Comparing automated and human ratings of photographic aesthetics

Ramakrishna Kakarala*, Todd S. Sachs, Vittal Premachandran
***School of Computer Engineering, Nanyang Technological University, Singapore**

Abstract

ACQUINE has made automated rating of photographic aesthetics available on the web. We compare the influence of various factors, such as color, sharpness, and contrast, on the ratings that it gives. We also compare results with those given by human judges.

Introduction

Photographic aesthetics has long been considered the domain of humans, and far outside the scope of computer vision other than for easily quantifiable aspects such as sharpness, noise, and contrast. As even people disagree on the merits of a single photograph, it is not possible for computer vision algorithms to discern the aesthetic appeal in a reliable manner. However, given that there are ratings of photographs available on photographic enthusiast websites, it becomes possible to mine that information to build image analysis algorithms that can rate a photograph with good correlation to human ratings. Of several approaches to automated aesthetic rating that have been tried [3][4], the most successful has been ACQUINE [5]. In this paper, which builds on an earlier work of ours [1], we describe experiments that measure the relationship between ACQUINE and human ratings, and also show the influence on ACQUINE scores through common image enhancements on color, sharpness, and contrast.

A review of the literature shows that various aspects of aesthetics, including colorfulness, sharpness, and composition, have been considered. Savakis, Etz & Loui [2] determined experimentally that the most important attribute to deciding which pictures deserve emphasis in a photo album is composition. Specifically, their study found that composition is more important by at least a factor of 3 than either colourfulness or sharpness, two traditional measures of image quality. It is important to note that the photos used in their paper are from ordinary consumers, rather than from professionals. In contrast, Tong *et al.* [3] consider both amateur and professional photographs, and attempt to classify the groups using computer vision techniques. Their methods rely on quantitative measures of sharpness, colourfulness, contrast, and saliency. Though their classifier correlates well (coefficient of 0.85) with rankings given by a group of 16 human observers, they do not consider composition as an attribute, nor possible equipment differences between professionals and amateurs. Ke, Tang, & Jing [4] also 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.* [3]. Therefore, Ke *et al.* use photos obtained from the website `dpchallenge.net` for testing, and find that the sharpness attribute is the most discriminative in distinguishing between the top 10% most highly-rated photographs from the bottom 10% in their test set. The simplicity measure of [4] measures composition to some extent, though, of course, composition means much more than simplicity. Composition as an attribute is also considered by Luo & Tang [7], who, like previous researchers, develop methods for classifying expert and amateur photographs, but provide the novel step of extracting subjects from the background using sharpness as a cue. Specifically, they measure composition geometry by distance of the subject centroid to the "rule-of- thirds" points[1], in addition to using texture, and familiarity (measured by similarity to a group of standard images). Their method outperforms that of Ke *et al.* [4] on the same data set obtained from `dpchallenge.net`.

The computer vision literature pays considerable attention to sharpness and colourfulness as attributes of photograph aesthetics, perhaps because those attributes are quantifiable. However, the study of Savakis *et al.* [2] shows that composition is far more important. Obviously, an image can be appealing even without being sharp or colourful; for example, the black-and-white photographs of the master photographer Henri Cartier-Bresson are often slightly defocused and lack contrast, but are nevertheless powerful due to their composition.

The most comprehensive computer vision study of aesthetics to date, by Datta, Joshi, Li, & Wang [5], uses a machine learning approach to provide numerical ratings of aesthetic appeal of photographs. Like the previously-mentioned studies, Datta *et al.* use the attributes of colourfulness, sharpness (depth of field), and also include consideration for composition by using the rule of thirds. Their system relies on 56 features extracted from each image, with a significant number of those features obtained after transforming into *HSV* color coordinates. Datta *et al.* compare their system's ratings with those given by human observers on `photo.net` and find en error variance of 0.69 on a 7.0 scale.

Most importantly for our paper, Datta & Wang [6] make their rating method, named ACQUINE (Aesthetic Quality Inference Engine) available online on the site `acquine.alipr.com`. For every picture uploaded to that site, ACQUINE returns a score between 0 to 100, with higher scores indicating greater aesthetic appeal. A histogram of scores of over $240,000$ pictures uploaded as of the writing of this paper is unimodal and asymmetrical, with a peak for the bin of scores between $20-30$.

While there is no single model of automated aesthetic analysis, ACQUINE is perhaps the most developed, tested, and accessible model at the present time. Its photo ranking algorithm is,

[1] An adage of photographic composition is to place the subject at one-third or two-third the height or width to draw the user's attention into the scene.

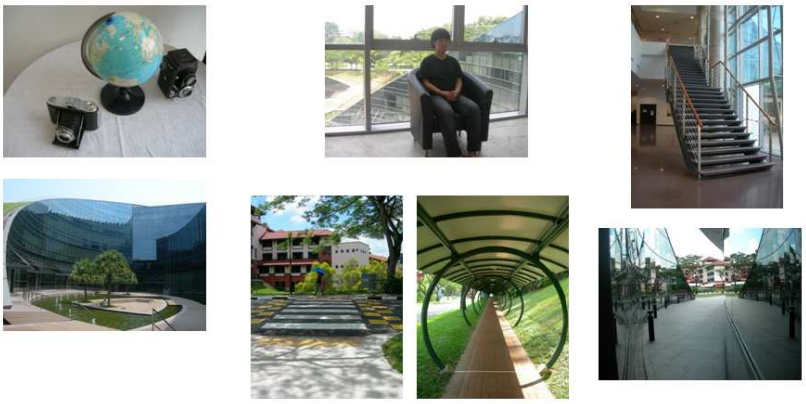

Figure 1. *Example photos of the seven scenarios are shown. From left to right, top to bottom: still life of objects; portrait of person against window; indoor staircase; outdoor fountain; road crossing; covered walkway; buildings mirrored-glass corner. Each of the photos shown is taken by a different photographer. See [1] for more details.*

in some ways, like the page ranking algorithms used by search engines. Though the principles have been published by Datta *et al.*, ACQUINE's algorithm uses numerous parameters obtained through a learning process and which may evolve over time as more feedback is obtained. Therefore, it is not possible to answer simple questions such as "how much will the ACQUINE score change if the sharpness of a photograph is increased?" It is the purpose of this paper to analyze the key factors in such automated ratings, and to compare those ratings to those obtained by human judges.

Experimental methods

Our starting point is the database obtained as described in an earlier paper [1], in which a total of 221 photos are collected from 33 unpaid subjects who used identical point-and-shoot cameras to take photos in each of 7 different scenarios. The scenarios are chosen by two professional photographers[2] to represent a variety of challenges in composition. To make sure photos taken by different volunteers are comparable, the following steps are taken: (1) the vantage points of the volunteers were limited and indicated by masking tape placed on the floor; (2) the cameras were set to fully-automatic mode, and the subjects were instructed not to change to other modes; (3) the camera zoom function was disabled; (4) for the portrait scenario, the model could not be asked to pose. Of the $7 \times 33 = 231$ photos taken, 10 are removed for violating one or more of those rules, leaving 221 for the study. Further details of the experiment are provided in the earlier paper [1]. Example photographs of each of the scenarios are shown in Figure 1.

In order to judge how typical the photos in our study group are compared to those which have been rated by ACQUINE, we uploaded each of the 221 photos to get ACQUINE scores. Figure 2 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

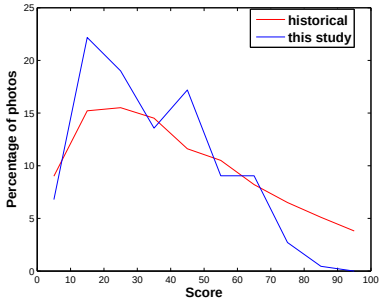

Figure 2. *Histogram of scores given by ACQUINE historically to* $240,000$ *photos, compared to the histogram of scores given to the* 221 *photos in our database. We see that our sample is not atypical.*

for which is obtained from the website[3]. This suggests that our sample is not atypical when compared to the types of pictures that the system has seen, and may in part have learned from, in the past.

In our previous study, we compared the scores given by ACQUINE to those given by 8 human judges and found a low correlation between the scores (no more than 0.27 between any of the judges and ACQUINE). The judges were asked to make their judgement based on composition, whereas ACQUINE considers a number of factors including color, sharpness, contrast, and rule-of-thirds composition. Therefore, we attempt in this study to find out the effect of color, contrast, and sharpness as individual factors in the ACQUINE rating.

The effect of color

An examination of the pictures from the 5 photographers most highly rated by ACQUINE in its history of operation (covering more than 240,000 pictures) shows that roughly a fifth are in black and white. Hence, it seems natural to wonder whether ACQUINE would prefer the pictures in our study equally in color or

[2] We thank Prof. S. Castleman and Dr. Shahidul Alam for their help

[3] http://acquine.alipr.com/stat.php

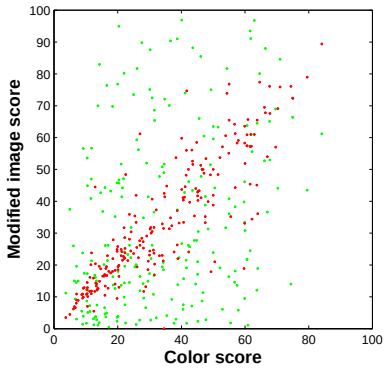

Figure 3. *Scatter plot of ACQUINE scores given to color images vs those given to gray converted images (green dots). The red dots show the scatter plot for the same input images, but which are automatically adjusted for contrast and tone using Picasa's tools as described in the text. We see that the 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.*

in black-and-white. To study that, we converted each of the pictures to gray using the "rgb2gray" function of MATLAB, which uses the formula

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

We uploaded each of the gray pictures using a Python script to ACQUINE and obtained the scores. As Figure 3 demonstrates, there is a low correlation of 0.3 between the scores of color and gray-converted images. On roughly 57% of our 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. One simple method, used as an option in the GIMP software[4], is lightness conversion as described below

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

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 rlights. Besides lightness, another and more sophisticated method of grayscale conversion is "decolorize", proposed by Grundland & Dodgson [8]. Decolorize analyzes 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. We used the authors' MATLAB implementation of decolorize in our experiments.

[4] `www.gimp.org`

We compared the lightness and decolorization methods to the default method (1) in terms of influencing ACQUINE scores, using the same 221 images. 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. Lightness received the highest scores overall, outscoring (1) approximately 61% of the time, and outscoring decolorize 53% of the time. Figures 4 and 5 respectively show examples of cases where lightness received the highest score, and similarly for decolorize.

The effect of automatic photo adjustment

The most highly rated pictures by ACQUINE in its history of operation, which are shown on its website, show excellent contrast and tonal range. Typically, such images are obtained after postprocessing by adjusting levels both locally and globally, in a digital version of what in film photography is referred to "dodge and burn". We explored whether automated adjustment produces a benefit that is measurable by ACQUINE using the following procedure. We used the "I'm feeling lucky" © enhancement feature of Google's Picasa software and applied it in batch-mode to all of the 221 pictures in our study group. This enhancement requires no user intervention, and typically adjusts both the contrast and color using proprietary methods. We then uploaded the enhanced pictures and obtained the corresponding scores. As Figure 3 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 3.

The effect of sharpness

There is as yet no known rule for determining when a photograph is sharp enough, or whether increasing its sharpness through post-processing adds to its aesthetic appeal. Sharpness is moreover both a global attribute and a local one, where local depth-of-field effects such as "bokeh" may be employed to improve the presentation of a subject. In a study of demosaicing algorithm performance, Longere *et al.* [9] 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.

We explored whether that is also true with ACQUINE ratings in the following way. We used Picasa in batch mode to sharpen each of the 221 images in our study by varying amounts, controlled by adjusting the slider position in the sharpening menu. We uploaded every batch of sharpened photos and obtained the ACQUINE score, and compared the results to the scores of the original data. As expected, we observed an overall decrease in scores as the amount of sharpening increased: though some images received higher scores after sharpening, most received lower scores. Table 1 shows how the mean score and the fraction of images receiving lower scores varies with the amount of sharpening. The results suggest that ACQUINE has an optimum sharpening

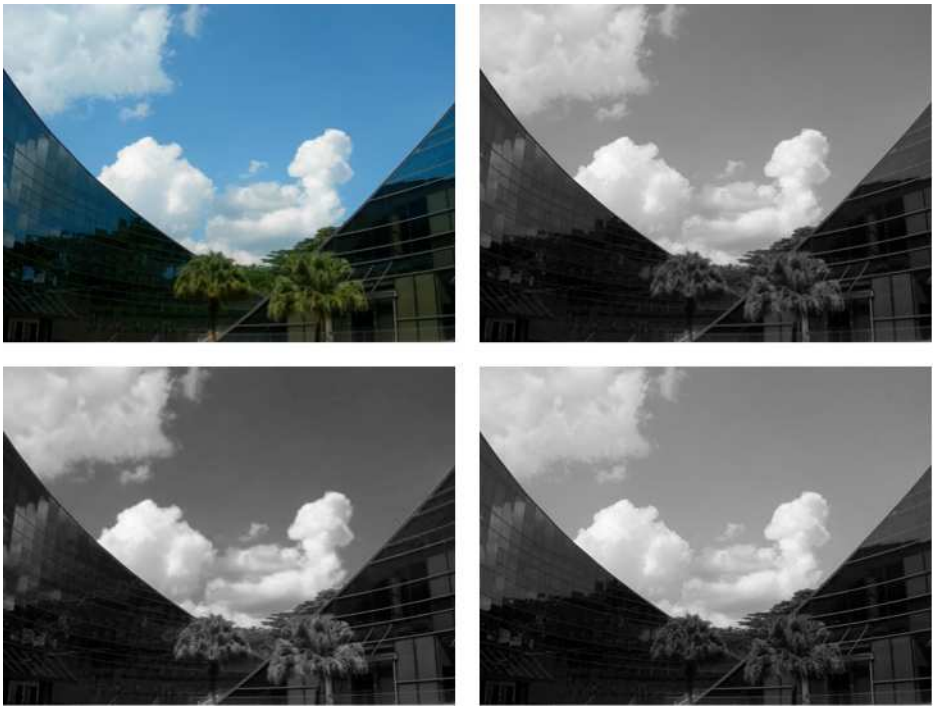

Figure 4. *Example of a gray scale conversion of a color image using three different methods, clockwise from top right: rgb to gray, lightness, and decolorization. Of the* 3 *methods, ACQUINE gives the highest score of* 60 *to decolorization (lower left), compared to* 31 *to rgb to gray (upper right) and* 15 *to lightness (lower left). Note the lighter tone of the green palm fronds in the decolorized image.*

Table 1: Effect of sharpening on ACQUINE score, showing that not only the mean score decreases with the amount of sharpening, but the percent of photos whose score decreases over the original also grows.

Sharpening amount	Mean scores	Percent of images decreasing score
0%	32.9	-
25%	31.3	57%
50%	27.6	59%
75%	24.4	66%
100%	20.8	77%

level that is image content dependent, but that it generally found the images in our set sufficiently sharp to begin with.

User studies

In order to compare ACQUINE ratings with those given by human judges, we designed a two-alternative forced choice experiment where observers chose the preferred image in pair presented as in Figure 6. The format of the experiment is as follows. For the group of approximately 30 photos in each of the 7 scenarios shown in Figure 1, observers viewed pairs of photos and selected which of the pair they preferred. They then repeated the selection on the "winners" of the previous round of pairwise comparisons, and continued the process in a pyramid fashion until the overall winner is found. This "bracket" selection process is then repeated on the next scenario. A score is given additively to each of the photos according to how many of the rounds of comparison it survived, so that the winners of the first round received 1, the next round received $1+2=3$, and so on. The weighting is chosen so that each higher round has twice the score of the lower one. Six subjects participated in the experiment, which took approximately 30 minutes to complete. We obtained a corresponding score for ACQUINE in the experiment by repeating the comparisons and choosing the image whose ACQUINE score is higher.

Table 2: Mean \ Maximum Spearman correlation in judge groups

R	R vs NR	H vs A
0.21\0.32	0.14\0.28	0.12\0.24

We examined the correlation between scores given by the six human judges and ACQUINE using Spearman rank correlation, which allows testing for nonlinear relationships between scoring methods[11]. Of the six human subjects, four are image processing researchers, a group we denote R, and two are non-researchers, a group we denote NR. If we let $H = R \cup NR$ denote the human judges, and A denote ACQUINE, then Table 2 show the mean and maximum Spearman correlation between the groups.

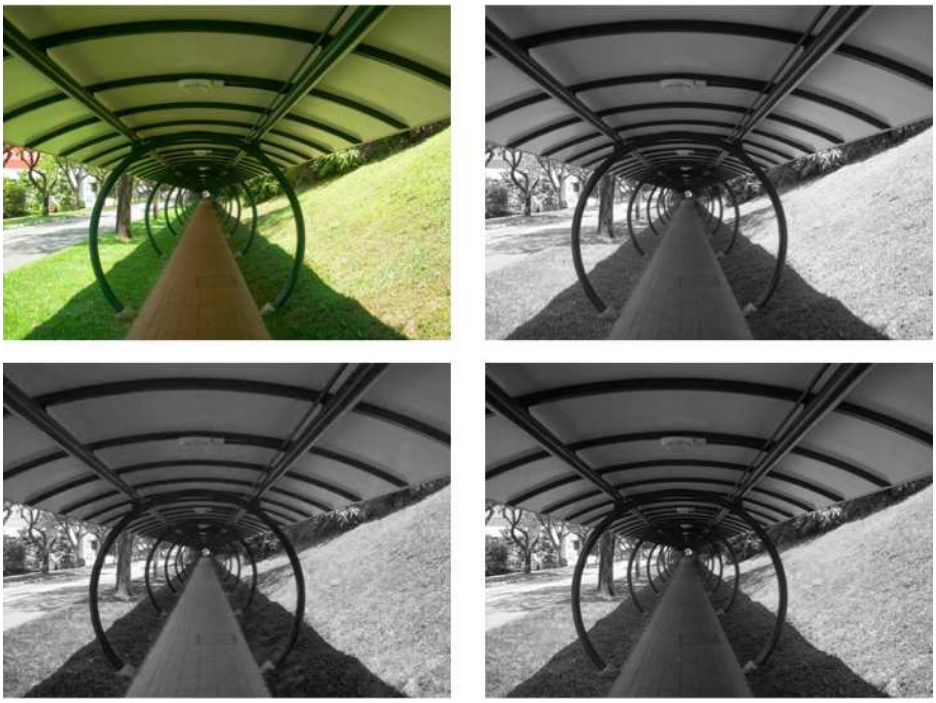

Figure 5. *Another example of a gray scale conversion of a color image using three different methods, clockwise from top right: rgb to gray, lightness, and decolorization. In this case, ACQUINE gives the highest score of* 59 *to the lightness conversion (bottom right) of this example, compared to* 24 *for the rgb to gray (upper right), and* 55 *for the decolorization (bottom left). Note the higher contrast of the grass, and the lighter tone of shadow in the lightness image.*

Figure 6. *Screenshot showing how observers observe and then choose one of two images. The process is repeated between winners of a given round until an overall best choice is obtained.*

Conclusions and future work

There is no doubting the value of an automated aesthetic rating system which correlates well with human rating. Though ACQUINE has made major progress in that regard, the problem is still far too complex to be settled by the current version. In this paper, we have explored the effect of various factors, such as color, sharpness, and contrast, on the rating given by ACQUINE. In future, we plan to repeat the user studies to assess whether color or gray-converted images are preferred, and compare the results to that of ACQUINE. We will also repeat the user studies with contrast enhancement and sharpness.

Acknowledgments

We thank the judges who volunteered their time to rate photographs: PK, VP, RH, RK, LRR, and SEK.

References

[1] Sachs, T., Kakarala R., Castleman S., Rajan D.: A data-driven approach to understanding skill in photographic composition. In: ACCV Workshop on Computational Photography and Aesthetics (2010), Queenstown, New Zealand.

[2] Savakis, A., Etz, S., Loui, A.: Evaluation of image appeal in consumer photography. In: SPIE Human Vision and Electronic Imaging V. (2000)

[3] Tong, H., Li, M., Zhang, H.J., He, J., Zhang, C.: Classification of digital photos taken by photographers or home users. In: Proceedings of Pacific Rim Conference on Multimedia, Springer (2004) pp 198–205

[4] Ke, Y., Tang, X., Jing, F.: The design of high-level features for photo quality assessment. In: CVPR (1), IEEE Computer Society (2006) pp 419–426

[5] Datta, R., Joshi, D., Li, J., Wang, J.Z.: Studying aesthetics in photographic images using computational approach. ECCV (2006) pp 288–301

[6] Datta, R., Wang, J.Z.: Acquine: aesthetic quality inference engine - real-time automatic rating of photo aesthetics. In Wang, J.Z., Boujemaa, N., Ramirez, N.O., Natsev, A., eds.: Multimedia Information Retrieval, ACM (2010) 421–424

[7] Luo, Y., Tang, X.: Photo and video quality evaluation: Focusing on the subject. In Forsyth, D.A., Torr, P.H.S., Zisserman, A., eds.: ECCV (3). Volume 5304 of Lecture Notes in Computer Science., Springer (2008) pp 386–399

[8] Grundland, M., Dodgson, N. A., Decolorize: fast, contrast enhancing, color to grayscale conversion. In: Pattern Recognition (2007), Vol. 40, pp 2891-2896.

[9] Longere, P., Zhang, X., Delahunt P.B., Brainard, D.H.: Perceptual assessment of demosaicing algorithm performance In: Proceedings of the IEEE (2002), Vol 90, pp 123-132.

[10] Zakia, R.: Perception and imaging: photography–a way of seeing. 3rd edn. Elsevier Science Ltd (2007)

[11] Maritz, J.S.: Distribution-free statistical methods. Chapman and Hall (1991)

[12] Snavely, N., Seitz, S.M., Szeliski, R.: Photo tourism: Exploring photo collections in 3d. In: SIGGRAPH Conference Proceedings, New York, NY, USA, ACM Press (2006) 835–846