The Good, the Bad, and the Ugly: Predicting Aesthetic Image Labels

Yaowen Wu[1,2], Christian Bauckhage[1,2]
[1]B-IT, University of Bonn
53113 Bonn, Germany
http://www.b-it-center.de

Christian Thurau[2]
[2]Fraunhofer IAIS
53754 St. Augustin, Germany
http://www.iais.fraunhofer.de

Abstract

Automatic classification of the aesthetic content of a picture is one of the challenges in the emerging discipline of computational aesthetics. Any suitable solution must cope with the facts that aesthetic experiences are highly subjective and that a commonly agreed upon theory of their psychological constituents is still missing. In this paper, we present results obtained from an empirical basis of several thousand images. We train SVM-based classifiers to predict aesthetic adjectives rather than aesthetic scores and we introduce a probabilistic postprocessing step that alleviates effects due to misleadingly labeled training data. Extensive experimentation indicates that aesthetics classification is possible to a large extent. In particular, we find that previously established low-level features are well suited to recognize beauty. Robust recognition of unseemliness, on the other hand, appears to require more high-level analysis.

1. Introduction

Automatic appraisal of the aesthetic or emotional content of images has been identified as a research topic already two decades ago [6]. However, research efforts have recently intensified, because in times of omnipresent digital photography and exploding amounts of image data computational aesthetics promises to improve quality of service and usability in many areas. Aesthetic content analysis may help to identify genres and epochs of paintings [11], it is used to match pictures and music [4], or may distinguish the work of professional photographers from that of amateurs [12, 8]. Most work on aesthetic image analysis, however, aims at improved performance in image retrieval [1, 2, 3, 13].

Scientific interest in the topic has been spawned by the new possibilities offered by platforms such as *flick*r, *picasa*, or *photo.net*. These web services provide access to billions of pictures that are labeled, rated, and commented on by dedicated communities of (semi)professional photographers. Given this wealth of data, a growing number of contributors proposes the use of statistical learning for aesthetic content classification. All of the works cited above consider low-level image features and train classifiers from labeled data.

Figure 1. Examples of pictures that users labeled 'beautiful' or 'awful'.

Figure 2. Are these 'beautiful' or 'awful'?

To the best of our knowledge, so far is has been largely ignored that people tend to assign inconsistent aesthetic labels [1]. In fact, aesthetics is known to depend on cultural context and to be a subjective experience and an intuitive concept that eludes quantification [7]. Figures 1 and 2 illustrate how this may hamper automatic classification. All six images were retrieved from the *most interesting* category on *flickr*. While the two examples on the left of Fig. 1 resulted from searches for 'beautiful' pictures, the other four pictures did in fact result from searching for 'awful' pictures.

In this paper, we examine ways of dealing with the issue of purposely mislabeled image data. We rely on established low-level features from the related literature [2, 4]. In contrast to earlier work, however, we consider a much larger database of labeled pictures. We train support vector machines and find them to perform com-

1051-4651/10 $26.00 © 2010 IEEE
DOI 10.1109/ICPR.2010.392

parable to the state of the art. In addition, we propose a simple yet effective postprocessing step that maps SVM responses into values that can be interpreted as probabilities. Our recognition rates after postprocessing consistently exceed previously reported results. Moreover, we empirically find that pictures that are commonly agreed to be pleasant and beautiful can easily be recognized as such just from considering low-level features. Pictures considered to be awful and disturbing on the other hand, often appear that way because of a degree of uncanniness that cannot be recognized without a semantic understanding of the scene depicted in an image.

2. Feature Extraction

The visual features which we consider are a subset of those introduced by Datta et al. [2] and Dunker et al. [4]. They were applied successfully in other work [3, 13]. and characterize local and global image properties on a low level of abstraction. Note that we do not compute any features to describe an image on the object level.

We follow the proposal in [2] and transform an RGB image of $M \times N$ pixels into the HSV color space. In the following, $\mathbf{H}$, $\mathbf{S}$, and $\mathbf{V}$ refer to $M \times N$ matrices containing the corresponding color information.

Global hue, saturation, and value are used to characterize chromatic purity, dominant color, and intensity of an image. For the saturation, we compute $\frac{1}{MN}\sum_{m=1}^{M}\sum_{n=1}^{N} S_{mn}$ and likewise for hue and value.

Central hue, saturation, and value are features also meant to account for the *rule of thirds* in photography which states that interesting image content is close to one of the intersections of four imaginary lines superimposed over the image (see Fig. 3). For the saturation, we compute $\frac{9}{MN}\sum_{m=M/3}^{2M/3}\sum_{n=N/3}^{2N/3} S_{mn}$ and likewise for hue and value.

The **position of the main object** is, again according to the rule of thirds, often close to the central rectangle of a picture (see Fig.3). Following [4], we compute Gabor filter responses for central rectangle to represent its content. This heuristic reliably captures the gist of a picture as it yields different characteristics for photos of natural scenes than for pictures of human-made artifacts [9]. Our filter bank contains 21 filters of different orientation and phase.

Colorfulness is an immediately apparent global property of a picture and colorful pictures are usually perceived to be pleasing [5, 10]. Following [2], we compute a color histogram of a picture and determine its deviation from a given set of histograms derived from prototypic colorful and less colorful images.

In total, we thus derive a 39 dimensional feature vector to represent an image and to train and test the classifiers described in the next section.

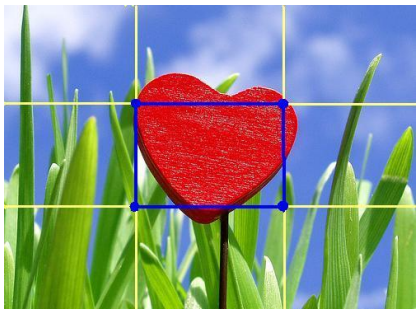

Figure 3. Rule of Thirds: Four imaginary lines are superimposed over the frame to produce nine rectangular sub-images.

3. Classification

The results reported in this paper were obtained from experimenting with large sets of labeled images. Given a set of opposing adjectives, we retrieved several thousand corresponding images from *flickr* and used them to train classifiers for two class classification. That is, in each experiment, the training data was split into two classes, where the images in the one class carried labels of *positive* connotation while the images in the other class were labeled with *negative* adjectives.

Experimenting with support vector machines for binary classification (class labels +1 or -1), we achieved satisfactory predictions of the aesthetic label of an image for the majority of our tests cases. However, for several subsets of our test images, the classification accuracy was rather disappointing. A closer inspection of these cases revealed that a substantial number of the images in our data appeared to be labeled misleadingly.

We frequently observed that for different images of visually similar content different people assigned rather incoherent labels. For example, we found many portraits of women labeled to be 'ugly' or 'awful' though objectively this assertion appears without merit (see the example in Fig. 2). Coquetting like this is but one example of the subjective nature of human behavior that has to be dealt with when analyzing data obtained from social web communities.

From the point of view of pattern recognition, these artifacts cause considerable class overlap in the feature space. Since a manual clean-up is infeasible given the extreme amounts of data in today's practical applications, we consider the use of informed postprocessing to remedy this situation.

The basic idea is to soften the binary decision function that is computed by an SVM classifier. If the SVM had been trained to regress examples of the *positive* class to +1 and examples of the *negative* class to -1,

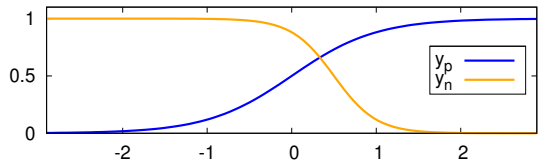

Figure 4. Examples of sigmoid functions used to soften the binary class label predictions produced by SVMs.

we thus compute sigmoid functions

$$y_p(x) = \frac{1}{1 + e^{-(x-\mu_p)\sigma_p}}$$
$$y_n(x) = \frac{1}{1 + e^{(x-\mu_n)\sigma_n}}$$

where x is the prediction value returned by the SVM. The parameters μ_p, μ_n, σ_p, and σ_n govern location and shape of the sigmoidal functions and are optimized using a verification set that is independent from the training and test data.

Using sigmoidal softening, the predictions of the SVM-based classifier are mapped to the interval $[0, 1]$. We can thus interpret the final results y_p and y_n as the probability of a pattern belonging to the *positive* or *negative* class, respectively. Note that this soft classification model does not assume the two cases to be mutually exclusive, i.e. in general $y_p + y_n \neq 1$ (see Fig. 4).

This approach allows us to determine classification accuracy with respect to classification confidence and thus alleviates the effect of overlapping feature space regions. While pictures in these regions will have low probabilities of belonging to either class, less ambiguous pictures will have higher probabilities of belonging to one class or the other. Figures 5 and 6 illustrate how the parameters σ and μ may steer the final decision. In these examples, our approach achieves a close to perfect recognition accuracy for pictures that are about 60% likely to be 'beautiful' or 'awful', respectively.

4. Experimental Result

The results presented here were obtained from experimenting with 10,800 pictures from the *most interesting* category on *flickr*. The set was gathered using English language queries for 3 adjectives of positive connotation ('beautiful', 'wonderful', and 'divine') as well as for 3 adjectives of negative connotation ('ugly', 'awful', and 'terrible'). After removing duplicates, there are 1,800 examples for each of these 6 classes.

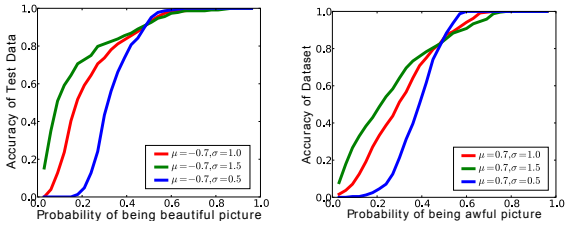

Figure 5. Varying σ for a fixed μ.

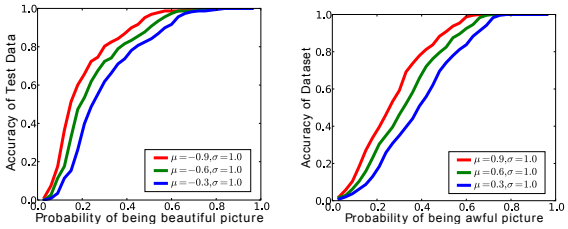

Figure 6. Varying μ for a fixed σ

In a first series of experiments, we explored if it is possible to distinguish pictures that evoke positive emotions from pictures evoking negative feelings. The three sets of positive images and the three sets of negative images were unified into two corresponding classes and we randomly subdivide the data into three independent subsets for training, verification and testing. After training, the verification phase was used to determine the parameters μ and σ of both classes such that accuracy exceeds 99% for a classification confidence of 60%.

In a second series of experiments, we evaluated if our approach distinguishes pictures from two classes of opposing adjectives ('beautiful' vs. 'ugly', 'wonderful' vs. 'awful', 'divine' vs. 'terrible'). Training and verification were done as in the first series of experiments.

Table 1 compares the classification accuracies of SVM-based classification only to those obtained from postprocessing the SVM results using sigmoidal smoothing. For the latter, the table lists the accuracy corresponding to a classification confidence of 55%. From the table, we can summarize our results as follows: i) predicting aesthetic labels using statistical classifiers works well to a large extent; ii) the proposed postprocessing step consistently improves recognition accuracy and copes well with the problem of incoherent labels; iii) for most of the 6 classes in our test, it appears beneficial, to train with larger superclasses of positive and negative images rather than with smaller sets of opposingly labeled images only; iv) pictures that evoke negative emotions are classified less reliably than pictures of positive content.

Table 1. Class specific accuracy. 1: training with two superclasses of positive and negative pictures; 2: training with three pairs of classes of opposing adjectives.

	experiment 1		experiment 2	
class	SVM	SVM+P	SVM	SVM+P
beautiful	77%	88%	73%	86%
ugly	56%	80%	52%	70%
wonderful	91%	97%	89%	96%
awful	40%	70%	59%	83%
divine	60%	80%	84%	93%
terrible	63%	84%	44%	77%

Figure 7. Pictures that were classified as 'beautiful', 'divine', and 'wonderful'.

Figure 8. Pictures that were classified as 'awful', 'terrible', and 'ugly'.

Figures 7 and 8 show examples of pictures that were classified correctly. Figure 9 shows examples of pictures our system deemed 'beautiful' but which actually labeled 'awful', 'terrible', and 'ugly'. From these prototypic examples, it appears that recognizing unseemliness requires semantic image understanding (e.g. in the case of a poisonous flower). Beauty, on the other hand, appears to be highly correlated with colorfulness, low frequencies, and local symmetries so that its recognition is possible from low-level image features only.

5. Conclusion and Future Work

We presented an improved approach to predicting the aesthetic quality of a picture. We extended SVM-based classifiers with an efficient postprocessing step that effectively copes with the problem of incoherent aesthetic labels due to subjective or cultural biases. Our results indicate that computational approaches to photographic aesthetics are indeed auspicious. In future work, we will extend the set of features towards descriptors of facial expression and lighting direction.

Figure 9. Examples of pleasant pictures that were labeled 'awful', 'terrible', and 'ugly' by the people who uploaded them.

References

[1] C. Bauckhage, T. Alpcan, R. Wetzker, and W. Umbrath. Image retrieval and web 2.0 – where can we go from here? In *Proc. ICIP*, 2008.

[2] R. Datta, D. Joshi, J. Li, and J. Wang. Studying aesthetics in photographic images using a computational approach. In *Proc. ECCV*, 2006.

[3] R. Datta, J. Li, and J. Wang. Algorithmic inferencing of aesthetics and emotion in natural images: an exposition. In *Proc. ICIP*, 2008.

[4] P. Dunker, S. Nowak, A. Begau, and C. Lanz. Content-based mood classification for photos and music: a generic multi-modal classification framework and evaluation approach. In *Proc. ACM MIR*, 2008.

[5] J. Hogg, editor. *Psychology and the Visual Arts*. Penguin Books, 1969.

[6] T. Kato. Database Architecture for Content-based Image Retrieval. In A. Jamberdino and C. Niblack, editors, *Image Storage and Retrieval Systems*, volume 1662 of *Proc. SPIE*, 1992.

[7] H. Leder, B. Belke, A. Oberst, and D. Augustin. A model of aesthetic appreciation and aesthetic judgements. *British J. of Psychology*, 95(4), 2004.

[8] P. Obrador and N. Moroney. Low Level Features for Image Appeal Measurement. In S. Farnand and F. Gaykema, editors, *Image Quality and System Performance*, volume 7242 of *Proc. SPIE*, 2009.

[9] A. Oliva and A. Torralba. Building the gist of a scene: the role of global image features in recognition. In *Progress in Brain Research*. Elsevier, 2006.

[10] R. Solso. *Cognition and the Visual Arts*. MIT Press, 1996.

[11] M. Spehr, C. Wallraven, and R. Fleming. Image statistics for clustering paintings according to their visual appearance. In *Int. Symp. Computational Aesthetics in Graphics, Visualization, and Imaging*, 2009.

[12] H. Tong, M. KLi, H.-J. Zhang, J. He, and C. Zhang. Classification of digital photos taken by photographers or home users. In *Pacific Rim Conf. Multimedia*, 2004.

[13] L.-K. Wong and K.-L. Low. Saliency-enhanced image aesthetic classification. In *Proc. ICIP*, 2009.