

Aesthetic Evaluation of Food Plate Images using Deep Learning

Veranika Mikhailava, Evgeny Pyshkin, Vitaly Klyuev

Software Engineering Lab, University of Aizu, Aizuwakamatsu, Japan
 m5231149@u-aizu.ac.jp, pyshe@u-aizu.ac.jp, vkluev@u-aizu.ac.jp

Abstract— This paper contributes to the experiments on using machine learning algorithms for image aesthetic evaluation, with particular focus on food plate image evaluation. Such an evaluation may be beneficial for professional and amateur cuisine makers, restaurant critiques, photographers, and travelers. We use a convolutional neural network at the training and recognition stages. The proof-of-concept experiments are arranged using our marked dataset of food plate images. Earth Mover's Distance algorithm is used for process validation. The analysis of results shows that the trained system predicts food image aesthetic value conforming to the collected human expert evaluations with good quality.

Keywords— image aesthetics, evaluating food plates, aesthetic assessment, deep learning, aesthetic scoring.

I. INTRODUCTION

The relationship between the Good and the Beautiful is a constant subject of philosophical examination since ancient times. Though Aristotle clearly differentiated these two concepts in his *Metaphysics* [1], currently, on the level of human perception of the surrounding world, many people often consider the beautifulness of a thing as an important contribution to its quality properties. Though sometimes the aesthetic value becomes a contemporary “surrogate for beauty” [2], in a broader sense, aesthetics remains to be an important aspect of the human critical reflection on art, beauty, and good taste. According to Kant, natural dualism of aesthetic judgments is conditioned by two important kinds of beauty introduced in *The Critique of Judgement* [3]: the first does not presuppose some concept what the object should be (according to the translation [4], it is called **free** or **self-subsisting beauty**), while the second is attached to such a concept (**dependent** or **conditioned beauty**). Thus, the dependent beauty is more connected to beauty objective nature (we can explain what makes an object beautiful), while the free beauty appeals to a viewer's interpretation based on one's taste, which is “subjective and varies according to class, cultural background and education” [5].

In recent decades, the analysis of aesthetics issues does not attract particular attention from philosophers only but also from the technical and engineering community. We know the works on analysis of software code aesthetics with respect to its quality characteristics [6], the efforts to introduce aesthetic software properties to learners of programming [7], the investigations on the role of aesthetics in creating successful engineering products [8], as well as the insights to aesthetics of computer system user interfaces [9], [10], to cite a few.

Attempts to go towards more formal ways to aesthetic evaluation are known since around a century ago, when Birkhoff (1928) defined the aesthetic measure M as the ratio between the order O and complexity C (hence, $M=O/C$) [11]. Though this equation could not give a direct implementable computational model for aesthetic judgments, it provides an interesting insight into our understanding of what makes the object aesthetically attractive. It is generally acknowledged that in many consumer industries, product beauty may significantly affect our decisions as clients. For example, travelers may prefer one hotel because of beautiful pictures posted on booking sites; they may select a restaurant to eat just because of beautiful pictures of served food; successful travel itineraries are often based on approaches to evaluate the attractiveness of the suggested points of interest. There could be many components affecting the place attractiveness, including its beauty, its sounds, and even its smells.

Similarly, in his book “Gastrophysics. The New Science of Eating”, Spense discusses many ways that might influence our perception of food, including our mood, surrounding sound and view, and, of course, how the plate looks like [12]. Thus, visual perception could not be underestimated. We have to admit that many users of social networks like posting the images of food they eat, regardless of the fact whether this practice characterizes society positively or not. Some of such pictures are beautiful, while some – definitely not. Of course, food plate aesthetic judgments may be as subjective as other aesthetic evaluations. However, it seems interesting to investigate, is there a practical possibility to rate and classify food plate images according to our taste of beauty.



Figure 1. Example of food plates

Applied to the context of food aesthetic value, the Kant's definition of the Beautiful sounds particularly impressive: "The beautiful is that which, apart from a concept, is cognized as an object of a necessary delight" [4]. Reinterpreting the list of traditional questions asked by aestheticians [5], we can "exactify" the possible research questions for the case of food plate aesthetic evaluation: Does the plate beauty make a dish more successful apart from its taste? Why do we find certain plates beautiful? Are aesthetic judgments merely subjective, and what could make them objective and explainable? Finally, can an aesthetic judgment be improved or trained because of technology-dependent solutions?

II. RELATED WORK

We deal with images, so it is quite natural to study the applicability of general approaches used for image classification and quality evaluation. Since 2012, when a convolutional neural network (CNN) AlexNet designed by Krizhevsky and Hinton showed the best results in the classification and object detection competition ImageNet Challenge organized by Stanford lab [13], the interest to applying deep learning algorithms to different areas of knowledge dramatically increased. Among these areas, the problems of image classification, localization, object detection, and instance segmentation are in the first places of machine learning applications.

Many researchers working on computer vision and image recognition algorithms need large public datasets that can help researchers to compare the effectiveness of their models against others. ImageNet is one of the standard collections used for research purposes [14]. It contains about 14 million images organized in 20,000 categories. What is particularly interesting is that the features of models trained on ImageNet for classification tasks can be fine-tuned to serve a wide variety of possible tasks. Thus, the model can be adjusted to be more relevant to a particular problem. However, analysis of image aesthetics needs special efforts, since different image categories might require completely different approaches to decide on objects' beauty: the models used to evaluate the aesthetic quality of landscape compositions do not fit well the task of food plate aesthetic assessment, though there are some common elements in process workflow organization. In [15], the authors present a comprehensive experimental survey summarizing different approaches and image aesthetic assessment evaluation criteria. The authors evaluate the conventional approaches with hand-crafted features against the recent deep learning-based models. With a growing amount of available training data (that can be used due to currently available powerful computational resources), deep learning algorithms win over conventional approaches both by efficiency and accuracy.

In [15], the authors pointed out an important, challenging issue of image aesthetics assessment, which is about learning category-specific image aesthetics (such as food images, in our case), which requires more specific auxiliary information about the data. Within the context of presenting an approach to define structural cues responsible for image aesthetic properties, the exhaustive study [16] presents a number of important

application areas where photo aesthetic evaluation may be required. These areas include image-based information retrieval and searching, photo post-processing, photography education, visual art, and design. Particularly, design issues are not limited by only graphic design but include the environmental and digital design as well (such as selecting images to decorate an interior or to include into a web-based travel diary for sharing memorable travel experience). Good, aesthetically attractive, and responsive images make a strong impact on successful design solutions for both public and private spaces.

In turn, food plate images (though they seem to belong to a simplified type of photography) have a number of important specific properties: the photos are usually taken from specific angles, much attention is focused on details, a certain number of symmetrical patterns are used to plate the elements.

We aren't pioneering the works on food image processing. However, existing research is mostly focused on food detection [17], making possible further digitalization of approaches to better consumption and nutrition management [18] [19], but not on food image aesthetic properties.

Concerning image aesthetics, we have to mention the Aesthetic Visual Analysis (AVA) dataset containing over 250,000 images, which is one of the known datasets used for image aesthetic assessment tasks [20]. Though this dataset is much smaller compared to ImageNet, it contains important meta-data, for example, for each image, there is its aesthetic score rated by professional and amateur photographers. In addition, this dataset provides three types of image annotations:

- aesthetic annotations based on expert votes,
- semantic annotations based on textual tags,
- photography style annotations based on shooting modes and camera configurations.

The AVA photos have been scored by around 200 experts per each image, in response to nearly 900 photography contests. As a result, every image from this collection is associated with a single challenge theme, and the image ranks range between 1 to 10, the latter being the highest score. Mean ratings are concentrated around the overall mean score (≈ 5.5). Among all images, AVA includes about 7,000 food and drink images. However, these images are very general and do not fit well our problem definition of food plates.

Some researchers on image aesthetics try to solve an easier task of evaluating an image using the binary score (beautiful or not) [15]. For the AVA dataset, they decide the borderline of the mean score: e.g., if the score is greater than 5, the image is supposed to be beautiful, otherwise, it is not.

In a strong connection to the approach used in our work, we have to mention the notable paper "NIMA: Neural Image Assessment" [21]. NIMA relies on the success of state-of-the-art deep neural networks for object recognition. Instead of experimenting with custom multicolumn network architectures, the authors suggest focusing on the evaluation criteria (e.g., which loss function is used) and the use of pre-trained models. The authors used deep CNN trained with the AVA data to predict the distribution of aesthetic value ratings (ranged between 0 to 10) for specific classes of images (landscapes, in

their experiments). The experiments demonstrated a high correlation to human perception, making it possible to use this model for developing automated tools for resource-intensive and substantially subjective tasks such as photo editing and visual quality optimization. In the NIMA approach, instead of classifying images a low/high score or by regressing to the mean score, the model produces a distribution of ratings for any given image. The result of the evaluation process is not only the average score but an assignment of probabilities to each of the possible ten scores. It is more directly in line with how training data is typically captured, and it turns out to be a better predictor of human preferences when measured against other approaches.

To sum up, existing approaches to image aesthetics evaluation based on deep learning algorithms seem very promising within the scope of representing subjective human perception, and we can rely on models similar to NIMA in our work for a specific context of food plate aesthetic evaluation.

III. APPROACH

A. Problem Statement and Constraints

In our research, we limit the food images by the pictures representing the plate with a dish (preferably viewed from above). Since there is no existing dataset containing meta-data with rated food plate images, for early design proof-of-concept assessment, we collected a small dataset consisting of 3497 food plate images gathered from social networks, friends' feeds, and public resources. To mark up the dataset, we asked three experts to evaluate the images on the 5-star scale, which is probably the easiest and the most popular human-oriented evaluation scale with the following grades: 1 – Poor; 2 – Fair; 3 – Good; 4 – Very Good; 5 – Excellent. Table 1 summarizes a number of pictures falling into each category based on average evaluation score rounded to the closest integer.

Table 1 helps to see the general distribution of scores. However, the neural network uses a normalized distribution of votes per each image, not simply a single number label. Thus, the distribution based label is, in fact, a vector representing normalized probabilities for each score: e.g., if two experts evaluate the image as 4, and one expert gives 3, the label vector is $[0, 0, 0.33, 0.66, 0]$.

TABLE 1. NUMBER OF IMAGES FALLING INTO EACH CATEGORY BASED ON AVERAGE EVALUATION SCORE

Categories	Pictures
1 - Poor	647
2 - Fair	585
3 - Good	788
4 - Very Good	825
5 - Excellent	652
Total:	3497

The problem we want to investigate is as follows. Assume we have a labeled dataset with normalized scores, as described earlier. We want to train a convolutional neural network that

would predict the distribution of votes as close as possible to human opinions for an image that the network has not seen before. Then we developed a web application where we integrated the evaluation system on the client-side in the browser. The system scores the images according to the pictures represented in the dataset.

B. Deep CNN Training

We randomly divided our data into training (80%) and test set (20%), which is a popular choice for small to medium datasets. We used two architectures to compare: MobileNet [22] and Inception-v3 [23], both initialized with ImageNet weights. These architectures are known as recommended for image classification tasks because of their top results on the Large Scale Visual Recognition Challenge [24]. Most of the CNNs resize the image to 224x224 pixels. That is why our web application automatically resizes all the uploaded images to 224x224. During training, we randomly allocated 10% of the training set for validation purposes. As suggested in [21], we use Earth Mover's Distance (EMD) as the loss function to compare the similarity between the prediction and human evaluation, and we confirm that it gives better performance. We used Keras and Tensorflow as our main backend for training the models. We trained the model using NVIDIA TITAN Xp GPU and 32 GB of RAM.

What is particularly important about CNN is that, unlike traditional machine learning approaches, a CNN could not be considered as a black box, since it allows visualizing the intermediate layers and, therefore, getting some feedback that can be used to explain the decisions and the result of the prediction. However, we have to admit that the problem of representing this feedback to end-users in a user-friendly and understandable form is challenging and needs further investigations and design efforts. Generally speaking, the problem of tailored feedback is one of the serious issues of digitalization, when unobvious technology solutions are applied to new domains of human-centric computing [25].

C. Research Software

To arrange a convenient experimentation environment, we developed a web application enabling uploading food plate images to the evaluation system to get the aesthetics scores. Figure 2 illustrates the output of our experimentation environment after the user uploaded a photo for the evaluation. The evaluation is done on the client-side in the browser. Once the model is loaded into the memory, the processing time for one picture is about 100 milliseconds on the mid-range laptop with Intel integrated graphics.

It is important to mention that before sending the picture to the aesthetics assessment module, the application checks whether there is a food plate on the picture. To train the recognition model (a binary classification task), we combined the collected food plate images with an equal number of pictures without food plates from public datasets (ImageNet, Faces, Landscapes). In total, we had 6994 images for the recognition task.



Figure 2. Example of food plates scored by the developed system

IV. EXPERIMENTAL RESULTS

Figure 3 illustrates the results of aesthetic predictions using MobileNet architecture, chosen for the reason that it can be directly used on end-user devices (which are not equipped with expensive, powerful GPUs). Figure 4 represents the EMD loss achieved using Inception-v3 architecture. From the diagrams, we can see that they both show similar performance, although Inception-v3 has slightly better results. However, let us note that the resulting model for the MobileNet model's file size is four times smaller and decreases network load.

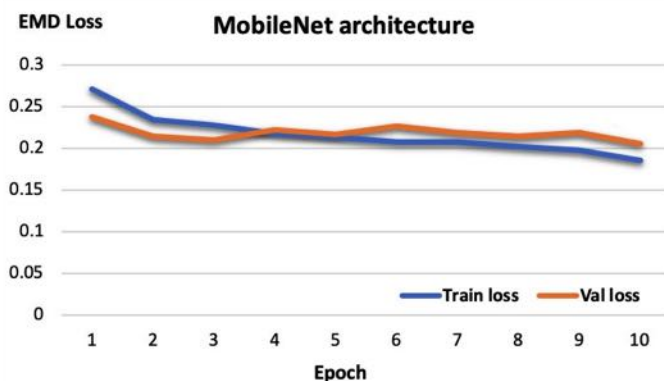


Figure 3. Earth Mover's Distance loss using MobileNet architecture

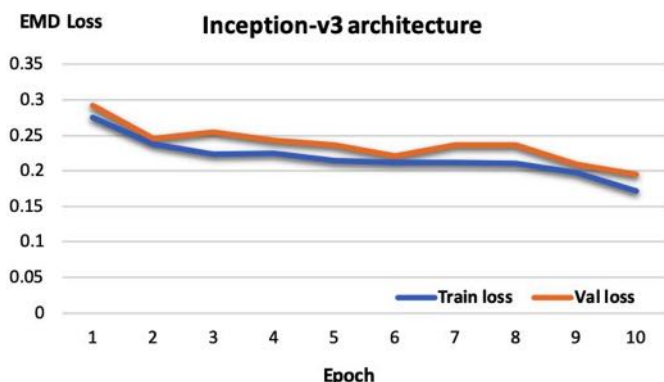


Figure 4. Earth Mover's Distance loss using Inception-v3 architecture

V. CONCLUSIONS

For the question of whether aesthetic judgments can be improved or trained because of technology-dependent solutions, we can go to the following conclusions. Currently, our experiments only demonstrate that using CNN deep learning may help in predicting the food image aesthetic value, which is in good correlation with the human judgments used for CNN training, whatever the initial quality of these judgments is. As soon as the publicly available dataset annotation is made with involving high-quality professional experts, we may expect that the tools for food plate aesthetic evaluation may help interested users to improve and evaluate their taste and experience against the expert judgments.

We believe that the algorithms and experimental results described in this contribution may be helpful in a wider range of human-centric applications, including information systems for travelers, interior design, and education.

Future work should include developing an open environment for collecting expert knowledge and facilitating end-user experiments. The data collected with such an environment may provide us with a better understanding of what makes food image aesthetic judgments objective and explainable, and how to proceed to improve them.

ACKNOWLEDGMENT

This work is partially supported by the research grant 17K00509 of Japan Society for the Promotion of Science (JSPS).

REFERENCES

- [1] J. Sully, "Aesthetics, Encyclopedia Britannica," [Online]. Available: <https://www.1902encyclopedia.com/A/AES/aesthetics-09.html>. [Accessed 10 10 2019].
- [2] N. Pappas, "Plato's Aesthetics," 2008. [Online]. Available: <https://seop.illc.uva.nl/entries/plato-aesthetics/>. [Accessed 10 10 2019].
- [3] I. Kant, "Kritik der Urteilskraft [Critique of judgment]," *Lagarde und Friederich, Berlin*, 1790.
- [4] I. Kant, "Critique of Judgement, translated by James C. Meredith," *Wadsworth Philosophy Source*, vol. 3, 2007.

- [5] M. L., "The Basics of Philosophy," 2009. [Online]. Available: https://www.philosophybasics.com/branch_aesthetics.html. [Accessed 10 10 2019].
- [6] A. P. Ershov, "Aesthetics and the human factor in programming," *Communications of the ACM*, vol. 15, pp. 501-505, 1972.
- [7] E. Pyshkin, "Liberal arts in a digitally transformed world: revisiting a case of software development education," in *Proceedings of the 13th Central & Eastern European Software Engineering Conference in Russia*, 2017.
- [8] R. A. Faste, "The role of aesthetics in engineering," *Japan Society of Mechanical Engineers (JSME) Journal*, vol. 28, p. 385, 1995.
- [9] M. Vogel, "Temporal evaluation of aesthetics of user interfaces as one component of user experience," in *Proceedings of the Fourteenth Australasian User Interface Conference-Volume 139*, 2013.
- [10] A. N. Tuch, S. P. Roth, K. Hornbæk, K. Opwis and J. A. Bargas-Avila, "Is beautiful really usable? Toward understanding the relation between usability, aesthetics, and affect in HCI," *Computers in Human Behavior*, vol. 28, pp. 1596-1607, 2012.
- [11] F. Akiba, "Preface: Natural Computing and Computational Aesthetics," in *Natural Computing and Beyond*, Springer, 2013, pp. 117-118.
- [12] C. Spence, *Gastrophysics: The new science of eating*, Penguin UK, 2017.
- [13] A. Krizhevsky, I. Sutskever and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012.
- [14] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248-255, 2009.
- [15] Y. Deng, C. C. Loy and X. Tang, "Image aesthetic assessment: An experimental survey," *IEEE Signal Processing Magazine*, vol. 34, pp. 80-106, 2017.
- [16] L. Zhang, Y. Gao, R. Zimmermann, Q. Tian and X. Li, "Fusion of multichannel local and global structural cues for photo aesthetics evaluation," *IEEE Transactions on Image Processing*, vol. 23, pp. 1419-1429, 2014.
- [17] S. Trambadia and H. Mayatra, "Food detection on plate based on the HSV color model," in *Online International Conference on Green Engineering and Technologies (IC-GET)*, Coimbatore, 2016.
- [18] R. C. Post, J. Haven and S. Maniscalco, "Setting the Table with a Healthy Plate: Make Half Your Plate Fruits and Vegetables," *Journal of the Academy of Nutrition and Dietetics*, vol. 111, no. 11, pp. 1644-1647, 2011.
- [19] S. B. Bush, K. S. Karp, L. Popelka and V. Miller Bennett, "What's on Your Plate? Thinking Proportionally," *National Council of Teachers of Mathematics*, vol. 18, no. 2, pp. 100-109, 2012.
- [20] N. Murray, L. Marchesotti and F. Perronnin, "AVA: A large-scale database for aesthetic visual analysis," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2012.
- [21] H. Talebi and P. Milanfar, "Nima: Neural image assessment," *IEEE Transactions on Image Processing*, vol. 27, pp. 3998-4011, 2018.
- [22] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov and L.-C. Chen, "MobileNetV2: Inverted Residuals and Linear Bottlenecks," *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018.
- [23] C. Szegedy, V. Vanhoucke, J. Shlens and Z. Wojna, "Rethinking the Inception Architecture for Computer Vision," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
- [24] O. Russakovsky, J. Deng, H. Su, J. Krause and L. Fei-Fei, "Imagenet large scale visual recognition challenge," *International journal of computer vision*, vol. 115 (3), pp. 211-252, 2015.
- [25] E. Pyshkin, J. Blake, A. Lamtev, I. Lezhnenin, A. Zhuikov, and N. Bogach, "Prosody Training Mobile Application: Early Design Assessment and Lessons Learned," in *10th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS)*, Metz, France, 2019.
- [26] A. Oram and G. Wilson, *Beautiful code* Leading programmers explain how they think, O'Reilly Media, Inc, 2007.



Veranika Mikhailava is a master student in the Department of Computer Science and Engineering at the University of Aizu, Japan. She is a member of IEEE and AAAI. Her research area is machine learning and software engineering. She has several publications in referred conference proceedings.



Evgeny Pyshkin is a senior associate professor with the School of Computer Science and Engineering, University of Aizu. Besides his major research interests in the areas of software engineering, software education, and human-centric computing, he contributed to the papers and related talks referring to the research on cross-cultural communication and exploration of the links between technology disciplines and humanities. He is a program committee member of international conferences in the domains of information systems and software engineering sponsored by ACM, IEEE, and IARIA.



Vitaly Klyuev is a professor at the University of Aizu, Japan. He received a Ph.D. degree in Physics and Mathematics from St. Petersburg State University, Russia in 1983. His research domain includes information retrieval, software engineering, and analysis of computer algorithms. He has more than 100 publications in referred journals and conference proceedings, three co-authored, and eight co-edited books. Dr. Klyuev is a member of the editorial board of several academic journals and a program committee member of more than 20 conferences sponsored by ACM, FTRA, IEEE, ISCA, and IARIA.