Aesthetic Image Classification for Autonomous Agents

Mark Desnoyer and David Wettergreen
Carnegie Mellon University
markd@cmu.edu, dsw@ri.cmu.edu

Abstract

Computational aesthetics is the study of applying machine learning techniques to identify aesthetically pleasing imagery. Prior work used online datasets scraped from large user communities like Flikr to get labeled data. However, online imagery represents results late in the media generation process, as the photographer has already framed the shot and then picked the best results to upload. Thus, this technique can only identify quality imagery once it has been taken. In contrast, automatically creating *pleasing imagery requires understanding the imagery present earlier in the process. This paper applies computational aesthetics techniques to a novel dataset from earlier in that process in order to understand how the problem changes when an autonomous agent, like a robot or a realtime camera aid, creates pleasing imagery instead of simply identifying it.*

1 Introduction

Prior studies in computational aesthetics developed and evaluated various features that correspond to aesthetic quality. In order to perform this evaluation, labeled datasets were scraped from online photography enthusiast websites. The first studies identified image-wide features that weakly correspond to aesthetic quality [4, 7]. This was later refined by Luo et al. who achieved a significant improvement by using focus to segment the subject in an image from its background and calculating new features between these two regions [9]. However, these studies are limited by their datasets. Online imagery represents a product late in the media generation process; the picture has already been framed during the shot, edited by the artist and chosen for upload. In this paper, we analyze imagery earlier in this process to determine how the data changes and how the techniques perform if an autonomous agent were to aid in the *creation* of beautiful imagery.

Other groups have considered creating imagery using automated cameramen. In one early example, Drucker et al. created a framework that allows a virtual scene director to specify high level constraints on the camera motion while resulting motion is determined using constrained optimization [5]. In a real-world application, Lampi et al. developed an automated camera system for capturing lectures in an engaging, but inexpensive way [8].

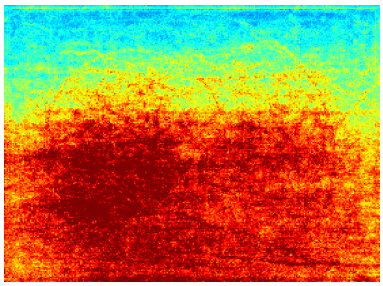


Figure 1: The average laplacian response to landscape imagery. A strong laplacian response correlates with local complexity in an image. Red = high, Blue = low.

However, these techniques all use a bottom-up approach that codify explicit rules based on the techniques taught to professional cinematographers in film school. Though these techniques provide good rules of thumb for taking pleasing shots, beauty is a complex concept and the heuristics sometimes contradictory. Our approach, is to use top-down, data driven techniques, which can automatically adapt based on the context, to quantify aesthetic value.

2 Approach

In this work, features are extracted from images and then used to train a boosted classifier. Features are chosen that can both indicate aesthetic quality and be efficiently computed at runtime. Efficiency is important because we envision the classifier being used in an autonomous agent that must run in near real-time. The boosted classifier automatically selects features, so we can safely evaluate extra features without adding noise to the classification.

2.1 Features

Features selected are inspired by previous studies in human perception [11], analysis of film making guides

and previous work in computational aesthetics [4, 7]. These base features are then applied to a patch of pixels in the image or the image as a whole. Furthermore, since spatial structure and composition is important to photography, we have used four different techniques to segment the images before applying a base feature. The resulting feature/spatial combinations are listed in Table 1.

2.1.1 Spatial Structure

Whole Image This type of segmentation uses the whole image to calculate the base feature.

Spatial Blocking The "rule of thirds" provides a compositional guideline for photographers. It says that if an image is broken into thirds horizontally and vertically, areas of interest should lie on the intersections of the breaks. To encode this type of structure, we break up the image into 3x3 and 5x5 blocks and compute base features on those sub blocks.

Color Segmentations To get an estimate of objects in the scene we segment the image in the CIELAB color space using the mean shift algorithm [3] with a radius of 7. We calculate base features on the 5 largest segments found.

Spatial Ratios Composition of an image can be identified by looking at the relative structure in the scene. For example, Figure 1 shows an image created by applying a laplacian filter to a set of landscape images. The resulting frame shows that the top third of the image tends to have a simple structure (e.g. sky) while there is also a bias for placing main items in the image to the left of the scene. In order to capture this kind of spatial relationship, we use the rule of thirds again and break the image into 3x3 blocks and also into horizontal and vertical bands, creating 15 overlapping image regions. Base features are calculated on these regions and then final features are extracted by calculating all of the pairwise ratios between image regions.

2.1.2 Base Features

Base features are those that can be calculated from an arbitrary patch of pixels. They provide the foundation to identify components of the image that correlate with aesthetics and are calculated on spatial regions generated using the methods described above.

Color Color features represent the dominant colors as well as their distributions. Some features are created by converting the RGB colors into Ohta [10] and HSV color spaces. Two features are used from prior studies: GIST color features [12] and Colorfulness [4] because of their success. Finally, two novel features were created: Color Harmony and Hue Distribution. Color harmony uses the two color theory from [11] to calculate the harmony between the largest five segments of the image. Hue Distribution uses the color wheel from the HSV color space, where complementary colors are 180 degrees apart. The resulting feature is a histogram of the colors and shifted so that the dominant color is at 0 degrees.

Complexity Two features encode image complexity as simpler images are often more striking. The first feature is the number of large segments found when performing a mean shift segmentation. The second, calculated as in [7], measures the size of the bounding box with 75% of the edge energy.

Contrast Contrast is calculated using both the Michelson measure and the RMS of the intensity values in a region.

Texture/Blur Sharp areas in an image often correspond to areas of interest. Thus, we calculate four features based on the texture in the image: the FFT Blur feature from [7], the Gabor Filters from [12], the Spatial Saliency from [6] and the Laplcian response.

Uniqueness Humans tend to be interested in novel things. Therefore, we create a uniqueness measure that is calculated by creating the 16x16 GIST signature [12] for all the images. Then, uniqueness is defined as the mean distance to the three closest images in this high dimensional space.

2.2 Classification

Once features are extracted from the images, they are used to build a classifier. For each dataset, the images are randomly split 70/30 into training/test sets. Images whose ratings are greater than half of a standard deviation are considered to be "pleasing" while those below half of a standard deviation are considered to be "ugly". The images in between these thresholds are ignored as being too ambiguous.

Classification is done using AdaBoost performed on decision stumps as weak classifieries. Each stump thresholds on a single feature and applies a label of "pleasing" one one side and a label of "ugly" on the other. Thus, the classifier simultaneously trains on the data and performs feature selection. The number of boosting iterations is determined using 5-fold cross validation.

3 Evaluation Datasets

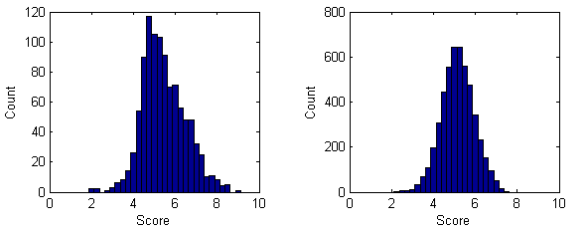
To evaluate this approach for use in image generation, we use two different datasets to represent imagery in different stages of the media generation process.

3.1 DPChallenge

The website dpchallenge.com is a photography forum where users upload their pictures and are ranked by other users of the site. The site was scraped randomly for 4955 images with at least 30 user ratings. This dataset was chosen to represent imagery at the end of the creation process and provide a reference with the

Table 1: List of base features extracted and the spatial structure operators used

	Whole Image	Spatial Blocking	Color Segmentation	Spatial Ratios
Color Harmony [11]			x	
Ohta Color [10]	x	x		
Hue Distribution	x			
HSV	x	x	x	x
Hue Edge Energy	x			x
Colorfulness [4]	x	x	x	
GIST Color [12]		x		
Segment Count	x			
Edge Energy Extents	x			
Contrast	x	x	x	x
Gabor Filters [12]		x		
FFT Blur [7]	x			
Laplacian	x	x	x	x
Spatial Saliency [6]		x		
Uniqueness	x			



(a) Apollo (after normalization) (b) DPChallenge

Figure 2: Human ranking distribution of the two ranked datasets.

work in [4], [7] and [9], all of whom used data from this site. The resulting characteristics of the dataset are shown in Figure 2b.

3.2 Apollo

This dataset consists of a random sample of images taken by the Apollo astronauts while on the Moon [1]. This dataset is biased by human input because humans took the pictures and were explicitly setting up their shots for either media or documentation purposes. However, no post filtering has been done to select the iconic images. Therefore, this dataset represents imagery that is in the middle of the creation process.

The Apollo images were labeled using a survey run on Amazon's Mechanical Turk [2]. Participants were shown a page with 8 random images from the dataset and asked to "rate each image on a scale of 1-10 how beautiful you think the image is" where 1 is "ugly" and 10 is "stunning". Once the scores were collected, the scores from each user are normalized to account for individuals' different scales. The final dataset consists of 1012 images with at least 10 ratings from unique users. The resulting distribution is shown in Figure 2a.

4 Results

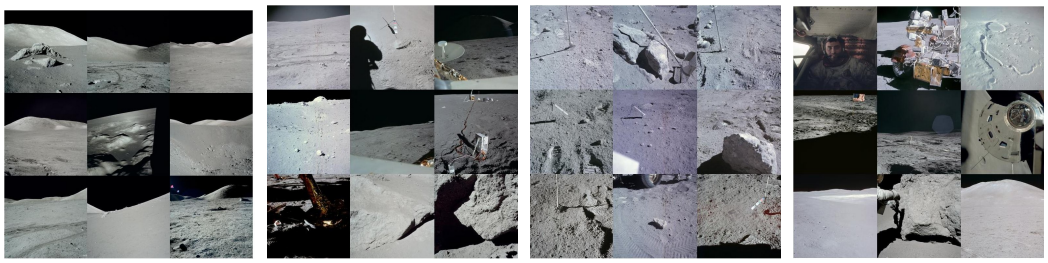
The datasets are evaluated by calculating the classification error (Figure 4) and the precision vs. recall (Figure 5). For the Apollo dataset, the boosted classifier has an 18.1% error rate once boosting is complete at 62 iterations. This compares to a baseline error rate of 20.5% using nearest neighbor on GIST features as described in [12]. In the DPChallenge dataset, the boosted classifier has a 38.2% error rate after 99 boosting iterations, while the GIST baseline was 47.4%.

These results provide insight into the characteristics of the data. In the DPChallenge dataset, the boosted classifier performes significantly better than the GIST baseline, while the improvement on the Apollo dataset is marginal. Earlier in the media pipeline, less filtering has been performed on the imagery so there will be more examples of similar shots. This is perfect for a nearest neighbor approach. In contrast, later in the pipeline, the imagery is more varied so nearest neighbor fails and a classifier that models aesthetics is necessary.

Figure 3 shows samples of the Appollo dataset and how they were classified. From this, we can see that the classifier was able to identify broad structures in the images so that scenes with a horizon and crisp lines were prefered. However, when the striking element of the picture is a smaller structure, the classifier had trouble. For example, the false negatives include scenes with human faces, an astronaut and a surface rift.

Interestingly, the most successful features in the boosted classifier are similar for both the Apollo and DPChallenge datasets. As these datasets are very different, this observation implies that some features, like the laplacian and color saturation are components of universal beauty and that this technique can be successfully trained for many different applications.

Therefore, when developing an agent to create great imagery, a boosted aesthetics classifier could be valuable. However, simpler techniques like nearest neighbor or tracking areas of high laplacian response are likely to



(a) True Positives (b) False Positives (c) True Negatives (d) False Negatives

Figure 3: Examples of classified Apollo images.

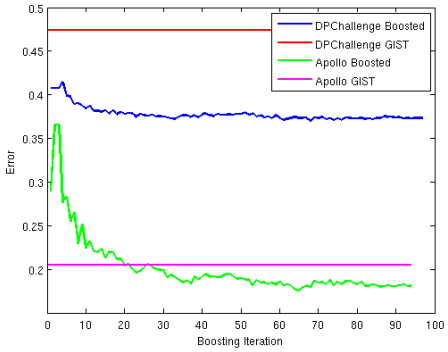


Figure 4: Error rate vs. the number of features selected for the two datasets. Baseline results from nearest neighbor GIST features are also shown.

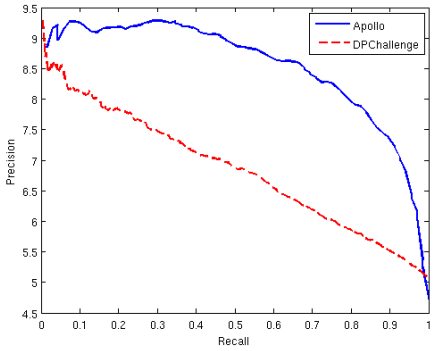


Figure 5: Precision-Recall curves for the two datasets

achieve acceptable performance without undue costs.

5 Conclusions and Future Work

This paper demonstrates that approaches in computational aesthetics can effectively characterize imagery that has not been heavily filtered by human users. However, simpler techniques also perform remarkably well and could more easily be used in autonomous agents that aid in the creation of great imagery. The immediate direction for this work integrates these results into an autonomous robotic cameraman that can create great imagery. Also, a logical extension would be to apply this approach to video where new features are available and there is potential larger amounts of poor quality footage that a user must handle.

References

[1] Project apollo archive. http://www.apolloarchive.com, Oct. 2009.

[2] Mechanical turk. http://www.mturk.com, 2010.

[3] Y. Cheng. Mean shift, mode seeking, and clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(8):790–799, 1995.

[4] R. Datta, D. Joshi, J. Li, and J. Z. Wang. Studying aesthetics in photographic images using a computational approach. *Lecture Notes in Computer Science*, 3953:288, 2006.

[5] S. M. Drucker. *Intelligent Camera Control for Graphical Environments*. Ph.D thesis, Massachusetts Institute of Technology, 1994.

[6] X. Hou and L. Zhang. Saliency detection: A spectral residual approach. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR07). IEEE Computer Society*, pages 1—8, 2007.

[7] Y. Ke, X. Tang, and F. Jing. 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-Volume 1*, pages 419–426. IEEE Computer Society Washington, DC, USA, 2006.

[8] F. Lampi, S. Kopf, M. Benz, and W. Effelsberg. An automatic cameraman in a lecture recording system. In *Proceedings of the international workshop on Educational multimedia and multimedia education*, pages 11–18. ACM Press New York, NY, USA, 2007.

[9] Y. Luo and X. Tang. Photo and video quality evaluation: Focusing on the subject. In *Proceedings of European Conference on Computer Vision (ECCV)*, volume 1, pages 386–399, 2008.

[10] Y. Ohta, T. Kanade, and T. Sakai. Color information for region segmentation. *Computer graphics and image processing*, 13(3):222–241, 1980.

[11] L. Ou and M. R. Luo. A colour harmony model for two-colour combinations. *Color Research & Application*, 31(3):191–204, 2006.

[12] C. Siagian and L. Itti. Rapid biologically-inspired scene classification using features shared with visual attention. *IEEE transactions on pattern analysis and machine intelligence*, 29(2):300, 2007.