Size Does Matter: How Image Size Affects Aesthetic Perception?

Wei-Ta Chu[†], Yu-Kuang Chen[†], and Kuan-Ta Chen[‡]

[†]Department of Computer Science and Information Engineering, National Chung Cheng University
[‡]Institute of Information Science, Academia Sinica

ABSTRACT

There is no doubt that an image's content determines how people assess the image aesthetically. Previous works have shown that image contrast, saliency features, and the composition of objects may jointly determine whether or not an image is perceived as aesthetically pleasing. In addition to an image's content, the way the image is presented may affect how much viewers appreciate it. For example, it may be assumed that a picture will always look better when it is displayed in a larger size. *Is this "the-bigger-the-better" rule always valid?* If not, *in what situations is it invalid?*

In this paper, we investigate how an image's resolution (pixels) and physical dimensions (inches) affect viewers' appreciation of it. Based on a large-scale aesthetic assessments of 100 images displayed in a variety of resolutions and physical dimensions, we show that an image's size significantly affects its aesthetic rating in a complicated way. Normally a picture looks better when it is bigger, but it may look worse depending on its content. We develop a set of regression models to predict a picture's absolute and relative aesthetic levels at a given display size based on its content and compositional features. In addition, we analyze the essential features that lead to *the size-dependent property* of image aesthetics. It is hoped that this work will motivate further research by showing that both content and presentation should be considered when evaluating an image's aesthetic appeals.

Categories and Subject Descriptors

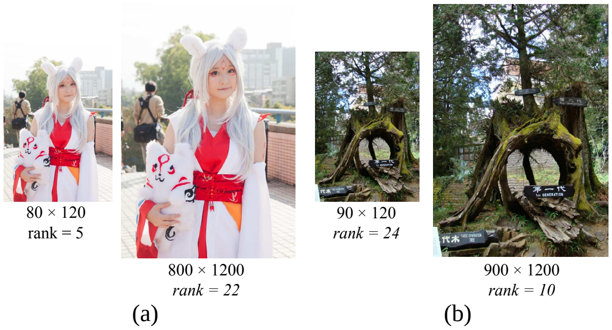
I.2.10 [**Artificial Intelligence**]: Vision and Scene Understanding—*Perceptual reasoning*

Keywords

Crowdsourcing; Human perception; Image aesthetics; Quality assessment; Size-dependent aesthetics

1. INTRODUCTION

Image aesthetic quality assessment has generated a great deal of interest in recent years because it is a fundamental component of many multimedia applications, such as image summarization and automatic photo editing. Previous works have shown that the pixel-level features and compositional features of an image may jointly influence how humans evaluate the image aesthetically [3, 7, 13]. The content of a picture is certainly an important factor in determining whether the picture looks appealing; however, very little is known about whether and how a picture's aesthetic level is affected by *the way it is presented* to the viewer. In this work, we focus on the relationship between image aesthetics and display size, and consider the factors that determine the relationship.


80 × 120
rank = 5

800 × 1200
rank = 22

(a)

90 × 120
rank = 24

900 × 1200
rank = 10

(b)

Figure 1: Examples of how display size affects aesthetic ratings in a heterogeneous way.

When people surf online auction websites and browse the thumbnail images of products, those images and the full-size images of the products sometimes create very different impressions. The thumbnail version may look attractive, while the full-size version is unattractive, or vice versa[1]. In addition, on photo sharing websites such as Flickr, pictures shown in the default size (i.e., small or medium) are often more appealing than their full-size counterparts. Based on these observations, we posit that *some complicated mechanisms, psychological and/or physiological, cause the same picture to have very different levels of aesthetic appeal when it is shown in different sizes.* This phenomenon is the motivation for our study.

First, we conducted a pilot study to verify the effect of display size on image aesthetics. In the study, 117 subjects were shown 25 high-resolution images at random in two different display scales. The large images used 1200x1200 pixels, while the thumbnail versions used 120x120 pixels. The subjects were asked to give an aesthetic rating between 1 and 5 for each displayed image. Then, we averaged the score to obtain the representative rating. Figure 1 shows examples of pictures that were given *different aesthetic ratings for different display scales.* The small image of the girl in Figure 1(a) was ranked 5 (out of 25), while the large image was

MM'13, October 21–25, 2013, Barcelona, Spain.

http://dx.doi.org/10.1145/2502081.2502102.

[1] We observe that the first scenario occurs more frequently when the product pictures are taken by amateur photographers.

ranked 22; that is, viewers preferred the thumbnail version. By contrast, the thumbnail image and large image of the tree in Figure 1(b) were ranked 24 and 10 respectively. This is reasonable because the thumbnail of the tree is too complex and unclear, while the large image is much clearer. It is apparent that *the changes in display size may impact images' aesthetic appeals images in distinct ways*.

The findings of our pilot study seem intuitive and unsurprising; however, we should point out that existing works on aesthetics modeling *do not* consider the impact of display size. We used ACQUINE [4], a state-of-the-art image aesthetics assessment engine, to evaluate the thumbnail and large versions of the 25 images in the pilot study. ACQUINE ranked the large and small images in Figure 1(a) at 21 and 22 respectively, and the large and small images in Figure 1(b) at 24 and 23 respectively. The results provide strong evidence that *the effect of display size has long been overlooked* and the issue is worth further investigation. To the best of our knowledge, this is the first work that considers the effect of display size on image aesthetic assessment.

In this paper, we investigated the effect of display size on image aesthetic assessment systematically. First, we conducted an Internet crowdsourced experiment to collect the participants' assessments of 100 high-resolution images shown in a variety of resolutions (pixels) and physical dimensions (inches). Our data analysis reveals that image scaling influences a user's aesthetic assessment significantly, and the impact is *heterogeneous* across different image categories. Subsequently, we developed a set of partial least square regression (PLSR) models to describe the absolute and relative aesthetic ratings of any image based on the image's content and compositional features.

The contributions of this paper are three-fold:

1. Based on a large set of crowdsourced user ratings, we confirm that the display size significantly affects image aesthetics assessment. We also show that the effect is not consistent for all images; instead, it is highly dependent on image content.
2. Through our aesthetics prediction model, we show that the physical dimensions (in inches) of an image on a screen is more important than its resolution (in pixels) in terms of aesthetics perception. We demonstrate that assessment of an image's aesthetic quality should consider the image's content and how it is presented.
3. The prediction model provides promising performance compared with earlier size-agnostic approaches, as well as clues about features that tend to be more effective in aesthetics prediction.

The remainder of this paper is organized as follows. In Section 2, we review related works; in Section 3, we describe how image ratings are collected from Internet users; and in Section 4, we consider how display size impacts humans' aesthetic perceptions of pictures. We introduce a set of regression models to describe and predict the effect of display size on image aesthetics in Section 5, and evaluate the performance of the models in Section 6. In Section 7, we discuss applications of our models and the implications for future research. Section 8 contains our concluding remarks.

2. RELATED WORK

Earlier studies of image aesthetics focused on distinguishing professional photos from amateur photos [3,7,12–14]. Tong et al. [13] extracted blurriness, intensity contrast, colorfulness, and saliency values, and constructed discriminative models to classify images. Based on interviews with professional photographers, Ke et al. [7] concluded that simplicity, realism, and basic photographic techniques are important factors. To model image aesthetics, they extracted the spatial distribution of edges, color distribution, hue count, blur, contrast, and brightness. In addition to the visual appearance of images, Datta et al. [3] proposed using compositional features, such as the rule of thirds and aspect ratio, to model image aesthetics. Subsequently, Datta and Wang [4] extended the approach and developed ACQUINE, the first public online image aesthetics assessment platform. Recently, Dhar et al. [6] posited that image cues, such as compositional attributes, content attributes, and sky-illumination attributes, are good indicators of how humans evaluate aesthetic quality. Using prior knowledge of art and a survey, Li and Chen [8] extracted various global and local features from paintings and designed a data-driven machine learning scheme to classify high-quality and low-quality paintings.

Based on the relationship between saliency and aesthetics, Sun et al. [12] proposed using the rate of focused attention to evaluate aesthetic quality. The approach uses statistics derived from human observers' inputs to measure the degree to which salient objects are located in the subject's (i.e., the topic's) mask. In [14], as well as global image features, exposure, sharpness, and textural features are extracted from salient regions. The results reported in [12, 14] and a number of follow-up works show that the inclusion of salient regions is a promising way to quantify image aesthetics. Luo et al. [10] devised and compared various global and regional features to categorize photos, such as animals, plants, buildings, and portraits. As aesthetics assessment is highly subjective, Wu et al. [15] proposed that distribution vectors of human assessment, rather than scalar values, should be used to describe images. In their approach, constructing a prediction function to output a distribution label for a test image is formulated as a structural output learning problem, which is solved by the proposed support vector distribution regression algorithm.

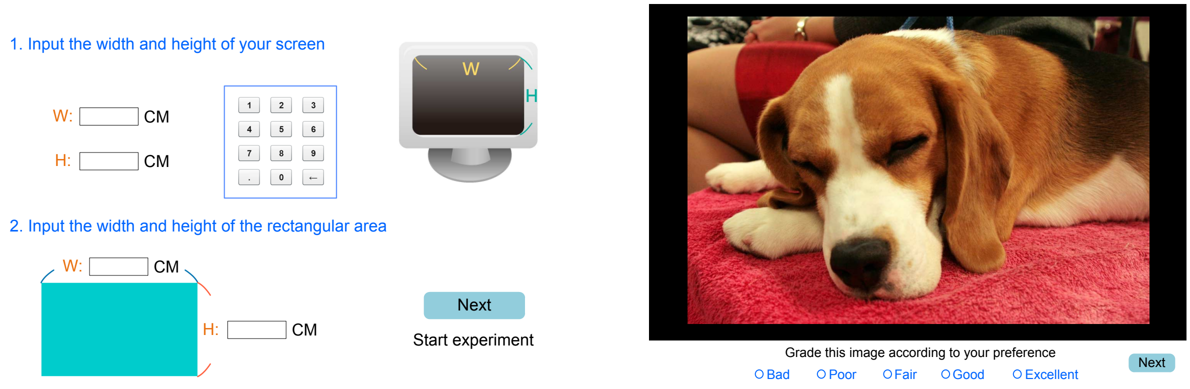
Our study differs from previous works because it focuses on the impact of display size on aesthetic quality assessment. We analyze the issue and its effects systematically. Rather than developing further image analysis and aesthetics evaluation techniques, we investigate how a picture's display size affects viewers' aesthetic perception of it.

3. TRACE COLLECTION

In this section, we describe how we collected users' aesthetics ratings for images presented in different resolutions and physical dimensions. We first explain the experiment design and then summarize the collected trace.

3.1 Experiment Design

To recruit subjects from the Internet for the aesthetics ratings study, we developed a web-based platform in Adobe ActionScript and Flash. When a potential subject enters the system, the platform presents a welcome screen followed by an instruction page, which describes the goal of our study and the tasks the subject must perform during the experiment. If the subject agrees to participate, we present a user interface in the full-screen mode as shown in Figure 2(a) to inquire about the physical dimensions of the subject's computer monitor. The subjects may not know the exact size of their monitors, so we ask them to provide the viewable area of their monitors in centimeters. Because some participants may give erroneous information [2], either carelessly or intentionally, a rectangle is displayed on the screen and the participants are asked to measure the size of the rectangle on their monitors with a ruler. By comparing both inputs, i.e., the dimensions of the monitor and those of


1. Input the width and height of your screen

2. Input the width and height of the rectangular area

Grade this image according to your preference

○ Bad ○ Poor ○ Fair ○ Good ○ Excellent

(a) The configuration interface where users are asked to input the physical dimensions of the viewable area of their computer monitors.

(b) The image rating interface where users are asked to rate each presented image on an MOS scale.

Figure 2: The user interfaces in our experiments

Table 1: Image Categories

Category	# Images	Category	# Images
Animal	15	Object	12
Architecture	16	Scenery	18
Art	9	Sport	10
Human	20		

the rectangle, we can detect and filter out erroneous inputs in the analysis stage.

Next, the system enters the image rating phase, which comprises 60 rounds. In each round, we randomly pick an image that has not been used in the experiment from an image pool and present it on the screen with a random scaling factor between 0.1 and 1.0. A scaling factor of 1.0 corresponds to a 1000x1000 display area; that is, an image is shrunk[2] so that it fits the display area. Similarly, an image presented with a scaling factor of 0.1 indicates that the image exactly fits a 100x100 display area. During the above scaling operations, we preserve the images' aspect ratios to ensure that the semantics of the images remain intact when they are presented in various scales.

When the experiment participants looked at a random image with a random display size, they were asked to grade their aesthetic perceptions of the picture on a five-point MOS (Mean Opinion Score) scale according to the options of Bad, Poor, Fair, Good, and Excellent. The participants were allowed unlimited time to rate each image. After the image had been rated, the round terminated and the system advanced to the next round until all 60 rounds in the experiment were completed. We consider that 60 rounds allows each subject to provide sufficient and consistent ratings without losing concentration.

We collected 100 high-resolution images from Flickr.com[3] to compile an image pool. The images were chosen arbitrarily based on the following criteria: 1) both the width and height of each image was at least 2000 pixels; 2) the images were released under a Creative Commons license[4]; and 3) the set of images covered a variety of subjects, such as humans, animals, scenery, and paintings.

Table 2: Trace Summary

# Subjects	230
# Images	100
# Ratings	13,800 (60 ratings each subject)
Screen resolutions	1024x600 (min) to 1920x1200 (max)
Display resolutions	100x100 (min) to 1000x1000 (max)
Rating time per image	2.9 seconds (std. dev. 2.6 seconds)
Rating time per exp.	178.7 seconds (std. dev. 68.5 seconds)
Overall rating time	685 minutes

Figure 3 shows thumbnails of the collected images and Table 1 lists the categories of the images in the image pool.

3.2 Trace Summary

For the experiments, we recruited 230 subjects from the Internet community. Each subject was given a reward of virtual currency equivalent to 0.1 USD; and we collected $230 \times 60 = 13,800$ image ratings in 230 experiments. The screen resolutions used by a subject varied between 1200x600 and 1920x1200, depending on the subject's hardware and configurations. To accommodate the limits due to screen resolutions, our platform does not display images at a scale that does not fit the screen resolution. In other words, if a subject uses a screen resolution of 1280x800, we only show images at scales between 0.1 and 0.8, which correspond to 100x100 and 800x800 display areas respectively.

During the experiments, we also recorded the time taken to complete each image rating round. On average, the subjects spent 3 seconds in each round with a distribution spread over 1–8 seconds. As each experiment comprised 60 rounds, it took approximately 3 minutes for a subject to complete an experiment. This indicates that rating images aesthetically was not a difficult task and the experiments did not place a heavy workload on the subjects. Table 2 shows a summary of the collected trace.

The histogram in Figure 4(a) shows the distribution of all ratings given by the subjects. Most of the images were rated as Fair or Good, while smaller proportions were deemed Poor or Excellent. The findings imply that, generally, the quality of the images used in the experiments was high. The lower user ratings (Bad and Poor) may indicate that some pictures were regarded as unattractive or the display size was too small. We consider these issues in the next section.

[2] We use the bicubic interpolation function provided by the GD Graphics Library to perform image shrinking.

[3] http://www.flickr.com/

[4] http://creativecommons.org/

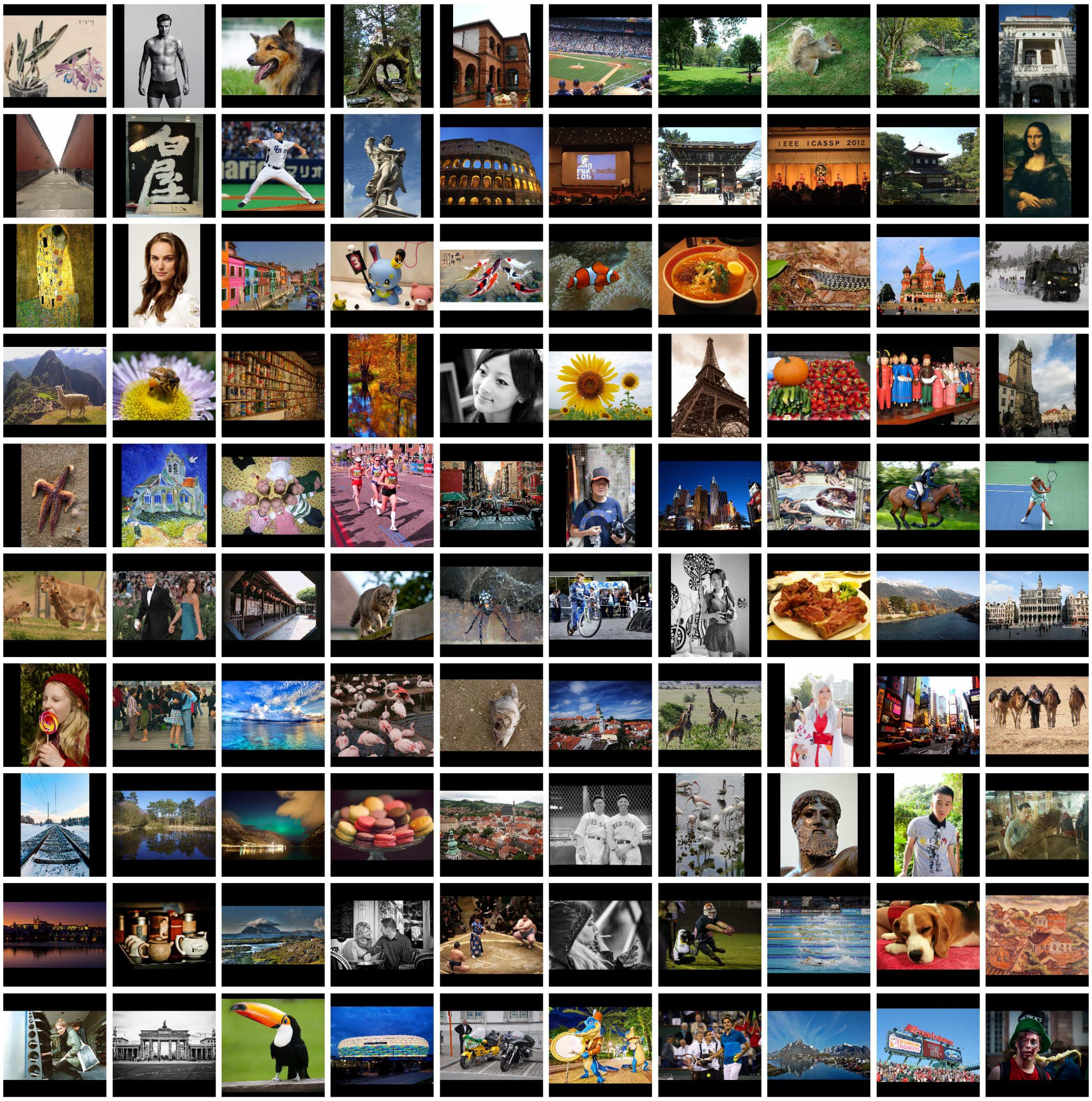
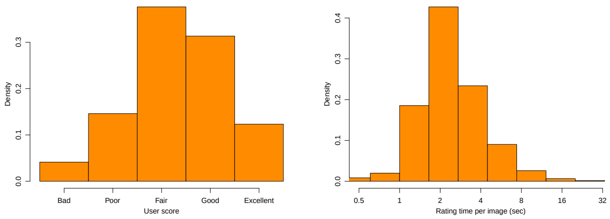


Figure 3: The 100 images used in the experiments



(a) User rating histogram (b) Rating time histogram

Figure 4: Histograms of the users' ratings and the time required to make rating decisions

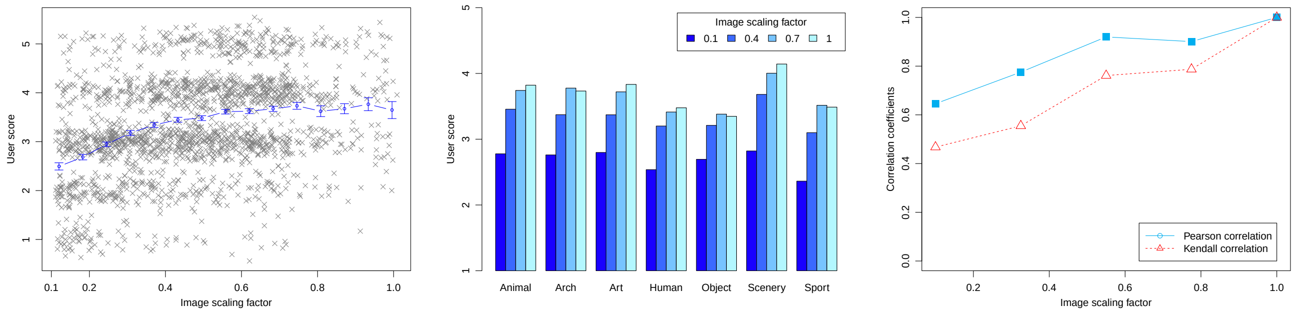
4. DATA ANALYSIS

In this section, we investigate whether the display size of an image influences viewers' aesthetic perception of the image. If the size is important, in which way is the influence made? Does the display size affect viewers' perception of all images in the same way? In other words, is this effect content-dependent and a non-trivial matter to describe and predict?

For ease of discussion, we use the term "image scaling" to refer to shrinking an image with a specific scaling factor (or scale for short), which is between 0.1 (100x100 pixels) and 1.0 (1000x1000 pixels) in this study.

4.1 Image Size Does Matter

Our first question is *Does image scaling influence users' aesthetic perception of an image*? We analyze the collected traces to determine the relationship between image scaling factors and users' scores. The results are shown in Figure 5. In the scatter plot of



(a) The scatter plot of user scores vs. scaling factors. The blue circles denote the average user scores across different scales.

(b) The average user scores at different scales for images in the seven categories.

(c) The correlation coefficients between the average image ratings at 1.0 and those at other scales.

Figure 5: Evidence of the impact of image scaling on aesthetics ratings

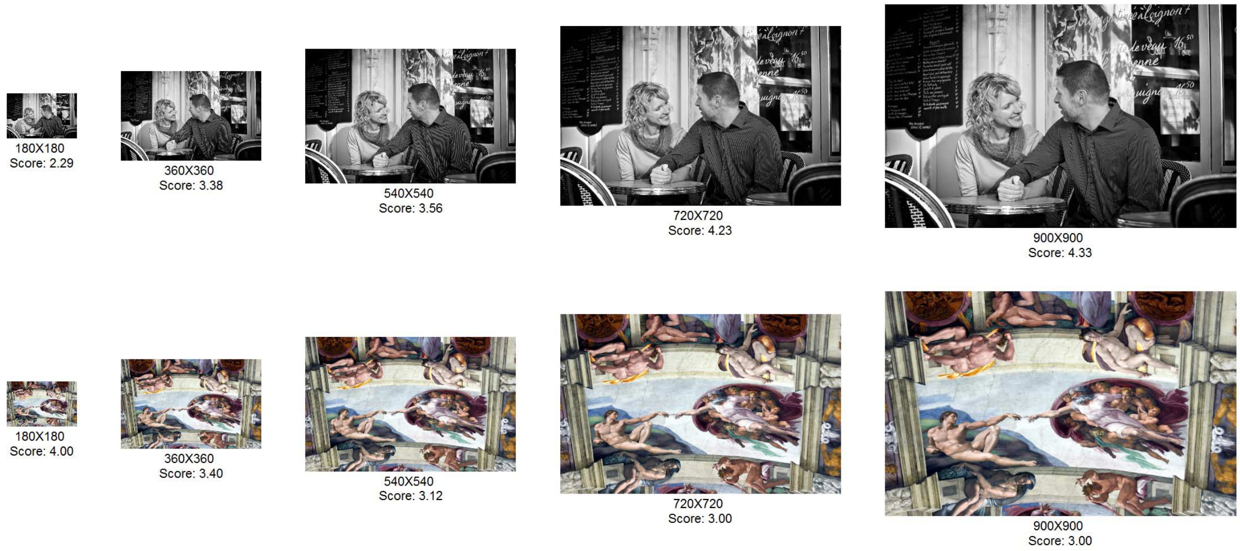


Figure 6: The impact of image scaling on users' scores can vary significantly. The top image (in the Human category) is given a higher rating when presented at a larger scale, while the bottom image (in the Art category) is rated higher when presented at a smaller scale.

the scaling factors and user scores (Figure 5(a)), each gray cross[5] corresponds to a user rating for an image presented at a particular scale; and the y-locations of the crosses (i.e., the user ratings) are randomly dispersed so that the crosses do not overlap. The blue circles indicate the average user scores across different image scale ranges, while the horizontal bars represent the 95% confidence intervals of the averaged scores. The trend of the blue circles indicates that, generally, the user rating increases with the image size; however, the effect diminishes when the scaling factor is larger than 0.7. The average user score at scale 0.1 is 2.5; while at scale 1.0, it is 3.5. Taken together with their narrow confidence intervals, the scores indicate that *the image size does influence users' aesthetic perception of the image*.

From a different perspective, in Figure 5(b), we plot the average user ratings at different scales for images in each of the seven categories listed in Table 1. The graph clearly shows that image scaling impacts the aesthetics ratings of images in all categories. Besides the two highest scales, in all the categories, the scores are significantly higher when the images are displayed at larger scales.

The above findings motivate the following question: *Is the impact of image scaling the same for all types of images*? To the answer the question, in Figure 5(c), we plot the correlation coefficients between the average user scores at scale 1.0 and those at a variety of other scales. Let s_j^i be the average user score of an image i displayed at scale j, and let $\mathbf{S_j}$ denote the sequence $(s_j^1, s_j^2, \ldots, s_j^{100})$. Then, the lines in Figure 5(c) correspond to cor($\mathbf{S_j}$,$\mathbf{S_{1.0}}$), where $0.1 \leq j \leq 1.0$ and cor($\cdot$) denotes the Pearson and Kendall correlation coefficients respectively. Both lines in the graph indicate that image scaling has a heterogeneous effect on viewers' perceptions of different images in terms of their aesthetic quality. In other words, while some images are deemed slightly worse at smaller scales, others may be rated significantly worse or better at smaller scales, depending on the images' content.

4.2 Heterogeneous Effect of Image Scaling

The examples in Figure 6 illustrate how image scaling impacts user ratings in distinct ways. The top image, which is in the Human

[5] The gray crosses are less dense at scales larger than 0.8 because some participants used screen resolutions with less than 800 vertical scan lines, such as the widely used 1024x768 resolution.

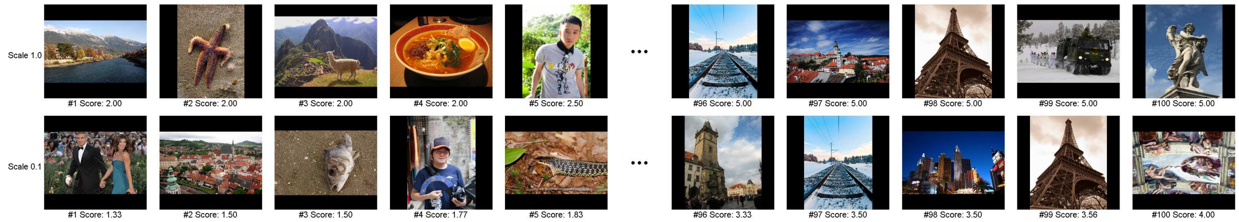


Figure 7: The top and bottom five images that receive the highest and lowest user ratings at scales of 1.0 and 0.1 respectively.

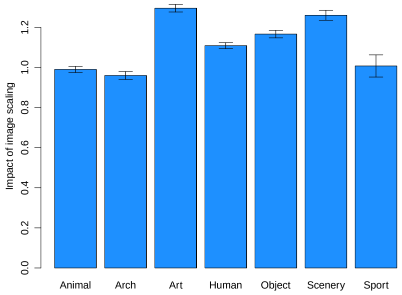


Figure 8: The differences in the average user scores between scale 0.1 and scale 1.0 for images in different categories.

category, receives a high score (4.33) at 900x900 resolution and a relatively low score (2.29) at 180x180 resolution. In contrast, the bottom image, which is in the Art category, receives a high score (4.00) at 180x180 resolution, but a relatively low score (3.00) at the 900x900 resolution. It seems that image scaling affects the aesthetics ratings of different images in a highly unpredictable and complicated way. Figure 7 shows the images that received the best and worst user ratings at the highest scale (1.0) and lowest scale (0.1). The sets of images are almost completely different; only the pictures of the Eiffel Tower and the Siberian Railway appear in both sets. The results further confirm that the aesthetics rankings of images are content-dependent and vary at different scales.

To determine whether the scaling effect is content-dependent and varies across the categories, we quantify the impact of scaling as the difference between the average score at the smallest scale (0.1) and that at the largest scale (1.0), as shown in Figure 8. The results indicate that the impact is content-dependent and varies across the categories. In general, the ratings of Art and Scenery pictures are more affected by image scaling than Animal, Architecture, and Sport pictures.

We also found that scaling affects perception of an image's aesthetic quality. Figure 9 shows the relationship between scaling and the ratings given to images at scale 0.1 and scale 1.0. Generally, images rated as good at a large scale tend to be regarded less favorably at small scales. On the other hand, images rated favorably at a small scale are generally less sensitive to image scaling in that they tend to receive a small "aesthetic bonus" when presented in a larger size.

5. AESTHETIC QUALITY PREDICTION

So far, we have shown that image scaling does impact the aesthetic ratings given to images; moreover, the impact is content-dependent and it is not homogeneous across all types of images. In this section, we investigate whether image content analysis can be exploited to model the impact of image scaling.

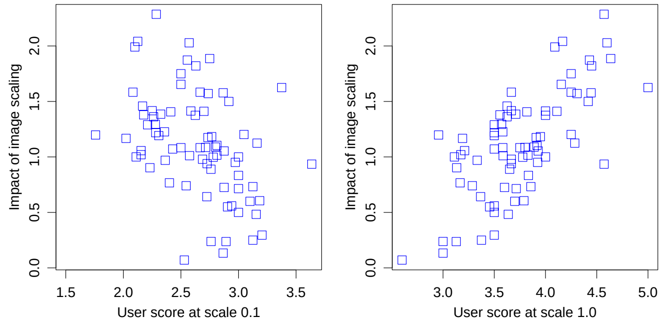


Figure 9: The impact of image scaling (quantified using the score difference between two extreme scales, 0.1 and 1.0) v.s. the user scores for the images presented at scale 0.1 and 1.0 respectively.

5.1 Model Construction

We utilize a data-driven approach to assess the influence of image scaling on aesthetic perception. The approach constructs a set of models to predict viewers' aesthetic perceptions based on features extracted from a set of images in various display sizes. We use the partial least square regression (PLSR) method to predict aesthetic perception. As an image's aesthetic quality can be described in terms of its *absolute* aesthetic value or in terms of its *relative* aesthetic ranking, we develop a *scoring model* and a *ranking model* respectively.

5.1.1 Scoring Model

Let $\boldsymbol{a}_1, ..., \boldsymbol{a}_n$ be m-dimensional feature vectors extracted from n training images to construct an $n \times m$ data matrix A. Each image is associated with an aesthetic score, denoted by b_i, and the combined scores of the images constitute a score column matrix $\boldsymbol{b} = (b_1, ..., b_n)^T$. The relationship between features and scores is then described as $A\boldsymbol{x} = \boldsymbol{b}$, where $\boldsymbol{x}$ is an $m \times 1$ column matrix that indicates how different columns are linearly combined to form the score matrix. Note that each column of A and $\boldsymbol{b}$ has a zero mean.

We treat different sizes of the same image separately. For example, an 800×600 image I_0 and its resized version 400×300 image I_1 are treated as different images. Feature vectors $\boldsymbol{f}_0$ and $\boldsymbol{f}_1$ are extracted from I_0 and I_1 and added into the data matrix A respectively.

The PLSR model uses latent variables to describe the relations between sets of observed variables. First, the data matrix A and the score matrix $\boldsymbol{b}$ are decomposed into $A = VP^T + E$ and $\boldsymbol{b} = U\boldsymbol{q}^T + \boldsymbol{f}$, respectively, where V and U are $n \times g$ matrices that contain g extracted latent vectors. The $m \times g$ matrix P and the $1 \times g$ vector $\boldsymbol{q}$ represent the loadings; and the $n \times m$ matrix E and the $n \times 1$ vector $\boldsymbol{f}$ are residuals. With the decomposed partial least components, the SIMPLS algorithm [5] is used to determine a set of weighting vectors $W = (\boldsymbol{w}_1, ..., \boldsymbol{w}_g)$ such that

$$[cov(\boldsymbol{v}_i, \boldsymbol{u}_i)]^2 = \max_{|\boldsymbol{w}_i|=1} [cov(A\boldsymbol{w}_i, \boldsymbol{b})]^2, \qquad (1)$$

where $\boldsymbol{v}_i$ and $\boldsymbol{u}_i$ are the ith columns of matrix V and matrix U respectively, and $cov(\boldsymbol{v}_i, \boldsymbol{u}_i)$ is the covariance between $\boldsymbol{v}_i$ and $\boldsymbol{u}_i$. The matrices A and $\boldsymbol{b}$ are then deflated by subtracting their rank-one approximations based on $\boldsymbol{v}_i$ and $\boldsymbol{u}_i$. The process is repeated until the desired number of latent vectors is extracted.

The $m \times 1$ regression coefficient $\boldsymbol{\beta}$ is then estimated by $\boldsymbol{\beta} = W(P^T W)^{-1} V^T \boldsymbol{b}$. Given a test image represented by a vector $\boldsymbol{a}_q$, the corresponding aesthetic score can be estimated by $b_q = \bar{b} + \boldsymbol{a}_q^T \boldsymbol{\beta}$, where $\bar{b}$ represents the average of $\boldsymbol{b}$.

5.1.2 Ranking Model

In many applications, it is not necessary to determine an image's absolute aesthetic rating. For example, it may be sufficient to decide if a large picture is preferable to a scaled-down thumbnail version. Therefore, we also developed a ranking model that predicts the *relative ranking of a particular image among a set of images* at a given display size.

Assume an image I_0 has a scaling factor of 0.5. The feature vector that describes the scaling from I_0 to I_1 is derived by $\boldsymbol{a}^T_{I_0 \rightarrow I_1} = (\boldsymbol{f}_1^T, \boldsymbol{f}_1^T - \boldsymbol{f}_0^T, d_s)$, where $\boldsymbol{f}_1^T - \boldsymbol{f}_0^T$ represents the dimension-wise feature difference. In this setting, d_s denotes the scaling factor difference. A positive d_s indicates that I_1 is a scaled-up version of I_0, while a negative d_s indicates that I_1 is a scaled-down version of I_0.

The matrix $\boldsymbol{b}$ represents the difference in rankings between multiple scales. Let there be n images with the same resolution as the image I_0, e.g., 800×800. The images are sorted in descending order according to their associated aesthetic scores. The rank of an image I_j is $r_j = 1/n$ if it has the largest score, and $r_j = n/n = 1$ if it has the lowest score. Thus, the rank difference in $\boldsymbol{b}$ that corresponds to $\boldsymbol{a}^T_{I_0 \rightarrow I_1}$ is equal to $r_1 - r_0$.

The ranking model describes how an image is ranked aesthetically when it is shrunk to a particular scale. Given a test image I_q to be scaled down to $I_{q'}$ and that $\boldsymbol{a}^T_{I_q \rightarrow I_{q'}}$ is the feature vector, the predicted rank difference is $b_{I_q \rightarrow I_{q'}} = \bar{b} + \boldsymbol{a}^T_{I_q \rightarrow I_{q'}} \boldsymbol{\beta}$, where $\bar{b}$ is the mean value of $\boldsymbol{b}$. Therefore, the predicted aesthetic rank of $I_{q'}$ is equal to $r_q + b_{I_q \rightarrow I_{q'}}$, where r_q is the rank of I_q at the original resolution.

5.2 Display Size: Pixels or Inches?

Image scaling can be characterized by the changes of an image's resolution (in pixels) or by its physical dimensions (in inches) on a display device. To predict the aesthetic quality of images from the two perspectives, the scoring and ranking models are constructed and evaluated respectively in two modes: the *pixel mode* and the *physical dimension mode* (*dimension mode* for short). Note that the number of pixels cannot be converted to inches directly because the conversion depends on the characteristics of the display device and the current display mode. For example, an 1136x640 image can be shown on a 4" iPhone 5 screen as well as on a 27" computer monitor as a full-screen image. Although the number of pixels displayed is exactly the same, the pixels' physical dimensions are vastly different and may result in significantly different ratings of the aesthetic quality. This is the reason we provide two modes in the aesthetics prediction models.

In our dataset, the image resolutions range from 100×100 pixels to 1000×1000 pixels. Compared with the maximum display size of 1000×1000 pixels, a resized image $I_j^{(s)}$ comprised of $n \times n$ pixels is said to be scaled with a factor $s = n/1000$. Image variants and the associated aesthetic assessments are then categorized into 9 groups based on s, namely, [0.1, 0.2), [0.2, 0.3), . . . , [0.9, 1]. For example, The *ground* aesthetic score for the image I_j with a scaling factor s between 0.1 and 0.2 is obtained by averaging all the scores given to $I_j^{(s)}$, $0.1 \leq s < 0.2$. Let $g_j^{(1)}, g_j^{(2)}, ..., g_j^{(9)}$ denote, respectively, the average aesthetic scores for nine groups of variants derived from I_j. In the test phase, suppose a model running in the pixel mode predicts an aesthetic score for the image $I_j^{(0.13)}$, which is then compared with $g_j^{(1)}$ because images with scaling factors between 0.1 and 0.2 belong to the first group (out of 9).

Assume that the display size of images in the physical dimension mode ranges from $N_1 \times N_1$ inches to $N_2 \times N_2$ inches. Comparing with the maximum size $N_2 \times N_2$ inches, a resized variant $I_j^{(s)}$ of $n \times n$ inches is said to be scaled with a factor of $s = n/N_2$. The image variants and their aesthetic assessments are also categorized into 9 groups. Note that a 450×450-pixel image I_k, for example, would be categorized in the fourth group in the pixel model, but it is not necessarily categorized into the fourth group in the dimension mode.

In the experiments, we evaluate the performances of the models in the pixel mode and the dimension mode, and use a five-fold cross-validation scheme for model training and testing. We also calculate the Pearson correlation and Spearman correlation between the predicted results and the ground truth. The Pearson correlation indicates a model's prediction accuracy; while the Spearman correlation indicates the consistency between the rankings of the predicted results and the rankings of the ground truth.

5.3 Feature Selection and Fusion

We categorize the features proposed in the literature into compositional attributes and content attributes [6]. Table 3 shows the features and their corresponding dimensions, types, and references. Compositional attributes, content attributes, and a combination of them are denoted as **S**, **C**, and **B** respectively. The last feature (f_{27}) comprises the width and the height of an image, and is categorized with **O** (others). We consider that [1] and the Gabor wavelet texture [11] are important because details of textural details are clearly expected to change when images are scaled up/down. Hence, we include them in the feature vectors even though, to the best of our knowledge, they have not been used in image aesthetics modeling previously.

To determine how the features affect aesthetics prediction, we evaluate the performance of individual features based on the scoring model in the dimension mode. The results are presented as bar charts in Figure 10. We observe that f_5, f_9, f_{12}, f_{15}, f_{16}, f_{25}, and f_{27} are more effective for making predictions. Surprisingly, simple features like the intensity histogram (f_{15}), the hue histogram (f_{16}), and the physical dimensions of width and height (f_{27}) are effective predictive features.

The results in Figure 10 raise the following question: *Is there a guideline that we could use to estimate the effectiveness of feature vectors without conducting a large-scale test?* From f_9, f_{15}, and f_{16}, it seems that high-dimensional features play more important roles in making predictions. To verify the point, we calculate the ratio of the energy of an individual feature to that of all the features. Specifically, we divide the L2 norm of each feature by the total norm of all the features. The Pearson correlation between the prediction performance and the energy ratios is only 0.47, which means they are moderately correlated. It also appears that features that vary significantly as the image size changes are more effective for making predictions. To verify this conjecture, we calculate the L2 distance between an individual feature extracted from images

Table 3: The features used in this paper and their corresponding dimensions, types, and references

	Meaning	D	Tp	Rf
f_1	area of the centric region containing 70% of edge pixels	1	**C**	[8]
f_2	contrast, mean, and standard deviation of intensity, hue, and saturation	9	**C**	[8]
f_3	hue contrast, hue count, and hue frequency	3	**C**	[8]
f_4	mean values of hue, saturation, and intensity in the three largest regions	9	**C**	[8]
f_5	blurriness	1	**C**	[8]
f_6	mean and standard deviation of intensity, hue, saturation in the center region	6	**C**	[3]
f_7	mean Daubechies wavelet coefficients of hue image, saturation image, and value image, at three levels	9	**C**	[3]
f_8	average hue distance between the five largest regions	1	**C**	[3]
f_9	histogram of average RGB	256	**C**	[7]
f_{10}	dark mean	1	**C**	[10]
f_{11}	depth of field of hue image, saturation image, and value image	3	**C**	[6]
f_{12}	level 1 pyramid histogram of oriented gradients (PHOG)	40	**C**	[1]
f_{13}	Gabor wavelet texture	480	**C**	[11]
f_{14}	mean values of hue, saturation, and intensity in ROIs	3	**C**	
f_{15}	intensity histogram	256	**C**	
f_{16}	hue histogram	360	**C**	
f_{17}	ratios of areas of the three largest regions to the whole image	3	**S**	[8]
f_{18}	coordinate vector of the largest region	2	**S**	[8]
f_{19}	spatial variance of the three largest regions	3	**S**	[8]
f_{20}	number of regions whose areas are larger than 10% of the whole image	1	**S**	[3]
f_{21}	ratio of the total area of the five largest region to the whole image	1	**S**	[3]
f_{22}	ROI weighted by rule of thirds	1	**S**	[9]
f_{23}	diagonal edges	1	**S**	[9]
f_{24}	mean positions and slopes of horizontal lines and vertical lines	4	**S**	[10]
f_{25}	sum of edge magnitudes weighted by rule of thirds	1	**B**	[7]
f_{26}	visual symmetry evaluated by difference between PHOGs in the left and right halves of an image	1	**B**	
f_{27}	width and height of the image	2	**O**	

with a scale of 0.1 and that extracted from images with a scale of 1.0. The Pearson correlation between the prediction performance and the feature difference is only 0.34, which means they are also moderately correlated.

The above observations show that it is difficult to explicitly estimate the effectiveness of features merely based on the feature vectors. Fortunately, we found the feature weights discovered by the PLSR process implicitly indicate a feature's effectiveness. The process estimates a weight for each dimension of each feature vector. For example, f_{12} has 40 dimensions, so it has 40 corresponding weights after model fitting. First, we take the absolute values of all the weights, and use the maximum weight of all dimensions of a feature vector to construct a line graph, as shown in Figure 10. The reason for finding the maximum is that the PLSR process only uses the most important feature vectors to construct partial least square components. In Figure 10, the peaks of the line graph are consistent with effective features to some extent. Quantitatively, the Pearson correlation between the weights and performance is 0.61. The positive correlation coefficient shows that the weights learned by PLSR, though not perfect, are good indicators for estimating the features that are more effective. This characteristic strengthens the motivation for using PLSR rather than a nonlinear regression method like support vector regression (SVR). In fact, we did construct an SVR-based prediction model and found that its performance was similar to that of PLSR. We utilize the PLSR-based model in this study because it is more efficient, and the discovered feature weights can be used to estimate the features' effectiveness.

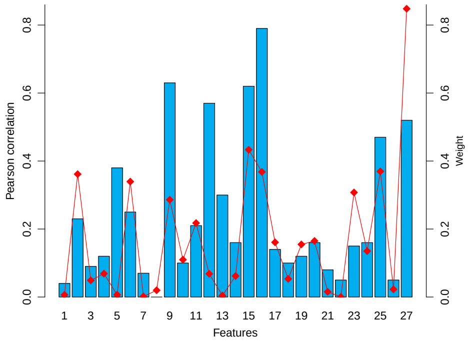


Figure 10: The prediction performance and feature weights obtained from the scoring model based on individual features

Table 4: The performance of the scoring model based on all the features and the selected features in the dimension mode

	All features	Selected features
Pearson	0.86	0.857
Spearman	0.87	0.867

Based on the feature effectiveness analysis, we compare the performance achieved by fusing all the features described in Table 3 with that derived by only fusing features that have a Pearson correlation larger than 0.3. Table 4 shows the prediction performance of the scoring model in the dimension mode based on all the features and the selected features. We observe that the performance derived by fusing the selected features is almost the same as that achieved by fusing all the features. Many features proposed in the literature were designed for different datasets, but they embed similar information and redundant information would be discarded by the learning process. Adding more features only yields a marginal improvement unless they provide novel information. Nevertheless, in the following evaluation, all features are fused to construct models in order to faithfully capture the maximum attainable prediction performance.

6. EVALUATION

6.1 Score Prediction vs. Ranking Prediction

Given a test image, the scoring model predicts its absolute aesthetic score; therefore, we can see how the aesthetic rating changes when the image is displayed in different sizes. Meanwhile, the ranking model predicts the changes in the relative ranking when an image is scaled up/down. In this subsection, we evaluate the performance of both models in the dimension mode. Then, we compare the pixel and dimension modes in the next subsection (Section 6.2).

Table 5 shows the prediction performance of the scoring model based on three categories of features: content-based features, composition-based features, and a combination of them. We make two observations from the table. First, irrespective of the model, content-based features play an important role in predicting aesthetic perception. This may be because the dimensionality of content-based features is much higher than that of compositional features. The latter are usually based on photographic rules and conventions, and can be described by a few scalars. The second observation is that the scoring model is more accurate than the ranking model in predicting

Table 5: The performance of the scoring and ranking models based on different feature categories in the dimension mode

Scoring model	**C**	**S**	**C+S**	All
Pearson	0.83	0.33	0.83	0.86
Spearman	0.84	0.35	0.84	0.87
Ranking model	**C**	**S**	**C+S**	All
Pearson	0.65	0.23	0.65	0.65
Spearman	0.64	0.24	0.64	0.64

Table 6: The performance of the scoring models

Scoring model	Pixel mode	Dimension mode
Pearson	0.84	0.86
Spearman	0.84	0.87

how aesthetic scores change in different scales. There is a trend in that *larger images tend to receive higher aesthetic ratings*, which is evident in Figure 5. On the other hand, the aesthetic ranking of an image in different scales seems to be more complicated, and it is harder to find a simple, general rule for rank prediction in a particular scale.

It may not be necessary to determine exactly how the ranking changes over multiple scales. In many situations, it is sufficient to assess whether an image is more aesthetically pleasing when it is scaled up (or down) to a certain size. Thus, we evaluate the ranking model from the perspectives of binary-decisions and ternary-decisions. From the binary-decision perspective, images in the ground truth dataset are categorized into two groups: *promoted images* and *demoted images*. An image is classified as promoted if gets a higher rank when it is scaled up (i.e, to a larger display size); otherwise, it is classified as demoted. From the ternary perspective, the ground truth dataset is divided into three classes: *promoted over 10%*, *demoted over 10%*, and *neither promoted or demoted over 10%*. The overall precision rates of evaluating the ranking model from the binary perspective and the ternary perspective are 0.79 and 0.58 respectively. Thus, the ranking model achieves a satisfactory performance in predicting promotion and demotion behavior.

6.2 Physical Dimensions Are More Important

We also investigate whether an image's resolution or its display size is more important in predicting perceptions of aesthetic quality. We compare the performance of both prediction models running in the pixel mode and the dimension mode. The Pearson correlation and Spearman correlation results in Table 6 show that both models achieve higher accuracy in the dimension mode than in the pixel mode. This is reasonable because the physical size of an image reflects how humans perceive it more directly, even though the number of pixels also matters.

7. DISCUSSION

7.1 Comparison with Human Assessors

To compare the prediction models with individual human assessors, we calculate the Pearson correlation and Spearman correlation between individual humans' assessments and the corresponding ground truths, which are obtained by averaging assessments over all the subjects. Surprisingly, the average Pearson correlation and the average Spearman correlation are 0.55 and 0.54 respectively. Comparing these results with those in Table 7 reveals two facts. First, image aesthetic assessment is more subjective than expected, such that the consistency between each individual's assessment and the ground truth is only moderate. Second, the proposed prediction model is slightly better than individual human assessors. One reason is that humans usually give one discrete value as the score, while the prediction models can output a real-valued score that more accurately reflects the general assessment of a group of people.

Table 7: The Pearson correlations of the scoring and ranking models in both modes

	Pixel mode	Dimension mode
Scoring model	0.84	0.86
Ranking model	0.62	0.65

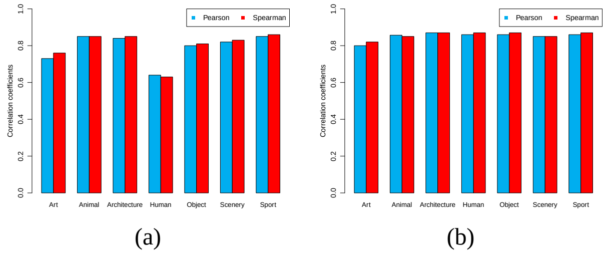


Figure 11: Pearson correlations and Spearman correlations for different image categories. Left: the pixel mode; right: the dimension mode.

7.2 Content-dependent Prediction

The collected trace is categorized into seven types of images: art, animals, architecture, humans, objects, scenery, and sport. We investigate the prediction performance variations of the above categories. A five-fold cross validation scheme is used to construct and evaluate the scoring models for each image category. The performance variations are shown in Figure 11(a). The performances of the art and human categories are relatively worse than those of other categories. Figure 11(b) shows that the performance of art images in the dimension mode is also the worst among all the categories, but performance variations in this mode are smaller. These figures may indicate that humans are more likely have diverse individual interpretations for art and human images. Further investigations are required for more evidences of this observation.

7.3 Existence of "Best Display Size"?

We have verified that the aesthetic perception of an image is influenced by its display size and resolution. We now consider another question: *Is there a rule for finding the "best size" to display an image*?

Generally, higher resolution images are perceived as aesthetically better than lower resolution images. That is, the aesthetic score increases with an image's resolution; thus, the *the-bigger-the-better rule* is generally valid when displaying images. However, we should also consider the following question: When a set of images is displayed, are larger images always *ranked* higher than smaller images? We posit that, although an image's aesthetic score increases with the resolution, the degree of increase is not consistent for all images. In other words, some images may be ranked relatively higher when they are displayed in smaller sizes, while others may be ranked higher when they are displayed in larger sizes. To verify this point, we ranked the 100 images in each scale based on their user-rated aesthetic scores, and collected the rankings of specific images at different scales. Figure 12 shows the evolution of the normalized aesthetic rankings of three sample images. Note that a smaller normalized ranking score means the corresponding

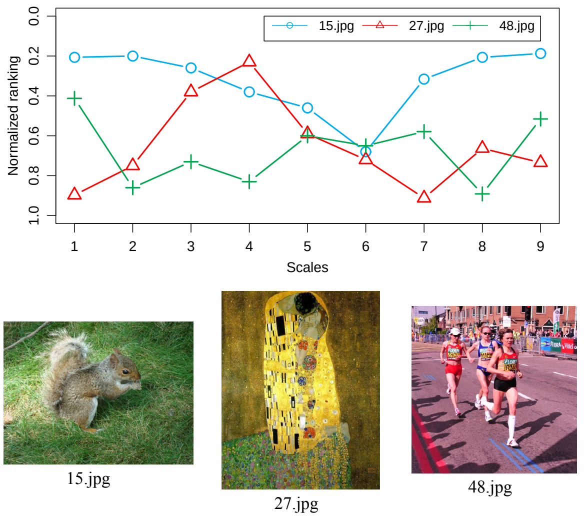


Figure 12: The evolution of the aesthetic rankings of three sample images in multiple scales

image is ranked higher (i.e., perceived more appealing), as defined in Section 5.2. Interestingly, the squirrel image (15.jpg) is ranked lower at scale 6, and higher at scales 1 and 9, i.e., the aesthetic quality of the medium-sized squirrel image is perceived as poorer than that of the smaller and larger images. By contrast, the medium-sized runner image (48.jpg) is ranked better than the smaller and larger scales. In most cases, aesthetic rankings are like the ranking of the painting image (27.jpg), which fluctuates across scales. Manual inspection of the image dataset revealed that 80% of the evaluation images did not follow a clear trend in terms of ranking evolution. This finding suggests that the mechanism humans use to compare the aesthetic quality of images is complex, and further study is needed to resolve this challenging issue.

7.4 Application of Findings

The proposed method predicts the aesthetic scores/rankings of images displayed in various sizes, and relevant information can be integrated into existing algorithms from various perspectives. First, the feature weights discovered by the PLSR process can be utilized by other algorithms to achieve better feature fusion. Second, the results of the scoring model can be used to decide the best way to present an image (i.e., the number of pixels or the physical dimension size) on different display peripherals, such as PC monitors and mobile phones, so that better aesthetic quality is guaranteed. Third, the results of the ranking model can be exploited in aesthetics-based image re-ranking as the display size of an image changes.

8. CONCLUSION

In this paper, we have demonstrated that an image's resolution and physical dimensions affect humans' aesthetic perception of it. Users' ratings were collected via a large-scale crowdsourced experiment, and a set of regression models was used to predict the aesthetic scores and aesthetic rankings of images. We show that the impact of image scaling is not consistent across images and is highly dependent on the image content. Through careful selection of features from a pool of content-based features and compositional features, the scoring model and the ranking model are constructed to predict the absolute aesthetic values and relative aesthetic rankings. The results of experiments conducted in the dimension mode demonstrate that estimating an image's aesthetic quality should consider its content as well as how it is displayed (i.e. its size).

We also discuss a number of potential future directions, including the fact that the *the-bigger-the-better* rule may not be valid in aesthetic ranking. To the best of our knowledge, this is the first work that investigates the impact of image resolution and display size on aesthetic quality assessment. Image features considering display size may be needed, and studies from various perspectives are highly encouraged in order to further the understanding of humans' visionary mechanisms for evaluating the aesthetic quality of images.

Acknowledgement

The work was partially supported by the National Science Council of Taiwan under the grants NSC101-2221-E-194-055-MY2 and NSC101-2221-E-001-012-MY3. The authors are in debt to Derec Wu and Ming-Hung Wang for their support on the design and implementation of crowdsourcing experiments.

9. REFERENCES

[1] A. Bosch, A. Zisserman, and X. Munoz. Representing shape with a spatial pyramid kernel. In *Proceedings of CIVR*, pages 401–408, 2007.

[2] K.-T. Chen, C.-C. Wu, Y.-C. Chang, and C.-L. Lei. A crowdsourceable QoE evaluation framework for multimedia content. In *Proceedings of ACM Multimedia 2009*, 2009.

[3] R. Datta, D. Joshi, J. Li, and J. Wang. Studying aesthetics in photographic images using a computational approach. In *Proceedings of ECCV*, pages 288–301, 2006.

[4] R. Datta and J. Wang. ACQUINE: Aesthetic quality inference engine — real-time automatic rating of photo aesthetics. In *Proceedings of MIR*, pages 421–424, 2010.

[5] S. de Jong. SIMPLS: an alternative approach to partial least squares regression. *Chemometrics and Intelligent Laboratory Systems*, 18:251–263, 1993.

[6] S. Dhar, V. Ordonez, and T. Berg. High level describable attributes for predicting aesthetics and interestingness. In *Proceedings of CVPR*, pages 1657–1664, 2011.

[7] Y. Ke, X. Tang, and F. Jing. The design of high-level features for photo quality assessment. In *Proceedings of CVPR*, pages 419–426, 2006.

[8] C. Li and T. Chen. Aesthetic visual quality assessment of paintings. *IEEE Journal of Selected Topics in Signal Processing*, 3(2):236–252, 2009.

[9] L. Liu, R. Chen, and D. Cohen-Or. Optimizing photo composition. *Eurographics*, 29(2):469–478, 2010.

[10] W. Luo, X. Wang, and X. Tang. Content-based photo quality assessment. In *Proceedings of ICCV*, pages 2206–2213, 2011.

[11] B. Manjunath and M.-Y. Ma. Texture features for browsing and retrieval of image data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18(8):837–842, 1996.

[12] X. Sun, H. Yao, R. Ji, and S. Liu. Photo assessment based on computational visual attention model. In *Proceedings of ACM Multimedia*, pages 541–544, 2009.

[13] H. Tong, M. Li, H. Zhang, J. He, and C. Zhang. Classification of digital photos taken by photographers or home users. In *Proceedings of PCM*, 2004.

[14] L.-K. Wong and K.-L. Low. Saliency-enhanced image aesthetics class prediction. In *Proceedings of ICIP*, pages 997–1000, 2009.

[15] O. Wu, W. Hu, and J. Gao. Learning to predict the perceived visual quality of photos. In *Proceedings of ICCV*, pages 225–232, 2011.