proceedings

MDPI

Proceedings

Dataset for the Aesthetic Value Automatic Prediction †

Nereida Rodriguez-Fernandez *, Iria Santos and Alvaro Torrente

Department of Computer Science, Faculty of Computer Science, University of A Coruña, 15071 A Coruña, Spain; iria.santos@udc.es (I.S.); alvaro.torrente@udc.es (A.T.)

* Correspondence: nereida.rodriguezf@udc.es

† Presented at 2nd XoveTIC Conference, A Coruña, Spain, 5–6 September 2019.

Published: 1 August 2019

Abstract: One of the most relevant issue in the prediction and classification of the aesthetic value of an image is the sample set used to train and validate the computational system. In this document the limitations found in different datasets used to classificate and predict aesthetic values are exposed, and a new dataset is proposed with images from the DPChallenge.com portal, with evaluations of three different populations.

Keywords: dataset; aesthetics; quality; prediction; classification; artificial intelligence; assessment

1. Introduction

Different research groups have tried to create computer systems capable of learning the aesthetic perception of a group of human beings as part of a generative system, with the intention of being used in the selection or automatic ordering of images. Due the subjective nature of the aesthetic problem, the selection of the dataset with which the system is trained is especially relevant. After analyzed, in previous research [1,2], the generalization degree of some datasets, it has been concluded that it is not enough to take them as a reference in the training of automatic image classification and prediction systems. In order to providing a solution to the problems detected, this paper describes the creation of a new dataset from the DPChallenge.com portal, with greater statistical coherence. In addition, this new dataset has been evaluated according to aesthetic and quality criteria by a human group in controlled experimental conditions and by another American group through online surveys.

2. Limitations Found in the Datasets Available

There are some datasets that have been used in several times for the images classification. Among them, Photo.net [3–5], DPChallenge.com [6,7] and the one created by Cela-Conde et al. [8–10] However, when its generalization capacity is studied, it has been detected that they cannot be considered as representative for the realization of image experiments. In some cases, the correlation is greater when the validation set belongs to the same data source as the training set, and this correlation drops markedly when the validation source set is different from that of the training. In addition, the sample sets trained with evaluations from the photographic portals have some defects: the evaluation system does not have the same control as a psychological test because it is not possible to obtain all the information about the evaluators or about the evaluation conditions; the number of images could be insufficient, since there is no justified reason to choose a sample size and there is a very high difference between the number of people who value each image; user ratings can be easily conditioned by personal tastes, personal relationships with the work creator, or by the momentary boom or popularity of certain styles. Lastly, in one of the cases [3] it has been shown that the users of these portals do not have sufficient grounds to differentiate between aesthetics and

originality criteria, with a Pearson correlation coefficient of 0.891. In the dataset created by Ke et al. [6] is another limitation: the web portal DPChallenge.com works as a photo contest and does not specify any criteria to evaluate the images with their own judgment and nothing related with that of other users. On the other hand, in the dataset created by Cela-Conde et al. [8] the number of images presented by category is not equitable, so the results obtained cannot be considered as representative of the set. In addition, it part of a considerable amount of subsets of images, which results in the dataset is eventually converted into several independent datasets, smaller and with less internal consistency.

3. A New Dataset

After the detection of the limitations described above, the construction of a new dataset for the aesthetic prediction of images has been carried out. This new creation method allows us to build a dataset with greater statistical coherence from the evaluation results collected on the DPChallenge.com photography website. Later, it is evaluated by two different population types. With this, we obtained the possibility of analyzing the correlation between the results obtained with subjects in controlled circumstances and those obtained through online surveys. First, a set of images has been compiled from the DPChallenge.com photo portal. This portal has been used previously to obtain data for aesthetic classification experiments [6,7]. Those images with a minimum of 100 ratings have been selected. In this way, it is intended that the average value that will be assigned to each image will be as little biased as possible. Once this selection is made, the images are organized in groups according to the average evaluation received in DPChallenge.com. The images of our selection have been classified in 9 scoring ranges, one for each whole value of evaluation allowed. Then, all groups are expected to have a minimum number of images, which in our case was 200. There are not sufficiently large groups of images with average evaluations lower than 3 or higher than 8, so the groups used were those collected in the range [≥3, <8]. Of these groups, the 200 images with the smallest standard deviation were selected, that is, those that present votes with greater internal coherence. This process provides a set of images with the same number of elements in each range and with high voting coherence.

4. Evaluation

The dataset proposed above was evaluated by a group of Spanish humans under controlled experimental conditions. The evaluations were carried out by student volunteers from the Universidade da Coruña, Spain. Ninety-nine participants (33 men and 66 women) were part of this study, with an average age of 18.7 years, in an age range of 18–30. Each participant evaluated at least 200 images in the members of the research group presence and under the same viewing conditions. For each image, users assessed their aesthetics and their quality independently. Later, another experiment was conducted through online surveys with the USA population. This experiment was carried out through the Amazon Mechanical Turk tool. 525 people evaluated the images, 39% men and 61% women, with an average age of 32.6 years, in a range of 18–70 years. The same images were used as in the on-site experiment and the evaluators had to score, in the same way, the aesthetics and quality criteria, independently.

5. Results

The correlation between the evaluations made in person and those recorded on the DPChallenge.com platform has been calculated. The Pearson correlation between the average score of DPChallenge.com and the average evaluation according to the aesthetic value is 0.692, and of 0.69 according to Spearman. The average correlation between DPChallenge.com and the average according to the quality value is 0.748 according to Pearson and 0.756 according to Spearman. Finally, the correlation between the two measurements obtained in the on-site experiment (aesthetics/quality) is 0.787 according to Pearson and 0.786 according to Spearman. When it analyzes the correlation between the on-site evaluations and the USA online survey, a correlation of 0.76 was

detected between the aesthetic criteria of both experiments, and of 0.85 between the quality criteria. The correlation between the aesthetic and the quality criteria in the USA evaluations is 0.89, the same correlation that exists between criteria in the experiment carried out by Datta et al. [3] With this new dataset, different models based on Machine Learning have been trained using different metrics for automatic prediction of aesthetic and quality value. The highest correlation obtained with these models is 0.58 using SVM [11].

6. Conclusions

The correlation results suggest that the evaluation of DPChallenge.com is closer to a quality criteria than aesthetics and that, in the same way, all evaluators coincide with greater precision when evaluating the quality criteria than aesthetics. In addition, it can be deduced that the evaluators better differentiate the criteria to be evaluated when the difference can be explained to them in person. It should be noted that the complex systems used predict better quality results than aesthetic ones, perhaps due to their lower subjective component and their greater relationship with the intrinsic characteristics of the images.

Acknowledgments: This work is supported by the General Directorate of Culture, Education and University Management of Xunta de Galicia (Ref. GRC2014/049) and the European Fund for Regional Development (FEDER) allocated by the European Union, the Portuguese Foundation for Science and Technology for the development of project SBIRC (Ref. PTDC/EIA– EIA/115667/2009), Xunta de Galicia (Ref. XUGA-PGIDIT-10TIC105008-PR) and the Spanish Ministry for Science and Technology (Ref. TIN2008-06562/TIN). We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan Xp GPU used for this research.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Carballal, A.; Castro, L.; Rodríguez-Fernández, N.; Santos, I.; Santos, A.; Romero, J. Approach to minimize bias on aesthetic image datasets. In *Interface Support for Creativity, Productivity, and Expression in Computer Graphics*; IGI Global: Hershey, PA, USA, 2019; pp. 203–219.
2. Carballal, A.; Castro, L.; Perez, R.; Correia, J. Detecting bias on aesthetic image datasets. *Int. J. Creat. Interfaces Comput. Graph.* **2014**, 5, 62–74.
3. Datta, R.; Joshi, D.; Li, J.; Wang, J.Z. Studying aesthetics in photographic images using a computational approach. In *European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2006; pp. 288–301.
4. Wang, W.; Cai, D.; Wang, L.; Huang, Q.; Xu, X.; Li, X. Synthesized computational aesthetic evaluation of photos. *Neurocomputing* **2016**, *172*, 244–252.
5. Wong, L.K.; Low, K.L. Saliency-enhanced image aesthetics class prediction. In Proceedings of the 2009 16th IEEE International Conference on Image Processing (ICIP), Cairo, Egypt, 7–10 November 2009; pp. 997–1000.
6. Ke, Y.; Tang, X.; Jing, F. The design of high-level features for photo quality assessment. In Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06) 2006, New York, NY, USA, 17–22 June 2006; Volume 1, pp. 419–426.
7. Tang, X.; Luo, W.; Wang, X. Content-based photo quality assessment. *IEEE Trans. Multimed.* **2013**, *15*, 1930–1943.
8. Cela-Conde, C.J.; Ayala, F.J.; Munar, E.; Maestú, F.; Nadal, M.; Capó, M.A.; del Río, D.; López-Ibor, J.J.; Ortiz, T.; Mirasso, C.; et al. Sex-related similarities and differences in the neural correlates of beauty. *Proc. Natl. Acad. Sci. USA* **2009**, *106*, 3847–3852.
9. Forsythe, A.; Nadal, M.; Sheehy, N.; Cela-Conde, C.J.; Sawey, M. Predicting beauty: Fractal dimension and visual complexity in art. *Br. J. Psychol.* **2011**, *102*, 49–70.
10. Nadal, M.; Munar, E.; Marty, G.; Cela-Conde, C.J. Visual complexity and beauty appreciation: Explaining the divergence of results. *Empir. Stud. Arts* **2010**, *28*, 173–191.

11. Carballal, A.; Fernandez-Lozano, C.; Rodriguez-Fernandez, N.; Castro, L.; Santos, A. Avoiding the inherent limitations in datasets used for measuring aesthetics when using a machine learning approach. *Complexity* **2019**, *2019*, 4659809.



© 2019 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).