Studying Aesthetics in Photographic Images Using a Computational Approach

Ritendra Datta*, Dhiraj Joshi, Jia Li, and James Z. Wang**

The Pennsylvania State University, University Park, PA 16802, USA

Abstract. Aesthetics, in the world of art and photography, refers to the principles of the nature and appreciation of beauty. Judging beauty and other aesthetic qualities of photographs is a highly subjective task. Hence, there is no unanimously agreed standard for measuring aesthetic value. In spite of the lack of firm rules, certain features in photographic images are believed, by many, to please humans more than certain others. In this paper, we treat the challenge of automatically inferring aesthetic quality of pictures using their visual content as a machine learning problem, with a peer-rated online photo sharing Website as data source. We extract certain visual features based on the intuition that they can discriminate between aesthetically pleasing and displeasing images. Automated classifiers are built using support vector machines and classification trees. Linear regression on polynomial terms of the features is also applied to infer numerical aesthetics ratings. The work attempts to explore the relationship between emotions which pictures arouse in people, and their low-level content. Potential applications include content-based image retrieval and digital photography.

1 Introduction

Photography is defined as the art or practice of taking and processing photographs. Aesthetics in photography is how people usually characterize beauty in this form of art. There are various ways in which aesthetics is defined by different people. There exists no single consensus on what it exactly pertains to. The broad idea is that photographic images that are pleasing to the eyes are considered to be higher in terms of their aesthetic beauty. While the average individual may simply be interested in how soothing a picture is to the eyes, a photographic artist may be looking at the composition of the picture, the use of colors and light, and any additional meanings conveyed by the picture. A professional photographer, on the other hand, may be wondering how difficult it may have been to take or to process a particular shot, the sharpness and the color contrast of the picture, or whether the "rules of thumb" in photography have been maintained. All these issues make the measurement of aesthetics in pictures or photographs extremely subjective.

* Corresponding author: R. Datta, datta@cse.psu.edu. More information: http://riemann.ist.psu.edu.

** This work is supported in part by the US National Science Foundation, the PNC Foundation, and SUN Microsystems.

A. Leonardis, H. Bischof, and A. Pinz (Eds.): ECCV 2006, Part III, LNCS 3953, pp. 288–301, 2006.

In spite of the ambiguous definition of aesthetics, we show in this paper that there exist certain visual properties which make photographs, *in general*, more aesthetically beautiful. We tackle the problem computationally and experimentally through a statistical learning approach. This allows us to reduce the influence of exceptions and to identify certain features which are statistically significant in good quality photographs.

Content analysis in photographic images has been studied by the multimedia and vision research community in the past decade. Today, several efficient region-based image retrieval engines are in use [13, 6, 21, 18]. Statistical modeling approaches have been proposed for automatic image annotation [4, 12]. Culturally significant pictures are being archived in digital libraries [7]. Online photo sharing communities are becoming more and more common [1, 3, 11, 15]. In this age of digital picture explosion, it is critical to continuously develop intelligent systems for automatic image content analysis.

1.1 Community-Based Photo Ratings as Data Source

One good data source is a large online photo sharing community, *Photo.net*, possibly the first of its kind, started in 1997 by Philip Greenspun, then a researcher on online communities at MIT [15]. Primarily intended for photography enthusiasts, the Website attracts more than 400,000 registered members. Many amateur and professional photographers visit the site frequently, share photos, and rate and comment on photos taken by peers. There are more than one million photographs uploaded by these users for perusal by the community. Of interest to us is the fact that many of these photographs are peer-rated in terms of two qualities, namely *aesthetics* and *originality*. The scores are given in the range of one to seven, with a higher number indicating better rating. This site acts as the main source of data for our computational aesthetics work. The reason we chose such an online community is that it provides photos which are rated by a relatively diverse group. This ensures generality in the ratings, averaged out over the entire spectrum of amateurs to serious professionals. While amateurs represent the general population, the professionals tend to spend more time on the technical details before rating the photographs. *One caveat:* The nature of any peer-rated community is such that it leads to unfair judgments under certain circumstances, and *Photo.net* is no exception, making our acquired data fairly noisy. Ideally, the data should have been collected from a random sample of human subjects under controlled setup, but resource constraints prevented us from doing so.

We downloaded those pictures and their associated metadata which were rated by at least two members of the community. For each image downloaded, we parsed the pages and gathered the following information: (1) average aesthetics score between 1.0 and 7.0, (2) average originality score between 1.0 and 7.0, (3) number of times viewed by members, and (4) number of peer ratings.

1.2 Aesthetics and Originality

According to the Oxford Advanced Learner's Dictionary, *Aesthetics* means (1) *"concerned with beauty and art and the understanding of beautiful things"*, and

(2) *"made in an artistic way and beautiful to look at"*. A more specific discussion on the definition of aesthetics can be found in [16]. As can be observed, no consensus was reached on the topic among the users, many of whom are professional photographers. *Originality* has a more specific definition of being something that is unique and rarely observed. The originality score given to some photographs can also be hard to interpret, because what seems original to some viewers may not be so for others. Depending on the experiences of the viewers, the originality scores for the same photo can vary considerably. Thus the originality score is subjective to a large extent as well.

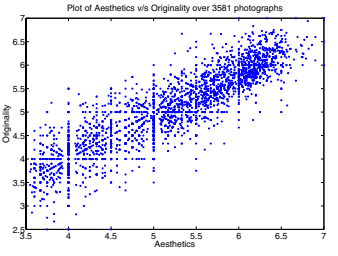


Fig. 1. Correlation between the aesthetics and originality ratings for 3581 photographs

One of the first observations made on the gathered data was the strong correlation between the aesthetics and originality ratings for a given image. A plot of 3581 unique photograph ratings can be seen in Fig. 1. As can be seen, aesthetics and originality ratings have approximately linear correlation with each other. This can be due to a number of factors. Many users quickly rate a batch of photos in a given day. They tend not to spend too much time trying to distinguish between these two parameters when judging a photo. They more often than not rate photographs based on a general impression. Typically, a very original concept leads to good aesthetic value, while beauty can often be characterized by originality in view angle, color, lighting, or composition. Also, because the ratings are averages over a number of people, disparity by individuals may not be reflected as high in the averages. Hence there is generally not much disparity in the average ratings. In fact, out of the 3581 randomly chosen photos, only about 1.1% have a disparity of more than 1.0 between average aesthetics and average originality, with a peak of 2.0.

Fig. 2. Aesthetics scores can be significantly influenced by the semantics. Loneliness is depicted using a person in this frame, though the area occupied by the person is very small. Avg. aesthetics: 6.0/7.0.

As a result of this observation, we chose to limit the rest of our study to aesthetics ratings only, since the value of one can be approximated to the value of the other, and among the two, aesthetics has a rough definition that in principle depends somewhat less on the content or the semantics of the photograph, something that is very hard for present day machine intelligence to interpret accurately. Nonetheless, the strong dependence on originality ratings means that aesthetics ratings are also largely influenced by the semantics. As a result, some visually similar photographs are rated very differently. For example in Fig. 2, loneliness is depicted using a man in the frame, increasing its appeal, while the lack of the person makes the photograph

uninteresting and is likely to cause poorer ratings from peers. This makes the task of automatically determining aesthetics of photographs highly challenging.

1.3 Our Computational Aesthetics Approach

A classic treatise on psychological theories for understanding human perception can be found in [2]. Here, we take the first step in using a computational approach to understand what aspects of a photograph appeal to people, from a population and statistical standpoint. For this purpose, we aim to build (1) a classifier that can qualitatively distinguish between pictures of *high* and *low* aesthetic value, or (2) a regression model that can quantitatively predict the aesthetics score, both approaches relying on low-level visual features only. We define *high* or *low* in terms of predefined ranges of aesthetics scores.

There are reasons to believe that classification may be a more appropriate model than regression in tackling this problem. For one, the measures are highly subjective, and there are no agreed standards for rating. This may render absolute scores less meaningful. Again, ratings above or below certain thresholds on an average by a set of unique users generally reflect on the photograph's quality. This way we also get around the problem of consistency where two identical photographs can be scored differently by different groups of people. However, it is more likely that both the group averages are within the same range and hence are treated fairly when posed as a classification problem.

On the other hand, the 'ideal' case is when a machine can replicate the task of robustly giving images aesthetics scores in the range of (1.0-7.0) the humans do. This is the regression formulation of the problem. The possible benefits of building a *computational aesthetics* model can be summarized as follows: If the low-level image features alone can tell what range aesthetics ratings the image deserves, this can potentially be used by photographers to get a rough estimate of their shot composition quality, leading to adjustment in camera parameters or shot positioning for improved aesthetics. Camera manufacturers can incorporate a 'suggested composition' feature into their products. Alternatively, a content-based image retrieval (CBIR) system can use the aesthetics score to discriminate between visually similar images, giving greater priority to more pleasing query results. Biologically speaking, a reasonable solution to this problem may lead to a better understanding of the human vision.

2 Visual Feature Extraction

Experiences with photography lead us to believe in certain aspects as being critical to quality. This entire study is on such beliefs or hypotheses and their validation through numerical results. We treat each downloaded image separately and extract features from them. We use the following notation: The RGB data of each image is converted to HSV color space, producing two-dimensional matrices I_H, I_S, and I_V, each of dimension $X \times Y$.

Our motivation for the choice of features was principled, based on (1) rules of thumb in photography, (2) common intuition, and (3) observed trends in ratings. In photography and color psychology, color tones and saturation play important roles, and hence working in the HSV color space makes computation more convenient. For some features we extract information from objects within the photographs. An approximate way to find objects within images is segmentation, under the assumption that homogeneous regions correspond to objects. We use a fast segmentation method based on clustering. For this purpose the image is transformed into the LUV space, since in this space locally Euclidean distances model the perceived color change well. Using a fixed threshold for all the photographs, we use the K-Center algorithm to compute cluster centroids, treating the image pixels as a bag of vectors in LUV space. With these centroids as seeds, a K-means algorithm computes clusters. Following a connected component analysis, color-based segments are obtained. The 5 largest segments formed are retained and denoted as $\{s_1, ..., s_5\}$. These clusters are used to compute *region-based features* as we shall discuss in Sec. 2.7.

We extracted 56 visual features for each image. The feature set was carefully chosen but limited because our goal was mainly to study the trends or patterns, if any, that lead to higher or lower aesthetics ratings. If the goal was to only build a strong classifier or regression model, it would have made sense to generate exhaustive features and apply typical machine-learning techniques such as boosting. Without meaningful features it is difficult to make meaningful conclusions from the results. We refer to our features as *candidate features* and denote them as $\mathcal{F} = \{f_i | 1 \leq i \leq 56\}$ which are described as follows.

2.1 Exposure of Light and Colorfulness

Measuring the brightness using a light meter and a gray card, controlling the exposure using the aperture and shutter speed settings, and darkroom printing with dodging and burning are basic skills for any professional photographer. Too much exposure (leading to brighter shots) often yields lower quality pictures. Those that are too dark are often also not appealing. Thus light exposure can often be a good discriminant between high and low quality photographs. Note that there are always exceptions to any 'rules of thumb'. An over-exposed or under-exposed photograph under certain scenarios may yield very original and beautiful shots. Ideally, the use of light should be characterized as normal daylight, shooting into the sun, backlighting, shadow, night etc. We use the average pixel intensity $f_1 = \frac{1}{XY} \sum_{x=0}^{X-1} \sum_{y=0}^{Y-1} I_V(x, y)$ to characterize the use of light.

We propose a fast and robust method to compute relative color distribution, distinguishing multi-colored images from monochromatic, *sepia* or simply low contrast images. We use the Earth Mover's Distance (EMD) [17], which is a measure of similarity between any two weighted distributions. We divide the RGB color space into 64 cubic blocks with four equal partitions along each dimension, taking each such cube as a sample point. Distribution D_1 is generated as the color distribution of a hypothetical image such that for each of 64 sample points, the frequency is 1/64. Distribution D_2 is computed from the given image

by finding the frequency of occurrence of color within each of the 64 cubes. The EMD measure requires that the pairwise distance between sampling points in the two distributions be supplied. Since the sampling points in both of them are identical, we compute the pairwise Euclidean distances between the geometric centers c_i of each cube i, after conversion to LUV space. Thus the *colorfulness* measure f_2 is computed as follows: $f_2 = emd(D_1, D_2, \{d(a,b) \mid 0 \leq a, b \leq 63\})$, where $d(a,b) = ||\text{rgb2luv}(c_a) - \text{rgb2luv}(c_b)||$.

Fig. 3. The proposed *colorfulness* measure. The two photographs on the *left* have high values while the two on the *right* have low values.

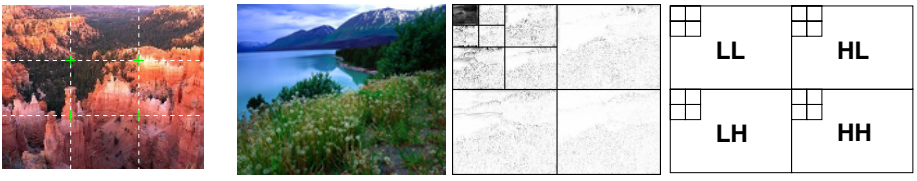
The distribution D_1 can be interpreted as the *ideal* color distribution of a 'colorful' image. How similar the color distribution of an arbitrary image is to this one is a rough measure of how colorful that image is. Examples of images producing high and low values of f_2 are shown in Fig. 3.

2.2 Saturation and Hue

Saturation indicates chromatic purity. Pure colors in a photo tend to be more appealing than dull or impure ones. In natural out-door landscape photography, professionals use specialized film such as the *Fuji Velvia* to enhance the saturation to result in deeper blue sky, greener grass, more vivid flowers, etc. We compute the average saturation $f_3 = \frac{1}{XY} \sum_{x=0}^{X-1} \sum_{y=0}^{Y-1} I_S(x,y)$ as the saturation indicator. Hue is similarly computed averaged over I_H to get feature f_4, though the interpretation of such a feature is not as clear as the former. This is because hue as defined in the HSV space corresponds to angles in a color wheel.

2.3 The Rule of Thirds

A very popular rule of thumb in photography is the *Rule of Thirds*. The rule can be considered as a sloppy approximation to the 'golden ratio' (about 0.618). It specifies that the main element, or the center of interest, in a photograph should lie at one of the four intersections as shown in Fig. 4 (a). We observed that most professional photographs that follow this rule have the main object stretch from an intersection up to the center of the image. Also noticed was the fact that centers of interest, e.g., the eye of a man, were often placed aligned to one of the edges, on the inside. This implies that a large part of the main object often lies on the periphery or inside of the inner rectangle. Based on these observations, we computed the average hue as $f_5 = \frac{9}{XY} \sum_{x=X/3}^{2X/3} \sum_{y=Y/3}^{2Y/3} I_H(x,y)$, with f_6 and f_7 being similarly computed for I_S and I_V respectively.



(a) (b) (c) (d)

Fig. 4. (a) The *rule of thirds* in photography: Imaginary lines cut the image horizontally and vertically each into three parts. Intersection points are chosen to place important parts of the composition instead of the center. (b)-(d) Daubechies wavelet transform. *Left:* Original image. *Middle:* Three-level transform, levels separated by borders. *Right:* Arrangement of three bands LH, HL and HH of the coefficients.

2.4 Familiarity Measure

We humans learn to rate the aesthetics of pictures from the experience gathered by seeing other pictures. Our opinions are often governed by what we have seen in the past. Because of our curiosity, when we see something unusual or rare we perceive it in a way different from what we get to see on a regular basis. In order to capture this factor in human judgment of photography, we define a new measure of *familiarity* based on the integrated region matching (IRM) image distance [21]. The IRM distance computes image similarity by using color, texture and shape information from automatically segmented regions, and performing a robust region-based matching with other images. Primarily meant for image retrieval applications, we use it here to quantify familiarity. Given a pre-determined *anchor* database of images with a well-spread distribution of aesthetics scores, we retrieve the top K closest matches in it with the candidate image as query. Denoting IRM distances of the top matches for each image in decreasing order of rank as $\{q(i)|1 \leq i \leq K\}$. We compute f_8 and f_9 as $f_8 = \frac{1}{20}\sum_{i=1}^{20} q(i)$, $f_9 = \frac{1}{100}\sum_{i=1}^{100} q(i)$.

In effect, these measures should yield higher values for uncommon images. Two different scales of 20 and 100 top matches are used since they may potentially tell different stories about the uniqueness of the picture. While the former measures average similarity in a local neighborhood, the latter does so on a more global basis. Because of the strong correlation between aesthetics and originality, it is intuitive that a higher value of f_8 or f_9 corresponds to greater originality and hence we expect greater aesthetics score.

2.5 Wavelet-Based Texture

Graininess or smoothness in a photograph can be interpreted in different ways. If as a whole it is grainy, one possibility is that the picture was taken with a grainy film or under high ISO settings. If as a whole it is smooth, the picture can be out-of-focus, in which case it is in general not pleasing to the eye. Graininess can also indicate the presence/absence and nature of *texture* within the image.

The use of texture is a composition skill in photography. One way to measure spatial smoothness in the image is to use Daubechies wavelet transform [10],

which has often been used in the literature to characterize texture. We perform a *three-level* wavelet transform on all three color bands I_H, I_S and I_V. An example of such a transform on the intensity band is shown in Fig. 4 (b)-(c). The three levels of wavelet bands are arranged from top left to bottom right in the transformed image, and the four coefficients per level, LL, LH, HL, and HH are arranged as shown in Fig. 4 (d). Denoting the coefficients (except LL) in level i for the wavelet transform on hue image I_H as w_i^{hh}, w_i^{hl} and w_i^{lh}, $i = \{1, 2, 3\}$, we define features f_{10}, f_{11} and f_{12} as follows:

$$f_{i+9} = \frac{1}{S_i}\left\{ \sum_x \sum_y w_i^{hh}(x,y) \; + \sum_x \sum_y w_i^{hl}(x,y) \; + \quad \sum_x \sum_y w_i^{lh}(x,y) \right\}$$

where $S_k = |w_i^{hh}| + |w_i^{hl}| + |w_i^{hh}|$ and $i = 1, 2, 3$. The corresponding wavelet features for saturation (I_S) and intensity (I_V) images are computed similarly to get f_{13} through f_{15} and f_{16} through f_{18} respectively. Three more wavelet features are derived. The sum of the average wavelet coefficients over all three frequency levels for each of H, S and V are taken to form three additional features: $f_{19} = \sum_{i=10}^{12} f_i$, $f_{20} = \sum_{i=13}^{15} f_i$, and $f_{21} = \sum_{i=16}^{18} f_i$.

2.6 Size and Aspect Ratio

The size of an image has a good chance of affecting the photo ratings. Although scaling is possible in digital and print media, the size presented initially must be agreeable to the content of the photograph. A more crucial parameter is the aspect ratio. It is well-known that $4:3$ and $16:9$ aspect ratios, which approximate the 'golden ratio,' are chosen as standards for television screens or $70mm$ movies, for reasons related to viewing pleasure. The 35mm film used by most photographers has a ratio of $3:2$ while larger formats include ratios like $7:6$ and $5:4$. While size feature is $f_{22} = X + Y$, the aspect ratio feature is $f_{23} = \frac{X}{Y}$.

2.7 Region Composition

Segmentation results in rough grouping of similar pixels, which often correspond to objects in the scene. We denote the set of pixels in the largest five connected components or *patches* formed by the segmentation process described before as $\{s_1, ...s_5\}$. The number of patches $t \leq 5$ which satisfy $|s_i| \geq \frac{XY}{100}$ denotes feature f_{24}. The number of color-based clusters formed by K-Means in the LUV space is feature f_{25}. This number is image dependent and dynamically chosen, based on the complexity of the image. These two features combine to measure how many distinct color *blobs* and how many disconnected significantly large regions are present.

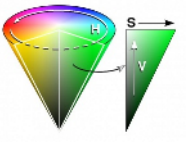


Fig. 5. The HSV Color Wheel

We then compute the average H, S and V values for each of the top 5 patches as features f_{26} through f_{30}, f_{31} through f_{35} and f_{36} through f_{40} respectively. Features f_{41} through f_{45} store the relative size of each segment with respect to the image, and are computed as $f_{i+40} = |s_i|/(XY)$ where $i = 1, ..., 5$.

The hue component of HSV is such that the colors that are 180° apart in the color circle (Fig. 5) are complimentary to each other, which means that they add up to 'white' color. These colors tend to look pleasing together. Based on this idea, we define two new features, f_{46} and f_{47} in the following manner, corresponding to *average color spread* around the wheel and *average complimentary colors* among the top 5 patch hues. These features are defined as

$$f_{46} = \sum_{i=1}^{5}\sum_{j=1}^{5}|h_i - h_j|, \quad f_{47} = \sum_{i=1}^{5}\sum_{j=1}^{5} l(|h_i - h_j|), \quad h_i = \sum_{(x,y)\in s_i} I_H(x,y),$$

where $l(k) = k$ if $k \leq 180°$, $360° - k$ if $k > 180°$. Finally, the rough positions of each segment are stored as features f_{48} through f_{52}. We divide the image into 3 equal parts along horizontal and vertical directions, locate the block containing the centroid of each patch s_i, and set $f_{47+i} = (10r + c)$ where $(r, c) \in \{(1,1), ..., (3,3)\}$ indicates the corresponding block starting with top-left.

2.8 Low Depth of Field Indicators

Pictures with a simplistic composition and a well-focused center of interest are sometimes more pleasing than pictures with many different objects. Professional photographers often reduce the depth of field (DOF) for shooting single objects by using larger aperture settings, macro lenses, or telephoto lenses. DOF is the range of distance from a camera that is acceptably sharp in the photograph. On the photo, areas in the DOF are noticeably sharper.

We noticed that a large number of low DOF photographs, e.g., insects, other small creatures, animals in motion, were given high ratings. One reason may be that these shots are difficult to take, since it is hard to focus steadily on small and/or fast moving objects like insects and birds. A common feature is that they are taken either by *macro* or by telephoto lenses. We propose a novel method to detect low DOF and macro images. We divide the image into 16 equal rectangular blocks $\{M_1, ...M_{16}\}$, numbered in row-major order. Let $w_3 = \{w_3^{lh}, w_3^{hl}, w_3^{hh}\}$ denote the set of wavelet coefficients in the high-frequency (level 3 by the notation in Sec. 2.5) of the hue image I_H. The *low depth of field indicator* feature f_{53} for hue is computed as follows, with f_{54} and f_{55} being computed similarly for I_S and I_V respectively:

$$f_{53} = \frac{\sum_{(x,y)\in M_6\cup M_7\cup M_{10}\cup M_{11}} w_3(x,y)}{\sum_{i=1}^{16}\sum_{(x,y)\in M_i} w_3(x,y)}$$

The object of interest in a macro shot is usually in sharp focus near the center, while the surrounding is usually out of focus. This essentially means that a large value of the low DOF indicator features tend to occur for macro shots.

2.9 Shape Convexity

It is believed that shapes in a picture also influence the degree of aesthetic beauty perceived by humans. The challenge in designing a shape feature lies in the understanding of what kind of shape pleases humans, and whether any such

Fig. 6. Demonstrating the *shape convexity* feature. *Left*: Original photograph. *Middle*: Three largest non-background segments shown in original color. *Right*: Exclusive regions of the *convex hull* generated for each segment are shown in white. The proportion of white regions determine the convexity value.

measure generalizes well enough or not. As always, we hypothesize that convex shapes like perfect moon, well-shaped fruits, boxes, or windows have an appeal, positive or negative, which is different from concave or highly irregular shapes. Let the image be segmented, as described before, and R patches $\{p_1, ..., p_R\}$ are obtained such that $|p_k| \geq \frac{XY}{200}$. For each p_k, we compute its convex hull, denoted by $g(p_k)$. For a perfectly convex shape, $p_k \cap g(p_k) = p_k$, i.e. $\frac{|p_k|}{|g(p_k)|} = 1$. We define the *shape convexity* feature as $f_{56} = \frac{1}{XY}\{\sum_{k=1}^{R} I(\frac{|p_k|}{|g(p_k)|} \geq 0.8)|p_k|\}$, allowing some room for irregularities of edge and error due to digitization. Here $I(\cdot)$ is the indicator function. This feature can be interpreted as the fraction of the image covered by approximately convex-shaped homogeneous regions, ignoring the insignificant image regions. This feature is demonstrated in Fig. 6. Note that a critical factor here is the segmentation process, since we are characterizing shape by segments. Often, a perfectly convex object is split into concave or irregular parts, considerably reducing the reliability of this measure.

3 Feature Selection, Classification, and Regression

A contribution of our work is the feature extraction process itself, since each feature represents an interesting aspects of photography. We now perform selection in order to (1) discover features that show correlation with community-based aesthetics scores, and (2) build a classification/regression model using a subset of strongly/weakly relevant features such that generalization performance is near optimal. Instead of using any regression model, we use a one-dimensional support vector machine (SVM) [20]. SVMs are essentially powerful binary classifiers that project the data space into higher dimensions where the two classes of points are linearly separable. Naturally, for one-dimensional data, they can be more flexible than a single threshold classifier.

For the 3581 images downloaded, all 56 features in $\mathcal{F}$ were extracted and normalized to the $[0, 1]$ range to form the experimental data. Two classes of data are chosen, *high* containing samples with aesthetics scores greater than 5.8, and *low* with scores less than 4.2. Only images that were rated by at least two unique members were used. The reason for choosing classes with a gap is that pictures with close lying aesthetic scores, e.g., 5.0 and 5.1 are not likely to have any distinguishing fea-

ture, and may merely be representing the noise in the whole peer-rating process. For all experiments we ensure equal priors by replicating data to generate equal number of samples per class. A total of 1664 samples is thus obtained, forming the basis for our classification experiments. We perform classification using the standard RBF Kernel ($\gamma = 3.7$, $cost = 1.0$) using the LibSVM package [9]. SVM is run 20 times per feature, randomly permuting the data-set each time, and using a 5-fold cross-validation (5-CV). The top 15 among the 56 features in terms of model accuracy are obtained. The stability of these single features as classifiers is also tested. We proceed to build a classifier that can separate *low* from *high*. For this, we use SVM as well as the classification and regression trees (CART) algorithm [8]. While SVM is a powerful classifier, a limitation is that when there are too many irrelevant features in the data, the *generalization performance* tends to suffer. Feature selection for classification purposes is a well-studied topic [5], with some recent work related specifically to feature selection for SVMs. *Filter-based methods* and *wrapper-based methods* are two broad techniques for feature selection. While the former eliminates irrelevant features before training the classifier, the latter chooses features using the classifier itself as an integral part of the selection process. In this work, we combine these two methods to reduce computational complexity while obtaining features that yield good generalization performance: (1) The top 30 features in terms of their one-dimensional SVM performance are retained while the rest of the features are *filtered* out. (2) We use *forward selection*, a wrapper-based approach in which we start with an empty set of features and iteratively add one feature at a time that increases the 5-fold CV accuracy the most. We stop at 15 iterations (i.e. 15 features) and use this set to build the SVM-based classifier.

Classifiers that help understand the influence of different features directly are tree-based approaches such as CART. We used the recursive partitioning (RPART) implementation [19], to build a two-class classification tree model for the same set of 1664 data samples. Finally, we perform linear regression on polynomial terms of the features values to see if it is possible to directly predict the aesthetics scores in the 1 to 7 range from the feature vector. The quality of regression is usually measured in terms of the *residual sum-of-squares error* $R^2_{res} = \frac{1}{N-1}\sum_{i=1}^{N}(Y_i - \hat{Y}_i)^2$ where $\hat{Y}_i$ is the predicted value of Y_i. Here Y being the aesthetics scores, in the worst case $\bar{Y}$ is chosen every time without using the regression model, yielding $R^2_{res} = \sigma^2$ (variance of Y). Hence, if the independent variables explain something about Y, it must be that $R_{res} \leq \sigma^2$. For this part, all 3581 samples are used, and for each feature f_i, the polynomials (f_i, f_i^2, f_i^3, $f_i^{\frac{1}{3}}$, and $f_i^{\frac{2}{3}}$) are used as independent variables.

4 Experimental Results

For the one-dimensional SVM performed on individual features, the top 15 results obtained in decreasing order of 5-CV accuracy are as follows: $\{f_{31}, f_1, f_6, f_{15}, f_9, f_8, f_{32}, f_{10}, f_{55}, f_3, f_{36}, f_{16}, f_{54}, f_{48}, f_{22}\}$. The maximum classification rate achieved by any single feature was f_{31} with 59.3%. With accuracy over 54%, they act as weak classifiers and hence show some correlation with the aesthetics.

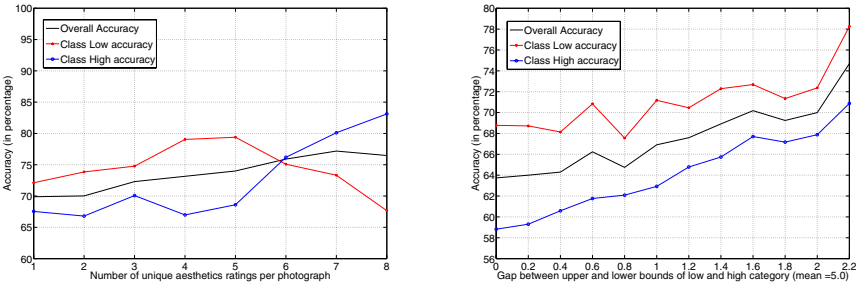


Fig. 7. *Left*: Variation of $5-CV$ SVM accuracy with the minimum number of unique ratings per picture. *Right*: Variation of $5-CV$ SVM accuracy with inter-class gap δ.

The combined filter and wrapper method for feature selection yielded the following set of 15 features:$\{f_{31}, f_1, f_{54}, f_{28}, f_{43}, f_{25}, f_{22}, f_{17}, f_{15}, f_{20}, f_2, f_9, f_{21}, f_{23}, f_6\}$. The accuracy achieved with 15 features is 70.12%, with precision of detecting *high* class being 68.08%, and *low* class being 72.31%. Considering the nature of this problem, these classification results are indeed promising.

The stability of these classification results in terms of number of ratings is considered next. Samples are chosen in such a way that each photo is rated by at least K unique users, K varying from 1 to 8, and the 5-CV accuracy and precision plotted, as shown in Fig. 7. It is observed that accuracy values show an upward trend with increasing number of unique ratings per sample, and stabilize somewhat when this value touches 5. This reflects on the peer-rating process - the inherent noise in this data gets averaged out as the number of ratings increase, converging toward a somewhat 'fair' score. We then experimented with how accuracy and precision varied with the gap in aesthetics ratings between the two classes *high* and *low*. So far we have considered ratings ≥ 5.8 as *high* and ≤ 4.2 as *low*. In general, considering that *ratings* $\geq 5.0 + \frac{\delta}{2}$, be *(high) and ratings* $\leq 5.0 - \frac{\delta}{2}$ *be (*low), we have based all classification experiments on $\delta = 1.6$. The value 5.0 is chosen as it is the *median* aesthetics rating over the 3581 samples. We now vary δ while keeping all other factors constant, and compute SVM accuracy and precision for each value. These results are plotted in Fig. 7. Not surprisingly, the accuracy increases as δ increases. This is accounted by the fact that as δ increases, so does the distinction between the two classes.

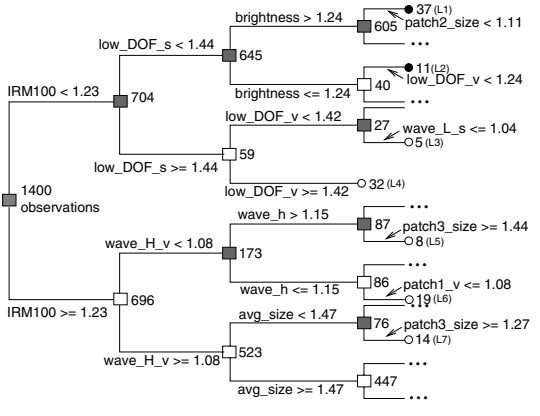


Fig. 8. Decision tree obtained using CART and the 56 visual features (partial view)

Fig. 8 shows the CART decision tree obtained using the 56 visual features. In the figures, the decision nodes are denoted by squares while leaf nodes are denoted by circles. The decisions used at each split and the number of observations which fall in each node during the decision process, are also shown in the figures. Shaded nodes have a higher percentage of *low* class pictures, hence making them *low nodes*, while un-shaded nodes are those where the dominating class is *high*. The RPART implementation uses 5-CV to prune the tree to yield lowest risk. We used a 5-fold cross validation scheme. With *complexity parameter* governing the tree complexity set to 0.0036, the tree generated 61 splits, yielding an 85.9% model accuracy and a modest 62.3% 5-CV accuracy. More important than the accuracy, the tree provides us with a lot of information on how aesthetics can be related to individual features. We do not have the space to include and discuss the entire tree. Let us discuss some interesting decision paths, in each tree, which support our choice of features. The features denoted by $IRM100$, i.e. f_9, and the low DOF indicators for S and V components, respectively (denoted by low_DOF_s, i.e. f_{54} and low_DOF_v, i.e. f_{55}), appear to play crucial roles in the decision process. The expected loss at L_3 and L_4 are 0% and 9%, respectively. A large numeric value of the low DOF indicators shows that the picture is focused on a central object of interest. As discussed before, taking such pictures requires professional expertise and hence high peer rating is not unexpected.

Finally, we report the regression results. The variance σ^2 of the aesthetics score over the 3581 samples is 0.69. With 5 polynomial terms for each of the 56, we achieved a residual sum-of-squares $R^2_{res} = 0.5020$, which is a 28% reduction from the variance σ^2. This score is not very high, but considering the challenge involved, this does suggest that visual features are able to predict human-rated aesthetics scores with some success. To ensure that this was actually demonstrating some correlation, we randomly permuted the aesthetics scores (breaking the correspondence with the features) and performed the same regression. This time, R_{res} is 0.65, clearly showing that the reduction in expected error was not merely by the over-fitting of a complex model.

5 Conclusions and Future Work

We have established significant correlation between various visual properties of photographic images and their aesthetics ratings. We have shown, through using a community-based database and ratings, that certain visual properties tend to yield better discrimination of aesthetic quality than some others. Despite the inherent noise in data, our SVM-based classifier is robust enough to produce good accuracy using only 15 visual features in separating *high* and *low* rated photographs. In the process of designing the classifier, we have developed a number of new features relevant to photographic quality, including a low depth-of-field indicator, a colorfulness measure, a shape convexity score and a familiarity measure. Even though certain extracted features did not show a significant correlation with aesthetics, they may have applications in other photographic image analysis work as they are sound formulations of basic principles

in photographic art. In summary, our work is a significant step toward the highly challenging task of understanding the correlation of human emotions and pictures they see by a computational approach. There are yet a lot of open avenues in this direction. The accuracy can potentially be improved by incorporating new features like dominant lines, converging lines, light source classification, and subject-background relationships.

References

1. Airlines.Net, `http://www.airliners.net`.
2. R. Arnheim, *Art and Visual Perception: A Psychology of the Creative Eye,* University of California Press, Berkeley, 1974.
3. ARTStor.org, `http://www.artstor.org`.
4. K. Barnard, P. Duygulu, D. Forsyth, N. -de. Freitas, D. M. Blei, and M. I. Jordan, "Matching Words and Pictures," *J. Machine Learning Research*, 3:1107–1135, 2003.
5. A. L. Blum and P. Langley, "Selection of Relevant Features and Examples in Machine Learning," *Artificial Intelligence*, 97(1-2):245–271, 1997.
6. C. Carson, S. Belongie, H. Greenspan, and J. Malik, "Blobworld: Color and Texture-Based Image Segmentation using EM and its Application to Image Querying and Classification," *IEEE Trans. on Pattern Analysis and Machine Intelli.*, 24(8):1026–1038, 2002.
7. C-c. Chen, H. Wactlar, J. Z. Wang, and K. Kiernan, "Digital Imagery for Significant Cultural and Historical Materials - An Emerging Research Field Bridging People, Culture, and Technologies," *Int'l J. on Digital Libraries*, 5(4):275–286, 2005.
8. L. Breiman, J.H. Friedman, R.A. Olshen, and C.J. Stone, *Classification and Regression Trees,* Wadsworth, Belmont, CA, 1983.
9. C.-c. Chang, C.-j. Lin, "LIBSVM : A Library for SVM", `http://www.csie.ntu.edu.tw/~cjlin/libsvm,` 2001.
10. I. Daubechies, *Ten Lectures on Wavelets*, Philadelphia, SIAM, 1992.
11. Flickr, `http://www.flickr.com`.
12. J. Li and J. Z. Wang, "Automatic Linguistic Indexing of Pictures by a Statistical Modeling Approach," *IEEE Trans. on Pattern Analysis and Machine Intelli.*, 25(9):1075–1088, 2003.
13. W. Y. Ma and B. S. Manjunath, "NeTra: A Toolbox for Navigating Large Image Databases," *Multimedia Systems*, 7(3):184–198, 1999.
14. B.S. Manjunath, W.Y. Ma, "Texture Features for Browsing and Retrieval of Image Data", *IEEE Trans. on Pattern Analysis and Machine Intelli.*, 18(8):837–842, 1996.
15. Photo.Net, `http://www.photo.net`.
16. Photo.Net Rating System, `http://photo.net/gallery/photocritique/standards`.
17. Y. Rubner, C. Tomasi, L.J. Guibas, "The Earth Mover's Distance as a Metric for Image Retrieval," *Int'l. J. Computer Vision*, 4(2):99–121, 2000.
18. A. W. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, "Content-Based Image Retrieval at the End of the Early Years," *IEEE Trans. on Pattern Analysis and Machine Intelli.*, 22(12):1349–1380, 2000.
19. T. M. Therneau and E. J. Atkinson, "An Introduction to Recursive Partitioning Using RPART Routines," *Technical Report, Mayo Foundation*, 1997.
20. V. Vapnik, *The Nature of Statistical Learning Theory,* Springer, 1995.
21. J. Z. Wang, J. Li, and G. Wiederhold, "SIMPLIcity: Semantics-Sensitive Integrated Matching for Picture Libraries," *IEEE Trans. on Pattern Analysis and Machine Intelli.*, 23(9):947–963, 2001.