ALGORITHMIC INFERENCING OF AESTHETICS AND EMOTION IN NATURAL IMAGES: AN EXPOSITION

Ritendra Datta, Jia Li, and James Z. Wang[*]

The Pennsylvania State University, University Park, PA 16802, USA

ABSTRACT

Initial studies have shown that automatic inference of high-level image quality or aesthetics is very challenging. The ability to do so, however, can prove beneficial in many applications. In this paper, we define the *aesthetics gap* and discuss key aspects of the problem of aesthetics and emotion inference in natural images. We introduce precise, relevant questions to be answered, the effect that the target audience has on the problem specification, broad technical solution approaches, and assessment criteria. We then report on our effort to build real-world datasets that provide viable approaches to test and compare algorithms for these problems, presenting statistical analysis of and insights into them.

***Index Terms*—** Aesthetics, emotion, learning, datasets

1. INTRODUCTION

The image processing and analysis community has, for long, attempted to quantify and rectify image quality at a low-level, given the original image [3] or without it [10]. At a higher level, the perception often affects our emotion and mood, but there has been little headway made in automatic inferencing of the quality in images that affect mood or emotion. What makes the latter problem hard is that low-level image properties are insufficient to characterize high-level perception of aesthetics. Furthermore, there is a lack of precise definitions, assessment metrics, and test data for this problem, despite being desirable for many applications, e.g., image search, photography, story illustration, and photo enhancement.

In this paper, we attempt to clear the cloud on the problem of natural image aesthetics inference from visual content, by defining problems of interest, target audiences and how they affect the problem at hand, assessment metrics, and introduce real-world datasets for testing. Insights are drawn from the handful of previous attempts [4, 5, 8, 11] at solving related problems. While facial attractiveness has been a theme for many popular Websites [7], and has led to work on automatic facial aesthetics inference [6] that make use of symmetry and proportion, here we concern ourselves with generic images.

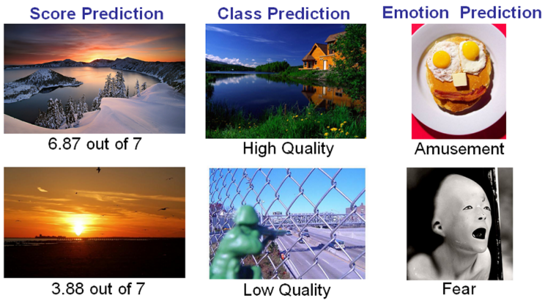

Fig. 1. Three aesthetics inferencing problems of significance.

2. QUESTIONS OF INTEREST

Being in its nascent stage, research on algorithmic aesthetics inference needs concretely defined tasks to solve, to start with. Aesthetics of natural images are, simply put, the emotions they arouse in people, which makes it relatively ill-defined. Contentious issues are 'emotion' and 'people'. Emotions are subjective across individuals, and they are of varied types (pleasing, boring, irritating, etc.). We leave aside subjectivity for now and consider aesthetic attributes to be a *consensus* measure over the entire population, such that they are meaningful to the average individual. Three data-driven aesthetics inference questions (Fig. 1) are discussed below.

2.1. Aesthetics Score Prediction

When a photograph is rated by a set of n people on a 1 to D scale on the basis of its aesthetics, the average score can be thought of as an estimator for its *intrinsic* aesthetic quality. More specifically, we assume that an image I has associated with it a true aesthetics measure $q(I)$, which is the asymptotic average if the entire population rated it. The average over the size n sample of ratings, given by $\hat{q}(I) = \frac{1}{n}\sum_{i=1}^{n} r_i(I)$ is an estimator for the population parameter $q(I)$, where $r_i(I)$ is the i^{th} rating given to image I. Intuitively, a larger n gives a better estimate. A formulation for aesthetics score prediction is therefore to infer the value of $\hat{q}(I)$ by analyzing the content of image I, which is a direct emulation of humans in the photo rating process. This lends itself naturally to a regression setting, whereby some abstractions of visual features act as predictor variables and the estimator for $\hat{q}(I)$ is the dependent variable. An attempt at regression based score prediction has been reported in [4], showing very limited success.

[*] J. Z. Wang is also with the Carnegie Mellon University. We acknowledge US NSF and Q.-T. Luong for their support.

Assessment: One method for assessing the quality of scoring prediction is to compute the rate or distribution of error [4].

2.2. Aesthetics Class Prediction

It has been observed both in [4] and [8] that score prediction is a very challenging problem, mainly due to noise in user ratings. Given the limited size rating samples, their averaged estimates have high variance, e.g., 5 and 5.5 on a $1-7$ scale could easily have been interchanged if a different set of users rated them, but there is no way to infer this from content alone, which leads to large prediction errors. To make the problem more solvable, the regression problem is changed to one of classification, by thresholding the average scores to create *high* vs. *low* quality image classes [4], or *professional* vs. *snapshot* image classes [8]. Suppose threshold values are ***HIGH*** and ***LOW*** respectively, then $class(I)$ is 1 if $\hat{q}(I) \geq HIGH$ and 0 if $\hat{q}(I) \leq LOW$. When the *band gap* $\delta = HIGH - LOW$ increases, the two classes are more easily separable, a hypothesis that has been tested and found to hold, in [4]. An easier problem but of practical significance is that of selecting a few representative high quality or highly aesthetic photographs from a large collection. In this case, it is important to ensure that most of the selected images are of high quality even though many of those not selected may be of high quality as well. An attempt at this problem [5] has proven to be more successful than the general HIGH/LOW classification problem described previously.

Assessment: The HIGH/LOW classification problem solutions can be evaluated by standard accuracy measures [4, 8]. On the other hand, the selection of high-quality photos need only maximize the *precision* in high quality within the top few photos, with *recall* being less critical.

2.3. Emotion Prediction

If we group emotions that natural images arouse into categories such as 'pleasing', 'boring', and 'irritating', then emotion prediction can be conceived as a multi-class categorization problem. These categories are fuzzily defined and and judgments are highly subjective. Consider K such emotion categories, and people select one or more of these categories for each image. If an image I gets votes in the proportion $\Pi_1(I), \ldots, \Pi_K(I)$, then two possible questions arise, none of which have been attempted in the past.

Most Dominant Emotion: We wish to predict, for an image I, the most voted emotion category $k(I)$, i.e., $k(I) = \arg\max_i \Pi_i(I)$. The problem is only meaningful when there is clear dominance of $k(I)$ over others, thus only these samples must be used for learning.

Emotion Distribution: Here, we wish to predict the distribution of votes (or an approximation) that an image receives from users, i.e., $\Pi_1(I), \ldots, \Pi_K(I)$, which is well-suited when images are fuzzily associated with multiple emotions.

Assessment: The 'most dominant emotion' problem is assessed like any standard multi-class classification problem. For 'emotion distribution', assessment requires a measure of similarity between discrete distributions, for which Kullback-Leibler (KL) divergence is a possible choice.

2.4. Context

In practice, any solution to the above problems can be tested either by user-generated feedback in online photo-sharing communities [9, 2, 1, 7], or by controlled user studies. Given this data-dependence, none of the models proposed will be fundamental or absolute in what they learn about aesthetics, but will be tempered to the given data acquisition setup, which we call the *context*. For example, what is considered 'interesting' (Flickr) may not be treated as being 'aesthetically pleasing' (Photo.net) by the population, and vice-versa. Therefore, we implicitly refer to it as aesthetics inference *under a given context* $\mathcal{X}$. Examples of key contextual aspects of test data are (a) the exact question posed to the users about the images, e.g., 'aesthetics' [9], 'overall quality' [2], 'like it' [1], (b) the type of people who visit and vote on the images, e.g., general enthusiasts [2, 9], photographers [9], and (c) The type of images rated, e.g., travel [12], topical [2]. Until fundamentals of aesthetics judgment are uncovered, contextual information is critical. The long-term goal is to have solutions that apply to as general a context as possible.

2.5. Personalization

While consensus measures and averaged-out ratings provide a generic learning setting, *personalized* models are of high relevance here due to the significant amount of subjectivity. In line with recommender systems, personalized models of aesthetics can potentially be learned, given sufficient feedback from a single user. In the absence of sufficient feedback from individuals, one solution is to consider *cliques* (groups or clusters of people with shared taste) instead of individuals, and make personalized inferences with respect to an user's parent clique, thus providing more data to learn. The cliques should ideally be determined automatically, may be overlapping, and an individual may belong to multiple cliques. There has been no reported attempt at personalized aesthetics.

3. TECHNICAL SOLUTION APPROACHES

Analogous to the concept of *semantic gap* that implies the technical limitations of image recognition, we can define the technical challenge in automatic inference of aesthetics in terms of the *aesthetics gap*, as follows: *The aesthetics gap is the lack of coincidence between the information that one can extract from low-level visual data (i.e., pixels in digital images) and the interpretation of emotions that the visual data may arouse in a particular user in a given situation.*

Past attempts [5, 8, 11] at aesthetics and quality inference have followed a logical series of steps, as discussed below.

Table 1. Datasets available for emotion/aesthetics learning.

Source	Feedback Type	Average Scores	Score Distribution	Individual Scores
Photo.net	1-7 (aesthetics)	Yes	Yes	Yes (partial)
DPChallenge	1-10 (quality)	Yes	Yes	No
Terragalleria	1-10 (liking)	Yes	Yes	No
Alipr.com	Emotion (8 types)	n/a	n/a	n/a

Feature Shortlisting: Possibly the most challenging part of the problem is conceiving meaningful visual properties that may have correlation with human ratings, and devising ways to convert them into numerical features. While feature shortlisting is largely ad-hoc in [11], the photography literature provides much of the intuitions for [4, 8]. The hypothesis there is that photographers follow principles (rule of thirds, complementary colors, etc.) that lead to aesthetically pleasing shots. The features proposed previously are limited, so there is scope for more comprehensive shortlisting.

Feature Selection: Once a feature set is decided, the hypothesis needs to be tested so as to eliminate those that in reality show no correlation with human ratings, given the data. For feature selection, [11] employs *boosting*, while [4] uses *forward selection*. There is further scope for effective exploitation of correlation across features in aesthetics modeling.

Statistical Learning and Inferencing: A suitable learning method, that makes use of the selected features to model aesthetics, is essential. Previous attempts have employed decision trees [4], Bayesian classifiers [5, 8, 11], SVMs [4, 11], boosting [11], and regression [4, 5], for answering one or more of the questions in Sec. 2. In general, we need some form of regression for score prediction (Sec. 2.1), a two-class classifier for class prediction (Sec. 2.2), and a multi-class discriminative or generative classifier for emotion prediction (Sec. 2.3). Because past efforts have yielded only limited success, a deeper exploration is needed to figure out if feature extraction alone is the performance bottleneck, or whether better learning method can also improve performance.

4. ANALYSIS OF DATASETS

Due to lack of theoretical grounding and controlled experimental data, there is heavy dependence on publicly available data for understanding, development, and validation for this problem, which include Web-based sources [1, 9, 12, 2] that solicit user feedback on image quality and aesthetics. A summary of some sources and the characteristics of available data is presented in Table 1. We collected large samples from each data source, drawing at random, to create real-world datasets (to be available at `http://riemann.ist.psu.edu/`) that can be used to compare competing algorithms. A description and preliminary analysis follows.

Photo.net: This Website [9] provides a platform for photography enthusiasts to share and get their shots peer-rated on a $1-7$ scale on their aesthetic quality. We collected a set of $14,839$ images, each rated by at least one user. The mean number of ratings per image is 12, with a std. dev. of 13. A smaller dataset from this source has been used before [4, 5].

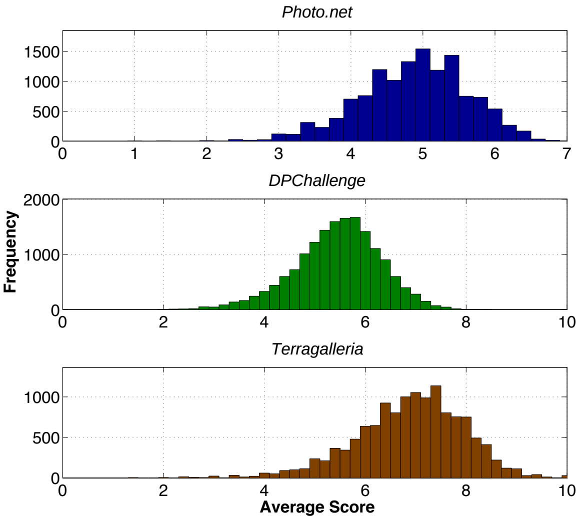

Fig. 2. Distributions of the average scores received.

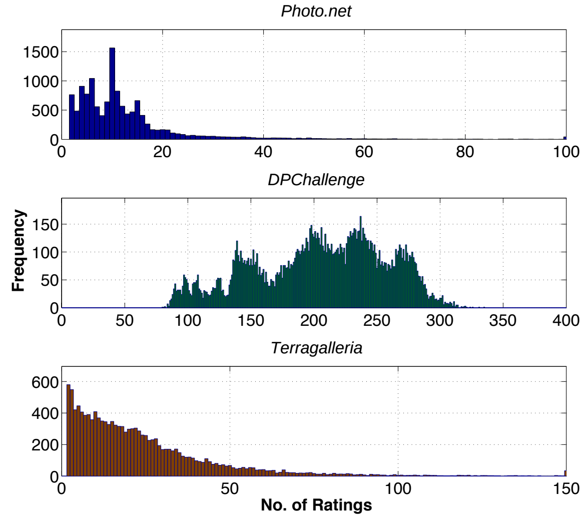

Fig. 3. Distributions of number of ratings received.

DPChallenge: This Website [2] allows users to participate in theme-based photography contests, and peer-rating on overall quality, on a 1-10 scale, determines winners. We collected $16,509$ images, each rated by at least one user. The mean number of ratings per image is 205, with a std. dev. of 53. A smaller dataset from this source has been before [8].

Terragalleria: This Website [12] showcases travel photography of Quang-Tuan Luong, and is one of the best sources of US national park photography. Thus, all photographs are taken by one person (unlike before), but multiple users rate them on overall quality, on a 1-10 scale. The mean number of ratings per image is 22, with a std. dev. of 23. We obtained

$14,449$ images from here, each rated by at least one user.

Alipr: This Website [1], primarily meant for image search and tagging, also allows users to rate photographs on the basis of 10 different emotions (See Fig.6). We collected $13,010$ emotion-tagged images (with repetitions).

4.1. Analysis

For the benefit of experimental design and dataset selection, we report on an analysis of each dataset, in particular the nature of user ratings received in each case (not necessarily comparable across the datasets). Figures 2 and 3 show the average score and score count distributions respectively, of sources [9, 2, 12]. Considering that the three scales are normalized to the same range, DPChallenge ratings are lower, on an average, which might reflect on the competitive nature. For the same reason, the number of ratings received per image are higher than the other two, which indicate that the averaged scores represent the consensus better.

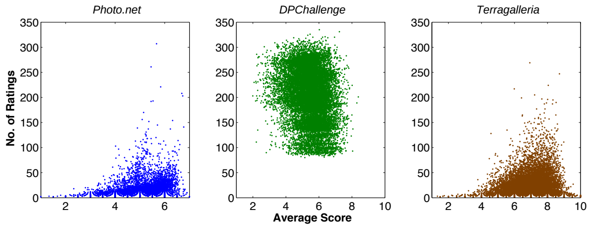

Fig. 4. Correlation plot of (avg. score, no. of ratings) pairs.

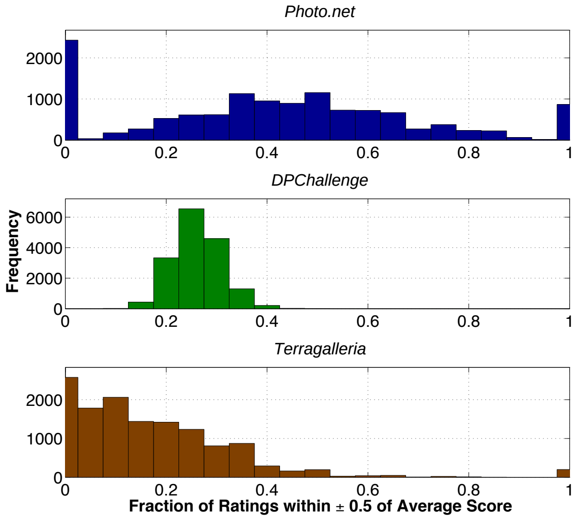

Fig. 5. Distribution of the level of consensus among ratings.

We then look at the correlation between the number of ratings and the average score for each image, by plotting the tuple corresponding to each image, in Fig. 4. Considering uniform random samples, the graphs indicate that in Photo.net and Terragalleria more users rate higher quality photographs, while this skewness is less prominent in DPChallenge. This must be carefully considered when designing inference methods. Another point of interest is consensus, i.e., the extent of agreeability in rating, among users. Let n be the number of ratings given by users, a be their average, and x be the number of ratings within $a \pm 0.5$, with greater value indicating greater consensus. The distribution of x/n over all images is shown in Fig. 5, which roughly indicates that Photo.net has better consensus over the ratings than the other two.

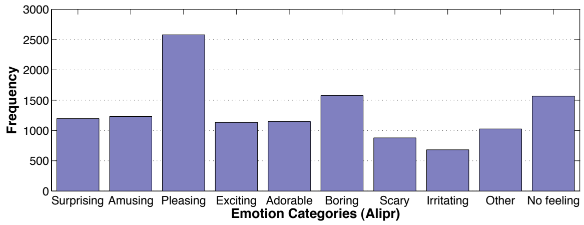

Fig. 6. Distribution of emotion votes given to images (Alipr).

Finally, we plot the distribution of emotion votes for the dataset sampled from Alipr [1]. Despite over $13,000$ votes, the number of them on a per-image basis is low. For higher reliability, we must wait till a greater number of votes are cast.

5. CONCLUSIONS

We have looked at key aspects of algorithmic inferencing of emotions that natural images arouse in people. While very limited work has been published so far, we hope that this exposition to the subtleties will encourage more contributions. We have built and analyzed a few datasets from uncontrolled Web-based sources. Still others, such as *Shutterpoint*, are mushrooming on the Web regularly and can help build more real-world benchmarks. A large, low-noise dataset based on controlled user studies will be a welcome addition.

6. REFERENCES

[1] Alipr, `http://alipr.com`.

[2] DPChallenge, `http://www.dpchallenge.com`.

[3] A.M. Eskicioglu and P.S. Fisher, “Image Quality Measures and their Performance,” *IEEE Trans. Communications*, 45(12):2959–2965, 1995.

[4] R. Datta, D. Joshi, J. Li, and J.Z. Wang, “Studying Aesthetics in Photographic Images Using a Computational Approach,” *Proc. ECCV*, 2006.

[5] R. Datta, J. Li, and J. Z. Wang, “Learning the Consensus on Visual Quality for Next-Generation Image Management,” *Proc. ACM Multimedia*, 2007.

[6] Y. Eisenthal, G. Dror, and E. Ruppin, “Facial Attractiveness: Beauty and the Machine,” *Neural Computation*, 18(1):119–142, 2006.

[7] Hot or Not, `http://www.hotornot.com`.

[8] Y. Ke, X. Tang, and F. Jing, “The Design of High-Level Features for Photo Quality Assessment,” *Proc. CVPR*, 2006.

[9] Photo.net, `http://photo.net`.

[10] H.R. Sheikh, A.C. Bovik, and L. Cormack, “No-reference Quality Assessment using Natural Scene Statistics: JPEG2000,” *IEEE Trans. Image Processing*, 14(11):1918–1927, 2005.

[11] H. Tong, M. Li, H. Zhang, J. He, and C. Zhang, “Classification of Digital Photos Taken by Photographers or Home Users,” *Proc. Pacific Rim Conference on Multimedia*, 2004.

[12] Terragalleria, `http://www.terragalleria.com`.