SAVOIAS: A Diverse, Multi-Category Visual Complexity Dataset

Elham Saraee, Mona Jalal, Margrit Betke
Department of Computer Science, Boston University, USA

esaraee@bu.edu, jalal@bu.edu, betke@bu.edu
https://www.bu.edu/cs/ivc/

Abstract

Visual complexity identifies the level of intricacy and details in an image or the level of difficulty to describe the image. It is an important concept in a variety of areas such as cognitive psychology, computer vision and visualization, and advertisement. Yet, efforts to create large, downloadable image datasets with diverse content and unbiased groundtruthing are lacking. In this work, we introduce SAVOIAS a visual complexity dataset that compromises of more than 1,400 images from seven image categories relevant to the above research areas, namely Scenes, Advertisements, Visualization and infographics, Objects, Interior design, Art, and Suprematism. The images in each category portray diverse characteristics including various low-level and high-level features, objects, backgrounds, textures and patterns, text, and graphics. The ground truth for SAVOIAS is obtained by crowdsourcing more than 37,000 pairwise comparisons of images using the forced-choice methodology and with more than 1,600 contributors. The resulting relative scores are then converted to absolute visual complexity scores using the Bradley-Terry method and matrix completion. When applying five state-of-the-art algorithms to analyze the visual complexity of the images in the SAVOIAS dataset, we found that the scores obtained from these baseline tools only correlate well with crowdsourced labels for abstract patterns in the Suprematism category (Pearson correlation r=0.84). For the other categories, in particular, the objects and advertisement categories, low correlation coefficients were revealed (r=0.3 and 0.56, respectively). These findings suggest that (1) state-of-the-art approaches are mostly insufficient and (2) SAVOIAS enables category-specific method development, which is likely to improve the impact of visual complexity analysis on specific application areas, including computer vision.

Figure 1: Sample images of the Art category of SAVOIAS (styles Suprematism, Socialist Realism, Biedermeier, and Cubo-futurism [47]) shown in increasing order of ground-truth visual complexity scores (10, 55, 71, and 83 out of 100).

1. Introduction

Visual complexity is a broad concept with decades of basic and applied research in a variety of areas such as psychophysics and cognitive psychology, product design, marketing, and computing. Various definitions can be found in the literature. One definition relates visual complexity to the level of intricacy and details in an image or the level of difficulty to describe an image [24, 45], which can be measured by extracting features from the image [8]. Another definition of visual complexity is based on the level of visual clutter and the amount of information conveyed in the image, which makes the study of visual complexity related to image compression and information theory [43].

Analysis of visual complexity facilitates assessment of how a human observer may perceive the image, for exam-

ple, its aesthetic appeal, and how the observer may interact with it. It has applications in a variety of research and technology fields. Computer vision, graphics, art, interior design, Web design, marketing and advertising are a few examples of the fields that can benefit from the analysis and computation of visual complexity. Visual complexity is connected to a variety of problems in the computer vision field. Visually complex images would need more sophisticated algorithms to perform visual search in the image or create a caption for it [32]. It is more challenging to detect or segment objects in an occluded scene where the object might be partially masked. Moreover, a visual question answering algorithm could benefit from analysis of the complexity of image regions – a visually complex region is likely to need more algorithm-generated questions and answers. In addition to computer vision problems, understanding visual complexity of images is beneficial in the context of computer graphics. For example, the more complex a 3D scene is, the more time would take for an algorithm to render it [41].

Visual complexity is a dominant factor in determining the pleasingness of a stimulus and is known to be related to aesthetic preference for artistic works [19]. In "optimal arousal theory," it is shown that the relation between visual complexity and preference follows an inverted U-curve in which images with intermediate level of visual complexity are most appealing [2]. In addition, it has been studied that the perception of appeal from a Web design has connection with the visual complexity of the Website [42], and thus understanding visual complexity can lead to a better subjective experience for users [31]. Furthermore, advertisement is the first step in eventually persuading consumers to adopt the brand; therefore, advertisers are always looking for ways to motivate consumer engagement with their work. Visual complexity, as one of the forms of complexity in an advertisement can increase arousal by increasing the cognitive demands in the customer [25].

For quantifying the visual complexity, various factors should be taken into account. The number of objects, colorfulness, edge density, luminance, patterns, and mirror symmetry are some of the examples. However, depending on the type of image, e.g., abstract patterns versus real-world scenes, the contribution of these factors are different. Due to the wide range of applications for the analysis of visual complexity, the study of visual complexity requires adequate amount of data for different types of images. In addition, the visual complexity ground truth for the data should be obtained in consistent set of experiments to facilitate the development of new objective algorithms for the analysis of visual complexity.

Studies of visual complexity has a long history in the literature [2, 3, 10, 16] and in the past few years, various datasets have been developed [10, 14, 21, 36]. These datasets, however, are insufficient for the applications mentioned above. They are either very small or lack diversity (both number and type of categories, as well as diversity of image content and appearance within each category). In addition, the methodology to obtain the ground truth for these datasets is not consistent.

In this paper, we make three contributions to the study of visual complexity in images:

1. We introduce SAVOIAS, a dataset for the analysis of visual complexity. SAVOIAS covers a variety of topics and provides a sufficient number of images per topic, therefore improving diversity and scale of publicly available datasets. Specifically, SAVOIAS consists of seven diverse categories and a total of 1,420 images. SAVOIAS is an acronym for **S**cenes, **A**dvertisement, **V**isualization and infographics, **O**bjects, **I**nterior design, **A**rt, and **S**uprematism (a category of art).

2. In order to minimize the potential bias from individual ground-truth contributors and limited rating scales, we obtained the ground-truth labels using a forced-choice crowdsourcing methodology on a [0,100] rating scale. Labels were obtained from 1,687 contributors on more than 37,000 pairs of images. The pairwise scores were then converted to absolute scores using the Bradley-Terry algorithm and matrix completion.

3. We explore the applicability of five state-of-the-art algorithms on the different categories of our dataset and report the gap between their performance and the ground truth. Our results highlight the need for the development of new, potentially category-specific techniques.

2. Related Work

Visual complexity has been previously studied extensively in the literature of psychophysics, cognitive science, and more recently in computer vision. While the temporal dimension of complexity is also an interesting topic of research [7, 33, 38]; in this work, we focus on the spatial dimension of visual complexity and algorithmic approaches to quantify it.

Applications of Visual Complexity. Existence of a linear relation between visual complexity and aesthetic beauty has been previously stated in the literature [16, 42]. Jacobsen and Höfel [28] argued that symmetry and visual complexity are two predictors for aesthetic judgment of beauty. Birkhoff came up with the formula $M = O/C$, where M refers to aesthetic measurement, O refers to aesthetic order, and C refers to complexity. Based on his formula, beauty decreases with increase in complexity [3].

Furthermore, studies show that in hedonic shopping experiences, the higher perceived complexity of mall interiors yields a higher satisfaction score while in utilitarian

Reference	# Images	Application Category	Ground-truth Process	Vis. Comp. scale	Open Source
Gartus *et al.* [20]	912	Black and white 8×8 abstract patterns	1-step	5-point	Shared
Nadal *et al.* [36]	120	Abstract & representational (artistic & non-artistic)	1-step	3-point	No
Olivia *et al.* [37]	100	Indoor scenes	3-step	8-point	No
Miniukovich *et al.* [35]	140	Webpage	1-step	5-point	No
Fan *et al.* [17]	40	Chinese ink painting (abstract and representational)	1-step	7-point	No
Schnur *et al.* [44]	9	Web maps	1-step	5-point	No
Corch *et al.* [14]	98	Real-world scenes	1-step	[0-100]	Yes
Corch *et al.* [14]	122	Textures	1-step	[0-100]	Shared
Ours	**1,420**	**Scenes, advertisement, visualizations, objects, interior design, art, Suprematism**	**Pairwise comparison**	**[0-100]**	**Yes**

Table 1: The datasets previously used in visual complexity studies, as well as our proposed dataset. A "1-step" groundtruthing process means that the users were asked to directly rate the visual complexity of a single image; "Shared" means the dataset may be shared with other researchers upon request.

shopping experiences, people prefer lower visual complexity [23]. For online shopping, the perceived visual complexity has been shown to negatively influence individuals satisfaction [46].

Visual impression of advertisement images plays a crucial role in engaging visitors. These first impressions further affect the mid- and long-term human behavior. Pieters *et al.* [39] have considered two different types of complexity for ads: feature complexity (depends on the visual features), and design complexity (depends on the creative design) both of which indicate perceptual complexity. They argued that feature complexity hurts attention to the brand, whereas design complexity can improve the consumer's attention.

Algorithms. In one of the early works on visual complexity, Chipman *et al.* explained the importance of two factors in the analysis of visual complexity, specifically, quantitative factor (related to amount of elements) which has positive correlation with visual complexity and structural element (determined by different forms of structural organization, but mostly by symmetry) which has negative correlation with visual complexity [10, 11]. In a similar approach, a more recent study explored the impact of these two factors on the visual complexity of more complex abstract patterns [20, 21].

The applicability of this method was further demonstrated by Nadal *et al.* on four categories: abstract artistic, abstract non-artistic, representational artistic, and representational non-artistic. The images used in their experiment contained 120 stimuli equally divided into three complexity levels: low, intermediate, and high [36].

Oliva *et al.* studied the impact of task constraint on representation of visual complexity by creating a visual complexity dataset of 100 indoor images. They argued that although the contribution of the perceptual dimensions are affected by task constraints, visual complexity can still be represented by perceptual dimensions such as quantity of objects, clutter, openness, symmetry, organization, and variety of colors [37]. In addition, five factors encapsulating visual complexity for web design and GUI applications are proposed [35]. Similarly, Fan *et al.* expanded this approach to Chinese ink paintings by introducing three new features, namely color richness, stroke thickness, and white spaces [17]. Furthermore, in a more recent study, linear combination of multiple features such as edge density, compression ratio, and number of objects was proposed. The images used in their experiments consists of real-world scene images and textures. They obtained ground truth on a scale in the range of 0 to 100.

Visual complexity can also be approximated using algorithmic information theory and compression algorithms [43]. In this case, the visual complexity is defined as the resulting file size when a compression algorithm such as JPEG or ZIP is applied on the given image. The stimuli used in these experiments were geographical maps and synthetic images of objects in sparse arrangements. In another study, visual complexity of three different online map providers (Google Maps, Bing Maps, and OpenStreetMaps) was explored with the objective of better understanding design decisions for Web maps. Their results implied that clutter, measured by feature congestion [43], is more important in perceived complexity than diversity of symbology [44].

The characteristics of the datasets mentioned in this section are summarized in Table 1. It shows differences in scales and image collection methods, as well as groundtruthing methodologies, and reveals that the diversity and number of samples in these datasets are inadequate for extensive analysis of visual complexity. It is also worth noting that not all the datasets mentioned above are publicly available or shared among researchers.

SAVOIAS, is introduced to address the lack of an appropriate dataset. It is a new dataset consisting of seven diverse categories and more than 1,400 sample images. In order to obtain ground truth for our dataset, instead of asking the

participants to rate the visual complexity of an image on a continuous scale, we incorporated a pairwise comparison between images to avoid any bias in the rating scale. The pairwise methodology also provides a more fine-grained range of scores compared to the common 3, 5, or 7-scale ratings used in the aforementioned datasets.

3. Dataset Description

Our dataset is the largest and most diverse open-source visual complexity dataset with 1,420 images in seven categories. In this section, we will first explain the image collection process and the different categories that we have used in our dataset. Next, we will discuss the data annotation using the Figure-Eight, a crowdsourcing platform.

3.1. Image Collection

In our dataset, we have used images from seven diverse categories. Examples from each category in the increasing order of visual complexity are shown in Figure 2. Despite the connection between the problem of visual complexity and other problems in computer vision, there exists a gap between these topics. To overcome this gap, we have collected images from commonly-used datasets with the following categories:

Advertisement: 200 images from the Advertisement dataset [26]. Visual impression of advertisement plays a crucial role in economical competitions [40]. This category is selected in order to give ad designers insight into what factors impact the perceived complexity of an advertisement. Examples of factors that contribute to visual complexity of an ad are amount of text, size of text, importance of brand logo, and number and composition of various elements. The images in this category include both advertisements for products, and ads that campaign for/against something, e.g. for preserving the environment.

Objects: 200 images from the MSCOCO dataset [32]. The purpose of this category is to understand how human perceives visual complexity of various objects and combination of objects. The number of objects is one of the leading factors contributing to the visual complexity of an image. This category can help researchers study the impact of characteristics of objects as well as the number of objects and their interaction with each other on visual complexity.

Scene: 200 images from the Places 2 dataset [48]. The purpose of this category is to understand how humans perceive visual complexity of various scenes. It may facilitate the study of the roles of the image foreground and background in visual complexity analysis.

Interior Design: We have collected 100 interior design images from the IKEA website [27]. This category is specifically selected to provide insight into how humans perceive the visual complexity of indoor spaces at home such as bedroom, living room, dining room, kitchen, and bathroom. Interior designers may want to understand how to design appealing interior spaces.

Visualization and Infographics: 200 images from the MASSViS dataset [4]. Visualization and infographics consists of charts, graphs, texts, and tables. Understanding the impact of each of these elements as well as their composition, thus, understanding the cognitive and perceptual processing of a visualization, can greatly influence the memorability, recognition, and comprehension of these designs.

Art: 420 Artistic images from the PeopleArt dataset [47]. This category consists of 10 images from each of the 42 categories of art styles and movements including Naturalism, Cubism, Socialist Realism, Impressionism, and Suprematism. Since the aesthetic beauty of an artistic image is directly influenced by the level of its visual complexity [16, 42], understanding the visual complexity of an artistic image can definitely help the artists to create more engaging artworks.

Suprematism: 100 images from the Suprematism category in the PeopleArt dataset for the analysis of geometric abstract art. The Suprematism category conveys various geometric shapes and objects in abstract form. This category enables studying the impact of various shapes, geometric objects, and composition on the perception of visual complexity.

3.2. Dataset Groundtruthing

3.2.1 Pairwise Comparison Algorithm

In order to obtain absolute ranking scores for an attribute of an image, in our case, visual complexity, one approach would be to ask users to assign a score to each image, where the score represents the ranking of the image relative to all other images. However, it has been shown that most people can only evaluate 5 to 9 options at a time. In addition, bias in the rating scale is a common problem in this type of groundtruthing [34].

The use of pairwise comparison and conversion of the pairwise ranking to global ranking is a better alternative [1, 15, 29]. Pairwise comparison is a relative measure that helps reduce bias from the rating scale. It is also invariant under monotone transformation of the rating values and depends only on the degree of relative difference between one option over the other in the pair [22].

Note that, for example, for a set of $n = 200$ images, including all of the pairs would result in $\binom{n}{2} = 19,900$ comparisons.

However, it is shown that for the pairwise comparisons, not all of the pairs are required in order to get the final global ranking, and information about a percentage of the pairs, $\ell \ll \binom{n}{2}$, is adequate [9].

In many practical applications with partially observed measurements or budget constraints (e.g. PIB-PET scan for Alzheimer's disease detection), it is possible to use ma-

Table 2: Sample images of the SAVOIAS dataset with increased visual complexity from left to right in each row.

trix completion methods in order to complement the results [6, 22, 30].

n this work, we follow the pairwise comparison approach. Our algorithm is iterative and selects two images randomly from the set of images in a particular category in each step. Images that have been selected in previous steps are less probable to be chosen in subsequent steps. The algorithm terminates once a target number of comparisons is reached. We decided on different target numbers for different categories, assuming that the visual complexity of images in some categories is easier to evaluate by human judgment than in others. For the categories scenes, advertisement, visualization, and objects we decided to run our algorithm until $\ell =$ 4,000 pairs are found, which results in 40 comparisons per each image, on average, given that the these categories have $n =$ 200 images each. For the interior design category, we ran the algorithm until $\ell =$ 2,000 pairs are found, also resulting in 40 comparisons per image, on average. For the art category, we obtain $\ell =$ 14,700 pairs, which results in 70 comparisons, per image, on average. Finally, for the Suprematism category, we used all possible $\ell = \binom{n}{2} = 4{,}950$ pairs, which resulted in 99 comparisons per image.

3.2.2   Crowdsourcing Methodology

To minimize potential bias caused by raters, we collected our human judgments via crowdsourcing. We used the Figure-Eight platform [18]. Each task was distributed to five contributors. Contributors (also known as crowdworkers in Amazon Mechanical Turk) were shown ten pairs of images per page and asked which of the two images in each pair is visually more complex. We explained visual complexity by attributes such as cluttered background, numerosity and variety of objects, people, textures, patterns, and shapes. We are using a forced-choice methodology, in which the contributors are supposed to select either image A or image B (Figure 2). In the case of similarly visually complex images, contributors are requested to select intuitively which image is more visually complex.

Our contributors are selected from Figure-Eight "level-3 contributors" in Figure-Eight, who have shown to produce accurate answers in previous work. Each contributor is shown 10 pairwise comparison task in a page and for each page they are paid $0.10 regardless of their choices. We did not allow any contributor to perform more than 300 tasks but did not select a lower bound for the number of tasks. Workers were not restricted by geographical locations.

Test questions, geared towards quality control, were distributed to contributors randomly throughout the entire job to make sure they are attentive to the task. While all the comparison tasks were paid, only the answers from the contributors who maintain a passing score of 90% or above on test questions were kept.
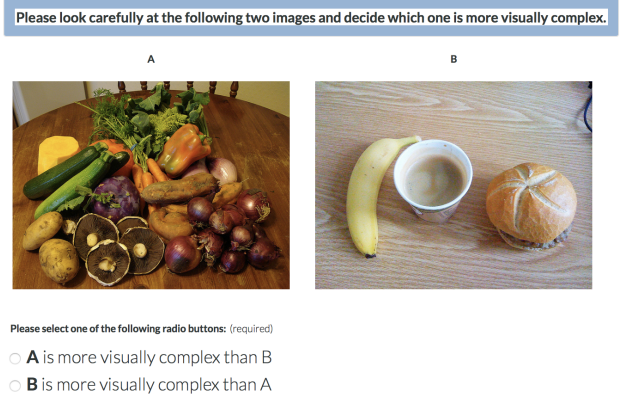

Figure 2: Screenshot of one of the pairwise comparisons shown to Figure-Eight contributors for the objects category.

3.2.3   Pairwise Ranking versus Global Ranking

In this work, we follow the pairwise comparison approach where the users on the crowdsourcing platform are asked to select whether image A or B is more visually complex. After the pairwise scores are collected, they need to be converted to absolute visual complexity scores. In order to convert pairwise ranking of images to global ranking, we applied two separate approaches, namely the Bradley-Terry method [5] and matrix completion. We then compare our results obtained from these two methods to validate our absolute scores.

We denote the pairwise comparison matrix as a count matrix $S = \{s_{i,j}\}$, where $s_{i,j}$ is the ratio of number of times that the contributors have selected image i as more visually complex compared to image j over the total number of times that image i and j have been compared. Thus, $s_{i,j} + s_{j,i} = 1$. The problem here is to find c_i, the absolute score of image i.

The **Bradley-Terry method** [5] describes the probability of choosing image I_i over image I_j as a Sigmoid function of the score difference between the two images,

$$P(I_i > I_j) = F(\Delta_{i,j}) = \frac{e^{\Delta_{i,j}}}{1 + e^{\Delta_{i,j}}}, \qquad (1)$$

where $\Delta_{i,j} = c_i - c_j$. The score parameter c can be estimated by solving a maximum a posteriori (MAP) problem, i.e., maximizing

$$\log Pr(S|c) = \Sigma_{i,j} s_{i,j} F(\Delta_{i,j}), \qquad (2)$$

where the prior is a uniform distribution. This optimization problem can be solved using gradient descent [9].

Matrix Completion method assumes, if $s_{i,j}$ is greater than 0.5 (image i is more visually complex than image j), $s_{j,k}$ is greater than 0.5, and the pairwise comparison between image i and k is missing, using image j as a link,
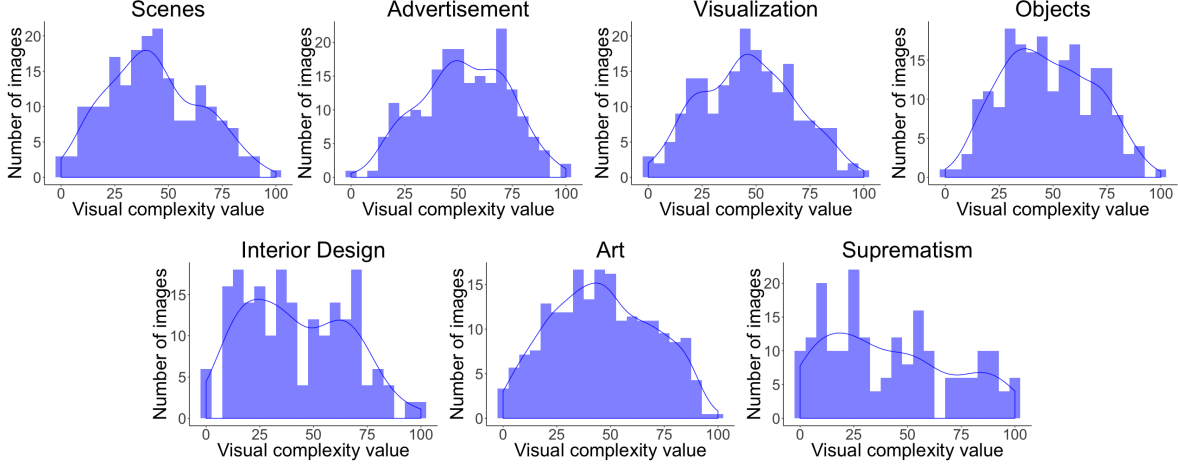
Figure 3: Distribution of absolute visual complexity scores per category for the SAVOIAS dataset.

we can infer that image $s_{i,k}$ is also greater than 0.5. Now we can create matrix $\hat{S}$ by filling the missing elements of matrix S:

$$\hat{s}_{i,k} = \begin{cases} s_{i,k} & \text{if } s_{i,k} \in S \\ \frac{1}{m}\Sigma_{j=1}^{m}\frac{s_{i,j}+s_{j,k}}{2} & \text{else if } s_{i,j} \in S, \ s_{j,k} \in S \\ & \text{and } s_{i,j} > 0.5, s_{j,k} > 0.5, \end{cases} \quad (3)$$

where m is the number of existing pairs of $s_{i,j}$ and $s_{j,k}$ in the count matrix S. For matrix completion, note the following points:

- As indicated in the formula, we only consider $s_{i,j}$ and $s_{j,k} \in S$ if they are greater than 0.5. Therefore, if $s_{i,j} > 0.5$ and $s_{j,k} < 0.5$, we will not make any judgments about the missing pair s_{ik}.

- For those pairs for which we have the result in one direction, we can fill the matrix in the other direction by this formula: $\hat{s}_{k,i} = 1 - \hat{s}_{i,k}$.

- For the rare case that a pair is not connected in either of directions, we use $\hat{s}_{i,k} = \hat{s}_{k,i} = 0.5$.

When the count matrix $\hat{S}$ is completed, the absolute score for each image is the mean of the pairwise scores for that image:

$$c_i = \frac{1}{n} \sum_{j=1, \ \hat{s} \in \hat{S}}^{n} \hat{s}_{i,j}. \quad (4)$$

To confirm the correctness of the two aforementioned methods and our final scores, we evaluated the correlation between the global ranking scores obtained from these two methods. For all image categories, we obtained correlations higher than 0.98 between the two methods. We will report the results based on the Bradley-Terry method in Section 4.

4. Results

In this section, we present an analysis of the visual complexity scores of SAVOIAS, discuss the results of the crowd-sourced groundtruthing procedure, and describe the performance of five state-of-the-art algorithms on SAVOIAS.

4.1. Distribution of the ground-truth scores

Initial analysis of the distribution of the absolute scores showed that the absolute scores are mostly distributed around zero. To mitigate this issue, we rescaled the range of pairwise scores, so that they are in the interval of $[0.33, 0.66]$ instead of $[0, 1]$, while still maintaining 0.5 as the score that represents equal visual complexity of an image pair.

Visual inspection of Figure 3, which presents the distribution of scores for the seven categories, shows that the rescaling step was successful – each histogram is well distributed among the range of visual complexity numbers.

4.2. Results on the Groundtruthing Methodology

The number of crowdplatform contributors who provided judgments on visual complexity for SAVOIAS was 1,687.

Their overall satisfaction score was 4.2 out of 5 (Table 4). The satisfaction score is a combination of "instructions clear," "test questions fair," "ease of job," and "pay."

Validity of partial matrix versus full matrix comparison: Here, we evaluate the accuracy of the absolute visual complexity scores as a function of ℓ, number of pairwise comparisons between images. Since we have the full matrix comparison for the Suprematism category, we can perform such an analysis. Recall the notations from the Section 3.2.3, where $S = \{s_{i,j}\}$ is the count matrix for the pairwise comparisons and $C = \{c_i\}$ is the list of absolute visual complexity scores, i.e., the output of the Bradley-Terry

Model	Edge Density [43]	Compression Ratio [13]	Number of Regions [12]	Feature Congestion [43]	Subband Entropy [43]
Scenes	0.16	0.30	**0.57**	0.42	0.16
Advertisement	0.54	**0.56**	0.41	**0.56**	0.54
Visualization	0.57	0.55	0.38	0.52	**0.61**
Objects	0.28	0.16	0.29	**0.30**	0.10
Interior Design	0.61	**0.68**	0.67	0.58	0.31
Art	0.48	0.51	**0.65**	0.22	0.33
Suprematism	0.18	0.60	**0.84**	0.48	0.39

Table 3: Baseline results for the SAVOIAS dataset. We use the Pearson correlation coefficient to measure the performance of five algorithms. The object category is the most challenging category in our dataset – the average correlation between ground-truth score and the score of the best performing algorithm (Feature Congestion) is only 0.30.

	Scenes	Ad.	Vis.	Objects	Interior	Art	Sup.	Avg.
Overall Satisfaction	4.2	4.3	4.1	4.5	4.1	4.3	**3.8**	4.2
Ease of Job	4.1	4	4	4.3	3.9	4.1	**3.6**	4

Table 4: Contributor satisfaction scores and ease of job scores for each of the crowdsourced datasets. Scores are reported out of 5 and provided by 237 internet workers.

Figure 4: For the category Suprematism, the correlation between C_ℓ and C_f is shown as a function of ℓ. High correlations can be achieved for some values of $\ell \ll \binom{n}{2}$.

algorithm. We define S_ℓ and C_ℓ as the count matrix and absolute scores where ℓ number of pairs have been selected for crowdsourcing.

We define S_f and C_f as the full count matrix and resulting absolute scores, where $\ell = \binom{n}{2}$.

The correlation between the visual complexity scores based on C_ℓ and C_f for ℓ in the range of $[100 - 4950]$ is shown in Figure 4, which highlights the trade-off between the accuracy and the number of pairwise comparisons. For example, if only 2,000 pairs had been chosen to define S_ℓ for the Suprematism category, the result would be close to the result of a full comparison, since the correlation between C_l and C_f is 0.96. Given this result for the Suprematism category, we hypothesize that high correlations can also be achieved for the other six categories if ℓ is selected to be much smaller than $\binom{n}{2}$.

4.3. Baseline Results

We evaluated the performance of five state-of-the-art algorithms on our dataset and explored which algorithm is more suitable to be used for images in a certain category. We used the Pearson correlation coefficient as the metric to evaluate the performance of each of these algorithms. As shown in Table 3, none of the algorithms can provide a correlation higher than 0.6 for the objects, scenes, and advertisement categories. This highlights the need for a more sophisticated algorithm for categories of images that contain human-identifiable regions such as objects or people.

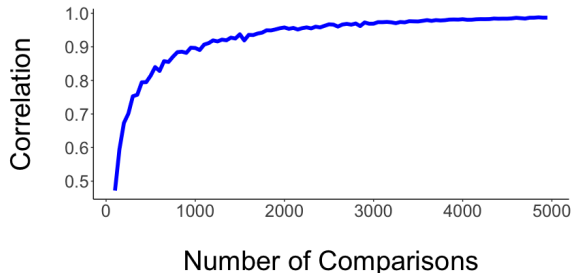Furthermore, the best-performing algorithm is different for each category of images. This can be attributed to the diverse characteristics of each category. A single algorithm that works well on all categories does not exist yet.

4.4. Discussion

Comparing the correlations between visual complexity scores based on crowdplatform contributors and the state-of-the-art algorithms, we observe that Suprematism, which is the most challenging category for the contributors, had the highest correlation for one of the algorithms. On the other hand, contributors found the object category to be the least challenging category, while all the algorithms performed poorly on this category (the highest correlation is only 0.3). Based on this observation, we postulate that the tested algorithms are more capable of making decisions based on the low-level features such as geometric shapes, textures, and patterns, found in the Suprematism category, than image features such as objects and people, which are easier for human contributors.

Furthermore, by analyzing the distribution of the ground-truth scores for each category visually (Figure. 3) and computationally (by estimating kurtosis), we observe that most of the categories have a distribution that resembles a Gaussian. The Suprematism category, however, has a more uniform distribution (kurtosis = 1.915). This obser-

vation shows that there exist some images in the Suprematism category that are either very visually complex or not visually complex at all, and the contributors were able to distinguish them. For the other images in this category, it was not a trivial task for the contributors to recognize minor differences visual complexity, and so the they evaluated the "ease of job" to be low (Table 4).

5. Conclusions

In this work, we introduced SAVOIAS, a new dataset for the analysis of visual complexity in images. SAVOIAS compromises of more than 1,400 images, which belong to seven diverse categories. The ground-truth values were obtained by processing the judgments of 1,687 crowdplatform contributors who compared the visual complexity of more than 37,000 pairs of images.

We suggest that the diverse range of image features found in SAVOIAS can support the development of new algorithms for quantifying visual complexity, and, moreover, be leveraged in research in other areas of computer vision, such as segmentation, object recognition and detection, visual search, image captioning, and visual question answering. The most promising route might be to design algorithms that try to evaluate human-identifiable image features. Furthermore, our proposed dataset can facilitate the research in the fields of psychophysics and cognitive science to find the underlying factors in the stimulus that affect the perception of visual complexity in humans. Lastly, artists, Web and graphic designers, interior designers, and advertisers may eventually benefit from our dataset once better algorithms are designed for the categories relevant to their fields.

References

[1] K. J. Arrow. A difficulty in the concept of social welfare. *Journal of political economy*, 58(4):328–346, 1950.

[2] D. E. Berlyne. *Aesthetics and Psychobiology*. New York: Appleton-CenturyCrofts, 1971.

[3] G. D. Birkhoff. *Aesthetic Measure*, volume 38. Harvard University Press Cambridge, 1933.

[4] M. A. Borkin, A. A. Vo, Z. Bylinskii, P. Isola, S. Sunkavalli, A. Oliva, and H. Pfister. What Makes a Visualization Memorable? *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2306–2315, 2013.

[5] R. A. Bradley and M. E. Terry. Rank Analysis of Incomplete Block Designs: I. The Method of Paired Comparisons. *Biometrika*, 39(3/4):324–345, 1952.

[6] E. J. Candès and B. Recht. Exact matrix completion via convex optimization. *Foundations of Computational mathematics*, 9(6):717–772, 2009.

[7] M. Cardaci, V. Di Gesu, M. Petrou, and M. E. Tabacchi. Attentional vs Computational Complexity Measures in Observing Paintings. *Spatial vision*, 22(3):195–209, 2009.

[8] M. Cardaci, V. Di Gesù, M. Petrou, and M. E. Tabacchi. A Fuzzy Approach to the Evaluation of Image Complexity. *Fuzzy Sets and Systems*, 160(10):1474–1484, 2009.

[9] H. Chang, F. Yu, J. Wang, D. Ashley, and A. Finkelstein. Automatic Triage for a Photo Series. *ACM Transactions on Graphics*, 35(4):148:1–148:10, 2016.

[10] S. F. Chipman. Complexity and Structure in Visual Patterns. *Journal of Experimental Psychology: General*, 106(3):269–301, 1977.

[11] S. F. Chipman and M. J. Mendelson. Influence of Six Types of Visual Structure on Complexity Judgments in Children and Adults. *Journal of Experimental Psychology: Human Perception and Performance*, 5(2):365–378, 1979.

[12] D. Comaniciu and P. Meer. Mean shift: A robust approach toward feature space analysis. *IEEE Transactions on pattern analysis and machine intelligence*, 24(5):603–619, 2002.

[13] S. Corchs, F. Gasparini, and R. Schettini. No reference image quality classification for jpeg-distorted images. *Digital Signal Processing*, 30:86–100, 2014.

[14] S. E. Corchs, G. Ciocca, E. Bricolo, and F. Gasparini. Predicting Complexity Perception of Real World Images. *PloS one*, 11(6):e0157986, 2016.

[15] H. A. David. *The method of paired comparisons*, volume 12. London, 1963.

[16] H. J. Eysenck. The Empirical Determination of an Aesthetic Formula. *Psychological Review*, 48(1):83–92, 1941.

[17] Z. B. Fan, Y. Li, J. Yu, and K. Zhang. Visual complexity of chinese ink paintings. In *Proceedings of the ACM Symposium on Applied Perception*, pages 9:1–9:8, 2017.

[18] https://www.figure-eight.com Machine Learning, Training Data, and Artificial Intelligence Solutions: Figure Eight. Last accessed on May 12, 2018.

[19] A. Forsythe, M. Nadal, N. Sheehy, C. J. Cela-Conde, and M. Sawey. Predicting Beauty: Fractal Dimension and Visual Complexity in Art. *British journal of psychology*, 102(1):49–70, 2011.

[20] A. Gartus and H. Leder. The small step toward asymmetry: aesthetic judgment of broken symmetries. *i-Perception*, 4(5):361–364, 2013.

[21] A. Gartus and H. Leder. Predicting Perceived Visual Complexity of Abstract Patterns using Computational Measures: The Influence of Mirror Symmetry on Complexity Perception. *PloS one*, 12(11):e0185276, 2017.

[22] D. F. Gleich and L.-h. Lim. Rank aggregation via nuclear norm minimization. In *Proceedings of the 17th ACM international conference on Knowledge discovery and data mining*, pages 60–68, 2011.

[23] D. L. Haytko and J. Baker. It?s All at the Mall: Exploring Adolescent Girls? Experiences. *Journal of Retailing*, 80(1):67–83, 2004.

[24] C. Heaps and S. Handel. Similarity and Features of Natural Textures. *Journal of Experimental Psychology: Human Perception and Performance*, 25(2):299–320, 1999.

[25] B. A. Huhmann. Visual Complexity in Banner Ads: The Role of Color, Photography, and Animation. *Visual Communication Quarterly*, 10(3):10–17, 2003.

[26] Z. Hussain, M. Zhang, X. Zhang, K. Ye, C. Thomas, Z. Agha, N. Ong, and A. Kovashka. Automatic Understanding of Image and Video Advertisements. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition CVPR*, pages 1100–1110, 2017.

[27] IKEA. https://www.ikea.com.

[28] T. Jacobsen and L. Höfel. Aesthetics Electrified: An Analysis of Descriptive Symmetry and Evaluative Aesthetic Judgment Processes using Event-related Brain Potentials. *Empirical Studies of the Arts*, 19(2):177–190, 2001.

[29] M. G. Kendall and B. B. Smith. On the method of paired comparisons. *Biometrika*, 31(3/4):324–345, 1940.

[30] W. H. Kim, M. Jalal, S. J. Hwang, S. C. Johnson, and V. Singh. Online Graph Completion: Multivariate Signal Recovery in Computer Vision. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pages 5019–5027, 2017.

[31] A. Krishen. Perceived versus Actual Complexity for Websites: Their Relationship to Consumer Satisfaction. *Journal of Consumer Satisfaction, Dissatisfaction and Complaining Behavior*, 21:104–123, 2008.

[32] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft COCO: Common Objects in Context. In *Proceedings of European Conference on computer vision ECCV*, pages 740–755. Springer, 2014.

[33] M. M. Marin and H. Leder. Effects of presentation duration on measures of complexity in affective environmental scenes and representational paintings. *Acta psychologica*, 163:38–58, 2016.

[34] G. A. Miller. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological review*, 63(2):81, 1956.

[35] A. Miniukovich and A. D. Angeli. Quantification of Interface Visual Complexity. In *Proceedings of International Working Conference on Advanced Visual Interfaces, AVI*, pages 153–160, 2014.

[36] M. Nadal, E. Munar, G. Marty, and C. J. Cela-Conde. Visual Complexity and Beauty Appreciation: Explaining the Divergence of Results. *Empirical Studies of the Arts*, 28(2):173–191, 2010.

[37] A. Olivia, M. L. Mack, M. Shrestha, and A. Peeper. Identifying the Perceptual Dimensions of Visual Complexity of Scenes. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 26, 2004.

[38] L. Palumbo, R. Ogden, A. D. Makin, and M. Bertamini. Examining visual complexity and its influence on perceived duration. *Journal of vision*, 14(14):3–3, 2014.

[39] R. Pieters, M. Wedel, and R. Batra. The Stopping Power of Advertising: Measures and Effects of Visual Complexity. *Journal of Marketing*, 74(5):48–60, 2010.

[40] L. Pilelienė and V. Grigaliūnaitė. Effect of Visual Advertising Complexity on Consumers? Attention. *Economics*, 3(8):489–501, 2018.

[41] G. Ramanarayanan, K. Bala, J. A. Ferwerda, and B. Walter. Dimensionality of Visual Complexity in Computer Graphics Scenes. In *Proceedings of Human Vision and Electronic Imaging XIII Conference*, page 68060E, 2008.

[42] K. Reinecke, T. Yeh, L. Miratrix, R. Mardiko, Y. Zhao, J. Liu, and K. Z. Gajos. Predicting Users' First Impressions of Website Aesthetics with a Quantification of Perceived Visual Complexity and Colorfulness. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 2049–2058. ACM, 2013.

[43] R. Rosenholtz, Y. Li, and L. Nakano. Measuring Visual Clutter. *Journal of vision*, 7(2):17, 2007.

[44] S. Schnur, K. Bektaş, and A. Çöltekin. Measured and Perceived Visual Complexity: A Comparative Study among Three Online Map Providers. *Cartography and Geographic Information Science*, 45(3):238–254, 2018.

[45] J. G. Snodgrass and M. Vanderwart. A Standardized Set of 260 Pictures: Norms for Name Agreement, Image Agreement, Familiarity, and Visual Complexity. *Journal of experimental psychology: Human learning and memory*, 6(2):174–215, 1980.

[46] S. Sohn, B. Seegebarth, and M. Moritz. The Impact of Perceived Visual Complexity of Mobile Online Shops on User's Satisfaction. *Psychology & Marketing*, 34(2):195–214, 2017.

[47] N. Westlake, H. Cai, and P. Hall. Detecting People in Artwork with CNNs. In *Proceedings of European Conference on Computer Vision (ECCV)*, pages 825–841. Springer, 2016.

[48] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba. Places: A 10 Million Image Database for Scene Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40:1452 – 1464, 2018.