IMAGE AESTHETICS DEPENDS ON CONTEXT

Florian Simond, Nikolaos Arvanitopoulos and Sabine Süsstrunk

School of Computer and Communication Sciences
École Polytechnique Fédérale de Lausanne (EPFL), Switzerland

ABSTRACT

We investigate the influence of low-level image features for aesthetics prediction. We show that the aesthetic quality of a photography depends on its context. Image features learned from a specific image category are not necessarily the same as features learned from a generic image collection. Experiments conducted on specific image categories show that specific features obtain statistically significantly better results than generic ones.

***Index Terms*—** aesthetic quality, feature extraction, classification.

1. INTRODUCTION

Given the ever increasing number of photographs taken, a rating of an image based on its aesthetic values can be very helpful in many different application scenarios such as personal image collections, image enhancements, photo book creation, and social media interactions. While "beauty" and "interestingness" are subjective criteria that can be very personal, certain correlations with measurable image content and quality features has been found [1, 2]. Consequently, automatically assessing the aesthetic value of a photograph based on low-level features is currently an active research topic [3, 4, 5, 6, 7, 8, 9, 10, 11].

What is common to all of these methods is that they evaluate the aesthetics of an image solely on the features of the image itself, which are either handcrafted or learned over a large database of images. As such, these methods are agnostic to the *context* of an individual image. Yet, one might wonder if the features that make a "good" or "bad" image are truly independent of what the image is meant to illustrate. For example, are features based on saliency as predictive of the aesthetics of an image when evaluating a city skyline as they are when evaluating an image containing animals? Can we just learn the correct features over a large corpus of images, or should we limit our training data to images that are labeled the same?

In this paper, we investigate if the context of an image has an influence on which low-level images features are best suited to evaluate aesthetic. The context is determined by the semantic label given to the image, i.e. its keywords. Deselaers and Ferrari [12] have shown, by analysing the images in ImageNet [13] that images with semantically similar annotations have more visual attributes in common than images with dissimilar annotations. Lindner et al. [14] have also found that images with the same keywords can have features that are statistically significantly different than images that are not annotated with that keyword. We thus investigate if this semantic induced difference in observed features relates also to the *aesthetics* of an image, or only to its content.

We first extract a large number of low-level features that may be useful in predicting image aesthetics. We then use the Sequential Forward Floating Search (SFFS) [15] to evaluate which features are most significant. Our ground truth image set is the large-scale AVA dataset [9] that contains 250K images with aesthetic scores and semantic labels. We select a subset of 80K images that have either very low or very high ratings to avoid evaluation reliability issues and unbalanced training sets. This resulting set contains arbitrary images and images that have the following labels: animals, nature, portraits, and city.

We then use SFFS to search for the best features over the global subset as well as over the individual categories determined by the semantic labels. We find that the best features are not exactly the same for the global and the labeled image sets, even though many are similar. Features based on saliency, for example, are only highly ranked for the categories that often contain a single object, such as animals or portrait. On the other hand, sharpness is an aesthetic attribute that is important for all images independent of their context.

To evaluate if the differences are statistically significant, we split each of the subsets into a training and testing set. We train each training set twice, once with the best features found for the global set, and once with the best features found for the specific semantic label that corresponds to that training set. We then test how accurately each model predicts the aesthetic ratings. We consistently obtain higher accuracy with the best label specific features than with the generic features, and the results are statistically significant.

2. STATE-OF-THE-ART

All of the state-of-the-art approaches for aesthetic quality prediction use the same two-step algorithmic pipeline: (1) image feature extraction and (2) classification of image quality according to the extracted features. The main focus of all the methods in the literature is on the design of features that will give accurate classification results. Several works have investigated intuitive, low-level image features for quality prediction [2, 3, 4, 5, 6]. They define features related to color, sharpness, relation between foreground and background, salient objects, etc. These features become the new image representation and are used as input to a classifier, such as the Support Vector Machine [16].

Another category of approaches [7, 8, 10] uses generic image descriptors, such as SIFT [17] together with image features for quality prediction. They report superior prediction accuracies compared to methods that exclusively use image features. A latest method [11] utilizes a deep convolutional neural network to automatically learn features from an artificially augmented version of the AVA dataset [9]. These methods obtain state-of-the-art results, however, they lack interpretation of the features in terms of their influence on the final prediction accuracy.

3. METHODOLOGY

Our proposed method consists of two main steps:

1. Feature extraction for aesthetic image prediction.
2. Feature selection using as a criterion the aesthetic prediction accuracy.

3.1. Feature extraction

We extract a set of 35 aesthetic quality features from each image. The features can be divided into three categories: (1) features that describe the whole image, (2) features that describe the salient region and (3) features that relate the main subject with the background (as defined in [4]). A brief description of each feature, associated with its category, can be found in Table 1.

Most of the features described in Table 1 are also used in other works [2, 18, 4]. However, we compute some of the features in a different way. The following features are computed differently with respect to the related work: (1) #Edges, (2) Sharpness, (3) Rule of Thirds (4) Salient region computation and (5) Color variance. It is worth mentioning the difference in the computation of the rule of thirds and the salient region computation, because they are very important features that photographers use to compose their photographs. In [2, 4], the rule of thirds is computed as the mean values, in the HSV color space, of the central region of the image. However, this modeling does not take into account the spatial relationship of the objects inside the composition. We approximate this rule by computing the shortest distance of the salient region to a power point. A power point is one of the intersection points of the two horizontal and the two vertical lines that split a photograph in nine equally sized regions. We believe that this modeling is more intuitive and represents the rule of thirds in a more robust and reliable way.

Name	**Description**
Brightness AVG and STD	(1) Average and standard deviation of the brightness, using the V channel in the HSV space.
Color Variance	(1) Variance of colors in the LAB space.
Contrast	(1) Width of the middle 96% mass of the histogram of the V channel in the HSV space.
#Edges and #Edges L, R, T, B, C	(1) We split the canny map into 16×16 blocks and we compute the number of blocks containing more than 10% of edges. We also compute this number on the left, right, top, bottom and center regions of the image.
Hue Count	(1) Approximation of the number of unique hues [18].
Saturation AVG and STD	(1) Average and standard deviation of the saturation.
Sharpness	(1) Variance of the Laplacian. [20]
Distance to the Center	(2) Distance of the salient region to the center of the image.
Rule of Thirds	(2) Shortest distance of the salient region to a power point.
Salient Hue, Brightness and Saturation	(2) Average hue, brightness and saturation of the salient region.
Salient Sharpness	(2) Sharpness of the salient region.
Salient Size	(2) Size of the salient region.
Salient LOC	(2) We split the image into nine equal parts, and compute the proportion of the salient region in each part. LOC can then take nine values: Top-Left, Middle-Left, Bottom-Right...
Color Difference	(3) Difference of colors in the LAB space between the salient object and the background.
Hue, Saturation and Brightness Difference	(3) Difference of hue, saturation and brightness between the salient region and the background.
Sharpness Difference	(3) Difference of sharpness between the salient region and the background.

Table 1: Description of the features used for aesthetic quality prediction.

Furthermore, we use a more recent algorithm for saliency detection [19]. More accurate salient object detection will result in a more robust computation of all the features of Table 1 that depend on the salient region.

Finally, we also propose the following new features: (1) Brightness STD, (2) Saturation STD, (3) Salient LOC, (4) Color Difference, (5) Distance to the center. Standard deviations and differences of color give a reliable indication of how spread the colors in a photograph are. High quality photographs usually contain few colors that are highly saturated. On the other hand, non-professional photographs tend to contain high variety of desaturated colors. The location of the salient region is another important feature that is highly correlated with the rule of thirds.

3.2. Feature selection

Feature selection is used frequently in classification problems (1) to avoid high-dimensionality, (2) to reduce the feature measurement cost and (3) to reduce computational complexity. Several feature selection methods exist in the literature and they can be divided into two main categories: (1) exhaustive (optimal) methods and (2) greedy (sub-optimal) methods.

Even though the exhaustive search methods are optimal for feature selection, they are computationally intractable in most of the real applications. In this work, we use a greedy and efficient method that can scale to high-dimensional problems: the *Sequential Forward Floating Search (SFFS)* algorithm [15]. We use SFFS to select the most significant features from Table 1 for aesthetic quality prediction. Our objective that guides the selection procedure is the cross-validation classification accuracy of a Support Vector Machine (SVM) classifier with a Radial Basis Function (RBF) kernel [16]. SFFS begins with an empty set of features X and it stops when the specified number of features n is reached. The main steps of SFFS are the following:

1. **Step 1: Inclusion.** Select the most significant feature with respect to X, and include it in X. Stop if n features have been selected, otherwise, go to step 2.
2. **Step 2: Conditional exclusion.** Select the least significant feature k in X. If k was the last one added in step 1, keep it and go to step 1. Otherwise, exclude it and go to step 3.
3. **Step 3: Continuation of conditional exclusion.** Again find the least significant feature in X. If its removal leaves X with at least two features, and the accuracy without this feature is better, then remove it and repeat step 3. Otherwise, return to step 1.

4. EXPERIMENTAL EVALUATIONS

4.1. Dataset

We apply our algorithm to the large-scale AVA dataset [9] of 250K images. Visual aesthetic scores and two semantic labels are associated to each image. These semantic labels will enable us to create subsets based on image categories.

We select the 40K images with the lowest ratings and the 40K with the highest ratings to form our dataset (see Fig. 1). The reason for selecting this split is twofold: (1) we avoid classification problems that come from heavily unbalanced datasets and (2) the gap between the low and high quality ratings enables us to reliably evaluate the performance of our algorithm. The maximum rating of the low quality class is 4.72 and the minimum rating of the high quality class is 6.06. To evaluate the prediction accuracy, we use the LIBSVM package [21], where we fix the parameters to $C = 1$ and $\gamma = 1/num\_features$.

Based on this dataset we use the semantic labels to built dataset subsets for four categories: animal (7k images), city (18k images), portrait (8k images) and nature (7k images). Sample images with high and low ratings from two categories are shown in Figure 2.

4.2. Generic Feature Learning

We use SFFS on our 80k dataset to extract the best features from Table 1 for this set. On Figure 3 we show the evolution of the accuracy over the number of features. We only keep the seven best features as the accuracy becomes stable once this number of features is reached. The best features for this dataset are shown in the first row of Table 2. We refer to these seven features as the *best generic features*.

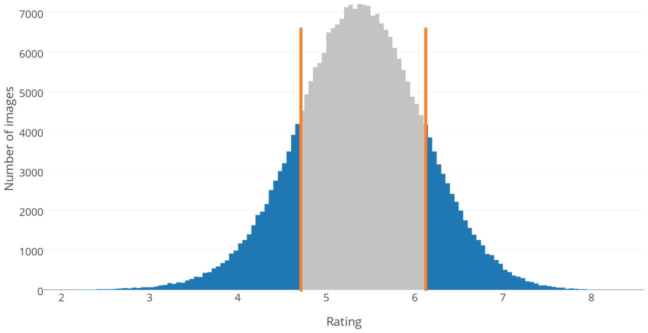
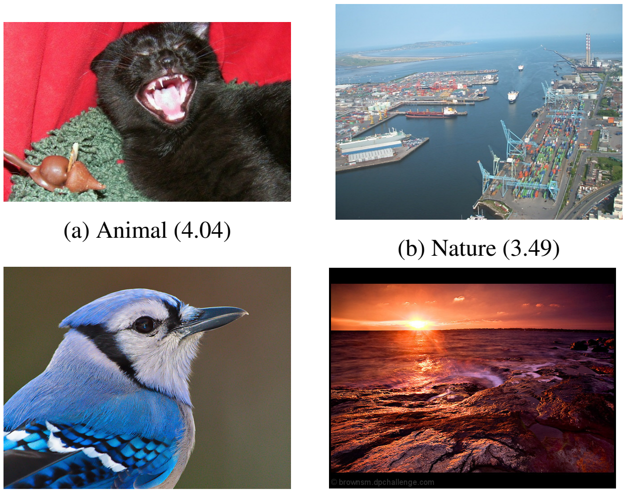


Fig. 1: Rating histogram of the AVA dataset. We use only the images that are in the tail-ends of the distribution (blue regions).



(a) Animal (4.04)

(b) Nature (3.49)

(c) Animal (7.01)

(d) Nature (7.02)

Fig. 2: Sample images for two categories with low and high ratings from the AVA dataset.

4.3. Specific Feature Learning

We use SFFS on the category subsets to extract the best features from Table 1 for each category in the same way as for the feature selection on the whole dataset. In Figure 3 we show the evolution of the accuracy over the number of features for each category. For the same reason as previously, and for consistency, we keep only the seven best features. The best features for each category are listed in Table 2. For each category, we refer to these seven features as the *best specific features*. From Figure 3 we can see that the accuracy between the different categories, except nature, is of the same magnitude. In most of the categories, the prediction accuracy saturates after seven features. The nature category behaves very similarly to the global dataset, where the saturation point happens after about 10-12 features. From the results of Table 2, we

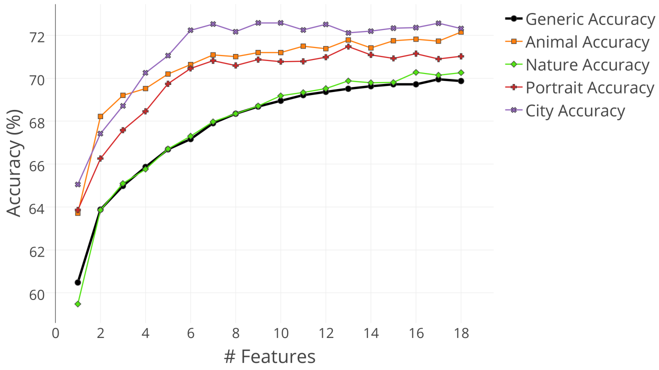


Fig. 3: Evolution of classification accuracy during the SFFS process.

Generic	Brightness AVG, Color Variance, #Edges, #Edges T, Hue Count, Saturation AVG, Sharpness
Animal	Contrast, #Edges R, #Edges T, Hue Count, Salient Saturation, Salient Sharpness, Sharpness
Nature	Brightness AVG, Color Variance, Contrast, #Edges, #Edges T, Saturation AVG, Sharpness
Portrait	Brightness AVG, #Edges, Hue Count, Salient Hue, Salient Sharpness, Saturation AVG, Sharpness
City	Brightness AVG, Color Variance, Contrast, #Edges T, Hue Count, Saturation AVG, Sharpness

Table 2: The seven features from Table 1 that were, for each category subset, selected by SFFS. The features are presented in alphabetical order.

observe that sharpness is a very important feature in a photograph, as well as the quantity of edges and the hue count. These last two features give an indication of the "simplicity" of a photo. From the results we observe that a "simple" photo, with few colors and few objects, is more appealing to the human eye. Features that describe the salient region, such as Salient Sharpness and Salient Saturation, are important for portrait and animal categories, where photos have a main subject. On the other hand, no salient features appear in the top seven for nature and city, which is an intuitive result, because there is usually no main subject in these categories. Compared to generic features, we observe a subtle difference in the extracted specific features: specific features are intuitively better tuned to a specific category, while generic features cover more general and holistic characteristics of the photo. It is important to note here that we cannot evaluate each one of the features of Table 2 independently from each other. The reason is that SFFS evaluates the performance of combination of features and not of individual features. It is not necessary that the best individual feature is part of the best combination of two features (see steps 2 and 3 of SFFS).

	Generic features	Specific features
Animals	$70.05\%(0.83)$	$\mathbf{70.35}\%(0.81)$
Nature	$66.65\%(0.61)$	$\mathbf{67.12}\%(0.61)$
Portrait	$68.34\%(0.92)$	$\mathbf{69.71}\%(0.84)$
City	$71.63\%(0.97)$	$\mathbf{71.81}\%(0.97)$

Table 3: Classification accuracy comparison on each category. In the first column we classify each category using the generic features extracted from the dataset (first row of Table 2). In the second column we classify each category using the learned features from the same category (remaining rows of Table 2).

4.4. Generic vs. Specific Features

To compare the efficiency of generic and specific features, we split the category subsets into a training and a test set. We train two models on the training set, one using the best generic features, and one using the best specific features. We then test those models on the test set and compare the results. We randomly split the subsets 250 times and repeat the same experiment. In Table 3 we show the average accuracies together with their standard deviations. Statistical significance tests were also performed in order to evaluate if the difference between the accuracies obtained is significant. The two-sided non-parametric Wilcoxon signed-rank test [22] gives a p-value equal to 0, therefore the accuracy differences between the two sets of features is statistically significant.

5. CONCLUSIONS

In this paper, we investigate the influence of image features for aesthetic quality prediction. We show that features learned from a specific image category are more accurate in predicting the aesthetic quality of images that belong to this category than generic features. Statistical tests show that this difference in accuracy is statistically significant. We show that predictive image features are dependent on the image's context. The code of our method can be found in our research page `http://ivrl.epfl.ch/research/aesthetics`.

Our method assumes that the semantic labels of a photograph are given, which is not always the case. However, in most of the cases, the semantic labels can be mined from a textual description of the photograph. Even if that is not possible, an intermediate step of image category classification can be implemented, so that semantic labels for new photographs can be added.

The way we divide the AVA dataset into categories is not the only one. There are several finer or coarser splits. In this work, we chose intuitive and highly general categories in order to investigate the potential differences in the most significant selected features. Further investigation and experiments are needed to generalize our findings to different categories and image collections.

6. REFERENCES

[1] Stefan Winkler, “Quality metric design: a closer look,” in *Proc. SPIE*, 2000, pp. 37–44.

[2] R. Datta, D. Joshi, J. Li, and J.-Z. Wang, “Studying Aesthetics in Photographic Images Using a Computational Approach,” in *Proc. European Conference on Computer Vision (ECCV)*, 2006, pp. 288–301.

[3] Y. Luo and X. Tang, “Photo and Video Quality Evaluation: Focusing on the Subject,” in *Proc. European Conference on Computer Vision (ECCV)*, 2008, pp. 386–399.

[4] L.-K. Wong and K.-L. Low, “Saliency-Enhanced Image Aesthetics Class Prediction,” in *Proc. IEEE International Conference on Image Processing (ICIP)*, 2009, pp. 997–1000.

[5] S. Bhattacharya, R. Sukthankar, and M. Shah, “A Framework for Photo-quality Assessment and Enhancement Based on Visual Aesthetics,” in *Proc. ACM International Conference on Multimedia*, 2010, pp. 271–280.

[6] W. Luo, X. Wang, and X. Tang, “Content-Based Photo Quality Assessment,” in *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2011, pp. 2206–2213.

[7] L. Marchesotti, F. Perronnin, D. Larlus, and G. Csurka, “Assessing the aesthetic quality of photographs using generic image descriptors,” in *Proc. International Conference on Computer Vision (ICCV)*, 2011, pp. 1784–1791.

[8] L. Marchesotti and F. Perronnin, “Learning beautiful (and ugly) attributes,” in *Proceedings of the British Machine Vision Conference*, 2013.

[9] N. Murray, L. Marchesotti, and F. Perronnin, “AVA: A large-scale database for aesthetic visual analysis,” in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012, pp. 2408–2415.

[10] L. Marchesotti, N. Murray, and F. Perronnin, “Discovering Beautiful Attributes for Aesthetic Image Analysis,” *International Journal of Computer Vision*, pp. 1–21, 2014.

[11] X. Lu, Z. Lin, H. Jin, J. Yang, and J.-Z. Wang, “RAPID: Rating Pictorial Aesthetics Using Deep Learning,” in *Proc. ACM International Conference on Multimedia*, 2014, pp. 457–466.

[12] T. Deselaers and V. Ferrari, “Visual and Semantic Similarity in ImageNet,” in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011, pp. 1777–1784.

[13] Jia Deng, Wei Dong, R. Socher, Li-Jia Li, Kai Li, and Li Fei-Fei, “ImageNet: A Large-Scale Hierarchical Image Database,” in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009, pp. 248–255.

[14] Albrecht Lindner, Appu Shaji, Nicolas Bonnier, and Sabine Süsstrunk, “Joint Statistical Analysis of Images and Keywords with Applications in Semantic Image Enhancement,” in *Proc. ACM International Conference on Multimedia*, 2012, pp. 489–498.

[15] P. Pudil, F.J. Ferri, J. Novovicova, and J. Kittler, “Floating search methods for feature selection with nonmonotonic criterion functions,” in *Proc. IEEE International Conferene on Pattern Recognition (ICPR)*, 1994, pp. 279–283.

[16] Bernhard Schölkopf and Alexander J. Smola, *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*, MIT Press, Cambridge, MA, USA, 2001.

[17] David G. Lowe, “Distinctive image features from scale-invariant keypoints,” *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, Nov. 2004.

[18] Y. Ke, X. Tang, and F. Jing, “The Design of High-Level Features for Photo Quality Assessment,” in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2006, pp. 419–426.

[19] Chuan Yang, Lihe Zhang, Huchuan Lu, Xiang Ruan, and Ming-Hsuan Yang, “Saliency Detection via Graph-Based Manifold Ranking,” in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013, pp. 3166–3173.

[20] J.L. Pech-Pacheco, G. Cristobal, J. Chamorro-Martinez, and J. Fernandez-Valdivia, “Diatom autofocusing in brightfield microscopy: a comparative study,” in *Proc. IEEE International Conference on Pattern Recognition (ICPR)*, 2000, pp. 314–317.

[21] Chih-Chung Chang and Chih-Jen Lin, “LIBSVM: A library for support vector machines,” *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 27:1–27:27, 2011, Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[22] Frank Wilcoxon, “Individual Comparisons by Ranking Methods,” *Biometrics Bulletin*, pp. 80–83, 1945.