Towards Beauty: Robot Following Aesthetics Gradients

Mathias Franzius

2019

Preprint:

This is an accepted article published in International Conference on Advanced Robotics (ICAR) . The final authenticated version is available online at: https://doi.org/[DOI not available]

Towards Beauty: Robot Following Aesthetics Gradients

Mathias Franzius[1] mathias.franzius@honda-ri.de

***Abstract*— Increasing numbers of devices are equipped with cameras generating large amounts of images. State of the art technologies allow to automatically identify relevant and aesthetically pleasing images after they were stored. Here, we demonstrate a robot that estimates the gradient of image aesthetics in its environment and actively navigates towards the maximum. Aesthetics navigation is integrated into a modified robotic lawnmower, switching online between tasks based on estimated aesthetics scores. This behavior generates higher aesthetics scores than offline selection of images captured during standard behavior. The proposed system extends robotic behavior from the purely functional towards a cooperative and empathic level.**

I. Introduction

"The soul cannot thrive in absence of a garden." (Thomas More, Utopia), yet often we spend less time in nature than we would like. Gardens continuously change, including lighting and weather and changing vegetation, offering different beautiful views only from specific angles at certain times. On the other hand, most views most of the time are less aesthetically pleasing.

A human garden enthusiast will typically not take random photos and later search for good ones. Rather he or she will continue gardening until something especially interesting catches the eye, move closer and consider taking a photo there. Here we introduce an outdoor service robot that mimics this behavior: the robot performs its standard task of mowing and simultaneously estimates the aesthetics of views. If the score is sufficiently high, it follows the gradient towards the most beautiful perspective and takes images. At the end of the day, the robot can present a digest of beautiful views to its owner. To our knowledge, this is the first instance of such robot navigation behavior in unconstrained environments. Our contributions include a fast online aesthetics score estimation in the wild, a simple method to follow its gradient, and the integration of aesthetics navigation with standard mowing behavior on a modified robotic lawn mower (see Fig. 1).

Fig. 1. **Prototype robot.** Research is demonstrated on a modified lawn mower robot with added camera and computational power. The system autonomously switches between mowing and following aesthetics gradients.

Cameras, especially in smartphones, have become ubiquitous and the total number of photos generated is estimated over one trillion per year [1]. Furthermore, autonomous cars, mobile robots and household appliances are increasingly equipped with cameras (e.g., [2]), giving the possibility of capturing images with high personal value from the specific environment of a user. However, exhaustive manual search for interesting or beautiful images becomes impossible and thus such autonomous systems should assist their users with selecting images, or preferably only generate relevant images.

Image aesthetics relates to image content, composition, appearance, and style – automatic prediction of aesthetics is a complex task. However, recent machine learning models have shown the feasibility to predict the aesthetic value of images (e.g., [3], [4], [5]). These approaches typically focus on learning from offline image databases like AVA [3] and estimate aesthetics on images after they were uploaded to a PC or the cloud. Such an approach is limited to the selection of fixed images, possibly yielding limited improvements by cropping promising parts [6] or postprocessing [7]. However, at this point in time the camera perspective that led to an image is obviously fixed and can not be improved anymore.

In contrast, our proposed system can actively drive towards a location yielding the highest scoring image. In Section III we show that image scores are significantly higher in this case than for recordings during standard mowing as a control task.

Google Clips [8] follows a similar approach, based on a highly integrated "intelligent camera" for automatic on-device detection of good images. It reduces the amount of captured imagery to limit human post-processing. However, the device has to be placed by the user at a reasonable position since it can not by itself navigate towards good views.

Cameras for "life logging" (e.g., [9], [10]) are devices worn by a user and typically generate an image digest of daily activities for the user. These systems also have to select relevant images from a large sequence and cope with unconstrained environments. Like the prior approach, though, they can not

[1]Honda Research Institute Europe GmbH, Offenbach, Germany

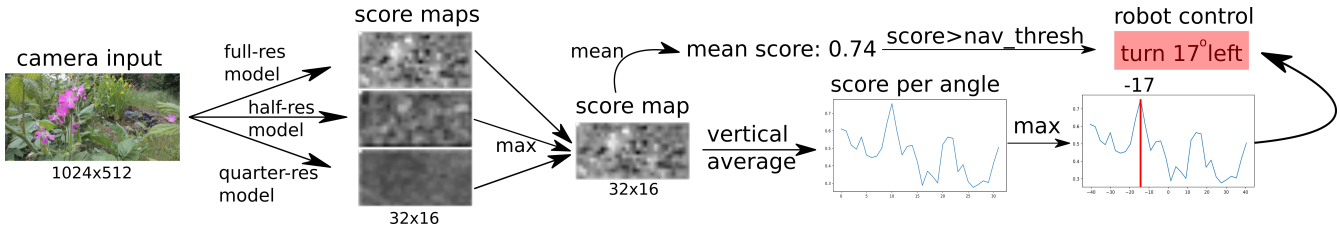

Fig. 2. **System overview.** Camera images are passed to aesthetics estimators, which were trained on three different resolutions. The resulting score maps are combined by selecting their maximum at each position. If the mean value of this map exceeds a threshold, aesthetics navigation can be activated (see further constraints in the text). The maximum of averages over score map columns determines the robot steering direction.

move independently and thus not actively find interesting or beautiful views.

The patent on a self-balancing robot [11] has been interpreted as a "personal paparazzi"[12]. Such a robot design has more degrees of freedom to adjust camera pose in 3D space and may follow humans around in larger environments.

An early implementation of a similar idea is a robot that detects interesting targets and actively navigates to a position to optimize photo quality [13]. The approach is limited to group photos of people and is based on hand-crafted rules for image composition.

II. Methods

We extended a commercial lawn mower robot (Honda Miimo 310) with a fixed frontal camera (Logitech Brio 4k with autofocus and a horizontal field of view of 81°). For the high computational load of online aesthetics estimation, the robot was additionally equipped with a notebook containing an NVidia Geforce GTX 1070. The mower's ECU is connected to an onboard embedded system via CAN bus as in [2], which is in turn connected to the laptop for aesthetics estimation. Low-level control is located on the ECU, whereas the embedded system monitors aesthetics score and switches robot behavior between standard mowing and aesthetics navigation. Navigation is initiated when the average image score exceeds a value of 0.6 (this setting was empirically determined so the robot spends approximately 1/3 of the time with aesthetics navigation and 2/3 with the default task of mowing). During navigation, the robot drives with constant speed while input images are filtered by the aesthetics estimator and the filter response is averaged vertically, allowing it to always turn towards the current angle with highest average score (see Fig. 2). Navigation terminates when (1) the score drops below 0.5, or (2) an obstacle is hit, or (3) navigation time exceeds three minutes, or (4) the robot drives over the border wire.

We applied two different models for outdoor aesthetics estimation and gradient following. The first model is the readily available ILGNet [14] with 13 layers trained on AVA [3], which achieves over 85% accuracy in classifying low-scoring vs. high-scoring samples in AVA2. The analysis of ILGNet in our scenario revealed a strong dependency on low-level image statistics (see Sec. III), as well as poor generalization to the unbiased image stream in the garden domain. Additionally, ILGNet's "fully connected" layer constrains the input image size to 227x227 pixels. For estimating the aesthetics gradient, however, we need to apply the estimator as a filter on a larger image to yield the individual subregion scores. The model's architecture would require a full and highly redundant computation pass on each image patch, leading to frame rates below 0.1Hz and thus unacceptably slow for robot navigation, even with the fastest available mobile GPU.

Consequently, we built and trained a second alternative model based on the fully convolutional VGG16 architecture [15], which can be scaled to arbitrary image dimensions. The final classification layer in VGG16 was replaced with a global averaging layer to allow direct regression to the AVA aesthetics labels during training. After training, the global averaging layer was removed, and the resulting 20-layer net could be used on larger images, yielding a spatial distribution of aesthetics scores. The number of trainable convolution layers was 14, with a total number of 14,714,688 trainable parameters. Initial parameters for VGG16 from pretraining on Imagenet [16] are available in Keras [17]. To improve performance in a garden domain, we extended the AVA training database with approximately 5% additional images from the Oxford Flower Database [18] as positive examples (regression target score 0.75) and clutter objects on grass [19] as negative examples (regression target score 0.25). From this combined set of 267,940 images, we generated 1 million training images using image augmentation and trained the network using batch normalization as a standard technique for improving model performance. After training, we remove the averaging layer in order to retrieve the aesthetics score estimate of each individual image subregion. See Fig. 3 for example images from the top 10 percentile (left) and bottom 10 percentile (right). The robot's onboard notebook with an NVidia Geforce GTX 1070 filters input images of size 1024x512 at approximately 4Hz.

The results of aesthetics gradient navigation strongly depend on the estimator's basin of attraction for finding high-scoring views, i.e., how likely it is that the robot can follow the gradient towards an aesthetics maximum for a given distance. Initial experiments showed that the above architecture allows the robot to reliably turns towards and approaches large and colorful scene elements, resulting in high scores at the end of the approach. For smaller and isolated objects, like a single blooming flower, the maximum distance where

Fig. 3. **Example images.** Left block: images from top-10 percentile score. Right block: images from bottom-10 percentile score.

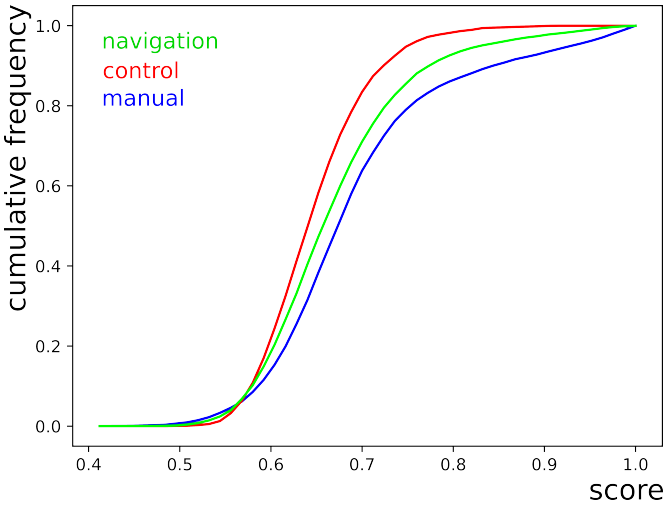

Fig. 4. **Cumulative aesthetics score distributions.** Red: results for random mowing as control condition; mean score is significantly lower than navigation mean and proportion of high-scoring images is much lower. Green: results for proposed behavior with automatic navigation towards aesthetics maximum, including non-navigation phases. Navigation mode scores are significantly higher than in control condition. Blue: results for hand-carried camera in garden, trying to capture beautiful views.

the approach likelihood is above 50% is below 0.5m. In this setting, for typical trajectories the gradient direction ofen changes erratically before converging towards a high-scoring target.

Image augmentation during training increases model invariances to small translations, rotations and rescaling of images. However, there is no mechanism to predict image scores for larger distances in scale space. Consequently, we extend our system with a multiscale estimator for computing the aesthetics gradient (see Fig. 2). We trained two additional similar network on half and quarter training image resolution, respectively. Applying these estimators to full resolution input images serves for estimating image scores in larger distances. For gradient estimation, the score values from full, half, and quarter scale models are combined by taking the maximum score of any scale at any given position. This approach increases the basin of attraction, i.e., the typical distance from which the robot reliably turns towards a local maximum, to approximately 2m, and the resulting frame rate is reduced to approximately 1Hz, which is still sufficient for our scenario (see Fig. 2 for a sketch of the full processing pipeline). Typical navigation trajectories in this setting have fewer changes of direction and converge more quickly onto a high-scoring region. Most often, high-scoring views are at the work area border, where navigation aborts when the robot reaches the area wire. Few trajectories lead from one high-scoring view to other distinct nearby regions with high scores. Cyclic trajectories never occured.

During any short time window of recording, images and computed scores are highly similar. Selecting only the highest scoring images for final presenting to the user then includes many redundant duplicates. Furthermore, the working area of a lawn mower robot is limited and similar views are likely to occur during operation. We implemented two mechanisms to cope with image redundancy: temporal selection and filtering by feature similarity. Temporal winner-takes all selection chooses the highest scoring image from each 2-seconds interval and discards lower-scoring images. This mechanism suppresses redundant views from within a single navigation episode. Scond, we want to remove overly similar images from a whole recording. Images from natural environments are never identical in practice. However, a meaningful similarity measure to identify and remove redundant images recorded over longer time spans is not trivial. Here, use a greedy heuristics based on the aesthetics estimator model's penultimate layer activation as a feature vector for each image and compare the norm of activations. Starting with the highest scoring image, all lower-scoring images with a feature L2 norm smaller than a threshold are discarded.

We collected data from four different gardens and one rural field in three different experimental conditions: "control mode" using standard random pattern mowing on the robot, "aesthetics navigation mode" as described above, and "manual mode", in which a webcam was manually carried through the gardens with the objective to capture beautiful views. We consider control mode for baseline performance and manual mode as an upper bound for result quality, because of full human control over the recording and more degrees of freedom in camera control than on the robot.

III. RESULTS

In the following, we first analyze results for the state of the art aesthetics estimation model ILGNet [4], discuss its shortcomings for aesthetics navigation and proceed with results of our adapted model based on VGG16.

ILGNet achieves very good results for estimating aesthetics scores on the AVA benchmark set. Does it represent high-level concepts such as content, style, or image composition, or does it rather correlate with trivial low-level image statistics? We analyzed the impact of image statistics, including image saturation, additive noise, contrast, gamma, hue angle, brightness, and JPEG block artifacts on aesthetics estimation with ILGNet. We identified strong influences of these manipulation on aesthetics scores. Figure 5 depicts means and variances of scores for changing image saturation (left) and additive white noise (right), as well as example images ('original') and their respective transformation yielding maximum score ('maximum'). Complete color desaturation, as well as color oversaturation, on average cause high scores, whereas low image saturation lead to lower scores. Human aesthetics preference positively correlates with image saturation [20], and, since AVA contains color and grayscale images, a high average score for grayscale images seems plausible. However, the model's strong score increase for extremely oversaturated images is unlikely to correlate with human preference. Second, ILGNet's scoring for noisy images is more surprising. For small amplitudes of additive white nose the average score decreases but increases for high noise amplitudes. Third, we tested the effect of small translations on ILGNet scores. As expected, scores of many images only changes minimally for translations of less than 10 pixels. For some images, the alignment of the image relative to the model's first layer filters resulted in score amplitude variations of up to 50%, whereas the human observer can hardly detect a difference in these images. We found that this effect is present in images that previously were encoded with JPEG, which leaves small 8x8 pixel block artifacts in images.
Very noisy images, extremely oversaturated images, and shifted JPEG images are probably not part of the AVA training database and thus the model predictions do not generalize well to such changes of statistics. It is, however, evident that low-level statistics has a very strong impact on model scores (also see similar findings in [21]). These results suggest that in our scenario the robot may follow a gradient score towards maximum saturation or contrast and thus fail at generating aesthetic images. In the following, we show that with the alternative model based on VGG16, despite such shortcomings of aesthetics prediction, a robot can follow the aesthetics gradient and find images that are preferred by human subjects over control recordings. Image augmentation during model training increases invariance towards small translations, noise, and other low-level image changes.

To show that aesthetics navigation yields more aesthetic images than control behavior, we first show that the average score is significantly higher in navigation mode. Second, we show that a majority of human subjects prefer high scoring images over medium and low-scoring images.
The mean aesthetics score under control conditions with default random mowing behavior is 0.65 (see Fig. 4 for cumulative score distributions). The mean scores of images during aesthetics navigation is 0.67, which is significantly higher than in control condition (p-value of permutation test for hypothesis "scores drawn from identical distributions" < 0.0001). Note that aesthetics navigation to a large degree also consists of default mowing behavior with random navigation whenever the aesthetics score threshold is not exceeded (see Sec. II above). Thus, the distribution of low scores is also similar between the two conditions. However, primarily the high scoring images are relevant for our scenario. The fraction of high scoring images in navigation mode (2.7% of images score higher than 0.9 and $1.4\% > 0.95$) are much higher than in control mode ($0.1\% > 0.9$ and $0\% > 0.95$; see Fig. 4). As expected, the mean score of 0.71 for manual recordings is higher than under navigation and control conditions.

The AVA image aesthetics database [3] is a popular benchmark for machine learning. It contains a wide variety of styles and domains, including outdoor imagery similar to what a robot mower could encounter. However, a number of issues put into question if a model learned on AVA will generalize to robotically generated garden images. Firstly, all of the images in AVA were taken and actively selected for publication by humans, whereas a robotic system does not have such a filtering step. Such a *capture bias* is a well-known phenomenon in machine learning [22]. Secondly, closer inspection of AVA reveals large amounts of images were strongly processed and/or unrelated to classical image aesthetics, e.g. jokes and memes. It is highly unlikely that such abstract, self-referential concepts can be learned and generalized by current machine learning methods. Finally, the robot's camera perspective strongly differs from the usual human perspective. Together, these observations may indicate that we achieved higher scores with aesthetic navigation but that the model may not generalize to images taken by the robot. Thus, we next validate that model scores agree with the human aesthetics preferences by performing the following test. The recorded garden images were binned by their estimated aesthetic score, and three bins (high: 90-100 percentile; medium: 45-55 percentile; low: 0-10 percentile) were selected. From each bin and for each test subject, we randomly sampled a set of 12 images. 17 test subjects were asked to order their respective sets by their perceived aesthetic value. 71% of subjects ranked the image set from the "high" group highest. No subject ranked a set from the "high" group lowest (see Table I). Note that four out of five subjects who preferred the medium-ranked images noted that, while they rated individual views from the high-ranking set higher, they preferred the medium ranking image set due to its higher diversity.

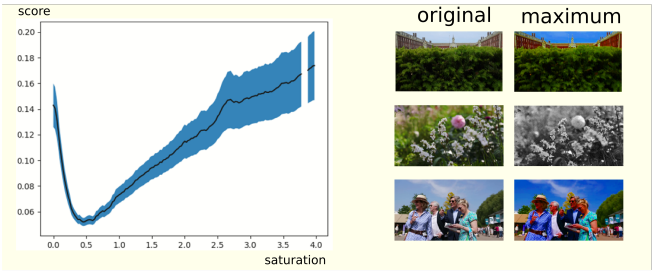

Fig. 5. **Impact of low-level image properties on aesthetics score.** Left box: average aesthetics score over variations of image saturation. Grayscale and oversaturated images typically score highly. Example images ("original") and their respective maximally scoring saturation transforms ("maximum"). Right box: average aesthetics score for additive white noise. Score typically increases with noise amplitude. Example images ("original") and their respective maximally scoring noisy transforms ("maximum").

	high human preference	medium human preference	low human preference
high model score	12	5	0
medium model score	4	9	4
low model score	1	3	13

TABLE I

Aggregate results of human aesthetics preferences versus model scores. 71% OF SUBJECTS RANKED THE IMAGE SET FROM THE HIGH MODEL SCORE GROUP HIGHEST. SEE TEXT FOR DETAILS.

IV. DISCUSSION

We have demonstrated an outdoor robot that integrates a standard service task of mowing with actively searching for and driving towards the most aesthetic views in its environment. Following the aesthetics gradient results in higher aesthetics scores and these scores agree with the preferences of a majority of subjects.

Robustness on unconstrained outdoor images and a smoothness in camera position space are necessary for following the aesthetics gradient. Our modified network architecture achieves these goals and allows fast online aesthetics estimation of each subregion. However, the model's final averaging layer removes most information on image composition, which is an important aspect of image aesthetics. Nevertheless, with our model architecture, we have shown that our system's aesthetics prediction is sufficient for aesthetics navigation. A very recent alternative model [23] may again allow to integrate information on image composition.

Models trained on AVA predict aesthetics based on training data from a large group of people. Can such an average aesthetics model be expected to produce reasonable results, or are personalized models necessary? Recent work [24] has shown a dependency on the visual domain: a relatively high degree of inter-person consistency in the assessment of aesthetics was found for natural landscape images, compared to images of artwork and architecture. Thus, gardens are a better domain for aesthetics navigation than, for example, a museum guidance robot. Nevertheless, personalization to a specific user's taste will certainly increase the perceived quality. Any sufficiently smooth scalar score can serve as an alternative to aesthetics, e.g., memorability [25], interestingness [26], image content, or a combination of these. A simple user-provided weighting of such scores may lead to personalized results in a much simpler way than requiring fine tuning of the deep learning aesthetics model.

With the current parametrization, the robot spends approximately one third of operation time in aesthetics navigation mode. Aesthetics navigation is more often triggered at the borders of the lawn than centrally, leading to a higher proportion of aesthetics navigation in smaller gardens. Since small gardens require less mowing, the combined behavior is unlikely to reduce mowing performance. Furthermore, both tasks can mostly be performed in parallel. During the testing period, both in standard mowing mode and in aesthetics navigation mode, we did not observe the mower to get stuck and require human intervention, nor did we observe systematically unmowed areas.

In order to increase user acceptance and to increase mowing efficiency, we plan to improve the approach for image diversity within recordings and over consecutive recordings.

Aesthetics navigation may be interesting for other robots than lawn mowers, e.g., cleaning robots or other (future) household robots. As a future outlook, replacing implicit models of aesthetics on 2D images with models based on 3D geometry and meaningfully segmented objects may lead to an "intelligent robotic photographer". Following the aesthetics gradients in such a representation may even incorporate environmental manipulations for composing better images ("move the garden chair to the left") by the robot itself or by human-robot cooperation.

V. ACKNOWLEDGEMENTS

Thanks to Helge Wagner for designing the robot cover, to Mark Dunn for support with the embedded control system, and to Oliver Schoen for the robot power system.

REFERENCES

[1] C. Cakebread, *People will take 1.2 trillion digital photos this year – thanks to smartphones*, 2017 (accessed July 2019). https://www.businessinsider.com/12-trillion-photos-to-be-taken-in-2017-thanks-to-smartphones-chart-2017-8.

[2] M. Franzius, M. Dunn, N. Einecke, and R. Dirnberger, "Embedded robust visual obstacle detection on autonomous lawn mowers," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 44–52, 2017.

[3] N. Murray, L. Marchesotti, and F. Perronnin, "Ava: A large-scale database for aesthetic visual analysis," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2408–2415, June 2012.

[4] X. Jin, L. Wu, X. Li, X. Zhang, J. Chi, S. Peng, S. Ge, G. Zhao, and S. Li, "ILGNet: inception modules with connected local and global features for efficient image aesthetic quality classification using domain adaptation," *IET Computer Vision*, vol. 13, no. 2, pp. 206–212, 2018.

[5] W. Wang, S. Yang, W. Zhang, and J. Zhang, "Neural aesthetic image reviewer," 2018.

[6] G. Guo, H. Wang, C. Shen, Y. Yan, and H.-Y. M. Liao, "Automatic image cropping for visual aesthetic enhancement using deep neural networks and cascaded regression," *IEEE Transactions on Multimedia*, vol. 20, no. 8, pp. 2073–2085, 2018.

[7] Y.-S. Chen, Y.-C. Wang, M.-H. Kao, and Y.-Y. Chuang, "Deep photo enhancer: Unpaired learning for image enhancement from photographs with gans," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[8] J. Lovejoy, "The ux of ai: Using google clips to understand how a human-centered design process elevates artificial intelligence," in *2018 AAAI Spring Symposium Series*, 2018.

[9] M. Bolanos, R. Mestre, E. Talavera, X. Giró-i Nieto, and P. Radeva, "Visual summary of egocentric photostreams by representative keyframes," in *2015 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, pp. 1–6, IEEE, 2015.

[10] F. Hu and A. F. Smeaton, "Image aesthetics and content in selecting memorable keyframes from lifelogs," in *International Conference on Multimedia Modeling*, pp. 608–619, Springer, 2018.

[11] S. Wiley, "Self-balancing robot," 2015. US Patent 9908573B2.

[12] https://futurism.com/facebook-wants-robot-personal-photographer.

[13] Z. Byers, M. Dixon, K. Goodier, C. M. Grimm, and W. D. Smart, "An autonomous robot photographer," in *Proceedings 2003 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2003)(Cat. No. 03CH37453)*, vol. 3, pp. 2636–2641, IEEE, 2003.

[14] X. Jin, J. Chi, S. Peng, Y. Tian, C. Ye, and X. Li, "Deep image aesthetics classification using inception modules and fine-tuning connected layer," in *2016 8th International Conference on Wireless Communications & Signal Processing (WCSP)*, pp. 1–6, IEEE, 2016.

[15] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[16] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255, Ieee, 2009.

[17] F. Chollet *et al.*, "Keras." `https://keras.io`, 2015.

[18] M.-E. Nilsback and A. Zisserman, "A visual vocabulary for flower classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, vol. 2, pp. 1447–1454, 2006.

[19] V. Losing, B. Hammer, and H. Wersing, "Interactive online learning for obstacle classification on a mobile robot," in *International Joint Conference on Neural Networks IJCNN, Killarney Ireland*, pp. 2310–2317, IEEE, July 2015.

[20] F. F. Ibarra, O. Kardan, M. R. Hunter, H. P. Kotabe, F. A. C. Meyer, and M. G. Berman, "Image feature types and their predictions of aesthetic preference and naturalness," *Frontiers in Psychology*, vol. 8, p. 632, 2017.

[21] H. Talebi and P. Milanfar, "Nima: Neural image assessment," *IEEE Transactions on Image Processing*, vol. 27, no. 8, pp. 3998–4011, 2018.

[22] A. Torralba, A. A. Efros, *et al.*, "Unbiased look at dataset bias.," in *CVPR*, vol. 1, p. 7, Citeseer, 2011.

[23] D. Liu, R. Puri, N. Kamath, and S. Bhattachary, "Composition-aware image aesthetics assessment," 2019.

[24] E. A. Vessel, N. Maurer, A. H. Denker, and G. G. Starr, "Stronger shared taste for natural aesthetic domains than for artifacts of human culture," *Cognition*, vol. 179, pp. 121–131, 2018.

[25] A. Khosla, A. S. Raju, A. Torralba, and A. Oliva, "Understanding and predicting image memorability at a large scale," in *The IEEE International Conference on Computer Vision (ICCV)*, December 2015.

[26] S. Rayatdoost and M. Soleymani, "Ranking images and videos on visual interestingness by visual sentiment features.," in *MediaEval*, 2016.