Rating Image Aesthetics Using a Crowd Sourcing Approach

Abhishek Agrawal*, Vittal Premachandran, and Ramakrishna Kakarala

School of Computer Engineering, Nanyang Technological University
Singapore, 639798
{aagrawal,ramakrishna}@ntu.edu.sg,
vittalp@pmail.ntu.edu.sg

Abstract. Any system that is able to reliably measure the aesthetic appeal of photographs would be of considerable importance to the digital imaging industry. Researchers have built automated rating systems using machine learning techniques applied to features extracted from images. In this paper, we study the effectiveness of ACQUINE, a comprehensive and publicly available rating system, using data obtained from voters in a crowd sourced manner. We analyze the effect of voting using a simple binary like/dislike rating in comparison to a numerical 10 point scale. We also show that global measures of image quality, such as contrast or colorfulness, do not correlate well with human ratings. The role of composition in determining human rating of aesthetics is discussed.

Keywords: Image quality, Image color analysis.

1 Introduction

Quantifying the aesthetic appeal of a photograph, automatically, interests many researchers in the image processing and computer vision communities [1][2]. Here, the term aesthetic refers to any aspect of a photograph that is appealing to many people, as opposed to beauty, which is a subjective impression that varies from person to person. Aesthetics is closely linked to perception, and human perception being a high-level task makes it very difficult for low-level image processing algorithms to mimic. Another aspect that makes aesthetic analysis difficult is that there are no well defined rules to follow while making a prediction about a photograph. That said, there have been attempts at building automated systems such as ACQUINE (Aesthetic Quality Inference Engine) [1], which tries to predict the aesthetic appeal of a photograph. While ACQUINE performs well in matching ratings on the photo-sharing website photo.net, there is such variation among photographic styles that one website is not a sufficient test. Indeed, a previous study [3], which obtained a database of photographs under controlled conditions, found that the ratings produced by ACQUINE do not

* Corresponding author.

F. Huang and A. Sugimoto (Eds.): PSIVT 2013 Workshops, LNCS 8334, pp. 24–32, 2014.

correlate well with the ratings given by human judges. This shows that a lot more understanding is needed of what humans find appealing in a photograph. Furthermore, it raises questions as to how to efficiently collect rating data from a large number of human observers for use in training automated systems.

The goal of this paper is, therefore, to try and understand the image characteristics that humans look into while making an aesthetic judgement, as well to explore the effects of different rating systems used by human observers. To those ends, we extend the results of previous studies [3][4]. In this paper, we compare the 10-point rating system used in [3] to results of a new study using a simple binary like/dislike system. It is more difficult for human judges to provide equitable ratings of images on 10-point scale, than it is for them to give simple binary like/dislike ratings. Hence, in a study using a larger number of judges, one would rather use the binary scale; however, the effectiveness of the binary scale has not been established for aesthetic ratings, nor the relation between binary and numeric ratings been determined for image aesthetics. We study those issues in this paper, making use of a database of photographs first described in [3] [4] to collect the human ratings using a crowd sourcing method. We compare the human ratings to ACQUINE, and examine whether global image properties such as contrast and colorfulness are useful predictors of those ratings.

2 Related Work

A review of the literature shows that various aspects of aesthetics, including colorful-ness, contrast, sharpness, and composition, have been considered. Savakis et al. [5] determined experimentally that the most important attribute to deciding which pic-tures 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 colorfulness or sharpness, two traditional measures of image quality. Ke et al. [6] 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. [7].

Studies of aesthetics by Datta et al. [1][8][9] describe ACQUINE, which uses a machine learning approach to provide numerical ratings of aesthetic appeal of photographs. ACQUINE relies on 56 features extracted from each image, with a significant number of those features obtained after transforming into HSV color coordinates. Unlike other studies, [1] 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. Sachs et al. [3] introduced a database of photographs designed to test the effect of photographic skill on composition, and found that ACQUINEs ratings had little correlation with those given by a panel of human judges. In [4], the same database is used and combined with various levels of image enhancement in color, contrast, and sharpness, with the purpose of exploring whether enhancing the image (using, for example, Google Picasas

Im feeling lucky automated enhancement) generally improves ACQUINE ratings. While ACQUINE ratings change considerably as a result of those enhancements, [4] found there was no enhancement that consistently improved the ratings.

In [3], 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 that reason, major companies such as Youtube have shifted from a 5-star rating system to the binary like/dislike rating system. Though no published study has analyzed the affect of the rating system on Youtube, the informal explanation is on the 5-star scale ratings peaked around 5 or around 1. This suggests that people find it difficult to quantify the likeability factor of videos. They are far more comfortable making a simple decision such as like/dislike without having to undergo the cognitive load that providing a rating on a range of values causes. With that result in mind, we describe in this paper a new set of data that uses binary like/dislike ratings on the same database as in [3].

Fig. 1. Sample images from each of 7 different scenarios used in the aesthetics study are shown. Clockwise from top left, the scenarios are Still Life, Building Corner, Fountain, Architectural Staircase, Portrait, Zebra Crossing and Open or Free Shot. In the database used in the study, 33 photographers provided one picture from each of the scenarios.

Other researchers have explored if there is any relation between N-point rating scale and binary scales. Bargagliotti et al. [11] study the mathematical relationship between a binary rating system and a N-point rating system, and show that

the two systems can in theory give consistent results if N is odd with N - 1 not divisible by 4. In an experimental study, 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. Their study is performed on movie ratings in the MovieLens database. The find a strong correlation of 0:59 between the 5-star ratings and binary ratings, a figure we come back to in Section 4, when discussing the results of our study of binary vs 10-point scales. In comparison to previous work, what is new in our paper is the exploration of rating systems for image aesthetics, and an analysis of effectiveness of global image measures.

3 Experimental Setup

The most extensive aesthetics database available is AVA [2]. However, it does not provide control over the shooting scenario, camera model and post processing, and moreover has far too many images to study the effect of rating scales. For those reasons, we use the database from [3]. This database consists of 221 images taken from 7 different scenarios (Fig. 1). The photographs were taken by 33 participants who used identical point-and-shoot cameras. All the cameras were set to automatic mode. Also, the participants were restricted to take the photographs only from a specified area (marked off with masking tape) around the scenarios. The shooting areas, however, were not extremely restrictive in that significant variation in composition was possible. Each participant was allowed to shoot multiple photographs of the 7 scenarios, and choose their best one for consideration. Of the 7 * 33 = 231 photographs collected, 10 were excluded for violation of the rules leaving the 221 used. The scenarios and shooting areas were selected by two professional photographers (university faculty in Photography).

In [3], ratings for each of the images were obtained from 8 human judges, who were asked to rate on a 10-point rating scale. The human judges were composed of both professional photographers and image processing researchers. Completion time for the task of rating all 221 images averaged around 4 hours. Though [3] reported that the judges were allowed to pause and resume the ratings, the cognitive load of providing fair ratings to such a large database of photos made fatigue a factor. It is difficult to extend this method to a large number of judges due to the large time re-quirement.

Since binary ratings are much simpler to collect, we extend [3] to provide a new set of data in this paper. Specifically, we use a crowd sourcing tool, Amazon Mechanical Turk (AMT), to collect binary like/dislike ratings on each image from a large group of people that we refer to below as AMT workers. We were able to collect independent like/dislike data on each image from a total of 168 AMT workers. Workers were al-lowed 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 this data may be more representative of the average person than the previous study [3] which used professionals and researchers.

We also obtained ratings for each of the 221 images from ACQUINE by uploading them to ACQUINEs public website.

4 Analysis

Figure 2 shows a histogram of the percentage of likes on each of the 221 images from the 168 AMT workers. The majority of the histogram is above 50%, which shows that workers tended to like more than dislike images. Moreover, there is a peak around 85%, which shows that there are many images on which most workers agreed.

Table 1. The table shows the correlations among rating schemes. For the 10-point data, the sum of all scores from 8 judges is used. $p < 0.01$ for all correlations where a number is shown, and x indicates a correlation where $p > 0.05$, meaning it is not significantly different from zero.

Scheme	ACQUINE	Like%	Colorfulness
10-point[3]	0.23	0.59	x
ACQUINE	1.0	0.19	x
Colorfulness	x	0.26	1.0

Next, we check for relations among the like percentage, ACQUINE scores, and the 10-point rating data from [3], using the Spearman rank correlation to allow for monotonic relationships. Table 1 summarizes the findings, showing only correlations that are significantly different from zero ($p < 0 : 01$). We see 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 in [12] between binary ratings and a 5-point rating. The agreement is striking given that [12] measured the correlation on a database of movie ratings, which is completely different than our photographic 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 [3], 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. It also suggests that experience does not play a major role in rating aesthetics.

The high correlation between AMT workers used in this study and the expert judges used in [3] makes us wonder what are the aspects that people, in general, find appealing in a photograph. Although aesthetics is very complex, it is worth exploring to the extent to which aesthetic ratings can be predicted by basic image statistics as might be used to evaluate image quality.

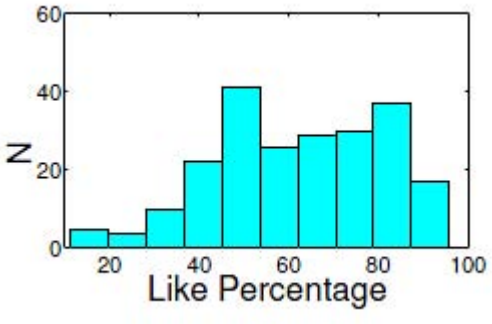

Fig. 2. Histogram of the percentage of likes given to the 221 images by the AMT workers. The peak around 80-90% shows that the voters tend to agree among themselves while liking a particular image.

4.1 Response to Global Image Statistics

We now check if the percentage of likes provided by AMT workers in our study is in any way related to global statistics measuring contrast, colourfulness, 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 * N image with mean luminance $\bar{Y}$ as:

$$C = \frac{1}{MN} \sum_{i=0}^{M-1} \sum_{j=0}^{N-1} [Y_{ij} - \bar{Y}]^2 \tag{1}$$

Colourfulness K is defined by the Earth Movers Distance (EMD) [13] between the uniform distribution D_u on the RGB colour cube with 4 bins on each axis (64 in all), colour distribution of the image D, and the distance between successive bins d i.e.,

$$K = EMD[D_u, D, d] \tag{2}$$

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 di-vided into thirds. This measure is defined as in [1] for the hue component H of the image in HSV coordinates as:

$$T = \sum_{i=\frac{M}{3}}^{\frac{2M}{3}} \sum_{j=\frac{N}{3}}^{\frac{2N}{3}} H_{ij} \tag{3}$$

Although T defined in this way is basically a regional average, we use it to determine whether a simple measure of hue distribution can predict the compositional appeal of our images. We measured the correlations among five variates:

C, K, T, the like percentage of images for the AMT workers, and the sum of ratings given by the 8 judges in [3]. Among these five variates, the only case where a correlation involving C, K, or T had significance $p < 0 : 01$ is shown in Table 1, where colorfulness K has a low but statistically-significant correlation of 0:26 with like percentage. The other remaining correlations, in particular those involving contrast C and the hue distribu-tion measure T, were such that they could not be distinguished from zero. We see that global statistics have little predictive value for the human ratings in our study.

4.2 Potential Features

Fig. 3. Most liked images in the entire database. Note the symmetrical structure, geometric regularity and reflectional symmetry in composition.

Since we have seen that global image statistics have little predictive value for image aesthetics, we led to ask what is it that the voters in our study really look for in an image? The question is undoubtedly complex. To understand the issues better, we looked at some of the images that received a large percentage of likes from the votes. In the top row of Figure 3, we show the images that more than 90% of the people liked. The well-liked images show composition consisting of geometric regularity of structure and also reflectional symmetry. Furthermore, we also examine the least liked images, to see what might need to

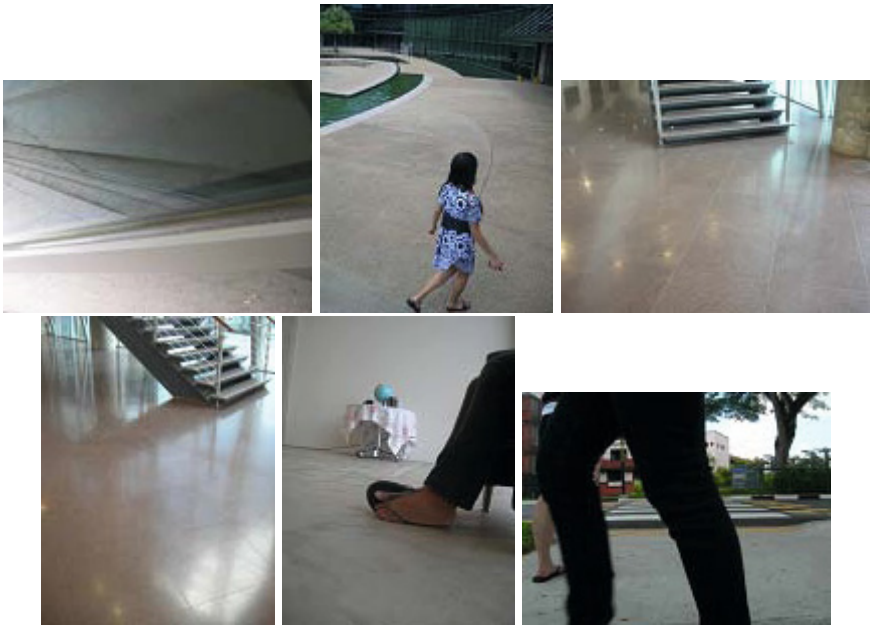

Fig. 4. Most disliked images in the entire database. Note the lack of symmetrical structure, geometric regularity and reflectional symmetry in composition.

be improved to make those images more appealing. The bottom row of Figure 3 shows examples of images that more than 80% of the people disliked. These are images that do not convey much about the scene, and can be said to lack identifiable composition elements such as structure or balance.

5 Conclusion

Quantifying image aesthetics is a difficult problem for humans, let alone computers. There is a strong demand for automatic aesthetic prediction systems in the digital imaging industry. To train such systems, we need methods to efficiently collect aes-thetic rating data from a large group of human subjects. We have seen in this paper that a binary rating scale is an effective tool for collecting such data, which is signifi-cant because binary ratings can replace the more difficult task of rating aesthetic content fairly on a numerical scale. Our experiments also show that simple global image statistics have little predictive value for aesthetic ratings, indicating the need for more research into measures of important factors involved in composition.

References

1. Datta, R., Zewang, J.: ACQUINE: aesthetic quality inference engine - real time automatic ratings of photo aesthetics. In: Multimedia Information Retrieval, pp. 421–424 (2010)
2. Murray, N., Marchesotti, L., Perronnin, F.: Ava: A large-scale database for aesthetic visual analysis. In: Computer Vision and Pattern Recognition (CVPR), pp. 2408–2415 (2012)
3. Sachs, T.S., Kakarala, R., Castleman, S.L., Rajan, D.: A data-driven approach to understanding skill in photographic composition. In: Koch, R., Huang, F. (eds.) ACCV 2010 Workshops, Part II. LNCS, vol. 6469, pp. 112–121. Springer, Heidelberg (2011)
4. Kakarala, R., Sachs, T.S., Premachandran, V.: Comparing automated and human ratings of photographic aesthetics. In: Proceedings of the 19th Color Imaging Conference, San Jose, CA (2011)
5. Savakis, A., Etz, S., Loui, A.: Evaluation of image appeal in consumer photography. In: SPIE Human Vision and Electronic Imaging (2000)
6. Ke, Y., Tang, X., Jing, F.: The design of high-level features for photo quality assessment. In: Computer Vision and Pattern Recognition, pp. 419–426 (2006)
7. Tong, H., Li, M., Zhang, H.-J., He, J., Zhang, C.: Classification of digital photos taken by photographers or home users. In: Aizawa, K., Nakamura, Y., Satoh, S. (eds.) PCM 2004. LNCS, vol. 3331, pp. 198–205. Springer, Heidelberg (2004)
8. Datta, R., Li, J., Wang, J.Z.: Algorithmic inferencing of aesthetics and emotion in natural images: An exposition. In: International Conference on Image Processing, pp. 105–108 (2008)
9. Datta, R., Joshi, D., Li, J., Wang, J.Z.: Studying aesthetics in photographic images using a computational approach. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006. LNCS, vol. 3953, pp. 288–301. Springer, Heidelberg (2006)
10. Marchesotti, L., Perronnin, F., Larlus, D., Csurka, G.: Assessing the aesthetic quality of photographs using generic image descriptors. In: International Conference on Computer Vision, pp. 1784–1791 (2011)
11. Bargagliotti, A.E., Li, Y.: Decision making using rating systems. When scale meets binary. MPRA Paper 16947. University Library of Munich, Germany (2006)
12. Cosely, D., Lam, S.K., Istvan, A., Konstan, J.A., Riedl, J.: Is seeing believing? How recommender system interfaces affect users' opinion. In: CHI, pp. 585–592 (2003)
13. Rubner, R., Tomasi, C., Guibas, L.J.: The earth mover's distance as a metric for image retrieval. International Journal of Computer Vision, 99–121 (2000)