Journal of Computing and Applied Informatics (JoCAI) Vol. 01, No. 1, 2017 | 13 - 20

DATA SCIENCE
Journal of Computing and Applied Informatics

Subject Bias in Image Aesthetic Appeal Ratings

Ernestasia Siahaan[1], Esther Nababan[2]

[1]Delft University of Technology, Delft, The Netherlands
[2]Universitas Sumatera Utara, Medan, Indonesia

Abstract. Automatic prediction of image aesthetic appeal is an important part of multimedia and computer vision research, as it contributes to providing better content quality to users. Various features and learning methods have been proposed in the past to predict image aesthetic appeal more accurately. The effectiveness of these proposed methods often depend on the data used to train the predictor. Since aesthetic appeal is a subjective construct, factors that influence the subjectivity in aesthetic appeal data need to be understood and addressed. In this paper, we look into the subjectivity of aesthetic appeal data, and how it relates with image characteristics that are often used in aesthetic appeal prediction. We use subject bias and confidence interval to measure subjectivity, and check how they might be influenced by image content category and features.

Keyword: Computational Aesthetics, Subjective Experiment, User Behavior.

Abstrak. *Prediksi otomatis dari suatu daya tarik estetika gambar merupakan bagian penting dalam penelitian multimedia dan computer vision, karena hal ini telah memberikan kualitas konten yang lebih baik kepada penggunanya. Beberapa tahun yang lalu, berbagai fitur dan metode pembelajaran telah diusulkan untuk memprediksi daya tarik estetika gambar secara lebih akurat. Efektifitas metode yang diusulkan acap kali bergantung pada data yang digunakan untuk melatih prediktor. Karena daya tarik estetik merupakan suatu kerangka subjektif, maka faktor-faktor yang mempengaruhi subjektivitas dalam data daya tarik estetik perlu dipahami dan ditangani. Pada tulisan ini, kami melihat pada subjektivitas data daya tarik estetik dan bagaimana kaitannya dengan karakteristik gambar yang sering digunakan dalam prediksi daya tarik estetika. Kami menggunakan subject bias dan confidence interval untuk mengukur subjektivitas dan memeriksa bagaimana parameter tersebut dipengaruhi oleh kategori konten gambar dan fitur.*

Kata Kunci: *Estetika Komputasi, Eksperimen Subyektif, Perilaku Pengguna.*

Received 18 April 2017 | Revised 15 May 2017| Accepted 19 June 2017

1. Introduction

Computational aesthetics is a field of computer vision that aims at understanding an image's (or visual media's) aesthetic appeal or aesthetic quality, such that diverse applications can predict image aesthetic appeal, and when applicable, suggest enhancements accordingly. A large part of research in computational aesthetics has focused on constructing features and training better machine learning algorithms to predict aesthetic appeal quality. This has resulted in diverse sets

*Corresponding author at: Multimedia Computing Group, Intelligent Systems Department, Delft University of Technology, Postbus 5031 2600 GA, Delft, The Netherlands
E-mail address: ernestasia.s@gmail.com

ISSN: 2580-6769 | e-ISSN: 2580-829X | DOI: 10.32734/jocai.v1.i1-63
Journal Homepage: https://talenta.usu.ac.id/JoCAI

of proposed handcrafted features, such as those related with color, composition, texture, or subject region (semantics) [1,2,3,4]. With the emergence of deep networks [5], the research community has also developed various neural networks to tackle the problem of aesthetic appeal prediction in images [6,7,8].

Another part of research in the field is related with collection of images and aesthetic appeal data that can be used as ground truth when creating features and training prediction algorithms. An aesthetic appeal dataset is usually constructed by asking a group of users to provide aesthetic appeal ratings or scores on a set of images, and finally assigning an aesthetic appeal score to each image by averaging the scores over all users for the particular image. One of the most well-known image aesthetic appeal dataset is AVA [9], which contains 250000 images with aesthetic appeal ratings, and semantic labels. Other datasets include those by Schifanella et al. [10], and Geng et al. [11].

Building aesthetic appeal datasets is not a trivial matter, as it raises several issues. In [12], the authors pointed out that the subjective nature of aesthetic appeal data might influence its reliability, as there may be large individual differences. They then conducted an experiment to look into subjective methodologies that would yield more reliable aesthetic appeal data, or minimum individual differences. The authors in [13] raises the issue that most aesthetic appeal datasets do not represent the images that users encounter naturally. Most existing datasets are built based on images collected through photography websites, and thus consist only of high aesthetic appeal and professionally taken photographs.

In this paper, we look into the problem of subjectivity in image aesthetic appeal. Subjectivity refers to the level of consensus that a group of users have on an image's attribute (in this case, aesthetic appeal). A larger subjectivity results in higher uncertainty in the subjective scores attributed to the image, and when machines learn from such scores, it produces inaccurate prediction. Our research question is whether or not subjectivity of image aesthetic appeal is influenced by certain image characteristics. Understanding this problem would give insights to help interpret and improve aesthetic appeal predictions on various images.

This paper is organized as follows. In Section 2, we briefly describe some related work. Section 3 describes the dataset that we use in this study. Section 4 gives the analysis of our data, and we discuss and conclude our results in Section 5

2. Related Work

In this section, we touch upon several work that look into subjectivity and aesthetic appeal features.

Subjectivity in data. Research on perception of visual stimuli (i.e. images and videos) often look into the problem of subjectivity, in order to a) design better subjective methodologies to collect user ratings on diverse stimuli, or b) to better interpret subjective scores data in order to build better models based on said data. We give the following examples of research that has been conducted in this direction. Related with the design of better subjective methodologies, the authors in [12], [14] and [15], looked into the effect of different rating methodologies, and experiment environment on subjective scores of image quality and aesthetic appeal. The study in [16] is an example of work that contributes to better interpreting subjective data. The authors developed a metric that models the relationship between standard deviation of opinion scores (SOS), and mean opinion scores (MOS), as a way to check the level of subjectivity. In the paper, the authors compared the level of subjectivity of quality scores collected for different application domains, which showed that users rate the construct "quality" differently for different application domains.

For this study, we use measures of subjectivity to check how subjectivity relates with different image characteristics. We use the following measures to represent subjectivity in our paper. The first is subject bias, which was proposed in [17]. The term subject bias characterizes user rating behavior, expressing how far off is a user's rating from the true image score value (estimated as MOS). The second measure that we use is confidence interval. We explain the formula used to calculate these terms in Section 4.

Aesthetic appeal features. As mentioned in Section 1, various features have been proposed to use in aesthetic appeal prediction; most relate to the image's color, texture, composition, or subject region [1,2,3,4]. Deep networks have also been explored to predict image aesthetic appeal [6,7,8], and in this case do not require prior designing of features. The networks will automatically learn discriminative features from the data that can predict image aesthetic appeal more accurately. In this paper, we do not use features constructed through a neural network, as we would like to maintain interpretability of the features we use in our analysis.

3. Data Set

In this section, we describe the experiment with which we collected our ground truth data, i.e. aesthetic appeal ratings. Figure 1 shows a subset of the images we used in our experiment. 79 pristine natural photographs were taken from the LabelMe image dataset [18], spanning a wide range of semantic content category (indoor and outdoor scenes, as well as humans and non-human objects). The images were 1024 x 768 in size. We then conducted a subjective experiment in

which we asked a group of users to rate the images based on their aesthetic appeal. At the beginning of the experiment session, each user was given a brief explanation of their task. We instructed them to rate each image based on their aesthetic appeal, and defined aesthetic appeal as how beautiful or pleasing an image is to the eyes. Users were then given a training session, in which they could give a rating to one example image, to familiarize themselves with the interface and rating scale used for the task. After the training session, users could then continue with the main task. One image was presented at a time, at random order, and users could observe the image without any limit in viewing time. When the user is ready to rate the image, he/she could indicate their rating of the image's aesthetic appeal on a discrete 5-point ACR-labeled rating scale below the image. In total, we had 19-20 individual ratings of aesthetic appeal for each image in the set.

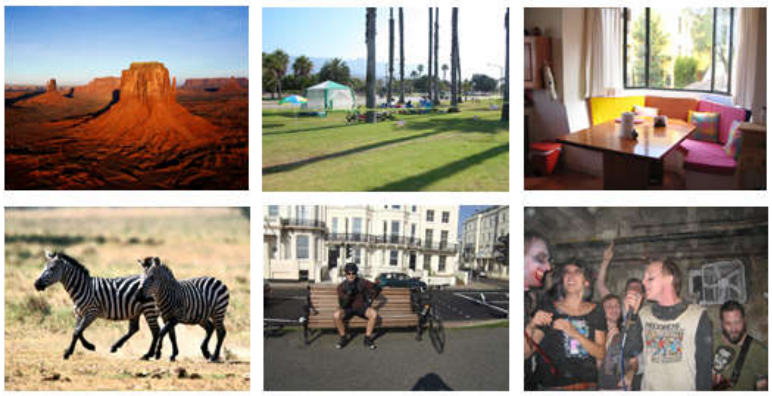

Figure 1. Example of images used in our experiment, taken from the publicly available LabelMe image dataset []

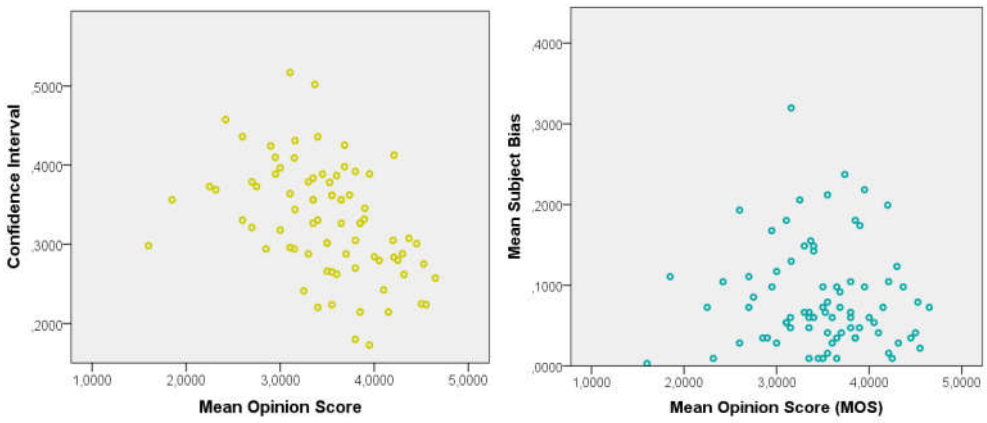

Figure 2. Scatter plots of mean opinion scores (MOS) against subjectivity measures (left: confidence interval, right: subject bias)

4. Data Analysis

In this section, we present the results of our analysis on the data in order to answer our research question. Before performing any analysis, we firstly calculated the mean opinion score (MOS) of image aesthetic appeal for all images in the dataset. We do this by taking the average of the

individual ratings given for each image. We then calculated the mean subject bias for each image, as follows.

$$OS(i,n) = MOS(i) + \Delta_n + \varepsilon_{i,n} \quad (1)$$

OS(i,n) indicates the opinion score or rating that user n gives to image or stimuli i. MOS(i) indicates the mean opinion score or average rating across all users for image or stimuli i. Δ_n represents the term subject bias of user n, and $\varepsilon_{i,n}$ is the error term for image i by user j.

After calculating the mean subject bias for each image, we also calculate the confidence interval for each image as follows.

$$CI(i) = MOS(i) \pm (1.96 * (SOS(i) / \sqrt{N})) \quad (2)$$

Here, CI(i) indicates the confidence interval for image or stimuli i, while MOS(i) and SOS(i) indicate the mean opinion score and standard deviation of opinion score for stimuli i, respectively. N refers to the total number of users that rated stimuli i.

We describe the analysis that we performed on our data as follows:

4.1 Aesthetic Appeal Rating and Subjectivity

In Figure 2, we show scatter plots of the aesthetic appeal ratings against the subject bias or confidence interval. The figures show that subjectivity does not depend on the position of the stimuli in the aesthetic appeal scale. This is different from subjectivity in quality assessment, as typically subjectivity becomes smaller at the end of the quality scale ([16,17]), showing that users have higher agreement in determining what is a low quality and bad quality stimuli, but lower agreement in determining what constitutes a medium quality stimuli. For aesthetic appeal, such is not the case, as the confidence interval and subject bias values can also be low even for images at the mid-range of the scale. Given this observation, we are interested to see whether or not the values of subjectivity depend on other factors related with the image content itself, for example, the image's semantic content category, colorfulness or texture.

4.2 Semantic Category and Subjectivity

The first image characteristic that we looked into in relation with subjectivity is semantic category. The 79 images in our dataset were annotated by 5 annotators based on their scene and object. The annotators were asked to categorize each image into one of three scene categories: indoor, outdoor natural, or outdoor manmade; and one of two object categories: animate, or inanimate. Animate object refers to human and animals, while inanimate objects refer to those outside of the animate category.

Figure 3 plots the mean subject bias of images belonging in different scene and object categories. We can see from the plots that the subjectivity differs across the different semantic content

categories. People seem to have more agreement in the aesthetic appeal of outdoor natural images compared to other scene categories. Furthermore, people seem to have more agreement in the aesthetic appeal of images with inanimate objects in it.

Table 1 Level of statistical significance for each feature on predicting subject bias and confidence interval through multiple linear regression

	Statistical Significance in Predicting Subjectivity (p value)														
	f_c	f_{t1}	f_{t2}	f_{t3}	f_{t4}	f_{t5}	f_{t6}	f_{t7}	f_{t8}	f_{t9}	f_{t10}	f_{t11}	f_{t12}	f_{t13}	f_{t14}
Subject bias	0.816	0.192	-	0.187	-	0.911	0.542	0.514	0.54	0.42	0.934	0.485	0.637	0.06	0.2
Confidence Interval	0.004	0.737	-	0.024	-	0.024	0.78	0.748	0.42	0.3	0.11	0.219	0.313	0.16	0.44

We conducted some statistical tests to check whether or not the differences shown through these plots are statistically significant. We conducted a one-way ANOVA test for the subject bias across scene categories, and a Student T-test for the subject bias across object categories. Our ANOVA test shows us that there is no significant difference of subject bias among the three different scene categories (df=2, F=1.344, p=0.267). Our T-test also shows no significant difference between the two object categories (t=1.687, p=0.096). These results show that although there is a difference in subject bias across different semantic content categories, these differences are not statistically significant. Interestingly, when we perform the statistical tests for confidence interval measure, our results show that there is significant difference between the different scene categories.

4.3 Aesthetic Features and Subjectivity

In this section, we check how subjectivity may be influenced by image features that are commonly associated with aesthetic appeal. The first feature that we look into is color features. We calculate the mean hue, saturation and brightness (HSV) across the whole image, and observe how it influences subject bias. We refer to this feature as f_c. The second feature that we look into is image texture. We choose to use Haralick features [19] to represent texture in images, and refer to this feature as $f_{t1\text{-}14,}$ with numbers 1-14 indicating the 14 Haralick features extracted for each image.

To check whether or not these features have an influence on subjectivity of user image aesthetic appeal ratings, we conducted a multiple linear regression, in which these features act as independent variables, and subjectivity (subject bias or confidence interval) as dependent variables. Table 1 shows the statistical significance of each feature in predicting subjectivity. From the table, we see that none of the features are statistically significant in predicting subject bias. However, the color feature f_c (mean HSV), and texture feature f_{t5} (inverse different moment/homogeneity) are shown to be statistically significant in predicting confidence interval, with p<0.05.

5. Discussions and Conclusion

This paper looks into the relationship of image characteristics with the level of subjectivity on the image's aesthetic appeal. We define subjectivity as the level of consensus among users regarding a certain construct related with a stimuli/image (in this case, aesthetic appeal). To measure subjectivity, we use the terms subject bias [17], and confidence interval. We look into whether or not the level of subjectivity for image aesthetic appeal depends on the image's semantic content category (the scene and object category of the image), hue, saturation and brightness level, and texture.

Our analysis shows that although there seems to be a difference in subject bias across scene and object categories, these differences are not statistically significant. However, the difference in confidence interval is statistically significant across scene categories. For color and texture features, we obtain similar observations. For subject bias, the influence of color features and texture features are not statistically significant. However, for confidence interval, the color feature (mean HSV) and homogeneity texture feature are shown to be statistically significant predictors. This raises the question of which measure should we use to represent subjectivity of image aesthetic appeal, if we want to check image characteristics which gives higher discrepancy in user agreement of aesthetic appeal assessment.

The limitation of this paper is that it still does not give a complete picture what creates subjectivity in image aesthetic appeal assessments, as it only looks into a few characteristics, namely semantic content, color and texture. There are other characteristics that can be looked into, such as image composition, depth, etc. Moreover, the features that we use to look into the semantic content, color, and texture here do not cover all aspects of semantics, color and texture. These may be looked into in future work.

REFERENCES

[1] R. Datta, D. Joshi, J. Li, and J. Z. Wang, "Studying aesthetics in photographic images using a computational approach", in European Conference on Computer Vision, pp. 288–301, 2006.

[2] X. Tang, W. Luo, and X. Wang, "Content-Based Photo Quality Assessment", in IEEE Trans. on Multimedia, vol. 15 no. 8, pp. 1930-1943, December 2013.

[3] C. Li, A. C. Loui, and T. Chen, "Towards aesthetics: a photo quality assessment and photo selection system", in Proc. of 18th ACM International Conference on Multimedia, pp. 827-830, October 2010.

[4] Y. Luo, and X. Tang, "Photo and video quality evaluation: focusing on the subject", in European Conference on Computer Vision, pp. 386-399, 2008.

[5] Y. LeCun, and Y. Bengio, "Convolutional networks for images, speech, and time series", in The Handbook of Brain Theory and Neural Networks, vol. 3361, no. 10, 1995.

[6] X. Lu, Z. Lin, H. Jin, J. Yang, and J.Z. Wang, "RAPID: Rating pictorial aesthetics using deep learning", in Proc. of the 22nd ACM International Conference on Multimedia, pp. 457-466, 2014.

[7] L. Mai, H. Jin, and F. Liu, "Composition-preserving deep photo aesthetics assessment", in Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 497-506, 2016.

[8] Y. Kao, R. He, and K. Huang, "Deep aesthetic quality assessment with semantic information", in IEEE Trans. on Image Processing, vol. 26, no. 3, pp. 1482-1495, January 2011.

[9] N. Murray, L. Marchesotti, and F. Perronnin, "AVA: A large-scale database for aesthetic visual analysis", in Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2408-2415, 2012.

[10] R. Schifanella, M. Redi, and L. M. Aiello. "An image is worth more than a thousand favorites: surfacing the hidden beauty of Flickr pictures", in Proc. of the 9th International AAAI Conference on Web and Social Media, 2015.

[11] B. Geng, L. Yang, C. Xu, X.-S. Hua, and S. Li. "The role of attractiveness in web image search", in Proc. of the 18th ACM International Conference on Multimedia, pp. 63-72, 2011.

[12] E. Siahaan, A. Hanjalic, and J. Redi, "A reliable methodology to collect ground truth data of image aesthetic appeal", in IEEE Trans. on Multimedia, vol. 18, no. 7, pp. 1338-1350, 2016.

[13] M. Redi, F. Liu, and N. O'Hare, "Bridging the gap: the wilde beauty of web imagery", in Proc. of the 2017 ACM International Conference on Multimedia Retrieval, pp. 242-250, 2017.

[14] M. H. Pinson, L. Janowski, R. Pepion, et al, "The influence of subjects and environment on audiovisual subjective tests: an international study", in IEEE Journal of Selected Topics in Signal Processing, vol. 6, no. 6, October 2012.

[15] Q. Huynh-Thu, M.-N. Garcia, F. Speranza, P. Corriveau, and A. Raake, "Study of rating scales for subjective quality assessment of high-definition video", in IEEE Trans. on Broadcasting, vol. 57, no. 1, March 2011.

[16] T. Hossfeld, R. Schatz, and S. Egger, "SOS: the MOS is not enough!", in IEEE 3rd International Workshop on Quality of Multimedia Experience (QoMEX), pp. 131-136, 2011.