Photo Aesthetics Ranking Network with Attributes and Content Adaptation — Supplementary Material

Shu Kong[1], Xiaohui Shen[2], Zhe Lin[2], Radomir Mech[2], Charless Fowlkes[1]

[1]UC Irvine {skong2, fowlkes}@ics.uci.edu
[2]Adobe Research {xshen, zlin, rmech}@adobe.com

1 Introduction

In this supplementary material, we first present in detail on collecting our AADB dataset, in Section 2, and analyze the dataset w.r.t its aesthetics attributes in Section 3. Then we carry out the consistency analysis of the dataset in Section 4 to show the annotations are reliable that the raters have consistently labeled the images. Furthermore, in Section 6, we visually demonstrate the effectiveness of our model for aesthetics rating and analysis w.r.t the attributes. To address the effectiveness of content-aware model described in the paper, we analyze performance of different methods in utilizing this information in Section 5. Lastly, we attach the instruction used for teaching the raters to pass qualification test. The instruction and qualification test can filter out spammers to a large extent.

2 Aesthetics and Attributes Database (AADB)

2.1 Attributes in AADB

We select eleven attributes that are highly related to image aesthetics after consulting professional photographers, which are

1. "balancing element" – whether the image contains balanced elements;
2. "content" – whether the image has good/interesting content;
3. "color harmony" – whether the overall color of the image is harmonious;
4. "depth of field" – whether the image has shallow depth of field;
5. "lighting" – whether the image has good/interesting lighting;
6. "motion blur" – whether the image has motion blur;
7. "object emphasis" – whether the image emphasizes foreground objects;
8. "rule of thirds" – whether the photography follows rule of thirds;
9. "vivid color" – whether the photo has vivid color, not necessarily harmonious color;
10. "repetition" – whether the image has repetitive patterns;
11. "symmetry" – whether the photo has symmetric patterns.

These attributes span traditional photographic principals of color, lighting, focus and composition, and provide a natural vocabulary for use in applications, such as auto photo editing and image retrieval. To visualize images containing these attributes, please refer to the attached our AMT instruction in the end of this supplementary material. The instruction is used for teaching raters to pass the qualification test.

2.2 Data Collection By Amazon Mechanical Turk

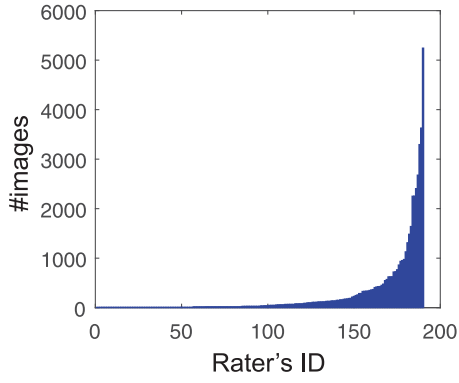

Fig. 1. Long tail distribution of AMT workers: number of rated images vs. each worker.

To collect a varied set of photographic images, we download images from Flickr website[1], which carry a Creative Commons license. We manually curate the dataset to remove non-photographic images (*e.g.* cartoons, drawings, paintings, ads images, adult-content images, etc.). We have multiple workers independently annotate each image with an overall aesthetic score and the eleven meaningful attributes using Amazon Mechanical Turk[2].

For each attribute, we allow workers to click "positive" if this attribute conveyed by the image can enhance the image aesthetic quality, or "negative" if the attribute degrades image aesthetics. The default is "null", meaning the attribute does not effect image aesthetics. For example, "positive" vivid color means the vividness of the color presented in the image has a positive effect on the image aesthetics; while the counterpart "negative" means, for example, there is dull color composition. Note that we do not let workers tag negative repetition and symmetry, as for the two attributes negative values do not make sense.

We launch a task consisting of 10,000 images on AMT, and let five different workers label each image. All the workers must read instructions and pass a qualification exam before they become qualified to do our task. The images are split into batches, each of which contains ten images. Therefore, raters will annotate different numbers of batches. There are 190 workers in total doing our AMT task, and the workers follow long tail distribution, as demonstrated by Figure 1. Figure 2 shows the interface of our AMT task.

Note that even though judging these attributes is also subjective, the averaged scores of these attributes indeed reflect good information if we visualize the ranked images

[1] www.flickr.com
[2] www.mturk.com

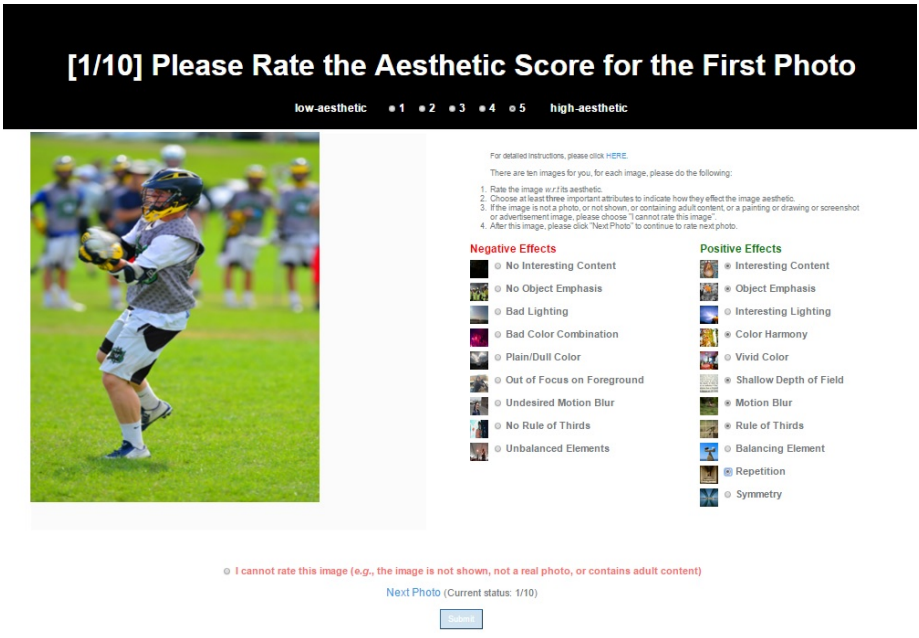

Fig. 2. Interface of data collection by AMT.

w.r.t averaged scores. Therefore, we use the averaged score as the ground truth, for both aesthetic score and attributes. Furthermore, we normalize aesthetic score to the range of $[0, 1]$, as shown by Figure 3, from which we can see that ratings are well fit by a Gaussian distribution. This observation is consistent with that reported in [1]. In our experiments we normalize the attributes' scores to the range of $[-1, 1]$. The images are split into testing set (1,000 images), validation set (500 images) and training set (the rest).

3 Statistics of AADB

The final AADB dataset contains 10,000 images in total, each of which has aesthetic quality ratings and attribute assignments provided by five different individual raters. Therefore, we have rating scores for attributes as well, which is different from AVA dataset [1] in which images only have binary labels for the attributes. Figure 4 shows the distribution of each attributes.

4 Consistency Analysis

As there are five individuals rating each image, one may argue that the annotations are not reliable for this subjective task. Therefore, we carry out consistency analysis. We

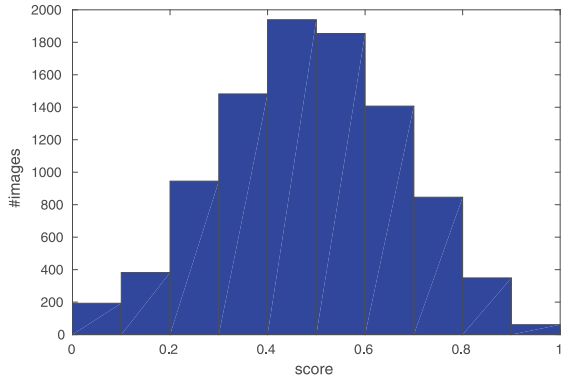

Fig. 3. The distribution of rated image aesthetic scores by the AMT workers follows a Gaussian distribution.

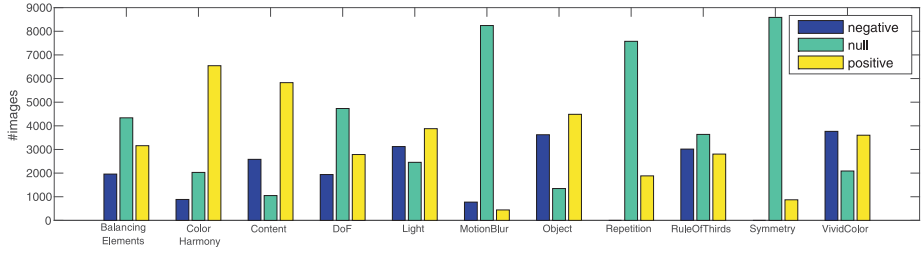

Fig. 4. The distributions of all the eleven attributes. Note that for attributes repetition and symmetry, we do not let AMT workers annotate negative labels, as these attributes are of neutral meaning. Instead, we only allow them to point out whether there exist repetition or symmetry. To solve the data imbalance problem in training attribute classifiers, we adopt some data augmentation tricks to sample more rare cases.

use both Kendall's W and Spearman's ρ for the analysis. Kendall's W directly measures the agreement among multiple raters, and accounts for tied ranks. It ranges from 0 (no agreement) to 1 (complete agreement). Spearman's ρ is used in our paper that compares a pair of ranking lists.

First, we conduct a permutation test over global W to obtain the distribution of W under the null hypothesis. We plot the curve of W: $p(W)$ vs. W in Fig. 5 and $p(W < t)$ vs. t in Fig 6. We can easily see that the empirical Kendall's W on our AADB dataest is statistically significant.

Then, for each batch, we can also evaluate the annotation consistency with Kendall's W, which directly calculates the agreement among multiple raters, and accounts for tied ranks. As there are ten images and only five possible ratings for each image, tied ranks may happen in a batch. The average Kendall's W over all batches is 0.5322. This shows significant consistency of the batches annotated by the AMT workers. To test the

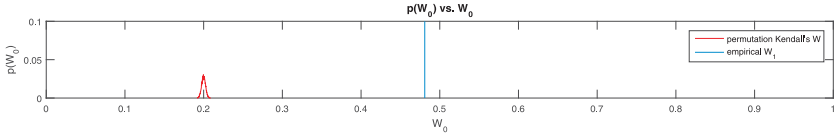

Fig. 5. Permutation test on Kendall's W: $p(W)$ vs. W.

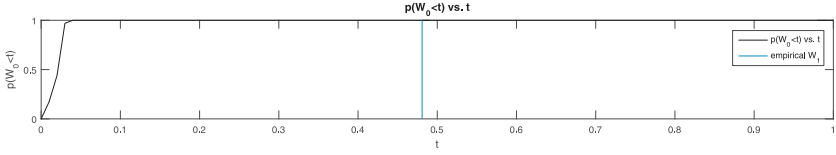

Fig. 6. Permutation test on Kendall's W: $p(W < t)$ vs. t.

statistical significance of Kendall's W at batch level, we adopt the Benjamini-Hochberg procedure to control the false discovery rate (FDR) for multiple comparisons [2]. At level $Q = 0.05$, 99.07% batches from $1,013$ in total have significant agreement. This shows that almost all the batches annotated by AMT workers have consistent labels and are reliable for scientific use.

Furthermore, we can also test the statistical significance w.r.t Spearman's ρ at batch levels using Benjamini-Hochberg procedure. The p-values of pairwise ranks of raters in a batch can be computed by the exact permutation distributions. We average the pairwise p-values as the p-value for the batch. With the FDR level $Q = 0.05$, we find that 98.45% batches have significant agreement. This further demonstrates the reliability of the annotations.

5 Analysis of Content-Aware Model

Table 1. Analysis of content-aware model on AVA dataset.

method	concatGT	concatPred	avg.	weightedSum	weightedSum_FT
Spearman's ρ	0.5367	0.5327	0.5336	0.5335	0.5426
accuracy(%)	75.41	75.33	75.39	75.33	75.57

To show the effectiveness of utilizing content information as a weights for output scores by different content-specific aesthetics rating branches, we report the performance on AVA dataset of different methods in Table 1. Our first method is named "concatGT", which means we use the ground-truth content label of an image, and get the estimated aesthetic score by the content-specific branch; then we put all the estimated scores together to get the global Spearman's ρ and classification accuracy. In method "concatPred", we use the predicted content label to choose which category-specific branch to use for estimating aesthetic score, then use the same procedure as in "concatGT". In method "avg.", we use all the content-specific aesthetics rating branches

to get multiple scores, and average them to a single score as the final estimation. In "weightedSum", we use the classification confidence score output by softmax of the content classification branch to do weighted sum for the final score. In "weightedSum_FT", we fine-tune the whole network but freezing the classification branch, and use the fine-tuned model to do weighted sum on the scores for the final aesthetics rating. From this table, we can clearly observe that "weightedSum_FT" performs the best, which is the one described in the paper.

6 Demonstration of Our Model

In this section, we test our model on personal photos qualitatively, in which these photos are downloaded online and not part of our AADB dataset. As our model can predicts all the eleven attributes, we show the attributes' estimation as well as the rated aesthetic scores. For better visualization, we simple set thresholds as (-0.2) and (0.2) to characterize "negative", "null" and "positive" attributes, respectively. Figure 7 – 9 show the results for images with high, low and medium estimated scores. We can see, in general, our model reasonably captures attributes and gives aesthetic scores.

References

1. Murray, N., Marchesotti, L., Perronnin, F.: Ava: A large-scale database for aesthetic visual analysis. In: Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on, IEEE (2012) 2408–2415
2. Benjamini, Y., Yekutieli, D.: The control of the false discovery rate in multiple testing under dependency. Annals of statistics (2001) 1165–1188

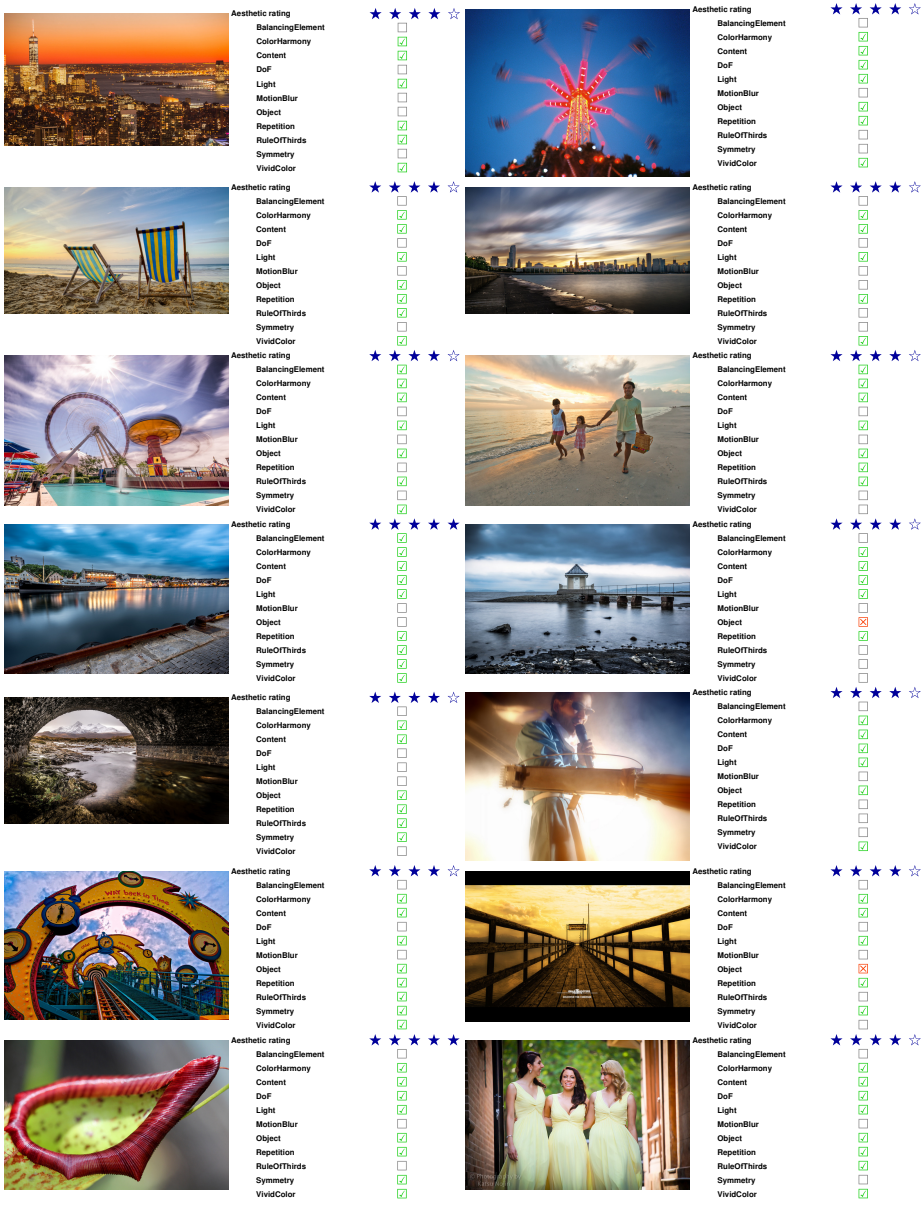

Fig. 7. Some images outside our database with high estimated scores.

Fig. 8. Some images outside our database with low estimated scores.

Fig. 9. Some images outside our database with medium estimated scores.