Comparing humans to automation in rating photographic aesthetics

Kakarala, Ramakrishna; Agrawal, Abhishek; Morales, Sandino

2015

Kakarala, R., Agrawal, A., & Morales, S. (2015). Comparing humans to automation in rating photographic aesthetics. Proceedings of SPIE - Imaging and Multimedia Analytics in a Web and Mobile World 2015, 9408, 94080C-. doi:10.1117/12.2084991

<https://hdl.handle.net/10356/88344>

<https://doi.org/10.1117/12.2084991>

© 2015 Society of Photo-optical Instrumentation Engineers (SPIE). This paper was published in Proceedings of SPIE - Imaging and Multimedia Analytics in a Web and Mobile World 2015 and is made available as an electronic reprint (preprint) with permission of Society of Photo-optical Instrumentation Engineers (SPIE). The published version is available at: [<http://dx.doi.org/10.1117/12.2084991>]. One print or electronic copy may be made for personal use only. Systematic or multiple reproduction, distribution to multiple locations via electronic or other means, duplication of any material in this paper for a fee or for commercial purposes, or modification of the content of the paper is prohibited and is subject to penalties under law.

Downloaded on 26 Apr 2023 19:02:14 SGT

Comparing humans to automation in rating photographic aesthetics

Ramakrishna Kakarala, Abhishek Agrawal and Sandino Morales

School of Computer Engineering, Nanyang Technological University, Singapore

ABSTRACT

Computer vision researchers have recently developed automated methods for rating the aesthetic appeal of a photograph. Machine learning techniques, applied to large databases of photos, mimic with reasonably good accuracy the mean ratings of online viewers. However, owing to the many factors underlying aesthetics, it is likely that such techniques for rating photos do not generalize well beyond the data on which they are trained. This paper reviews recent attempts to compare human ratings, obtained in a controlled setting, to ratings provided by machine learning techniques. We review methods to obtain meaningful ratings both from selected groups of judges and also from crowd sourcing. We find that state-of-the-art techniques for automatic aesthetic evaluation are only weakly correlated with human ratings. This shows the importance of obtaining data used for training automated systems under carefully controlled conditions.

Keywords: Aesthetics, photography, machine learning

1. INTRODUCTION

The “aesthetics” of a photograph is a term that refers to those aspects of the photograph which appeal to many people. Aesthetic appeal is surely complex, being based on many variables and factors including viewing conditions, prior training, and cultural conditioning, and also many factors which may not even be known. Despite that, when asked to rate the aesthetic appeal of a photograph on a scale, for example from 1 to 10, or to make a simpler binary decision whether or not a photograph is appealing, people show a consistency that suggests the possibility of machine learning of aesthetics. Many researchers have applied machine learning to large databases of photographs to develop automated rating of aesthetics.^{1,2} Such learning is usually based on extracting a variety of quantities from image attributes, and training classifiers to use those quantities to match mean ratings of a group of human observers.

This paper explores the methodology used both to train automated aesthetic rating systems and also to assess their performance. We begin by considering the databases used for training, and then explore the predictive natures of various quantifiable image attributes. Our goal is to contribute to the development of tools that help novice and amateur photographers improve their skills.

Traditionally, the training of automated rating systems has been on photographs extracted from online photo sharing sites that allow rating. Sites such as www.dpchallenge.com or www.photo.net allow registered members to view and rate photographs by category, such as flowers, portraits, or sports, or by challenges with rules, such as the level of enhancement allowed. Rating is on an integer scale such as from 1 to 10, and there are no instructions given to guide the user. Statistics of ratings given by users are available. Data from photo.net is used by Datta *et al.*¹ and data from dpchallenge.com is used to produce the AVA dataset of over 250,000 images by Murray *et al.*² Another popular site is www.flickr.com, which allows registered users to add photos to their favorites list. The number of users that select an image as their favorite serves as a score of the appeal of the photograph, though this number is not normalized. Data from this site is used in the ImageCLEF retrieval challenge.³

Following Sachs *et al.*,⁴ we may divide photographers into roughly four categories: Novices, who rarely take photographs; Amateurs, who often take photographs but do not pay particular attention to achieve aesthetically pleasing results; Enthusiasts, who frequently take photographs, share them, and invest time and money to

Contact author email: r.kakarala@ieee.org

Imaging and Multimedia Analytics in a Web and Mobile World 2015, edited by Qian Lin,
Jan P. Allebach, Zhigang Fan, Proc. of SPIE-IS&T Electronic Imaging, Vol. 9408, 94080C
© 2015 SPIE-IS&T · CCC code: 0277-786X/15/\$18 · doi: 10.1117/12.2084991

improve results; and Experts, whose work is so widely appreciated that it is exhibited in museums. With this taxonomy, we examine the data currently used for training automated rating systems.

Data obtained from photographic sharing sites is subject to several confounding factors, which have been commented on in the literature.² First, photos are predominately taken by Enthusiasts for rating by Enthusiasts, which reinforces certain tendencies. Among these are extensive enhancement of images, leading to textures and tones that are not possible in real world objects and scenes. It is not clear whether Novice viewers or, at the other end, Expert photographers would appreciate the same photographs that Enthusiasts do. Second, the data is not categorized by equipment level, which has an effect on such photographic attributes as depth of field and dynamic range. It is possible that equipment, not skill, drives the ratings that automated systems try to mimic. Third, subject matter is not tightly controlled; while categories such as “portraits” control the subject somewhat, it is possible that ratings are for the aesthetics of the subject (e.g. an attractive model) rather than the photograph. Fourth, the data includes ratings on an integer scale of 1 to 10, and numerical scales are difficult for human judges to maintain absolutely; there may be a bias to compare the present photograph to the one previously seen. Binary ratings, such as like/dislike votes, are an alternative that requires less memory and effort.

Given these confounding factors, we critique and propose alternatives to current techniques based on our previous work.⁴⁻⁶

2. RELATED WORK

Various aspects of aesthetics, including colorfulness, contrast, sharpness, and composition, have been considered in the literature. Savakis *et al.*⁷ found that composition is the most important attribute in user decisions over which pictures deserve emphasis in a photo album, dominating both colorfulness and sharpness, two traditional measures of image quality, by at least a factor of 3. Ke *et al.*⁸ explore attributes that distinguish between experts and amateurs, and argue that high level semantic features such as simplicity, which they measure using the spatial distribution of edges, are more important than the bag of low-level features approach of Tong *et al.*⁹

Studies of aesthetics by Datta *et al.*¹ describe ACQUINE, which uses a machine learning approach to provide numerical ratings of aesthetic appeal of photographs. ACQUINE relies on over 50 features extracted from each image, with a significant number of those features obtained after transforming into HSV color coordinates. ACQUINE¹ was available for public testing by means of a website that accepted image uploads and gave corresponding ratings. Marchesotti *et al.*¹⁰ describe a system that outperforms,¹ though it is not publicly available for testing. Instead of using predefined features, Lu *et al.*¹¹ use deep learning techniques to derive features and apply them to the AVA database.²

Sachs *et al.*⁴ introduced a database of photographs designed to test the effect of photographic skill on composition, and found that, on this database, ACQUINE’s ratings had little correlation with those given by a panel of human judges. In a followup study,⁵ the same database is used to exploring whether enhancing the image in color, contrast, or sharpness, generally improves ACQUINE ratings.

In Sachs *et al.*,⁴ the rating given by human judges to each image is on a 10-point scale. While this scale provides the opportunity for differentiation between similar images, there is significant cognitive load in choosing where exactly in the scale to position an image. For similar reasons, large providers of content such as YouTube have shifted from a 5-star rating system to the binary like/dislike rating system. One possible explanation is that histogram of 5-star ratings given by millions of viewers peaks at both extremes of 1 or 5, suggesting that people rarely use the middle ground in assessing videos.

Other researchers have explored if there is any relation between N -point rating scale and binary scales. Cosley *et al.*¹² check if there is any consistency among users performing a 5-star rating, and repeating those ratings using a simple thumbs up/down rating, on the MovieLens database. They find a significant correlation of 0.59 between the 5-star ratings and binary ratings. Agrawal *et al.*¹³ used a crowd-sourcing approach, as also used in some previous studies,¹⁴ to explore whether binary ratings were consistent with 10 point ratings of Sachs *et al.*⁴ They recruited 168 workers using Amazon Mechanical Turk (AMT) to serve as judges. Workers were allowed to pause and resume, and also to revise their ratings if necessary before a final submission. There were no restrictions on the imaging knowledge of the workers, and hence AMT data may be more representative of

the “average person” than the professional photographers and researchers used by Sachs *et al.*⁴ To check for frivolous voting, Agrawal *et al.* embedded two duplicate images in of the 7 categories used in the previous study, raising the total number of images from 221 to 235.

3. DATABASE

Figure 1: Example images from each of 7 different scenarios used in the study by Sachs *et al.*⁴ are shown. Two examples of each scenario are shown. From left to right, top to bottom, the scenarios are: walkway, still life, building corner, fountain, architectural staircase, portrait, street crossing.

The database used in Sachs *et al.*⁴ controls for scene, camera, subject skill level, and expertise of judges. They recruited 33 unpaid volunteers, including university students (some of whom were undergraduate photography majors), university staff, as well as professional photographers. The subjects identified their own skill level on

Table 1: Distribution of the 33 photographers in the CSE database by skill level⁴

Novices	Amateurs	Enthusiasts	Experts
4	15	12	2

a questionnaire. Through their answers, subjects were assigned to one of the four categories (Novice, Amateur, Enthusiasts, and Experts) described above. Table 1 shows the number of subjects in each group. There are few novices, as perhaps could have been predicted from the nature of such an experiment and the fact that there was no payment. Furthermore, the number of true Experts is low, but within the Enthusiast class there were 3 photography majors each with at least 3 years of formal training. The photo majors arguably have skill somewhere in between the formally untrained Enthusiasts and the Experts.

In the study,⁴ subjects were asked to submit one photograph of each of the 7 scenes shown in Fig. 1. The scenes were chosen by two professional photographers to test photographic skill: (1) a still life of manmade objects; (2) a portrait of a person seated in a chair, deliberately backlit against a bright window, and whom the subjects were not allowed to ask to pose; (3) a two-story high indoor staircase; (4) an outdoor fountain; (5) a striped road crossing with traffic; (6) a covered walkway, often seen in Singapore; and (7) a corner of a building with reflecting glass on both sides. The subjects were allowed to take as many photos as they liked of each scene, but were required to select only one of each as their submission.

Equipment was controlled in this study. As described in Sachs *et al.*, subjects were given the same model of consumer point-and-shoot camera with 7 megapixel resolution. The cameras were set in full automatic mode, to ensure that picture quality was not affected by the mode chosen. For each of the 7 scenes, the subjects were allowed to move around within a demarcated area (indicated on the ground by masking tape), selected by two professional photographers. The objectives of marking the area were both to allow freedom of composition and also to provide comparable photos. To further ensure comparable photos, zoom on the cameras was disabled by taping over the zoom button. The experiment was carried out over two separate days, between the hours of 11am-2pm to ensure similar light conditions. Of the $33 \times 7 = 231$ photographs that were submitted, 10 were eliminated from consideration for violating announced rules; for example, photos were removed if the subjects used zoom (despite the zoom button being taped over), asked the portrait model to pose, or focused on bystanders rather than the assigned scenes. Hence, the remaining 221 photos are referred to as the Controlled Scene and Equipment (CSE) database.

The CSE photos were judged in a double-blind manner by a panel of 8 judges, who were a separate group from the photographers. The panel included 3 professional photographers, and 5 imaging science and technology researchers from a leading digital image sensor company. The researchers can be considered to have similar characteristics to the Enthusiasts in the subject group. The judging was done online by selecting a rating on a 10 point scale. Since the process of judging all 221 images may produce fatigue, judges were allowed to save their ratings, log in and out in order to break up the rating process according to their convenience, and also to revise their ratings if necessary. The judges were instructed to rate the photograph only on the composition, rather than on focus, lighting or color. They viewed images of resolution 600×450 .

Sachs *et al.* compared the judge rating to those given by ACQUINE.¹ Figure 3 shows a distribution of the scores, with the peak occurring at the bin centered around 15. The distribution looks visually similar to the overall distribution of scores given by ACQUINE to the more than 240,000 photos uploaded since its inception, the data for which is obtained from the website*. This suggests that the CSE data is not unusual when compared to the types of pictures that ACQUINE has seen, and may in part have learned from, in the past.

Figure 4 shows a histogram of the percentage of likes on the CSE dataset from the AMT workers in the study by Agrawal *et al.*⁶ The majority of the histogram is above 50%, peaking at 85%, which shows that workers tended to like more than dislike images.

*<http://acquine.alipr.com/stat.php>

Figure 2: Two screenshots of the web-based rating process used by the judges for collecting ratings from the panel. The left hand image shows a thumbnail view of some of the photos taken of the “still life” in the CSE database, and the right hand side shows the 10-star rating system.

Figure 3: Histogram of scores given by ACQUINE historically to 240,000 photos, compared to the histogram of scores given to the CSE database.

4. ANALYSIS OF THE DATA

As a nonparametric method for analyzing scores, Sachs *et al.* used the Spearman ρ rank correlation [15, pg 206], defined by computed by converting raw scores X_i, Y_i , into ranks x_i, y_i , and computing the correlation on ranks. For example, raw scores 1.0, 1.1, 3.4, 10.0 are converted to ranks 1, 2, 3, and 4. Unlike the Pearson correlation, the Spearman coefficient ρ is nonparametric and indicates the degree to which one variable is a monotonic (not necessarily linear) function of another. We found no significant difference in our results between the Spearman and the Kendall τ correlation, another widely used nonparametric measure.

We review the results of Sachs *et al.*, who evaluated consistency between 9 judges: the 8 human judges and ACQUINE as follows. Table 2 summarizes their results, and shows that each group is consistent to the same degree among themselves, but there is less consistency between groups. None of the correlations were found to be strong (i.e., $\rho > 0.7$), indicating that the judges are fairly independent. In fact, the correlations between human judges and ACQUINE are weak (not more than 0.27). Figure 5 shows scatter plots of the mean rating of the human judges against the rating from ACQUINE. Sachs *et al.* noted that ACQUINE takes into account many factors in assessing photo aesthetics, including, but not limited to, composition, whereas the human judges were instructed to pay attention only to composition.

Table 2: Minimum \ Maximum Spearman correlation in judge groups⁴

Professionals (P)	Researchers (R)	Mixed (P vs R)	P vs Acquine (A)	R vs A
0.25\0.43	0.21\0.46	0.02\0.38	0.01\0.22	0.02\0.27

Figure 4: Histogram of the percentage of likes given to the CSE images by the AMT workers. The peak around 80-90% shows that the voters tend to agree when there is an image that they like.

Figure 5: The leftmost scatter plot compares all human judges (sum of scores) to ACQUINE (A), the middle plot compares professionals (P) to A, and the rightmost plot compares researchers (R) to A. The scatter plots confirm visually that there is low correlation between human judges and the automated rating system used by ACQUINE. Plots are from Sachs *et al.*⁴

Although the groups had little correlation with each other, remarkably they found similar results for best images. Figure 6 shows that, for each group, the symmetry of the building corner was appealing. Interestingly, each group chose a slightly different picture of the corner as their best one.

Agrawal *et al.*¹³ checked for relations among the like percentage, ACQUINE scores, and the 10-point rating data from the previous study,⁴ using the Spearman rank correlation to allow for monotonic relationships. They found a weak correlation (0.23) between the 10 point ratings from the human judges and ACQUINE, and also between the like percentage of the AMT workers and ACQUINE (0.19). Hence ACQUINE is not able to predict the aesthetic preferences of either of those two different groups of human raters. However, there is a strong correlation of 0.59 between the 10-point rating and the binary like/dislike ratings in the form of like percentage. Interestingly, this correlation is the *same* value measured in¹² between binary ratings and a 5-point rating. The agreement is striking given that¹² measured the correlation on a database of movie ratings, which is completely different than the CSE database. The high correlation of 0.59 indicates that for rating aesthetics, the cognitively difficult task of giving a rating on a N -point scale may be replaced with the much easier task of a giving a binary like/dislike rating. Furthermore, it is notable that the strong agreement indicated by the 0.59 correlation is between the group of highly-experienced judges in,⁴ and the AMT workers, who had no verifiable experience in rating images. That agreement suggests that both groups were responding to the aesthetic content of a photograph in a similar way, and that experience does not play a major role in rating aesthetics.

4.1 The effect of color

Many top rated photographs on the ACQUINE website are in black and white. To study the effect of color, Kakarala *et al.*⁵ converted each of the pictures to gray using the “rgb2gray” function of MATLAB, which uses the formula

$$\text{Gray} = 0.299 * R + 0.587 * G + 0.114 * B. \quad (1)$$

Figure 6: From left to right, the top row are respectively the worst (lowest rated) pictures as selected by the professionals (P), researchers (R), and ACQUINE (A) in the study.⁴ The bottom row is the best (highest rated) for the respective groups. Remarkably, all three groups chose the building corner with its inherent symmetry as the best.

Figure 7: Scatter plot⁵ of ACQUINE scores given to color images vs those given to gray converted images (green dots) for the CSE dataset. The red dots show the scatter plot for the same input images, but which are automatically adjusted for contrast and tone using Picasa's tools as described in the text. Scatter is larger for gray conversion (green) than for auto-contrast adjustment (red), indicating that the former has much greater effect in changing scores than the latter.

As Figure 7 demonstrates, there is a low correlation of 0.3 between the ACQUINE scores of color and gray-converted images. On roughly 57% of the CSE image set, the color image is preferred to the gray-converted image, and the average change in score is roughly 21 points out of 100. This result suggests that ACQUINE is very sensitive to color in computing its scores, a fact supported by noting that several of the 56 features that it uses are based on the hue (H) coordinate of the image pixel values in *HSV* space.

The above results raise the question of whether the specific method (1) of converting from color to gray is a significant factor. There are arguably better ways of gray conversion, such as lightness conversion:

$$\text{Gray} = \frac{1}{2} \{ \text{Max} (R, G, B) + \text{Min} (R, G, B) \} \quad (2)$$

Lightness conversion ensures that in regions where red or blue dominate, but green is small, that it is still possible to obtain a high gray value; this would occur, for example, in regions of blue sky or neon red lights. Besides lightness, another and more sophisticated method of grayscale conversion is “decolorize”.¹⁶ Decolorize analyzes

Figure 8: Examples of gray scale conversion of a color image using three different methods, clockwise from top right: rgb to gray, lightness, and decolorization.⁵

color differences between pairs of pixels, chosen at random, and derives a “chromatic axis” for the image on which to project the chromatic content, which is then subsequently added to the achromatic content.

Kakarala *et al.* compared the lightness and decolorization methods to the default method (1) in terms of influencing ACQUINE scores on the CSE database. Neither had a high correlation (< 0.31) with the score given to the original color image. Lightness’ ACQUINE scores are highly correlated with the score given to (1)—correlation is 0.96—while decolorize is less correlated with (1) at 0.88. Figures 8 shows examples of gray scale conversion methods.

4.2 The effect of automatic photo adjustment

Many highly rated photos are enhanced, with both local and global adjustment. Kakarala *et al.* explored whether automated adjustment, using the “I’m feeling lucky” © enhancement feature of Google’s Picasa software, improves ACQUINE scores. This enhancement requires no user intervention, and typically adjusts both the contrast and color using proprietary methods. As Figure 7 shows, there is a high correlation of 0.87 between the original scores and the automatically-enhanced image scores. On average, there is a slight change of 6.3 points between the two scores, but only about 52% of the enhanced images are rated more highly than the originals, a fact visually confirmed by the equal distribution of the scatter about the diagonal axis in Figure 7.

4.3 The effect of sharpness

It is not clear whether global sharpness adds to a photograph’s aesthetic appeal, while local sharpness may be appealing if it is part of a “bokeh” effect. In a study of demosaicing algorithm performance, Longere *et al.*¹⁷ found that users preferred a Bayesian demosaicing algorithm that sharpened the image. They also found that while blurry images benefit from sharpening, the perceived image quality degrades once the sharpening goes past an optimum point. The optimum sharpening amount according to users varies in a non-trivial way with the image content.

Figure 9: The top row shows the 6 most-liked images for the AMT workers, and the bottom row shows the 6 most disliked images.⁶

Kakarala *et al.* explored whether that is also true with ACQUINE ratings, by using Picasa in batch mode to sharpen each of CSE database images by varying amounts. They observed an overall decrease in scores as the amount of sharpening increased: though some images received higher scores after sharpening, most received lower scores. The results suggest that ACQUINE has an optimum sharpening level that is image content dependent, but that it generally found the images in the CSE set sufficiently sharp to begin with.

4.4 Global statistics and preference

Agrawal *et al.*⁶ checked if the percentage of likes provided by AMT workers in their study is in any way related to global statistics measuring contrast, colorfulness, and the rule of thirds in composition. Those statistics are calculated as follows. Contrast C is defined as the RMS value in luminance Y channel, defined for a $M \times N$ image with mean luminance $\bar{Y}$ as

$$C = \sqrt{\frac{1}{MN} \sum_{i=0}^{M-1} \sum_{j=0}^{N-1} (Y_{ij} - \bar{Y})^2}. \quad (3)$$

Colorfulness K is defined as in¹ by the Earth Mover's Distance (EMD) between the uniform distribution D_u on the RGB color cube with 4 bins on each axis (64 in all), and the image's color distribution D , with the distance between bins $d(a, b)$ defined in LUV space as $\|LUV(a) - LUV(b)\|$, i.e.,

$$K = \text{EMD}\{D_u, D, d\}. \quad (4)$$

The third measure attempts to quantify composition of hues according to the “rule of thirds”, in which features of interest are placed along lines where the image is divided into thirds. This measure is defined as in¹ for the hue component H of the image in HSV coordinates as

$$T = \sum_{i=M/3}^{2M/3} \sum_{j=N/3}^{2N/3} H_{ij}. \quad (5)$$

Agrawal *et al.*⁶ found that colorfulness K has a low but statistically-significant correlation of 0.26 with like percentage, but the other remaining correlations, in particular those involving contrast C and the hue distribution measure T , were such that they could not be distinguished from zero. Hence global statistics have little predictive value for the human ratings in our study.

5. SUMMARY

Current work on developing automated ratings of photographic aesthetics uses large scale databases, but the data derived from those suffers from various confounding effects. This paper reviews works involving the CSE

database for understanding both human and automated ratings of photographic aesthetics. Unlike the large databases in use today, the CSE database controls for scene and equipment. Hence, CSE is suitable for exploring whether skill in composition is evident in a photographic portfolio.

ACKNOWLEDGMENTS

The support of the Singapore Ministry of Education under grant MOE-T2-1-010 is gratefully acknowledged. The authors thank T. Sachs, S. Castleman, D. Rajan and M. Constable for helpful discussions.

REFERENCES

- [1] Datta, R. and Wang, J. Z., “ACQUINE: aesthetic quality inference engine - real-time automatic rating of photo aesthetics,” in *[Multimedia Information Retrieval]*, 421–424 (2010).
- [2] Murray, N., Marchesotti, L., and Perronnin, F., “AVA: A large-scale database for aesthetic visual analysis,” in *[2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, June 16-21, 2012]*, 2408–2415 (2012).
- [3] Muller, H., Clough, P., Deselaers, T., and Caputo, B., eds., *[ImageCLEF: Experimental evaluation in visual information retrieval]*, Springer, Berlin (2010).
- [4] Sachs, T. S., Kakarala, R., Castleman, S. L., and Rajan, D., “A data-driven approach to understanding skill in photographic composition,” in *[ACCV Workshops (2)]*, 112–121 (2010).
- [5] Kakarala, R., Sachs, T. S., and Premachandran, V., “Comparing automated and human ratings of photographic aesthetics,” in *[19th Color and Imaging Conference, CIC 2011]*, 217–222 (2011).
- [6] Agrawal, A., Premachandran, V., and Kakarala, R., “Rating image aesthetics using a crowd sourcing approach,” in *[Image and Video Technology - PSIVT 2013 Workshops]*, 24–32 (2013).
- [7] Savakis, A., Etz, S., and Loui, A., “Evaluation of image appeal in consumer photography,” in *[SPIE Human Vision and Electronic Imaging V]*, (2000).
- [8] Ke, Y., Tang, X., and Jing, F., “The design of high-level features for photo quality assessment,” in *[IEEE Conference on Computer Vision and Pattern Recognition]*, 419–426 (2006).
- [9] Tong, H., Li, M., Zhang, H.-J., He, J., and Zhang, C., “Classification of digital photos taken by photographers or home users,” in *[Proceedings of Pacific Rim Conference on Multimedia]*, 198–205, Springer (2004).
- [10] Marchesotti, L., Perronnin, F., Larlus, D., and Csurka, G., “Assessing the aesthetic quality of photographs using generic image descriptors,” in *[International Conference on Computer Vision]*, 1784–1791 (2011).
- [11] Lu, X., Lin, Z., Jin, H., Yang, J., and Wang, J. Z., “RAPID: rating pictorial aesthetics using deep learning,” in *[Proceedings of the ACM International Conference on Multimedia, MM '14, Orlando, FL, USA, November 03 - 07, 2014]*, 457–466 (2014).
- [12] Cosley, D., Lam, S. K., Albert, I., Konstan, J. A., and Riedl, J., “Is seeing believing? how recommender system interfaces affect users’ opinions,” in *[CHI]*, 585–592 (2003).
- [13] Agrawal, A., Premachandran, V., Somavarapu, R., and Kakarala, R., “Can relative skill be determined from a photographic portfolio?,” in *[SPIE Proceedings Vol. 8651: Human vision and electronic imaging XVIII]*, (2013).
- [14] Ribeiro, F., Florencio, D., and Nascimento, V., “Crowdsourcing subjective image quality evaluation,” in *[Image Processing (ICIP), 2011 18th IEEE International Conference on]*, 3097–3100 (Sept 2011).
- [15] Maritz, J. S., *[Distribution-free statistical methods]*, Chapman and Hall (1991).
- [16] Grundland, M. and Dodgson, N. A., “Decolorize: Fast, contrast enhancing, color to grayscale conversion,” *Pattern Recognition* **40**(11), 2891–2896 (2007).
- [17] Longere, P., Zhang, X., Delahunt, P., and Brainard, D., “Perceptual assessment of demosaicing algorithm performance,” *Proceedings of the IEEE* **90**, 123–132 (Jan 2002).