# What Does it Take to Cross the Aesthetic Gap? The Development of Image Aesthetic Quality Assessment in Computer Vision

**Sam Goree**
Department of Informatics
Indiana University
Bloomington, IN
sgoree@iu.edu

### Abstract

Computer vision research into image aesthetic quality assessment seeks to use machine learning to measure the aesthetic quality of images, which bears resemblance to a variety of topics within computational creativity, but has not been discussed in our community. To foster a conversation around this research literature, we trace the development of computer vision algorithms for aesthetic judgment over the past fifteen years and critically consider whether these algorithms actually cross the "aesthetic gap" proposed by the first researchers in this space. We then build towards a more fundamental question regarding machine learning and subjectivity.

## Introduction

Over the past fifteen years, computer vision researchers have investigated algorithms for image aesthetic quality assessment (IAQA). This research area seeks to apply machine learning to measure the aesthetic quality of images, usually by classifying them as "beautiful" or "not beautiful." While it is tempting to dismiss such a task as hopeless, given the subjectivity of the problem, this research has applications in automatic photo editing and curation and relates to several topics within computational creativity, including Machado and Cardoso's computing aesthetics (1998), Greenfield's computational aesthetics (2005), and Fisher and Shin's computational criticism (2019).

An interesting concept which arises from this literature is the notion of the *aesthetic gap*. Roughly analogous to the semantic gap in information retrieval, which separates the low-level features of images like pixels and lines from the high-level features humans observe in images like objects and symbols (Hare et al. 2006), Datta, Li, and Wang (2008) define the aesthetic gap as separating "the information that one can extract from low-level visual data" and "the interpretation of emotions that the visual data may arouse in a particular user." Since aesthetics is central to the value of many creative artifacts, and such value is often seen as an essential component of creativity (Boden 2004), thinking about the aesthetic gap and whether our algorithms really cross it is of central importance to computational creativity.

In this "debate spark" paper, however, we turn the concept of the aesthetic gap back on IAQA and question whether recent progress in this field actually constitutes a cross over that gap. To explore this topic, we present a historical narrative tracing the development of computer vision methods to measure aesthetic quality and their relation to prior philosophical and psychological study of aesthetics and its measurement. By introducing this problem area to the ICCC community, we hope to generate interest in developing computational approaches to aesthetics which engage more substantially with the fundamental subjectivity of the problem.

## Quantifying Aesthetics Before Computing

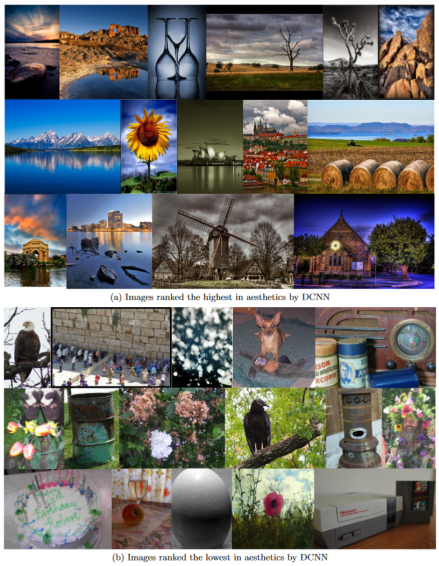
Before digging into the IAQA literature in computer vision, however, we will briefly define the term "aesthetics," give an example of how it is used in philosophy and discuss two influential historical attempts to measure the aesthetic qualities of stimuli.

In philosophy, aesthetics is the study of beauty, taste, experience and judgment. While many questions in philosophical aesthetics have a long history, even dating back to Plato, the study of aesthetics, as a "science of perception" is first named in the 18th century work of Alexander Gottlieb Baumgarten, whose work inspired responses from enlightenment philosophers (Guyer 2007). These philosophers understood that aesthetics is highly subjective, and developed methods for overcoming this subjectivity.

Immanuel Kant, in particular, had an influential approach. He claimed that judgments differed from person to person because they were bound up with interest, meaning that we make judgments based on feelings of pleasure, not just reason. However, if we remove our interests and make judgments which are completely *disinterested*, we can make judgments of taste, which should be universal among rational people. Kant is quick to clarify, though, that just because disinterested judgment is universal does not mean that such judgments can be made objectively, or based on the object alone: "aesthetic universality...does not unite the predicate of beauty with the concept of the object...and yet extends it to the whole sphere of judging persons," (Kant 1790).

Taste varies from person to person, across time and place and is even highly subject to influence, even in a laboratory setting (Bignardi, Ishizu, and Zeki 2020). Despite these challenges, aesthetics is one of the oldest topics of study in psychology, dating back to the 19th century work of the experimental psychologist Gustav Fechner. Fechner showed 347 subjects a series of rectangles and ellipses and asked

*Proceedings of the 12th International Conference on Computational Creativity (ICCC '21)*
*ISBN: 978-989-54160-3-5*




Lu et al. (Lu et al. 2014)


Lee et al. (Lee and Kim 2019)

Figure 1: Two figures from IAQA papers comparing high and low aesthetic quality images in their dataset.

them to choose the most appealing, and the rectangle with proportions drawn from the golden ratio was chosen the most frequently (Green 1995).

Fechner's work on aesthetics has been criticized by later psychologists and philosophers. For example, the 20th century Gestalt psychologist Rudolf Arnheim identifies a connection between Fechner's interest in measuring perception of beauty with his larger spiritual, cosmological and philosophical beliefs, and argues that Fechner's view of beauty as something which can be distilled down to one variable makes his findings related to art scientifically unreliable. "Just as Fechner's study does not tell us why people prefer the ratio of the golden section to others, so most of the innumerable preference studies carried out since his time tell us deplorably little about what people see when they look at an aesthetic object, what they mean by saying that they like or dislike it, and why they prefer the objects they prefer," (Arnheim 1985).

Inquiry specifically into aesthetic measures, like the ones put forward by contemporary computer vision researchers, starts with the work of the 20th century American mathematician George Birkhoff. Birkhoff's 1933 book, *Aesthetic Measure* puts forward a theory of aesthetic experience which divides it into three phases: first we recognize the complexity of a work, next we feel the sense that it is valuable, then finally we recognize the underlying order to which it adheres. Birkhoff claims these three properties: order ($O$), complexity ($C$) and value ($M$), can be related via an equation $M = \frac{O}{C}$.

Birkhoff's approach, like Fechner's, has been extremely influential, inspiring a century of computational approaches to aesthetics (e.g. Moon and Spencer's model of color harmony (Moon and Spencer 1944)), but it is poorly regarded by many philosophers. For example, Susanne Langer claims that the easily described nature of musical harmony has led to a great deal of hope that other aspects of art might be quantified and understood mathematically as well. However, "there is no use discussing the sheer nonsense or the academic oddities to which this hope has given rise, such as...the serious and elaborate effort of G.D. Birkhoff to compute the exact degree of beauty in any art work (plastic, poetic and musical) by taking the 'aesthetic measure' of its components and integrating these to obtain a quantitative value judgment," (Langer 1953). Langer goes on to argue that while musical sound is easy to describe, such description does not access the artistic qualities of music like motion, which exist in virtual space and time, rather than in the physical sound.

Langer's criticism of Birkhoff invokes a similar criterion to Datta, Li, and Wang: the difference between the explicitly measurable qualities of an object and the virtual and experiential qualities which inform its aesthetics are quite similar to the idea of a semantic or aesthetic gap. While rather simplistic mathematical models like those of Birkhoff likely lack the capacity to model something comparable to a human's aesthetic response, it is unclear whether more sophisticated computer vision models learned from data share that limitation.

## Early Machine Learning Approaches

Contemporary study of aesthetics in computer vision begins with the simultaneous work of Datta et al. and Ke, Tang, and Jing in 2006. Despite both working at the same time, and in the same US state (Pennsylvania), these two groups of authors arrived the problem area from different conceptual directions and take different approaches within the context of image classification.

Datta et al. are determined to automatically learn from

data which factors influence aesthetic value. They claim that, "in spite of the ambiguous definition of aesthetics...there exist certain visual properties which make photographs, *in general* more aesthetically beautiful." (Datta et al. 2006) Their concept of aesthetic value originates from their data: over 3000 images collected from the website `photo.net`, which allows users to upload their photos, and allows other users to rate them on "aesthetics" and "originality."[1] They cite two other sources on their understanding of aesthetics: the Oxford Advanced Learner's Dictionary and a book, Rudolf Arnheim's 1965 *Art and Visual Perception. A Psychology of the Creative Eye* (1965). Aesthetic quality assessment is framed in terms of image classification: they train decision trees and support vector machines to classify images into high and low aesthetics categories based on a variety of features extracted from images (e.g. measures of colorfulness, the photographic rule-of-thirds, image dimensions).

The decision to cite Arnheim pulls this approach towards psychological aesthetics, a field which exists in dialogue with both the work of earlier psychologists like Fechner, as well as the history of aesthetic philosophy. In a later survey paper (Joshi et al. 2011), the same authors cement that link. They discuss the approaches of analytic philosophers like Nelson Goodman and Richard Wollheim, as well as recent work in neuroaesthetics by Semir Zeki, who claims that aesthetic experience can be identified and explained by activity in specific brain regions.

To contrast, Ke, Tang, and Jing (2006) approach IAQA from the perspective of photo curation. Rather than psychological aesthetics, they ground their work in image quality assessment, an area of computer vision research concerned with measuring image noise and degradation (Kamble and Bhurchandi 2015). Rather than making claims about philosophy, Ke, Tang, and Jing argue that a well-designed set of features may be used to reason about the subjective aspects of image quality, like the difference between professional and amateur photos. Their method makes use of images and ratings from the photo challenge website `DPChallenge.com`, which they divide into "professional" and "amateur" categories based on ratings. They cite two popular photography books to justify their choices of features, which include edge and color histograms, as well as Fourier transform-based blur metrics, which they use to train a Naive Bayes classifier.

Over the next six years, a variety of other publications emerged proposing different combinations of image features for solving the aesthetic quality assessment problem. While other scholars used similar approaches at first (Datta, Li, and Wang 2008; Jiang, Loui, and Cerosaletti 2010), later authors shifted towards low-level features like GIST or SIFT descriptors due to an influential paper by Marchesotti et al. which made the case that hand-crafted features are ineffective because they are non-exhaustive, computationally expensive and rely on heuristic assumptions which may not generalize well (Marchesotti et al. 2011).

The relationship between Datta, Ke and both earlier and later aesthetic thought is at the heart of our claims about the aesthetic gap. The work of Datta et al. is framed as an approach to computational aesthetics, but like Ke, Tang, and Jing, they only measure how consistent a photograph is with common photography rules of thumb. Later work further conflates these two concepts of "aesthetic quality" by shifting to lower-level image features to better fit the dataset labels. However, inspection of the "high quality" and "low quality" images in these datasets makes it clear that the distinction between them is more of a stylistic difference than anything else. Figure 1 shows comparisons between high and low quality photos from two IAQA papers. The qualities shared by all of the photos labeled as "high quality" is evident: these are overwhelmingly photos of landscapes and flowers which prioritize color and emotionality over realism. We would argue, however, that this style is not the only way that photographs can be beautiful. Photography can be aesthetically pleasing in as many ways as other art forms, and many genres of art photography like candid photography or photojournalism do not prioritize the use of such dramatic visual effects. In other words, these papers and datasets seem to conflate explicit emotionality with the potential to arouse emotion.

## The AVA Dataset and Deep Learning

In 2012, two major events shifted the conversation around IAQA. First, in June, Murray, Marchesotti, and Perronnin (2012) released the Analysis of Visual Aesthetics (AVA) dataset, which contains over 250,000 photos from `DPChallenge.com`, an order of magnitude larger than any existing dataset, along with metadata, including rating distributions and category labels, where possible. Second, in October, Krizhevsky, Sutskever, and Hinton (2012) dramatically beat the benchmark on the ImageNet LSVRC using a deep convolutional neural network (CNN). While deep learning had profound effects on computer vision as a whole, these two contemporaneous changes produced a paradigm shift in the study of IAQA.

Lu et al. (2014) published the first paper applying deep learning to aesthetic image classification in 2014. They reiterate the argument from Marchesotti in favor of generic image features, and claim that deep features are even more generic, since they work with pixels directly. Lu et al. identify that the fixed input size of AlexNet makes it difficult to apply to images of many different dimensions in AVA, since cropping or warping might disrupt aesthetic quality, so they use a two-column model to learn from warped and cropped versions of the image simultaneously. Neither this work, nor the generation of papers which followed their lead in applying CNNs to the AVA dataset (Kao, Wang, and Huang 2015; Zhou et al. 2016; Lv and Tian 2016), make much reference to the problem statement and its context at all, aside from acknowledging its highly subjective nature.

While CNNs do not carry all of the assumptions of things like measures of colorfulness or edge histograms, they are not blank slates either. The connectivity structure of convolutional and max-pooling layers encode the as-

[1] `photo.net`, surprisingly, was not created by professional photographers, but by Philip Greenspun, a computer scientist at MIT interested in online communities.

| Paper | Year | Acc. |
|---|---|---|
| (Murray, Marchesotti, and Perronnin 2012) | 2012 | 67% |
| (Lu et al. 2014) | 2014 | 71% |
| (Mai, Jin, and Liu 2016) | 2016 | 77.1% |
| (Kong et al. 2016) | 2016 | 77.3 |
| (Zhou et al. 2016) | 2016 | 78.1% |
| (Wang et al. 2016) | 2016 | 76% |
| (Kao, He, and Huang 2017) | 2017 | 78% |
| (Ma, Liu, and Wen Chen 2017) | 2017 | 82.5% |
| (Ko, Lee, and Kim 2018) | 2018 | 82.2% |
| (Sheng et al. 2018) | 2018 | 83.3% |
| (Lee and Kim 2019) | 2019 | 91.5% |

Table 1: Accuracy benchmark results on the AVA dataset.

sumption that the salient features of an image are locally situated, translation-invariant and the presence of an activation is more significant than the absence, which are good assumptions for classifying between different types of objects or handwritten digits (Krizhevsky, Sutskever, and Hinton 2012), but are not necessarily good for aesthetic judgment, which at least in the eyes of psychologists like Arnheim (1965), is more holistic.

In the past five years, several trends have emerged in IAQA. First, Kong et al. (2016) suggest including user data to personalize image assessments, which Ren et al. (2017) formalize into an active learning task. Second, different objectives beyond classification have emerged, including pairwise comparison (Lv and Tian 2016) and distribution learning (Cui et al. 2017). Finally, the binary classification accuracy benchmark on the AVA dataset has steadily increased, reaching over 91% (see Table 1).

Additionally, a new claim for significance, related to curation and editing of photographs for social media, has emerged. Several recent authors make reference to the widespread popularity of social networking services (Wang et al. 2019), the exponential growth of online visual data (Sheng et al. 2018; Lee and Kim 2019) and the growing need for automatic photo editing tools (Wang et al. 2019). This claim for significance brings IAQA into the realm of AI-based creativity support tools, further increasing its relevance to the computational creativity community.

Our narrative in this section emphasizes the continuity between the current state of the art in IAQA and the long history of aesthetics in other disciplines. There is a direct continuity from classical to deep methods: Marchesotti et al. made their argument in favor of low-level image features before the advent of deep learning, and the first deep learning-based method of Lu et al. is framed as the natural extension of that argument. Even highly technical recent papers, which are quite distant from the philosophical motivations of authors like Datta et al., are implicitly weighing into a long conversation on the nature of art and beauty, which may have wide reaching implications. But do any of them really cross the aesthetic gap and reason about "the interpretation of emotions that the visual data may arouse in a particular user?"

## Discussion

So far, we have traced the evolution of IAQA in computer vision from its prehistory in the work of Fechner and Birkhoff, its origins in psychological aesthetics and photographic rules of thumb and its shift from hand-engineered features to deep learning. We saw how its two goals, rooted in computational aesthetics and image quality assessment, merged over time, and how performance on the AVA dataset, which arguably only captures a specific, popular photographic style, has been treated as a stand-in for an algorithm's ability to measure aesthetic quality more generally. With that continuity in mind, we find it difficult to point to a specific paper or accuracy level where these approaches cross the aesthetic gap introduced by Datta, Li, and Wang. However, such a claim raises other questions about the nature of this gap.

For example, it's possible that the success of recent deep learning models on the AVA dataset demonstrates that there is no such gap: the neuroscientific arguments indicate that our aesthetic responses exist in a lower level of the visual system than we might believe (Chatterjee and Vartanian 2016), and it's possible that we actually make judgments based on simple visual statistics and only use higher cognitive processes to explain those judgments. Such a finding would vindicate scholars like Birkhoff, who believed that a measure of aesthetics could be computed from measures of order and complexity, without regard for the emotions of the observer. On the other hand, if we assume that an aesthetic gap does exist, and making aesthetic judgments requires algorithms which understand meaning and emotional attachment, that would cast further doubt on whether IAQA models are actually measuring aesthetics, and whether accuracy on the AVA is a suitable measure of performance.

If deep learning models cannot overcome the aesthetic gap, how should we, as artificial intelligence researchers, proceed? It's not unreasonable to imagine a computationally creative agent which both interprets symbols and models emotional attachments enough to have something resembling an understanding of taste. But since taste is subjective, it is still unclear how to measure performance. Can a model have its own preferences, or should it merely predict the preferences of a human?

This last point reaches towards an important question regarding artificial intelligence and subjectivity. When we say that a task is subjective, who should be the subject? Is our goal to develop algorithms which have their own aesthetic experiences (for some definition of "own"), or merely predict the preferences of humans? If it is the former, is an algorithm a Kantian disinterested agent? If it is the latter, which humans' preferences should count? Versions of this question exist throughout computational creativity. For example, should an algorithmic musician create music that appeals to its own computational sense of taste, its creator's taste or an average of other humans' tastes? IAQA chooses to derive ground-truth labels from an average of many humans' aesthetic quality ratings, but such data risks conflating aesthetic quality with popularity. The theory and research methods for issues relating to aesthetics and subjectivity in machine learning demand more scholarly attention.

## References


Arnheim, R. 1965. *Art and visual perception: A psychology of the creative eye*. Univ of California Press.

Arnheim, R. 1985. The other Gustav Theodor Fechner. In Koch, S., and Leary, D. E., eds., *A century of psychology as science*. American Psychological Association.

Bignardi, G.; Ishizu, T.; and Zeki, S. 2020. The differential power of extraneous influences to modify aesthetic judgments of biological and artifactual stimuli. *PsyCh Journal*.

Boden, M. A. 2004. *The creative mind: Myths and mechanisms*. Psychology Press.

Chatterjee, A., and Vartanian, O. 2016. Neuroscience of aesthetics. *Annals of the New York Academy of Sciences* 1369(1):172–194.

Cui, C.; Fang, H.; Deng, X.; Nie, X.; Dai, H.; and Yin, Y. 2017. Distribution-oriented aesthetics assessment for image search. In *ACM SIGIR RDIR*, 1013–1016.

Datta, R.; Joshi, D.; Li, J.; and Wang, J. Z. 2006. Studying aesthetics in photographic images using a computational approach. In *ECCV*, 288–301. Springer.

Datta, R.; Li, J.; and Wang, J. Z. 2008. Algorithmic inferencing of aesthetics and emotion in natural images: An exposition. In *ICIP*, 105–108. IEEE.

Fisher, D. H., and Shin, H. 2019. Critique as creativity: Towards developing computational commentators on creative works. In *ICCC*, 172–179.

Green, C. D. 1995. All that glitters: A review of psychological research on the aesthetics of the golden section. *Perception* 24(8):937–968.

Greenfield, G. R. 2005. Computational aesthetics as a tool for creativity. In *Proceedings of the 5th conference on Creativity & cognition*, 232–235.

Guyer, P. 2007. 18th century german aesthetics. *Stanford Encyclopedia of Philosophy*.

Hare, J. S.; Lewis, P. H.; Enser, P. G.; and Sandom, C. J. 2006. Mind the gap: Another look at the problem of the semantic gap in image retrieval. In *Multimedia Content Analysis, Management, and Retrieval 2006*, volume 6073, 607309. International Society for Optics and Photonics.

Jiang, W.; Loui, A. C.; and Cerosaletti, C. D. 2010. Automatic aesthetic value assessment in photographic images. In *2010 IEEE International Conference on Multimedia and Expo*, 920–925. IEEE.

Joshi, D.; Datta, R.; Fedorovskaya, E.; Luong, Q.-T.; Wang, J. Z.; Li, J.; and Luo, J. 2011. Aesthetics and emotions in images. *IEEE Signal Processing Magazine* 28(5):94–115.

Kamble, V., and Bhurchandi, K. 2015. No-reference image quality assessment algorithms: A survey. *Optik* 126(11-12):1090–1097.

Kant, I. 1790. Critique of judgment. In Ross, S. D., ed., *Art and its Significance: An Anthology of Aesthetic Theory*. SUNY Press.

Kao, Y.; He, R.; and Huang, K. 2017. Deep aesthetic quality assessment with semantic information. *IEEE Transactions on Image Processing* 26(3):1482–1495.

Kao, Y.; Wang, C.; and Huang, K. 2015. Visual aesthetic quality assessment with a regression model. In *ICIP*, 1583–1587. IEEE.

Ke, Y.; Tang, X.; and Jing, F. 2006. The design of high-level features for photo quality assessment. In *CVPR*, volume 1, 419–426. IEEE.

Ko, K.; Lee, J.-T.; and Kim, C.-S. 2018. Pac-net: pairwise aesthetic comparison network for image aesthetic assessment. In *ICIP*, 2491–2495. IEEE.

Kong, S.; Shen, X.; Lin, Z.; Mech, R.; and Fowlkes, C. 2016. Photo aesthetics ranking network with attributes and content adaptation. In *ECCV*, 662–679. Springer.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. *NIPS* 25:1097–1105.

Langer, S. 1953. *Feeling and Form*. Scribner.

Lee, J.-T., and Kim, C.-S. 2019. Image aesthetic assessment based on pairwise comparison a unified approach to score regression, binary classification, and personalization. In *ICCV*, 1191–1200.

Lu, X.; Lin, Z.; Jin, H.; Yang, J.; and Wang, J. Z. 2014. Rapid: Rating pictorial aesthetics using deep learning. In *ACM Multimedia*, 457–466.

Lv, H., and Tian, X. 2016. Learning relative aesthetic quality with a pairwise approach. In *ICMM*, 493–504. Springer.

Ma, S.; Liu, J.; and Wen Chen, C. 2017. A-lamp: Adaptive layout-aware multi-patch deep convolutional neural network for photo aesthetic assessment. In *CVPR*, 4535–4544.

Machado, P., and Cardoso, A. 1998. Computing aesthetics. In *Brazilian Symposium on Artificial Intelligence*, 219–228. Springer.

Mai, L.; Jin, H.; and Liu, F. 2016. Composition-preserving deep photo aesthetics assessment. In *CVPR*, 497–506.

Marchesotti, L.; Perronnin, F.; Larlus, D.; and Csurka, G. 2011. Assessing the aesthetic quality of photographs using generic image descriptors. In *ICCV*, 1784–1791. IEEE.

Moon, P., and Spencer, D. 1944. Geometric formulation of classical color harmony. *JOSA* 34(1):46–59.

Murray, N.; Marchesotti, L.; and Perronnin, F. 2012. Ava: A large-scale database for aesthetic visual analysis. In *CVPR*, 2408–2415. IEEE.

Ren, J.; Shen, X.; Lin, Z.; Mech, R.; and Foran, D. J. 2017. Personalized image aesthetics. In *ICCV*, 638–647.

Sheng, K.; Dong, W.; Ma, C.; Mei, X.; Huang, F.; and Hu, B.-G. 2018. Attention-based multi-patch aggregation for image aesthetic assessment. In *ACM Multimedia*, 879–886.

Wang, Z.; Chang, S.; Dolcos, F.; Beck, D.; Liu, D.; and Huang, T. S. 2016. Brain-inspired deep networks for image aesthetics assessment. *arXiv preprint arXiv:1601.04155*.

Wang, L.; Wang, X.; Yamasaki, T.; and Aizawa, K. 2019. Aspect-ratio-preserving multi-patch image aesthetics score prediction. In *CVPR Workshops*, 0–0.

Zhou, Y.; Lu, X.; Zhang, J.; and Wang, J. Z. 2016. Joint image and text representation for aesthetics analysis. In *ACM Multimedia*, 262–266.