Can relative skill be determined from a photographic portfolio?

Abhishek Agrawal*[a], Vittal Premachandran[a], Rajesh Somavarapu[b], Ramakrishna Kakarala[a]
[a]Nanyang Technological University, 50 Nanyang Avenue, Singapore 639798;
[b]University of Texas, Dallas, 800 W Campbell Rd, Richardson, TX, USA, 75080

ABSTRACT

In this study, our primary aim is to determine empirically the role that skill plays in determining image aesthetics, and whether it can be deciphered from the ratings given by a diverse group of judges. To this end, we have collected and analyzed data from a large number of subjects (total 168) on a set of 221 of images taken by 33 photographers having different photographic skill and experience. We also experimented with the rating scales used by previous studies in this domain by introducing a binary rating system for collecting judges' opinions. The study also demonstrates the use of Amazon Mechanical Turk as a crowd-sourcing platform in collecting scientific data and evaluating the skill of the judges participating in the experiment. We use a variety of performance and correlation metrics to evaluate the consistency of ratings across different rating scales and compare our findings. A novel feature of our study is an attempt to define a threshold based on the consistency of ratings when judges rate duplicate images. Our conclusion deviates from earlier findings and our own expectations, with ratings not being able to determine skill levels of photographers to a statistically significant level.

Keywords: Image Aesthetics, Computational Photography, Composition, Photography, Crowd-sourcing, Aesthetics, Image Quality, Photographic Skill.

1. INTRODUCTION

Our aim is to understand the role that skill plays in photographic composition and aesthetics. Few scientific studies have been carried out in this domain and even fewer address the supposition that skill is apparent in photography, even without considering the quality of the instrument used to take the photograph. Composition is defined as the placement or arrangement of visual elements or ingredients in a work of art or a photograph, as distinct from the subject of a work [1]. This definition makes composition it amenable to a scientific enquiry, which we carry out through a data driven approach. Our work primarily differs from previous studies and other known literature in this domain in terms of analyzing photographic composition and its impact on appeal from a scientific perspective. Our work is also different from photo quality, which has been studied in detail earlier, as pointed out in [2].

Although there are many photographs for which the skill of the photographer is immediately apparent, those are usually photographs for which the photographer has full control over the subject, composition, lighting etc. But we intend to probe, whether relative skill is decipherable among subjects for whom the setup in which the pictures are taken is controlled. A number of studies have been conducted in the recent past to evaluate various aspects of image aesthetics. These studies have used different subjective descriptors such as "appeal" or "interestingness" to define what we surmise are the components that describe aesthetics. We begin with describing the relevant work in reverse chronological order. For instance, Berg et al. [3] evaluated how compositional factors such as object placement and size were important in determining human perception of importance within an image. Their findings indicate that the objects that are larger and closer to the center of an image are perceived to be more important and are likely to be mentioned in the subjects' description of the image. Another study [4] demonstrated the use of a set of compositional attributes in selecting high quality aesthetic images and predicting interestingness from large image collections. A unique aspect of this study was that it introduced the concept of using a set of high level describable compositional attributes such as the "rule of 3rds", "opposing colors" or "depth of field" to estimate the aesthetic appeal and interestingness of the image. Their results prove that classifiers trained in these high level attribute features are much more effective than those trained in purely low level features. Their finding on the accuracy of these compositional attributes motivates us to look for similar cues while evaluating the judges' preference of the images used in our experiment, where a large number of human judges evaluate their aesthetic appeal.

*abhishekagrawal.bits@gmail.com; phone: +65 – 94677497, +91 – 9981350116

Human Vision and Electronic Imaging XVIII, edited by Bernice E. Rogowitz, Thrasyvoulos N. Pappas, Huib de Ridder,
Proc. of SPIE-IS&T Electronic Imaging, SPIE Vol. 8651, 86510X • © 2013 SPIE-IS&T
CCC code: 0277-786X/13/$18 • doi: 10.1117/12.2008065

The studies conducted by [5] and [6] are the primary motivation behind our current work, in which the same set of images are rated, albeit with different group and number of human judges. A key finding of [5] is that expert judges are able to discern photographic skill to a statistically significant level. We test this result on a large number of experimental subjects which we enroll using a crowd-sourcing platform (more details in Section 2).

We also establish correlations between judges across experiments to evaluate how the ratings compared between judges whose photographic skill was known based on their domain expertise (photographic experts and imaging researchers) and the judges whose photographic skill was known based on self-evaluation. Other studies such as [7] and [8] have tried to define visual properties that make photographs aesthetically more appealing and establish a relationship between emotions which are aroused by the low level features present in them. These low level features include exposure of light and colorfulness, saturation and hue, rule of thirds, texture, size, aspect ratio and depth of field indicators etc. They have successfully used these features to design a machine-learning based image classification engine called ACQUINE which rates images according to their aesthetic quality.

This online engine predicts an aesthetic score for user-uploaded images between 0 and 100. An interesting study conducted by [5] compares the score predicted by ACQUINE to the scores given by a set of human judges whose photographic skill was previously known on the same set of experimental images. Their results show little correlation between the evaluation of those images by the human judges and ACQUINE. This leads us to the important question: Are the numerous low level features used in the online machine learning based prediction engine enough to accurately determine aesthetic value of an image? The evidence given by [5] certainly negates this, and the author of [7] points out in a separate exposition [9] that low-level image properties are insufficient in describing the quality of image aesthetics which should include descriptions of high-level perception.

Apart from the above studies, a few others such as [10] and [11] have used a set of high-level semantic features to achieve high precision rates of up to >90% for use in web image search and retrieval applications. Another attempt at classification of pictures taken by home users and photographers is made in [12] in which a set of low-level features explicitly related to high level semantic concepts are investigated together with a set of general-purpose low-level features. These studies [10], [11] and [12] are different from our current work in that they aim at measuring image quality which can be more readily determined through well established quantitative measures such as image sharpness, contrast, noise etc.

Finally, the study [2] empirically determines that image appeal (which we consider as an important component of the overall aesthetic value of an image) is related to image quality only with respect to the influences in the category of objective metrics defined above. Further, major factors that decide image appeal to a large extent belong to non-objective categories such as people, composition and subject. The influences of these non-objective measures are what we aim to measure in this current study.

2. EXPERIMENTAL METHODS AND DESIGN

We have developed our experimental design considering the design made by an earlier study [5] as a benchmark, but changing the rating scales, introducing a test for evaluating photographic skill and using a crowd sourcing method (Amazon Mechanical Turk interface) for collecting data. We have collected experimental data from a large number of subjects (total 168) and have categorized the subjects by their skill, knowledge and prior experience in photography. We now discuss the specifics of the experiment in the below sections.

2.1 Dataset Used

For our experiment, we use the database described in [5] which consists of photographs of 7 different commonplace scenes like a portrait, zebra crossing, building corner etc. These were taken by 33 photographers who varied in their photographic skill, i.e. Novices, who rarely take photographs (4), Amateurs, who often take photographs (15), Enthusiasts, who invest time and resources on photography and have probably been trained and possess knowledge of photography (12) and Experts, who are professional photographers with previously published or exhibited work (2). Each photographer took one photograph of the 7 dfferent scenes using the same model of point-and-shoot camera with 7 megapixel resolution, set to full automatic mode. This ensured that picture quality was not affected by the mode chosen. With the help of two professional photographers, we also defined the area in which photographers were allowed to maneuver. This also made sure that individual photographers had enough room to adjust the composition of the photographs and also click pictures having similar subjects. Out of the 231 photographs that were obtained, 10 of them were discarded for violation of announced rules, bringing the total number of orginal images used to 221. Examples of the images from each of these 7 categories are listed below.

Fig. 1. Sample images from seven categories (from top left) are Open or Free Shot, Still life, Building Corner, Portrait, Staircase, Zebra crossing, Fountain

2.2 Rating Scales

The study [13] lists a number of rating scales that have been extensively used in prior studies and how they affect users' opinion. Their findings indicate that users rate consistently across different rating scales. However, when taking users' opinion on their preference of rating scales, they found that the half-star rating scale (used in [5]) was the most preferred scale and the binary scale (used in the current study) was the least preferred.

In a previous study [5], a graduated 10 point scale was used which is purported to put a higher cognitive load on decision-making of a human judge than a binary rating scale. We map the findings of the previous work [5] and the current study to evaluate whether the results of the study [13] are consistent. With this in mind, we also try to provide possible explanations on how changing rating scales affects users' opinions, and whether there is any inherent bias associated with a particular rating scale.

(a)

(b)

Fig. 2. Parts (a) and (b) show the rating system used in the current and previous studies. (a) Graduated Scale based voting system (b) Like/Dislike based voting system

2.3 Embedding Duplicates

Our method is similar to the study [2] in that we both have embedded duplicate images in respective image categories. While they identified in their pilot study that the duplicate images were treated differently by judges and adjusted the scores to account for it, we embedded duplicates to devise a method to detect frivolous voting by non-serious voters. The main difference between the two experimental setups is that our current setup had an online voting system whereas the former setup [2] had an offline method where they gave hard copies of the experimental photographs to the human judges.

We inserted two duplicate images in each of the seven categories, bringing the total number of images from 221 to 235. Finally, we evaluate the distribution of number of duplicate images that were rated differently by each human judge and present our results in Section 3.

2.4 Use of Amazon Mechanical Turk (AMT) Interface

The AMT has been used in [16] to obtain ratings of image quality. In this paper we use AMT to obtain ratings of aesthetics rather than quality. In AMT, the experimenter as a "requester" can design a set of human intelligence tasks or HITs that can be used to enrol "workers" with a given skill and qualification to perform a given task. The requester can create additional qualifications and tests to filter out and categorize the workers as per the needs of the experiment. Once the worker obtains the necessary qualifications, they are redirected to a HIT window where he/she performed the main task. The AMT interface also has a feature which allows the worker to complete a HIT in a span of time rather than in a single time window, which we have incorporated in our design.

We used two separate HITs for this experiment; the first one for obtaining registrations which redirected workers to our website; www.computationalphotography.in. Here, they filled out a basic questionnaire comprised of questions related to their prior experience, training and expertise in photography. Upon registration with a valid email id, the registered person received an email containing a code which he/she put into the AMT interface to complete the registration successfully. Thereafter, the worker was granted the necessary qualification (called "photography experiment test") which allowed them to participate in the second HIT and complete the voting on our website within a fixed time span of 7 days. Figure 3 below is a sample screenshot of the AMT interface used for registering the workers:

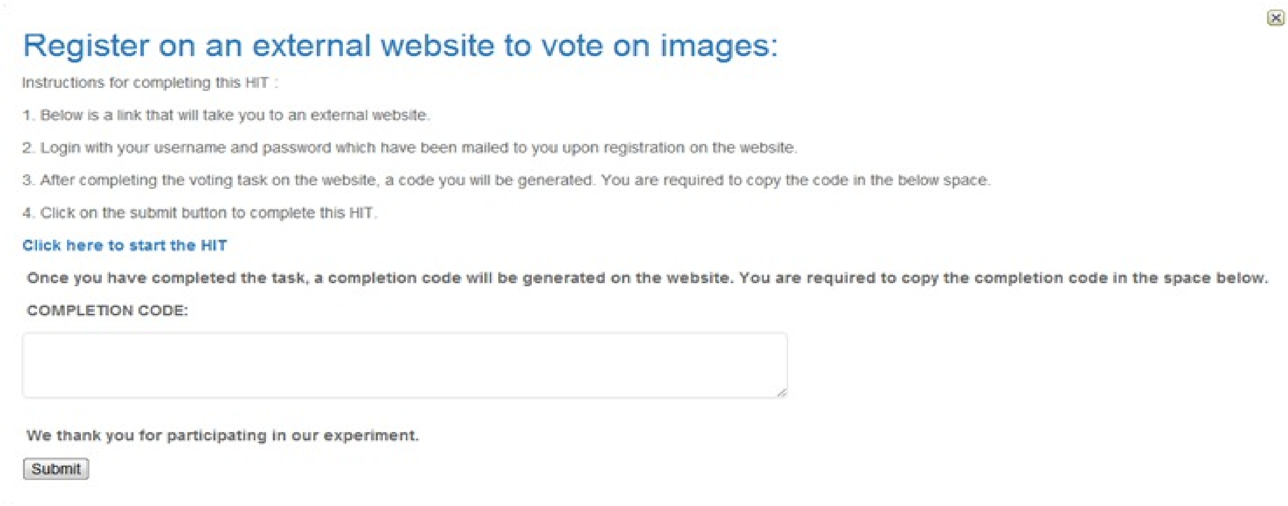

Fig. 3. Sample Human Intelligence Task (HIT) window presented to a worker on the Amazon Mechanical Turk platform

2.5 Photography test

We used the basic photography test available at [14] with two minor modifications to evaluate the photographic skill of human judges objectively and draw out a comparison with their own (non-objective, estimated) rating. We have omitted questions 3 and 4 from the test, as they were ambiguous and only applicable to judges using the Windows operating system environment. The total number of questions and score were both equal to 23, with a worker/judge getting one point of each correct answer and no marks for a wrong answer. We have also modified the scales used to fit our experimental needs as reflected in the table below:

Table 1. Relationship between test score and skill level

RANGE OF SCORE	SKILL LEVEL
Greater than 21	Expert
Between 16 and 21	Enthusiast
Between 10 and 15	Amateur
Less than 10	Novice

We divided the 168 human judges on the AMT interface into two batches: Batch-1, which did not take the test and had 101 judges and Batch-2, which took this test and had 67 judges. Our test serves multiple purposes: (a) Segregating users based on a standard well established measure of photographic knowledge; (b) For correlating the skill of the human judges and their ratings on photographs across experiments; (c) To evaluate the accuracy with which human judges rate their own photographic skill. We discuss the above results and correlations in the next section.

3. RESULTS AND DISCUSSION

There are several interesting parameters and correlations that can be measured using the data calculated using our method which provides valuable insights about the role of skill in image composition, users' behavior while rating on different scales across experiments and consistency of voting in the same user group. We discuss these results below.

3.1 Correlation between our findings and the 10-point scale

We use the Spearman correlation coefficient discussed in [15] to compare the findings of this study with that of earlier studies [5] and [7]. To establish this correlation, we calculate the like percentage given to a particular image that is defined as the percentage of the 168 judges that liked the image. We then calculate the average 10 point rating given to corresponding images by the 8 human judges in [5] and obtain the below statistically significant correlations:

Table 2. Statistically significant correlation coefficient across different experiments ($p<0.01$)

	10-POINT SCALE [5]	ACQUINE [7]
LIKE PERCENTAGE	0.59	.19

The high correlation between the 10-point scale and the Like/Dislike ratings indicates the consistency of subjective human evaluation of image aesthetics independent of the scale used to evaluate the ratings. This result probably eliminates the need to compensate for variation in rating scales in future studies. There is also a significant diversity in the group of judges who rated the study conducted using the 10-point scale (imaging researchers and photography experts) and the ones in our study. This indicates that there is no significant variation in the ratings given by judges who varied in their prior knowledge, skill and experience in photography. We see that ACQUINE scores have low correlation with like percentages indicating that the machine learning approach is still unable to capture the preferences of the human judges.

Table 3. Distribution of human judges as per the skill level

CATEGORY	BATCH – 1		BATCH – 2			
	Self – Evaluation	Percentage of total Judges	Self – Evaluation	Percentage of total Judges	Photography Test	Percentage of total Judges
Novices	33	32.7	19	28.4	34	50.8
Amateurs	40	39.6	35	52.2	16	23.9
Enthusiasts	26	25.7	13	19.4	6	8.9
Experts	2	2.0	0	0	11	16.4
Total	**101**	**100**	**67**	**100**	**67**	**100**

3.2 Judges' Evaluation of their Photographic Skill

We found that there is a major difference between judges' self-evaluation of their photographic skill and when it was evaluated using the photography test described in Section 2.5. Table 3 contains figures for the number of judges within each category in both the batches and Table 4 reveals the bias in their own rating of their photographic skill.

Table 4. Bias in self - evaluation of skill in Batch - 2

	Under Rated	Over Rated	Equal
No. Of Judges	23	28	16

A high proportion (41.8 %) of judges over-rated their photographic skill, most of them being Novices who rated themselves as Amateurs and Enthusiasts as reflected in Table 3. Most of the judges (34.3 %) who under-rated their photographic skill turned out to be Experts, as evaluated by the standard photography test and tabulated in the Table 3. Overall, the above findings indicate that judges have a tendency to over-rate or under-rate their skill.

3.3 Correlation between Judges' ratings across experiments

We evaluate correlation between the ratings given by judges in our experiment and the earlier study [5] described in section 2.2 which used the 10 point rating scale.

Table 5. Correlation among judges across experiments

	Like/Dislike Experts (Self – Evaluation)	Like Dislike Experts (Test Based)	Like/Dislike Enthusiasts (Self – Evaluation)	Like/Dislike Enthusiasts (Test Based)
10-point scale Experts	.253	.071*	×	×
10-point scale Imaging Researchers	×	×	0.635	0.365

*****statistically insignificant, since pval (.287) > 0.05, for all other values pval < 0.01

As presented in Table 5, we found statistically significant correlations between expert judges using the 10-point scale and the judges who rated their skill by themselves in our experiment. However, the ratings based on the test-based evaluation of the judges and experts failed to correlate. One factor that might contribute to this is that judges might have referenced online sources while taking the test, as all of them evaluated themselves as amateurs while submitting their opinion on their photographic skill, knowledge and experience separately on our website.

3.4 Analyzing the Like/Dislike Ratings

After calculating the percentage of likes and dislikes for each image, we have chosen the 6 best and worst images in the entire dataset from this experiment and the earlier study [5], to see the distribution of ratings across the two experiments.

Fig. 3. Top Liked and Disliked Images on the Like/Dislike Scale

Below are the percentage ratings obtained by the best and the worst images in the Like/Dislike scale by the human judges.

Table 6. Best and Worst Images in the Like/Dislike Scale

BEST IMAGES		WORST IMAGES	
CATEGORY	**LIKE PERCENTAGE**	**CATEGORY**	**LIKE PERCENTAGE**
Open or Free Shot	95.8	Building Corner	11.3
Fountain	95.2	Staircase	14.3
Open or Free Shot	95.2	Staircase	15.5
Open or Free Shot	94.6	Zebra Crossing	15.5
Open or Free Shot	94.6	Fountain	19.6
Open or Free Shot	94.6	Portrait	20.2

We also look at category-specific choices for the best and the worst images.

Table 7. Category Specific Best and Worst Images in the Like/Dislike Scale

	Still Life	Building Corner	Fountain	Staircase	Portrait	Zebra Crossing	Open or Free Shot
Best Images	86.9	82.7	95.2	88.1	75.6	89.3	95.8
Worst Images	29.2	11.3	19.6	14.3	20.2	15.5	28.5

We found that judges particularly liked the images in the Open and Free Shot category, with 5 images out of the top 6 receiving greater than 94 percent positive votes. We associate this with the symmetry attribute of the images in this category as is clear from Fig. 3. No such prominent category-specific preference was observed among the worst images. Overall, the images receiving a poor rating had out-of-focus or misplaced objects or asymmetric compositional attributes which could be responsible for them receiving poor ratings. It is also worth noting that none of the images among the best and the worst were used as duplicates to detect frivolous voting, so we can rule out any re-rating effects.

3.5 Distribution of judges' ratings for duplicate images

We now look at the distribution of ratings when the same set of judges voted on the duplicate images in the category, and the distribution of how much they differed from their original ratings. Our ultimate aim is to come up with a threshold of difference that could be used to segregate frivolous users in similar studies in future. A look at our data reveals that 55.4 % of the judges (93 out of 168) rated differently on 5 to 8 out of 14 duplicate images (please refer to the below figure).

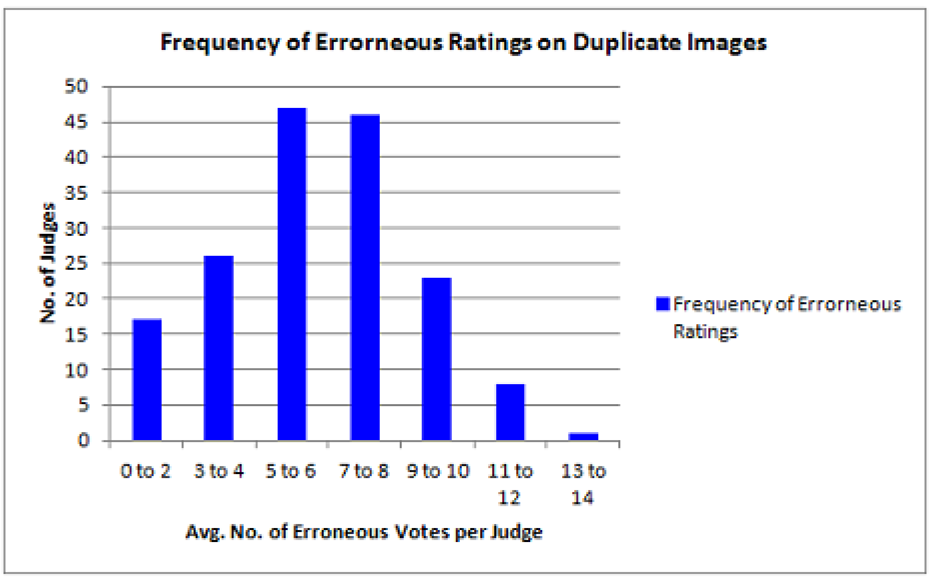

Fig.4. Frequency of Erroneous Rating on Duplicate Images

Based on this statistical distribution, we can put an estimated threshold for an expert or enthusiast to be around 4 and an estimated threshold for an amateur or novice to be around 8 images. Anything greater will probably reflect inconsistency, indecisiveness and probably lack of seriousness on the part of the judge while rating.

The study [13] also evaluated user consistency while collecting ratings on movies on recommender interface MovieLens, and found that the users re-rated their original ratings 60 % of time, or rated differently 40 % of time. Another study [2] found that judges tend to rate on images on the basis of their relative importance to other images in the group and assigned duplicates or near duplicates a zero score. Our findings indicate that while rating images, judges are more erroneous than [13]. But this high error rate in our study could be attributed to the fact that the images are less memorable than the movies in [13].

3.6 Correlation between Skill of the Photographers & Ratings of Judges

We now answer an important question pertinent to the immediate practical relevance of our findings. We applied the Spearman correlation to understand whether aesthetically appealing images in large image collections could be accurately segregated by an unknown, but significantly large and diverse group of judges. Surprisingly, our results provide evidence contrary to that of a similar earlier study [5], which found that the skill levels were decipherable in ratings given by professionals. In fact, we have found that there is no statistically significant correlation between the overall like percentage and the skill level of photographers, as well as between the like percentages of each of the judge groups (self-rated experts, enthusiasts, amateurs and novices) and the skill level.

The reason behind this lack of correlation could be limited understanding of general photography rules, and even more so, the lack of objective behavior when given a subjective task of rating on aesthetic appeal of images. These results also indicate that consideration and application of basic principles of photography is a rare trend in a general population mass while evaluating the aesthetic appeal of images. This observation follows from the results obtained in section 3.2 where we have found that a large number of judges over-rate their domain specific skill and knowledge.

Table 7. Overall like percentage mapped with skill level of photographers

	Novices	**Amateur**	**Enthusiasts**	**Experts**
Overall Like %	62.6	63.8	61.6	54.8

We also studied the rating pattern by the judges for each of the photographer groups separately, as reflected in Table 7. The ratings of photographers having different skill ratings were found to be fairly consistent, with the judges liking pictures clicked by individual photographers equivalently.

4. CONCLUSION AND FUTURE WORK

Our results confirm that there is a high correlation between aesthetic ratings obtained using a binary scale and those obtained using a 10-point scale. We also find that judges' evaluations of their skill levels are often biased and even a fairly significant performance-based inconsistency is observed. This is a useful finding for future experimenters contemplating an evaluation of skill of the participants using similar interfaces based on crowd-sourcing. On the judges' side, we have found that experts usually tend to disagree among themselves while rating similar set of images. This is reflected by the low (0.253, although statistically significant) and statistically insignificant values of correlation coefficients between judges rating on the 10-point scale and the binary scale. However, we can comment with certainty that enthusiasts tend to agree with each other more often as reflected by the high value (.653, statistically significant) of the correlation coefficient. Within our experiment, judges' opinions are fairly consistent with each other on the best and the worst images with 5 out of the top 6 images belonging to one particular category. When asked to rate the same images in a single blind manner, the judges will rate differently on 3.5 to 6 images out of 10 in similar experiments.

Finally, we observe a lack of correlation between skill of the photographers and the cumulative ratings of judges in our experiment. This leads us to suggest that general users are probably not aware and certainly not considerate of photographic principles, which contribute significantly to overall aesthetic appeal of images. We therefore propose to evaluate this statement thoroughly in our future studies, so that appropriate measures could be integrated in recommender interfaces and other such platforms to minimize the erroneous perceptions and inherent biases of judges.

REFERENCES

[1] http://en.wikipedia.org/wiki/Composition_ (visual_arts), accessed 30th January 2013.

[2] Savakis, Andreas E., Stephen P. Etz, and Alexander CP Loui. "Evaluation of image appeal in consumer photography." In *Electronic Imaging*, pp. 111-120. International Society for Optics and Photonics, 2000.

[3] Berg, Alexander C., Tamara L. Berg, Hal Daume, Jesse Dodge, Amit Goyal, Xufeng Han, Alyssa Mensch et al. "Understanding and predicting importance in images." In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pp. 3562-3569. IEEE, 2012.

[4] Dhar, Sagnik, Vicente Ordonez, and Tamara L. Berg. "High level describable attributes for predicting aesthetics and interestingness." In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pp. 1657-1664. IEEE, 2011.

[5] Sachs, Todd, Ramakrishna Kakarala, Shannon Castleman, and Deepu Rajan. "A data-driven approach to understanding skill in photographic composition." In *Computer Vision–ACCV 2010 Workshops*, pp. 112-121. Springer Berlin/Heidelberg, 2011.

[6] R. Kakarala, T. Sachs, and V. Premachandran, "Comparing automated and human ratings of photographic aesthetics", 2011 Color Imaging Conference, 7-11 Nov, San Jose CA.

[7] Datta, Ritendra, and James Z. Wang. "ACQUINE: aesthetic quality inference engine-real-time automatic rating of photo aesthetics." In *Proceedings of the international conference on Multimedia information retrieval*, pp. 421-424. ACM, 2010.

[8] Datta, Ritendra, Dhiraj Joshi, Jia Li, and James Wang. "Studying aesthetics in photographic images using a computational approach." *Computer Vision–ECCV 2006* (2006): 288-301.

[9] Datta, Ritendra, Jia Li, and James Z. Wang. "Algorithmic inferencing of aesthetics and emotion in natural images: An exposition." In Image Processing, 2008. ICIP 2008. 15th IEEE International Conference on, pp. 105-108. IEEE, 2008.

[10] Luo, Yiwen, and Xiaoou Tang. "Photo and video quality evaluation: Focusing on the subject." *Computer Vision–ECCV 2008* (2008): 386-399.

[11] Ke, Yan, Xiaoou Tang, and Feng Jing. "The design of high-level features for photo quality assessment." In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, vol. 1, pp. 419-426. IEEE, 2006.

[12] Tong, Hanghang, Mingjing Li, Hong-Jiang Zhang, Jingrui He, and Changshui Zhang. "Classification of digital photos taken by photographers or home users."*Advances in Multimedia Information Processing-PCM 2004* (2005): 198-205.

[13] Cosley, Dan, Shyong K. Lam, Istvan Albert, Joseph A. Konstan, and John Riedl. "Is seeing believing?: how recommender system interfaces affect users' opinions." In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pp. 585-592. ACM, 2003.

[14] http://www.canonblogger.com/top/basic.html.

[15] Maritz, Johannes S. *Distribution-free statistical methods*. Vol. 17. Chapman & Hall/CRC, 1995.

[16] Ribeiro, Flávio, Dinei Florencio, and Vítor Nascimento. "Crowdsourcing subjective image quality evaluation." In *Image Processing (ICIP), 2011 18th IEEE International Conference on*, pp. 3097-3100. IEEE, 2011.