IEEE International Conference on Image Processing, Beijing, China, Sep. 2017.

A DATABASE FOR PERCEPTUAL EVALUATION OF IMAGE AESTHETICS

Wentao Liu, and Zhou Wang

Dept. of Electrical & Computer Engineering, University of Waterloo, Waterloo, ON, Canada
Email: {w238liu,zhou.wang}@uwaterloo.ca

ABSTRACT

Objective image aesthetics assessment (IAA) is attracting an increasing amount of attention in recent years. One of the most critical issues that hampers IAA research is the lack of publicly available and reliable image databases that can be used to train and test IAA features and models, especially those databases that offer continuous-valued subjective opinion scores. In this work, we construct a Waterloo IAA database containing more than $1,000$ images, and carry out a lab-controlled subjective user study. There are several unique and desirable features of the new database as compared to existing ones – It helps us better understand the level of diversity of subject opinions; it provides continuous-valued IAA scores approximately evenly distributed from poor to excellent aesthetics levels; it also allows us to test the effectiveness of various aesthetics features on predicting continuous aesthetics scores. Using the new database as a benchmark, we test more than 1,000 IAA features. The results indicate that existing features are still weak at aesthetics estimation, and the effectiveness of aesthetics features are content dependent. Therefore, understanding and assessing image aesthetics remain a major challenge for future research. The database will be made publicly available.

Index Terms— image aesthetics assessment, subjective testing, image database

1. INTRODUCTION

As digital images becoming a dominant form of information in the modern world [1], objective image aesthetics assessment (IAA) is drawing a great deal of attention due to its potential use in a growing number of applications, including image recommendation, photo album management, and photo capturing suggestion. In principle, image aesthetics can be interpreted as the experience of beauty for subjects viewing an image. Scientific studies suggest that image aesthetics are mainly determined by the composition of semantic symbols uncovered in the image [2]. However, image semantics can be highly abstract. High-level features that may capture such semantic symbols include *simplicity*, *colorfulness*, *color combination*, *sharpness*, *image pattern*, and *object composition* [3, 4, 5, 6, 7, 8]. Recently, it has been shown that local image descriptors [9] and learned features [10, 11, 12, 13, 14] also demonstrate promises in predicting image aesthetics annotations. Despite various features being used, most IAA algorithms only produce a binary result, indicating whether an image is of very high or very low aesthetics [3, 4, 5, 6, 7, 8, 9, 10, 12, 13], but do not work well with images of mid-level aesthetics. However, the perceived aesthetics of real-world images can be much richer than only two levels. Continuous-valued IAA models are highly desirable, but are still lacking until now [11, 14].

A key problem that slows down the development of objective IAA is the lack of reliable image databases that could be used for training and testing IAA features and models. Considerable effort has been made, and several subject-annotated databases were constructed [5, 15, 16, 17]. Based on the type of the subjective annotations, existing databases can be classified into two kinds. The first kind is binary-annotated databases, which contain very beautiful and very undesired images only. The aesthetics labels can be collected in a lab-controlled environment, such as the CUHKPQ database [5], or from relevant on-line tags, such as the CLEF database [15]. Databases of this kind were built for binary-valued IAA algorithms, and do not easily facilitate the development of continuous-valued IAA models. The second kind of databases contain multi-level or continuous aesthetics annotations [16, 17]. Databases of this kind are generally *website-based*, where images are crawled from photo-sharing websites [16, 17]. In addition, the on-line score for each image is also downloaded and regarded as its aesthetic annotation. The advantage of website-based databases is that a large number of subject-rated images can be collected at a very low cost. However, there are three main drawbacks. First, the on-line scores can be affected by many factors other than image aesthetics, for example viewing conditions and user emotions, and there is no simple mechanism to properly align the scores and remove outliers. Second, it is difficult to gauge how much agreement is obtained between subjects. Third, the distribution of on-line scores often concentrates at the middle score range, making it hard to perform fair and meaningful evaluations of IAA models on these databases.

In this work, we build a new lab-controlled image aesthetics database, namely the Waterloo IAA database, with continuous subjective ratings. The main contributions of this

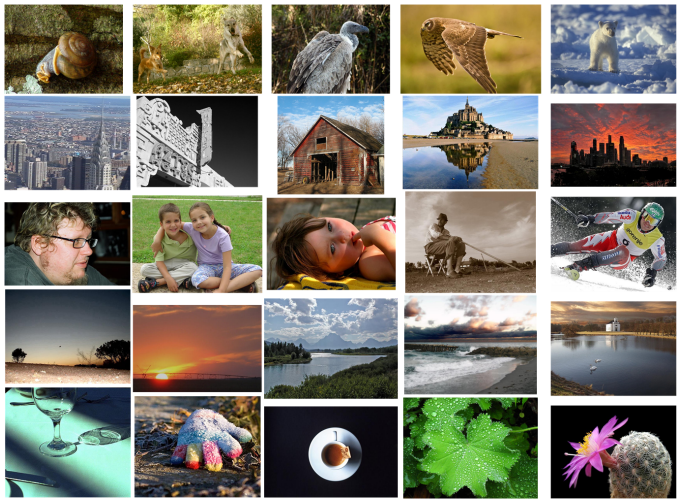

Fig. 1. Sample images for aesthetics assessment: from the top to the bottom row are Animal, Architecture/City Scenes, Human, Natural Scene, Still Object images; from the leftmost to the rightmost column are images from the lowest to the highest on-line score ranges.

database are as follows. 1) This is the first lab-controlled database, which provides a continuous-valued benchmark for objective IAA models. Through data analysis, we are able to have a better understanding about the level of agreement between humans on evaluating image aesthetics. 2) The images in this database are more uniformly distributed in the aesthetics spectrum than the databases directly crawled from the same website [17], where images are over-concentrated at the mid-range. 3) The database enables us to better investigate the effectiveness of aesthetics features in predicting subjective opinions of image aesthetics. Our results show that no existing aesthetics feature is significantly correlated with the continuous-valued IAA scores, and image aesthetics with different contents may be affected by different types of features.

2. DATABASE CONSTRUCTION AND SUBJECTIVE ASSESSMENT

$1,000$ images are selected from the well-known image sharing website *photo.net* according to their on-line ratings and contents. Specifically, we first determine five non-overlapping score ranges, uniformly spanning from the low to the high ends of the on-line score range $[2, 7]$ (On *photo.net*, the score range is from 1 to 7. However, score 1 is rarely used.), and five image content types, namely Animals (A), Architectures/City Scenes (C), Humans (H), Natural Scenes (N), and Still Object (S). For each of the five score ranges and each of the five manually labeled content types, around 30-50 images are selected so that the 1,000 images are roughly uniformly distributed over all aesthetics levels and content

Fig. 2. The GUI used for the subjective user study.

Table 1. Distribution of images in different score ranges

Subsets	A	C	H	N	S	All
< 4.06	33	31	39	54	43	200
$[4.5, 4.7]$	39	38	39	50	34	200
$[5.0, 5.2]$	40	40	40	40	40	200
$[5.55, 5.75]$	40	40	40	40	40	200
> 6.17	40	31	39	41	49	200
All	192	180	197	225	206	1000

types. The actual number of images in each subset are listed in Table 1. Fig. 1 shows sample images for the 25 subsets.

After image selection, we perform preprocessing to remove frames surrounding images and to unify image sizes. To evaluate whether a frame has an effect on perceived image aesthetics, we add back 80 images with frames. Moreover, 20 of the 1,000 images are duplicated for consistency check. Consequently, we have 1,100 images for the subjective study.

The subjective user study is conducted at the University of Waterloo in the Image and Vision Computing (IVC) laboratory, which has a normal lighting condition without reflecting ceiling walls and floor. A Truecolor (32 bits) LCD monitor of 27 inches with resolution of 1920×1080 pixels is used to display all images. We adopt the single-stimulus methodology recommended by the ITU-R BT.500 [18] in the study, and the monitor is calibrated accordingly. A customized MATLAB GUI (Fig. 2) is built to render one image at a time. The 1,100 images are displayed in a random order. A total of 33 observers, including 18 male and 15 female subjects aged between 22 and 33, participated in the subjective experiment. All the subjects have normal or corrected-to-normal vision, and viewed the images from a normal distance (around 60 cm). The length of the experiment is around 90 minutes. The subjects are asked to take a rest every 30 minutes to reduce fatigue effect.

Before the study, the participants are trained with 20 independent training images with various image contents and different on-line scores. The purpose of the training session is to help participants become familiar with the test environment,

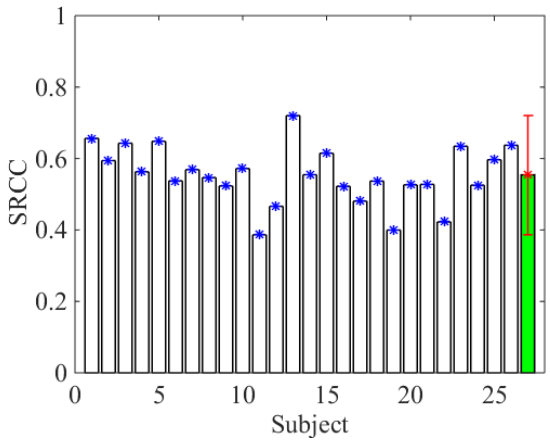

Fig. 3. SRCC between individual subject score against MOS. Rightmost column: average subject performance.

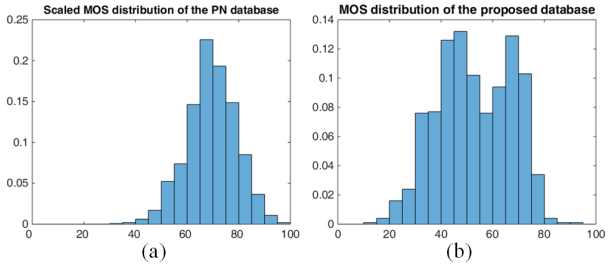

Fig. 4. MOS distributions of (a) the PN database [17] and (b) the proposed database. The MOS of the PN database have been scaled to the same score range $[0, 100]$ as the proposed database for better comparison.

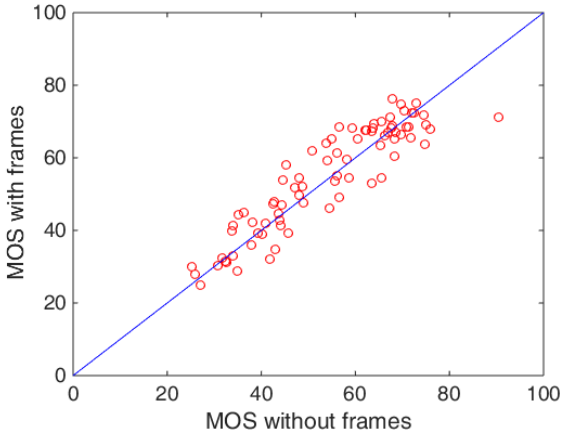

Fig. 5. Scatter plot of MOS with and without frames.

and build up their own criteria of scoring. During the training session, the instructor provides no opinion about which images should be given what scores. In the test session, the participants are asked to score the displayed image based on its aesthetics level using a sliding bar. The position of the slider is converted to an integer between $[0, 100]$. A higher score indicates that the subject considers the image more aesthetically appealing.

3. ANALYSIS AND DISCUSSION

3.1. Subjective Data Analysis

We first use consistency check to detect unreliable subjects. Note that 20 images are displayed twice in the subjective test. The mean absolute error (MAE) of the first and the second scores are calculated for each subject. If the MAE is greater than 25, then a subject is considered unreliable, and all of his/her scores are discarded. By doing so, 6 subjects are considered unreliable and rejected. We then perform the data alignment and outlier detection and removal schemes suggested in [18]. As a result, one more subject is rejected as an outlier.

The final aesthetic score, namely the Mean Opinion Score (MOS), is computed by averaging the aligned subjective scores from the remaining 26 subjects. Regarding the MOS values as the ground truth, we can evaluate the performance of individual subjects by calculating the correlation between individual subject's score and the MOS across the whole database. Pearson Linear Correlation Coefficient (PLCC) and Spearman's Rank Correlation Coefficient (SRCC) are employed as the evaluation criteria. Both criteria lie in $[0, 1]$ with a higher value indicating higher agreement with the MOS. The SRCC results are summarized in Fig. 3 (PLCC results are similar but not shown due to space limit), where the performance of an average subject is given at the rightmost column. It can be seen that there is a decent degree of agreement on image aesthetics among subjects. Although this has been previously noted and verified with small-scale subjective test [19, 20], this is the first time to quantitatively evaluate the extent of such agreement using a relatively large database. On the other hand, it is not surprising that different people have different understandings on image aesthetics, as the lowest individual SRCC values are below 0.4.

As mentioned earlier, a major issue with existing website-based databases is that the aesthetics scores are over concentrated at mid-levels. Fig. 4(a) shows the MOS histogram of the PN database [17] as an example. This is not a desirable feature. For example, a straw-man model that predicts any image to have the same score at the average level could result in a fairly low prediction error, but is indeed meaningless. By contrast, the MOS distribution of our proposed database is much more uniform as shown in Fig. 4(b).

A common practice of photographers to "enhance" image aesthetics is to add an artificial frame surrounding the original image. Our proposed database contains 80 images with

Table 2. SRCC between MOS and the best feature in each feature type for each content type.

Type	A	C	H	N	S	All
f_1	0.294	0.216	0.262	0.203	0.279	0.222
f_2	0.255	**0.314**	0.197	0.282	0.324	0.198
f_3	0.162	0.223	0.146	0.186	0.255	0.185
f_4	**0.329**	**0.387**	0.251	**0.364**	0.236	0.253
f_5	0.276	**0.305**	0.172	0.247	0.202	0.161
f_6	0.191	**0.357**	0.179	0.250	**0.316**	0.191
f_7	0.018	0.066	0.012	0.003	-0.170	-0.005
f_8	0.042	-0.090	0.179	-0.006	0.089	-0.003
f_9	-0.027	-0.049	-0.015	0.043	0.165	0.038

frames, together with their frame-removed versions. The scatter plots of MOS of these images with and without the frames are shown in Fig. 5, where all points are closely aligned along the diagonal line, and clear improvement by adding a frame is not observed. Our two-sample t-tests of all image pairs further confirm the observation. Therefore, adding a frame in order to enhance aesthetics level is not justified by our experimental results.

3.2. Effectiveness of Aesthetics Features

To test the effectiveness of the aesthetics features proposed in the literature [3, 8], we compute more than $1,000$ features for all images in the database, and categorize them into 9 types, each assessing an image from a different perspective. These include simplicity (f_1), colorfulness (f_2), color combination (f_3), sharpness (f_4), texture and symmetry pattern (f_5), object composition (f_6), luminance (f_7), aspect ratio (f_8), and low depth of field (DoF) indicator (f_9). We calculate the SRCC between these features and the MOS across the whole database, and draw a histogram of SRCC for each feature type, as shown in Fig. 6. Note that there is only one feature in the last 3 types, so their SRCC histograms have only one bin. It can be seen that the absolute values of SRCC of most features with MOS are smaller than 0.2, suggesting that it is difficult to predict human sense of aesthetics from a single factor.

To explore the potential of each type of features in predicting aesthetics scores, we list the overall SRCC value of the best feature in each feature type in Table 2, where we find that the most relevant aesthetics features turn out to be in the order of sharpness (f_4), object composition (f_6), simplicity (f_1), colorfulness (f_2) and color combination (f_3). This is somewhat consistent with our intuition: humans prefer sharp images with rich details, and are also attracted by simple and colorful images.

As humans tend to use different criteria to judge aesthetics for images with different contents [2, 21], we also list the SRCC of the best feature obtained by each feature type for the 5 content types in Table 2. The SRCC values greater than

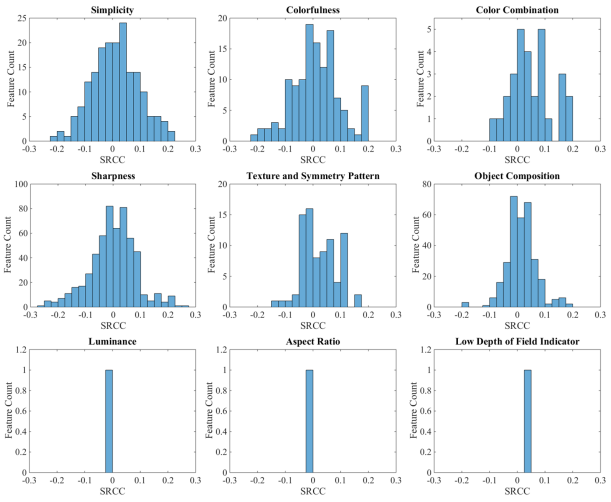

Fig. 6. SRCC histograms of features of 9 types. The name of each is indicated by each subfigure title.

0.3 are highlighted by boldface. It can be observed that aesthetics of Animal, Architecture/City Scene and Natural Scene images are best predicted by sharpness features (f_4), and object composition (f_6) features appear to be a strong factor for Architecture/City Scenes (C) and Still Objects (S). It is interesting to see that SRCCs of Human images are relatively low for all features. This may be because that more aesthetics cues are involved in Human images. For example, a beautiful or a familiar face may affect human opinions more than colorfulness or object composition in such images. Additionally, Human images show special preference to high aspect ratios (f_8) compared to the other content types. A possible reason is that portrait orientation is better at conveying the beauty of body shape than landscape. It is not surprising that global luminance (f_7) and low DoF indicator (f_9) are less relevant with image aesthetics in general. Nevertheless, professional photographers often use relatively dark background and reduce the DoF when shooting single object images, so f_7 and f_9 exhibit some correlation in the Still Object images.

4. CONCLUSION

We construct a new Waterloo IAA database, and conduct a lab-controlled subjective user study. The database contains more than $1,000$ images with continuous-valued aesthetics scores approximately evenly distributed from poor to excellent aesthetics levels. Using the database, we test more than $1,000$ features of 9 different types for aesthetics prediction. We find that all individual features are weak at aesthetics prediction, and the prediction effectiveness of different feature types varies for images of different content types. Our results suggest that understanding and automatically predicting image aesthetics remain a challenging problem. We will make the database publicly available to facilitate future research.

5. REFERENCES

[1] Cisco Corporation, "Cisco visual networking index: Global mobile data traffic forecast update, 2014-2019," Feb 2015, [Online]. Available: http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/white_paper_c11-520862.pdf.

[2] Dhirai Joshi, Ritendra Datta, Elena Fedorovskaya, Quang-Tang Luong, James Z. Wang, Jia Li, and Jiebo Luo, "Aesthetics and emotions in images," *IEEE Signal Processing Magazine*, vol. 28, no. 5, pp. 94–115, Sept 2011.

[3] Ritendra Datta, Dhiraj Joshi, Jia Li, and James Z. Wang, "Studying aesthetics in photographic images using a computational approach," in *European Conf. Computer Vision*, 2006, pp. 288–301.

[4] Yan Ke, Xiaoou Tang, and Feng Jing, "The design of high-level features for photo quality assessment," in *Proc. IEEE Int. Conf. Computer Vision and Pattern Recognition*, June 2006, vol. 1, pp. 419–426.

[5] Yiwen Luo and Xiaoou Tang, "Photo and video quality evaluation: Focusing on the subject," in *European Conf. Computer Vision*, David Forsyth, Philip Torr, and Andrew Zisserman, Eds., 2008, pp. 386–399.

[6] Wei Luo, Xiaogang Wang, and Xiaoou Tang, "Content-based photo quality assessment," in *Proc. IEEE Int. Conf. Computer Vision*. IEEE, 2011, pp. 2206–2213.

[7] Florian Simond, Nikolaos Arvanitopoulos, and Sabine Ssstrunk, "Image aesthetics depends on context," in *Proc. IEEE Int. Conf. Image Proc.*, Sept 2015, pp. 3788–3792.

[8] Eftichia Mavridaki and Vasileios Mezaris, "A comprehensive aesthetic quality assessment method for natural images using basic rules of photography," in *Proc. IEEE Int. Conf. Image Proc.*, Sept 2015, pp. 887–891.

[9] Luca Marchesotti, Florent Perronnin, Diane Larlus, and Gabriela Csurka, "Assessing the aesthetic quality of photographs using generic image descriptors," in *Proc. IEEE Int. Conf. Computer Vision*, Nov 2011, pp. 1784–1791.

[10] Xinmei Tian, Zhe Dong, Kuiyuan Yang, and Tao Mei, "Query-dependent aesthetic model with deep learning for photo quality assessment," *IEEE Trans. Multimedia*, vol. 17, no. 11, pp. 2035–2048, Nov 2015.

[11] Bin Jin, Maria V. Ortiz Segovia, and Sabine Ssstrunk, "Image aesthetic predictors based on weighted cnns," in *Proc. IEEE Int. Conf. Image Proc.*, Sept 2016, pp. 2291–2295.

[12] Xin Lu, Zhe Lin, Xiaohui Shen, Radomr Mech, and James Z. Wang, "Deep multi-patch aggregation network for image style, aesthetics, and quality estimation," in *Proc. IEEE Int. Conf. Computer Vision*, Dec 2015, pp. 990–998.

[13] Long Mai, Hailin Jin, and Feng Liu, "Composition-preserving deep photo aesthetics assessment," in *Proc. IEEE Int. Conf. Computer Vision and Pattern Recognition*, June 2016, pp. 497–506.

[14] Y. Kao, C. Wang, and K. Huang, "Visual aesthetic quality assessment with a regression model," in *Proc. IEEE Int. Conf. Image Proc.*, Sept 2015, pp. 1583–1587.

[15] Michael Grubinger, Stefanie Nowak, and Paul Clough, *Data Sets Created in ImageCLEF*, pp. 19–43, Springer Berlin Heidelberg, Berlin, Heidelberg, 2010.

[16] Naila Murray, Luca Marchesotti, and Florent Perronnin, "Ava: A large-scale database for aesthetic visual analysis," in *Proc. IEEE Int. Conf. Computer Vision and Pattern Recognition*. IEEE, 2012, pp. 2408–2415.

[17] Ritendra Datta, Jia Li, and James Z Wang, "Algorithmic inferencing of aesthetics and emotion in natural images: An exposition," in *Proc. IEEE Int. Conf. Image Proc.* IEEE, 2008, pp. 105–108.

[18] ITU, "Recommendation ITU-R BT.500-13 methodology for the subjective assessment of the quality of television pictures," Jan 2012, [Online]. Available: https://www.itu.int/dms_pubrec/itu-r/rec/bt/R-REC-BT.500-13-201201-I!!PDF-E.pdf.

[19] Denis Dutton, *The art instinct: beauty, pleasure, & human evolution*, Oxford University Press, USA, 2009.

[20] Ernestasia Siahaan, Alan Hanjalic, and Judith Redi, "A reliable methodology to collect ground truth data of image aesthetic appeal," *IEEE Trans. Multimedia*, vol. 18, no. 7, pp. 1338–1350, July 2016.

[21] Zihan Zhou, Siqiong He, Jia Li, and James Z. Wang, "Modeling photographic composition via triangles," *CoRR*, vol. abs/1605.09559, 2016.