard deviation of scores in each image is between 1.5 and 2.0. A Shapiro-Wilk test [16] performed on these distributions reveals that they are not Gaussian distributed.

For attributes, the distributions are shown in Figure 4. We can see that they are skewed. This may be due to the way images have been

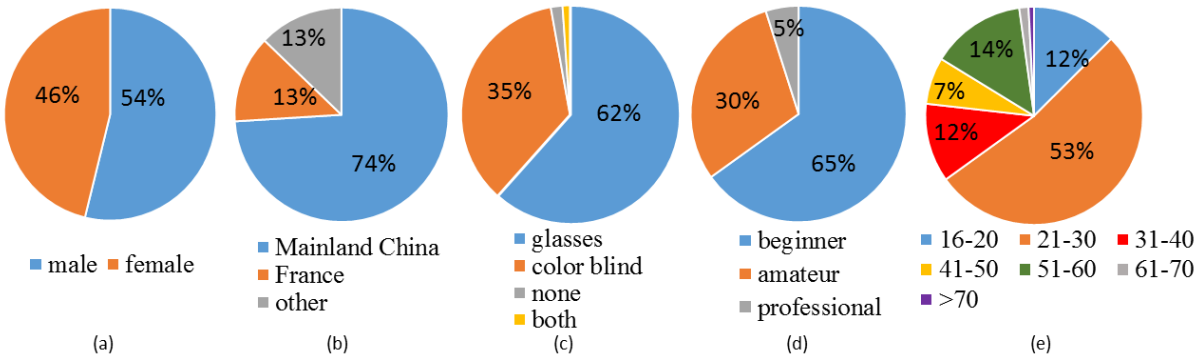

Figure 1: Statistics of votes in EVA dataset: (a) Gender (b) Region (c) Visual Status (d) Photography Experience (e) Age

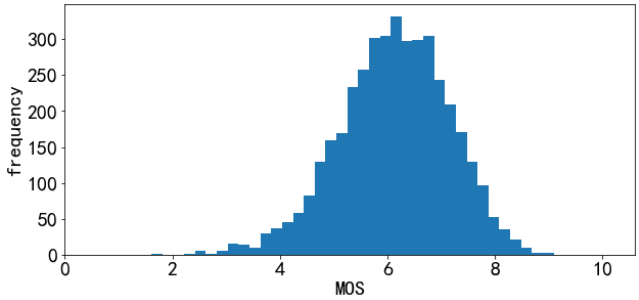

Figure 2: Distribution of Mean Opinion Score (MOS)

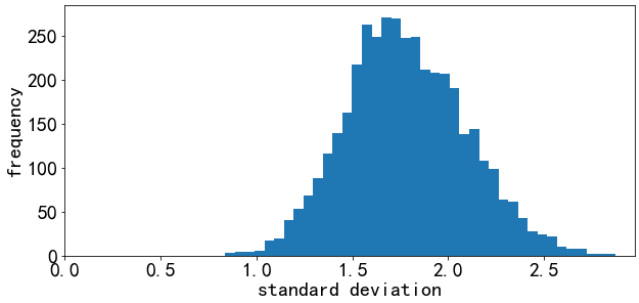

Figure 3: Distribution of Standard Deviation (STD) of scores

selected from the AVA dataset, i.e., images with very low quality have been discarded.

4.3 Data Analysis

To study the relation between aesthetic attributes and the overall aesthetic score, we report the Spearman's Rank Correlation Coefficient (SRCC) and Pearson Correlation Coefficient (PCC) of the whole samples, as well as for each content category, in Table 1. The average answer of the second question in the survey, interpreted as the average personal uncertainty in an image, is denoted as "difficulty".

The correlations are divided in four groups: 1) correlation between overall aesthetic score (MOS) and average (per image) magnitude of each attribute; 2) correlation between standard deviation of the overall aesthetic score (STD) and average (per image) magnitude of each attribute; 3) correlation between difficulty and average (per image) magnitude of each attribute; and finally 4) correlation between MOS, difficulty and STD. The peak value of each row is in bold, and the one of each column is underlined.

4.3.1 Relation between attributes and aesthetic score. All of the aesthetic attributes are significantly related to the general aesthetic score in a linear relationship (this is confirmed by a visual inspection of scatter plots, not reported due to space limitations). The correlation coefficients between attributes and global score from these data seem quite similar except for the quality. This may be explained by the fact that the image quality in EVA stimuli is generally good.

Composition and depth are the most correlated attribute with the global score in all content categories. It reaches 0.90 in PCC for all the images. Even though in "natural and rural scenes" the most linearly related attribute is light and color, with a PCC of 0.91, composition and depth still gets a very high correlation coefficient, with a PCC of 0.90. Semantic preference is the most correlated attribute in the "other" category (where the content variability is higher), and it is the second most linearly related attribute among all the images. Across the categories, it can be observed that "natural and rural scenes" have a more direct relation of visual and photographic attributes to the overall score. This is probably due to the larger variety of colors, brightness, etc. than the one in a portrait or still life.

4.3.2 Difficulty and subjectivity in the aesthetic evaluation. The results about the correlation between standard deviation and attributes' values show that the subjects' disagreement in aesthetics relates more to whether the subjects like the semantics than to the preference in low-level attributes, since the PCC in general scores' standard deviation and semantic gets -0.58. It is similar in categories "animals", "human", "still life" and "other". In "architecture and city scenes" and "natural and rural scene" images, composition and depth disagreement matters more than other attributes, getting -0.55 and -0.61 respectively.

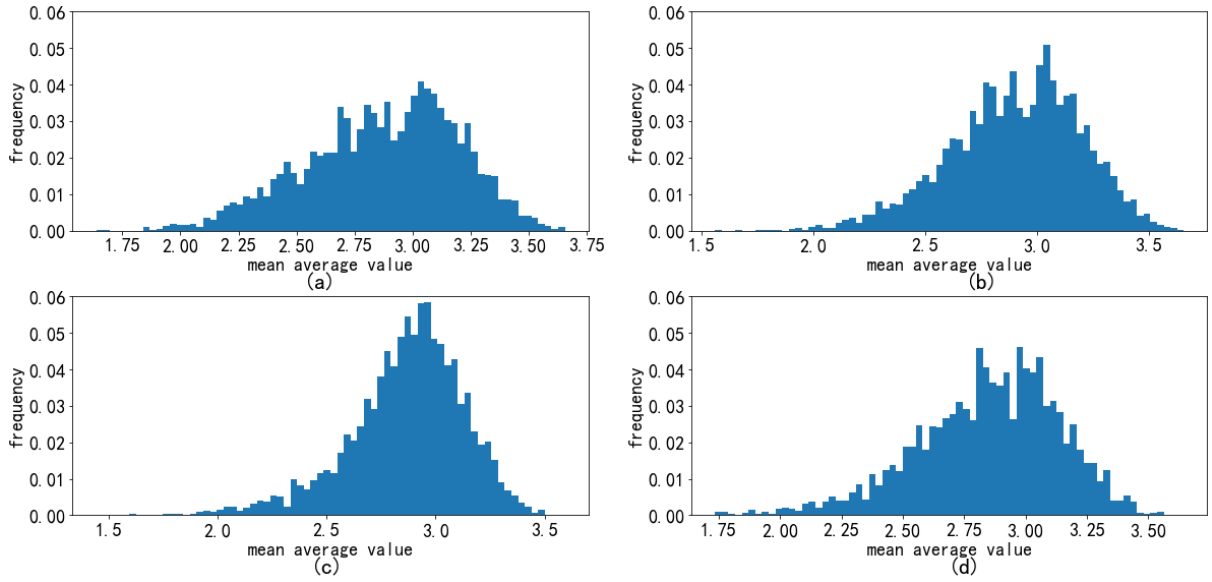

Figure 4: Distribution of attributes. (a) Light and color (b) Composition and depth (c) Quality (d) Semantic

Table 1: Correlation between Mean Score of Attributes and Mean Opinion Score (MOS). STD denotes the standard deviation of the global aesthetic scores per image. The numbers are SRCC/PCC.

item	general	animal	architecture and city scenes	human	natural and rural scenes	still life	other
MOS and light and color	0.85/0.85	0.80/0.81	0.85/0.85	0.83/0.84	**0.91/0.91**	0.83/0.83	0.82/0.85
MOS and composition and depth	0.89/**0.90**	0.87/0.89	0.88/0.89	0.88/0.88	**0.90/0.90**	0.88/0.89	0.89/**0.90**
MOS and quality	0.76/0.77	0.73/0.76	0.79/0.81	0.74/0.77	**0.88/0.88**	0.77/0.78	0.74/0.78
MOS and semantic	0.87/0.88	0.83/0.85	0.86/0.88	0.85/0.86	0.90/0.90	0.86/0.87	**0.92/0.92**
STD and light and color	-0.47/-0.47	-0.36/-0.37	-0.49/-0.50	-0.43/-0.44	**-0.56/-0.56**	-0.41/-0.41	-0.50/-0.49
STD and composition and depth	-0.54/-0.55	-0.53/-0.51	-0.52/-0.55	-0.45/-0.48	**-0.61/-0.61**	-0.48/-0.49	-0.55/-0.53
STD and quality	-0.45/-0.45	-0.25/-0.35	-0.52/-0.52	-0.34/-0.36	**-0.59/-0.59**	-0.43/-0.41	-0.45/-0.42
STD and semantic	-0.56/-0.58	-0.47/-0.59	-0.52/-0.54	-0.52/-0.56	-0.59/**-0.60**	-0.53/-0.54	**-0.62**/-0.59
difficulty and light and color	-0.62/-0.60	-0.58/-0.54	-0.60/-0.59	-0.56/-0.55	**-0.74/-0.71**	-0.51/-0.52	-0.52/-0.52
difficulty and composition and depth	-0.52/-0.47	-0.50/-0.44	-0.50/-0.45	-0.44/-0.38	**-0.65/-0.56**	-0.42/-0.37	-0.43/-0.43
difficulty and quality	-0.49/-0.43	-0.50/-0.39	-0.47/-0.43	-0.42/-0.36	**-0.67/-0.59**	-0.39/-0.33	-0.40/-0.38
difficulty and semantic	-0.53/-0.48	-0.47/-0.43	-0.47/-0.45	-0.44/-0.38	**-0.67/-0.59**	-0.43/-0.39	-0.47/-0.46
MOS and difficulty	-0.63/-0.61	-0.60/-0.55	-0.62/-0.59	-0.57/-0.53	**-0.74/-0.70**	-0.53/-0.52	-0.56/-0.57
MOS and STD	-0.61/-0.62	-0.61/-0.59	-0.60/-0.62	-0.56/-0.58	**-0.66/-0.66**	-0.56/-0.56	-0.62/-0.58
difficulty and STD	0.24/0.24	0.16/0.15	0.22/0.22	0.16/0.15	**0.37/0.35**	0.13/0.13	0.24/0.24

Difficulty has similar correlation coefficients, but light and color is the most correlated attribute, reaching -0.60 in PCC of all the images, and -0.71 for "natural and rural scenes" category. Difficulty and the attributes always get higher correlation in this category, and get the lowest correlation in "human" and "still life". In general, difficulty is negatively correlated with all the attributes, suggesting somehow that observers find easier to assign scores when they deem images being of high aesthetic quality. However, the small absolute values of the correlations make it difficult to draw precise conclusions at this stage.

Looking at the last group of the table, MOS has a slightly better correlation with the difficulty than STD, especially in "natural and rural scenes" category, with -0.70 and -0.66 in PCC respectively. However, difficulty and STD have weak correlation in both SRCC

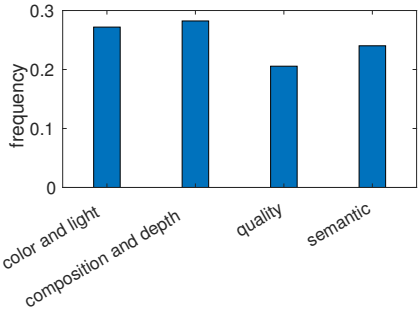

Figure 5: Average probability for one attribute to affect the overall aesthetic judgment.

and PCC, which is 0.24 in general. This implies that average personal difficulty to judge is quite uncorrelated to group disagreement for aesthetic values [8].

4.3.3 Relative importance of attributes in explaining the global aesthetic opinion. In EVA we directly elicit from observers the importance of each attribute in forming their overall aesthetic opinion. As discussed in Section 3.2, voters had to indicate which factor(s) influenced their overall aesthetic score, among the rated attributes. This provides valuable information to explain which features of an image lead to a certain aesthetic score. Figure 5 reports the average probability (over all the images in the dataset) of each attribute to be selected by observers as one of those affecting the overall aesthetic score. We observe that the relative importance of each attribute is related to the correlation between the magnitude of attributes and the overall aesthetic score (Table 1). Quality is the least important attribute for the voters. As mentioned above, this is probably due to the fact that the selected stimuli have a relatively good image quality when displayed on personal screens and phones. Composition and depth is the most influencing attribute, and light and color are slightly less voted. Semantic is more important than quality, but it is not generally deemed the most frequent factor of explanation of the MOS by human observers.

As showed above, there is a quite good linear relationship between the average rating of an attribute for an image and the aesthetic MOS. We can then model the overall aesthetic quality of an image as a weighted sum of the quality of its attributes, that is:

$$s_i = \sum_{j=1}^{4} a_j \cdot f_{ij} \tag{1}$$

where s_i is the MOS for image i, $a_j \geq 0$ is the weight for attribute $j \in \{1, 2, 3, 4\}$ where 1 is for color and light, 2 is composition and depth, 3 is quality, and 4 is semantic. $\sum_j a_j = 1$, and f_{ij} is the average rating of attribute j for image i. In practice, this model should include a bias term to account for the non-centered nature of the data (due to the different use of the rating scales). However, to make our analysis easier to interpret, we assume that both attributes and MOS are first normalized by removing their mean and dividing by standard deviation over the dataset.

It is possible to estimate a_j from data, by solving a constrained least-squares problem, yielding the following solution:

$$s_i \simeq 0.2877 \cdot f_{i1} + 0.2881 \cdot f_{i2} + 0.0821 \cdot f_{i3} + 0.3420 \cdot f_{i4} \tag{2}$$

This descriptive model fits very well our data: the root mean squared error (RMSE) of the MOS estimated by the model is 0.28, which is far below the average standard deviation of the global subjective aesthetic scores collected in the dataset (see Figure 3). This leads to two interesting observations about aesthetic quality assessment. First, despite its simplicity, the linearity assumption can explain effectively how aesthetics is formed. In particular, our model postulates that the weights a_j are *constant* over the dataset. This is generally not true in practice. However, even a simple zero-order approximation of these weights provides valid results: the weights in Equation (2) are coherent with the importance weights collected in the dataset (see Figure 5). Second, we conjecture that the goodness of fit of our linear model is partially due to our choice of attributes in the test design. Even if attribute scores are inter-correlated (a PCA on attribute ratings revealed that the first principal component accounts for almost 80% of the variance of the data), the well-defined nature of the attributes, which describe different qualities of the picture (from perceptual to photographic and semantic) somehow enables to easily disentangle the factors of variation of the aesthetic scores. Notice that a different selection of the attributes may have led to different models, e.g., with non-linear attribute interaction as in [1]. We believe this linear behavior is a valuable feature of EVA that might facilitate obtaining interpretable explanations of aesthetic quality.

We can also estimate importance weights from data for each image category, reported in Figure 6. We compare side by side the weights estimated by linear regression, with the average (normalized) importance weights collected in EVA. We observe that in general they follow a similar trend, with specific differences depending on the image category. In particular for semantics, the discrepancies are more pronounced for landscape/natural scenes and architecture, where perceptual and photographic attributes are predominant. Indeed, the impact of semantics appears to be rather complex and more difficult to describe — the definition of semantic in our dataset is quite broad and may include several co-occurring factors. Further study on this aspect is a promising research avenue for future work on aesthetic assessment.

Finally, by averaging and normalizing the binary votes over attributes, we can get a continuous, per image probability distribution of importance weights. It could then be possible to modify the linear model (1) to have *image-dependent* weights a_{ij}, where this time the weights are *not* computed from data, but directly obtained by eliciting them from voters. By plugging these weights into (1), we obtain MOS predictions with an RMSE of 0.29, just slightly worse than the global weights estimated through linear regression. This is a surprisingly good result, considering that these weights are not optimized to minimize the fitting error as in Equation (2). This validates the quality of the collected weights as a means to effectively explain aesthetics, and provides valuable ground-truth for future research on image aesthetics.

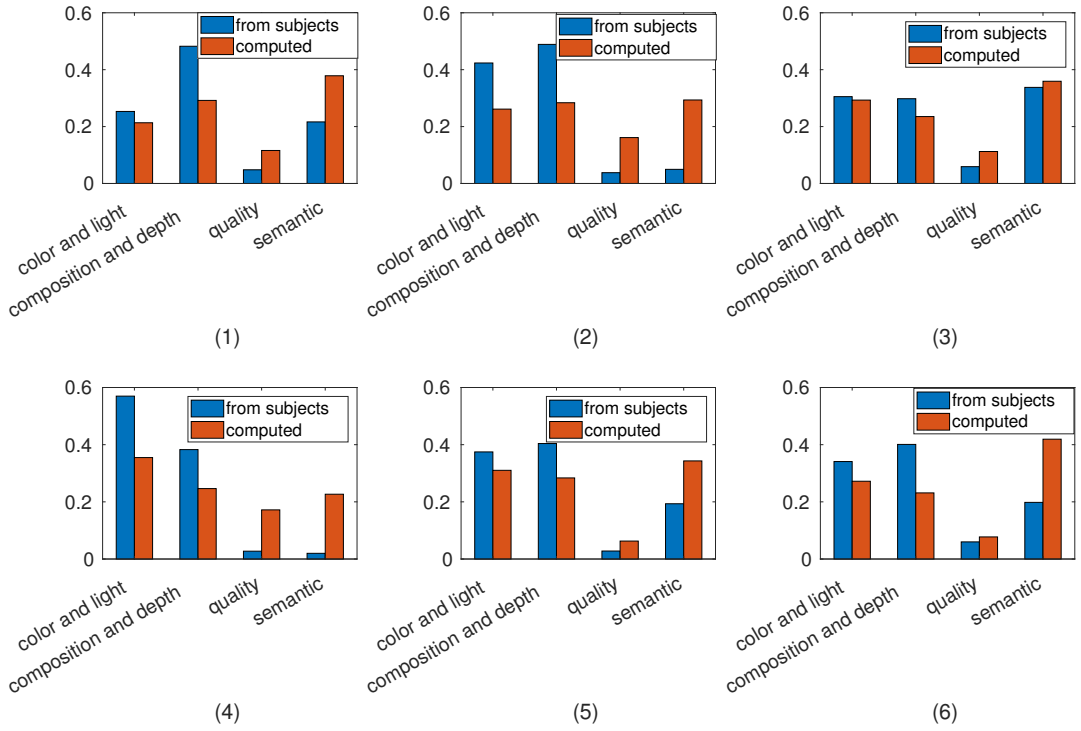

Figure 6: Average distribution of attributes importance per content category: (1) animals (2) architectures and city scenes (3) human (4) natural and rural scenes (5) still life (6) other

5 CONCLUSIONS AND FUTURE WORK

We propose EVA, the first annotated image dataset for explaining visual aesthetics. It contains 4070 annotated images with 30 to 40 votes per image, collected using a disciplined approach including subject training and unambiguous definition of aesthetic attributes inspired by traditional quality assessment guidelines. As a results, EVA overcomes the limitations of previously proposed datasets, in particular noisy labels due to misinterpretations of the tasks or limited number of votes per image. At the same time, it offers a number of novel features, including the degree of difficulty in judging the aesthetic level of a picture; the magnitude of 4 different aesthetic attributes spanning various level of the aesthetic appraisal (from perceptual to photographic and semantic aspects); as well as their relative importance in forming the overall aesthetic score.

Statistical analysis on the collected data shows that the chosen attributes are linearly related to the overall aesthetic score. This leads to proposing a simple, yet effective, linear model to explain aesthetic score formation. We find that the subjective importance weights expressed by observers provide a surprisingly good fit to data under this model, which demonstrates the goodness of the collected dataset. In particular, EVA enables to estimate the importance of each aesthetic factor *per image*, thus effectively enabling the explanation of aesthetic scores.

The data in EVA is made publicly available and opens up several new research avenues in aesthetics, including a better understanding of uncertainty in aesthetic evaluation and of the link between aesthetic assessment and demographic/cultural background of the observers, as well as the disentanglement of the factors of variation of aesthetic quality.

ACKNOWLEDGMENTS

This work is financially supported by China Scholarship Council (CSC). We thank all the participants who took part in the experiment.

REFERENCES

[1] Tunç Ozan Aydın, Aljoscha Smolic, and Markus Gross. 2014. Automated aesthetic analysis of photographic images. *IEEE transactions on visualization and computer graphics* 21, 1 (2014), 31–42.

[2] Huiwen Chang, Fisher Yu, Jue Wang, Douglas Ashley, and Adam Finkelstein. 2016. Automatic triage for a photo series. *ACM Transactions on Graphics (TOG)* 35, 4 (2016), 148.

[3] Mihai Gabriel Constantin, Chen Kang, Gabriela Dinu, Frédéric Dufaux, Giuseppe Valenzise, and Bogdan Ionescu. 2019. Using Aesthetics and Action Recognition-based Networks for the Prediction of Media Memorability. In *MediaEval Benchmarking Initiative for Multimedia Evaluation (MediaEval)*. MediaEval 2019 Workshop, Sophia Antipolis, France.

[4] Yubin Deng, Chen Change Loy, and Xiaoou Tang. 2017. Image aesthetic assessment: An experimental survey. *IEEE Signal Processing Magazine* 34, 4 (2017), 80–106.

[5] Yubin Deng, Chen Change Loy, and Xiaoou Tang. 2018. Aesthetic-driven image enhancement by adversarial learning. In *2018 ACM Multimedia Conference on*

Multimedia Conference. ACM, Seoul, Korea, 870–878.

[6] Kevin A Hallgren. 2012. Computing inter-rater reliability for observational data: an overview and tutorial. *Tutorials in quantitative methods for psychology* 8, 1 (2012), 23.

[7] Magzhan Kairanbay, John See, and Lai-Kuan Wong. 2018. Towards Demographic-Based Photographic Aesthetics Prediction for Portraitures. In *International Conference on Multimedia Modeling*. Springer, Bangkok, Thailand, 531–543.

[8] Chen Kang, Giuseppe Valenzise, and Frédéric Dufaux. 2019. Predicting Subjectivity in Image Aesthetics Assessment. In *2019 IEEE 21st International Workshop on Multimedia Signal Processing (MMSP)*. IEEE, Kuala Lumpur, Malaysia, 1–6.

[9] Yueying Kao, Ran He, and Kaiqi Huang. 2017. Deep aesthetic quality assessment with semantic information. *IEEE Transactions on Image Processing* 26, 3 (2017), 1482–1495.

[10] Shu Kong, Xiaohui Shen, Zhe Lin, Radomir Mech, and Charless Fowlkes. 2016. Photo aesthetics ranking network with attributes and content adaptation. In *European Conference on Computer Vision*. Springer, Amsterdam, Netherlands, 662–679.

[11] Wentao Liu and Zhou Wang. 2017. A database for perceptual evaluation of image aesthetics. In *2017 IEEE International Conference on Image Processing (ICIP)*. IEEE, Beijing, China, 1317–1321.

[12] Xin Lu, Zhe Lin, Hailin Jin, Jianchao Yang, and James Z Wang. 2015. Rating image aesthetics using deep learning. *IEEE Transactions on Multimedia* 17, 11 (2015), 2021–2034.

[13] Naila Murray, Luca Marchesotti, and Florent Perronnin. 2012. AVA: A large-scale database for aesthetic visual analysis. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, Providence, Rhode Island, 2408–2415.

[14] Kyung-Wha Park, JungHoon Lee, Sunyoung Kwon, Jung-Woo Ha, Kyung-Min Kim, and Byoung-Tak Zhang. 2019. Which Ads to Show? Advertisement Image Assessment with Auxiliary Information via Multi-step Modality Fusion. *arXiv preprint arXiv:1910.02358* (2019).

[15] Xueming Qian, Cheng Li, Ke Lan, Xingsong Hou, Zhetao Li, and Junwei Han. 2017. POI summarization by aesthetics evaluation from crowd source social media. *IEEE Transactions on Image Processing* 27, 3 (2017), 1178–1189.

[16] Nornadiah Mohd Razali, Yap Bee Wah, et al. 2011. Power comparisons of shapiro-wilk, kolmogorov-smirnov, lilliefors and anderson-darling tests. *Journal of statistical modeling and analytics* 2, 1 (2011), 21–33.

[17] Joseph Redmon and Ali Farhadi. 2018. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767* (2018).

[18] Flávio Ribeiro, Dinei Florencio, and Vítor Nascimento. 2011. Crowdsourcing subjective image quality evaluation. In *2011 18th IEEE International Conference on Image Processing*. IEEE, 3097–3100.

[19] Katharina Schwarz, Patrick Wieschollek, and Hendrik PA Lensch. 2018. Will people like your image? learning the aesthetic space. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, Lake Tahoe, United States, 2048–2057.

[20] Ernestasia Siahaan, Alan Hanjalic, and Judith Redi. 2016. A reliable methodology to collect ground truth data of image aesthetic appeal. *IEEE Transactions on Multimedia* 18, 7 (2016), 1338–1350.

[21] Hossein Talebi and Peyman Milanfar. 2018. NIMA: Neural image assessment. *IEEE Transactions on Image Processing* 27, 8 (2018), 3998–4011.

[22] Xiaoou Tang, Wei Luo, and Xiaogang Wang. 2013. Content-based photo quality assessment. *IEEE Transactions on Multimedia* 15, 8 (2013), 1930–1943.

[23] Chaoqun Wan and Xinmei Tian. 2017. A Small Scale Multi-Column Network for Aesthetic Classification Based on Multiple Attributes. In *International Conference on Neural Information Processing*. Springer, Guangzhou, China, 922–932.

[24] Wenshan Wang, Su Yang, Weishan Zhang, and Jiulong Zhang. 2019. Neural aesthetic image reviewer. *IET Computer Vision* 13, 8 (2019), 749–758.

[25] Stefan Winkler. 2009. On the properties of subjective ratings in video quality experiments. In *International Workshop on Quality of Multimedia Experience*. IEEE, 139–144.

[26] Ye Zhou, Xin Lu, Junping Zhang, and James Z Wang. 2016. Joint image and text representation for aesthetics analysis. In *Proceedings of the 24th ACM international conference on Multimedia*. ACM, Amsterdam, Netherlands, 262–266.