CHALLENGES OF FINDING AESTHETICALLY PLEASING IMAGES

João Faria[*‡], *Stanislav Bagley*[†], *Stefan Rüger*[†], *Toby Breckon*[*]

[*]Cranfield University, [†]The Open University, [‡]University of Porto
joao.faria@fe.up.pt, {stanislav.bagley, stefan.rueger}@open.ac.uk, toby.breckon@cranfield.ac.uk

ABSTRACT

We present an analysis of existing methods to automatic classification of photos according to aesthetics. We review different components of the classification process: existing evaluation datasets, their properties, most commonly-used image features, qualitative and quantitative, and classification results where comparable. We argue there are methodology gaps in the existing approaches to evaluating the classification results. We introduce the results of our experiments with Random Forest classification applied to image aesthetics classification and compare them to AdaBoost and SVM approaches.

1. INTRODUCTION

The problem of evaluating the aesthetics of photos is considered to be quite complex. Most photographers appear to be more or less skeptical on the ability of automatic aesthetics evaluation. The reasons they give usually refer to the complexity of the nature of photography. Despite this a number of automated methods of evaluating the quality of photos exist, e.g. [1-9].

To evaluate the quality of classification we need an evaluation dataset containing photos with associated scores or comments. Some of the datasets have already been created in previous studies [2, 6, 8]. Comparing the collections allows identifying the strong and weak points of the datasets and will be useful to inform future research.

The existing approaches to aesthetics classification [1-9] are based on their own sets of features that are believed to be the most effective. Some of the features refer to the global nature of an image while others utilize the information about the local areas within the image. These local features can include faces [9], humans [1, 9], image segments based on connected components [2] or other objects. There are also features describing the background of the objects and the relationship between the object and its background. Usually a certain set of features is selected for classification subjectively with scant justification [1, 2, 9].

In this study we explore the existing approaches and properties in the lifecycle of aesthetics classification process with the ultimate goal to find potential gaps in these areas and to define the prominent development directions that will make it possible to achieve a new level of classification quality. This includes all the aspects affecting classification, e.g., approaches to feature extraction, creation of evaluation datasets and selection of features.

To apply machine learning algorithms for the task of classification we need to interpret the notion of image aesthetics by means of computed image features. In the next section we give a list of those features. In Section 3 we discuss the advantages and drawbacks of the most well-known datasets for aesthetics evaluation and followed by a review and critique of previous works in Section 4. We also present the results of our own experiments in Section 5 and describe the potential future work on the task of aesthetics classification in Section 6.

2. FEATURES EXTRACTION

A process of classification requires the extraction of formal features within images. These features are usually selected according to the intuition of the authors [2, 7], the guidance of professional photographers [8, 9] or using some rules derived from the literature in the area of photography [2, 3, 6]. In the following we present features that showed processing results in competitive evaluation [2, 7, 8]. We classify these features denoted by f*i* according to their role into either global or local features. A global feature can be defined as a property characterizing the whole image while the values of local features are computed only using a certain region of the image.

The list of **global features** includes average pixel intensity and saturation (f1, f3 [2]); brightness contrast across the whole image (f2 [7]); blurring effect across the whole image (f4 [7]); Laplacian filter (f5 [4]); average saturation value in application to the rule of thirds (f6 [2]); level 1 wavelet transform on all three color bands for hue (f10 [2]); spatial distribution of edges (f12 [7]); composition geometry (f13 [8]); colour harmony (f14 [8]); level 3 wavelet transform on all three color bands for saturation (f15 [2]); level 1 and level 3 wavelet transform on all three color bands for intensity (f16, f70 [2]); hue count (f17 [4, 7]); histogram of oriented gradients 2x2 (f19 [1]); size (f22 [2]); dark channel (f25 [12]); hue contrast across the whole image (f46 [7, 8]); complimentary colours: rough position of the segment of the colour wheel (f48 [2]); low depth of field indicator for saturation and for intensity (f54, f55 [2]); average brightness (f65 [4]); luminance RMS (f66 [4]); sum of the average wavelet coefficients over all three frequency levels for saturation and for intensity (f71, f72 [2]); aspect ratio (f73 [2]); colour combination (f77 [8]).

Examples for **local features** are: familiarity based on the integrated region matching image distance (f9, f8 [2]); brightness contrast between segments (f20 [7]); horizontal coordinate for the mass centre of the whole image (f21 [7]); average saturation for the focus region and for the largest segment (f23, f24 [7]); region composition, an average hue value for the best five of the sets of pixels within the largest connected components (f26-f30 [2]); region composition, an average saturation value for the best five of the sets of pixels within the largest connected components (f31-f35 [2]); region composition, an average intensity value for the best five of the sets of pixels within the largest connected components (f36-f40 [2]); relative size of the set of pixels in the largest connected components with respect to the whole image (f43 [2]); the clarity of face regions (f44 [12]); average hue for the largest segment (f45 [7]); number of quantized hues presented in the image (f47 [7]); average saturation for the largest segment (f50-f52 [7]); average hue for the largest segment (f56-f58 [7]); vertical coordinate for the mass centre of the largest segment (f59-f61 [7]); average brightness for the largest segment (f62-f64 [7]); saturation squared difference (f67 [4]); Weber contrast (f68 [4]); Michelson contrast (f69 [4]); number of colour based clusters formed by K-Means in the LUV space (f74 [2]); simplicity by counts of quantized colours present in the background (f75 [8]); lighting, brightness difference between the subject and the background (f76 [8]).

Only features recognized as the most effective are selected (and detailed) above. The total number of features that were used in studies [2, 4, 7, 8] is much larger.

3. DATASETS: WHAT CAN BE IMPROVED?

The existing datasets are mainly based on data of popular photography websites that let users rate and and/or comment the photos of others. This approach allows collecting the data in a comparatively easy way.

The **DPChallenge** dataset is represented by a collection of images from DPChallenge.com [6]. The dataset contains user scores of 16,509 images in a grade of from 1 to 10. One advantage of the collection is that each photo has been evaluated by at least one hundred users. The collection contains for each image the number of aesthetics ratings received, the mean of ratings, and a distribution of quality ratings on a 1-10 scale.

The **Photo.net** collection [2] contains 3,581 images from a website, which is organized in a similar way to social networks. Website users assign independent aesthetics and originality scores in the range of 1 to 7 to each photo. The data contain their average scores, the number of times viewed by the members and the number of user ratings.

The **CUHKPQ** dataset [9] consists of 17,613 images selected from online resources for professional photography by amateur photographers. Images are divided into seven categories according to the contents of the photo: animal, plant, static, architecture, landscape, human and night. The scores were set by ten different peer experts.

MIRFlickr [11] is a dataset which is commonly used for the evaluation task in the community of multimedia information retrieval. It has two versions, one consisting of 25,000 images and the other of 1 million images from Flickr.com. The dataset contains user tags, EXIF and other metadata including an interestingness flag which may be interpreted as an aesthetics feature.

AVA [10] is a collection of images and metadata derived from DPChallenge.com. The dataset contains about 255,000 photos. Each image is associated with one of approximately 1,000 challenges selected from the contest site and a distribution of viewer scores. Also semantic annotations are provided for about 200,000 images, and 150,000 images contain no less than two tags.

Summarizing, we notice that most of the commonly used datasets contain photos which were provided by photographers who do not position themselves as professionals. Most of the popular collections are based on one of the three sources: dpchallenge.com, photo.net and flickr.com. DPChallenge.com claims that the goal of the project is creating a place where participants “could teach themselves to be better photographers”. The Photo.net website describes its audience as “consisting of photography enthusiasts ranging from newcomers to experienced”. Flickr.com focuses on simplifying the ways to collecting photos. The only dataset which has been claimed to contain professional photos is CUHKPQ.

The information of the aesthetics quality of the images above collections is based on the scores of the website users the only exception being CUHKPQ, which was created by amateur photographers who selected the works of professionals. All scores are considered to have equal weight. In reality this is not valid as the skillbase among assessors will vary substantially. Considering there is a small number of expert assessors, we assume that the existing approaches for creating the evaluation datasets contain an assessment gap. Potentially, this may lead to a lower quality of evaluation and ultimately to wrong conclusions for aesthetics ranking systems. How large can this problem be? One of the possible ways to discover this issue is to create a dataset that contains a certain part of photos taken by the professionals or the acknowledged photographers which will be selected by the experts in photography. This part of a collection can be established as a “gold standard” of high-quality photos. Having different types of collections, including and excluding the part containing “gold standard” we can evaluate the effect of human factor in the rating of photos. To date this has not been carried out.

The other property our dataset evaluation concerns the method of selecting photos for the collection. The DPChallenge dataset which was used in a variety of studies [3, 6, 8] contains just 20% of those photos that had a large number of user scores. One half of this 20% subset includes only the images with lowest scores and the other contains just the top-rated photos. As a result, the majority of the initial collection is excluded both from the training and test set. Of course, the absence of images with intermediate scores makes it much easier to predict the right class.

Furthermore, it is well known that one of the hardest classification tasks is to deal with objects that are close to the boundary between the classes. Considering this, the task clearly has been simplified to the extent that the evaluation results would be different when working with real-life data.

Summarizing, future evaluation datasets should be large-scale, contain rich annotations and semantic labels, have a uniform distribution of the data according to its quality, contain a broad range of scores for each image and have a high level of trust for the scores. One of the possible ways to utilize the expert opinions about the photos is to use online digital collections of photos of the high-ranked museums or art galleries where each photo goes through a selection of an expert community.

Most of the previous studies in the area use the existing datasets for evaluation tasks. This means the deficiencies of the datasets clearly affected the obtained results.

4. CHALLENGES OF AESTHETICS DISCOVERY

Using the Photo.net dataset in [2] the following feature set was identified as optimal for aesthetics classification: {f31, f1, f54, f28, f43, f74, f22, f70, f15, f71, f2, f9, f72, f73, f6}. Still the classification results remained relatively low with the accuracy achieved 0.701, with precision of detecting great photos being 0.681, and bad ones being 0.723 despite clearly separated positive and negative examples.

Feature f4 was discovered in [6] as the most discriminative between photos of high and low quality. The authors applied Naïve Bayes and AdaBoost algorithms for classification of the DPChallenge collection. The moderate level of precision 0.72-0.76 showed on a very small subset of data leaves space for making improvements.

In the approach [3] a boosted classifier was applied to the DPChallenge dataset and showed 0.52 precision rate at the level of recall 1 and 0.8 precision at the level of recall 0.81.

The Bayes classifiers were applied in [7] to estimate the importance of each individual feature using a unitary Gaussian model. A test dataset was represented by a small collection of manually selected paintings. The authors found the following features as those achieved the highest performance using the Naïve Bayes: {f20, f21, f2, f12, f24, f17, f45, f23, f4, f46}. Using of AdaBoost algorithm showed the most effective features are the following: {f20, f2, f17, f4, f21, f60, f63, f23, f47, f24}. To study the influence of separate features to the results of classification by AdaBoost the authors took error rates for each weak learner.

The study [5] showed the best results could be achieved by using both global features and object features in calculating the measure of memorability they introduced. The authors used large-scale scene recognition image database for experiments. The top ranked photos achieved about 0.85 of the proposed measure.

The CUHKPQ dataset was used in [9] for experiments which led the authors to evaluate the influence of particular local and global features to different predefined classes of photos organized according to genres. For example, they found a set of {f25, f46} shows the best performance for the static images, but for Architecture f2 is also an important feature and for Humans the most discriminative feature is f41. The approach showed the precision level of 0.86 when recall was 0.81, which is better than in previous approaches. This study shows that the nature of applicability of features is not universal; various features are better for work with different genres of photos.

The results presented in the previous works showed that aesthetics could be extracted automatically using machine learning algorithms [1-9]. Nevertheless, the results presented do not always allow making a clear comparison between the approaches. Firstly, some of the studies [5, 6] evaluate the performance of the complete set of features, which hampers discovering of the contribution of each feature to the quality of classification. Secondly, while selecting features for comparison authors do not necessarily replicate the features used in previous works. This impedes figuring out which features perform better under the same conditions [3, 7]. Thirdly, the newer results are not always being compared to the previous studies or verified on different datasets that reduces the clarity of comparison [8, 9]. Finally, different machine learning approaches were used in the studies [2, 3, 5, 6, 9]. Comparing the obtained results can lead to a wrong overall evaluation of the feature performance. We suggest including the features proposed in previous studies into the set for evaluation and to use the common dataset to evaluate different features or their combination.

Despite the deficiencies named above, we can learn from the classification quality results that there is considerable room for improvement. Considering the properties of the evaluation datasets used, the results seem fairly modest. We will show in the next experiment that gains can be relatively easily obtained by selecting a subset of features which show better performance and by using Random Forest classification for extracting aesthetics.

5. EXPERIMENTS

The idea of our experiment is to make a comparable evaluation with the results of the previous approaches [3, 6, 8]. Additionally, the goal was to find the features which perform better for our task. We based our experiments on the widely used DPChallenge image dataset, despite its shortcomings, so we can compare results with previous studies [3, 6, 8].

We used some part of the experience of previous studies implementing AdaBoost and SVM and accomplished the set of the classifiers by Random Forest classification [13], which has not been used before in this task. The reasons for implementing the Random Forest algorithm were the advantages of high accuracy, low sensitivity to outliers in training data and variable importance generated automatically. We estimate the average accuracy returned from the 10-fold cross validation.

	Accuracy	Precision	Recall	TNR
AdaBoost	0.776	0.797	0.741	0.811
SVM	0.78	0.841	0.689	0.87
R. Forest	0.863	0.896	0.822	0.905

Table 1. Classification evaluation on the whole feature set

For **AdaBoost** method was applied a weak decision tree classifier set of 110, weight trim rate 0.98 and maximum depth 3. We used **Support Vectors Machine** algorithm with sigmoid kernel trained by grid search over the parameter space, and for **Random Forest** maximum number of 100 trees in the forest and maximum depth value of 32 was taken.

The full set of features used for classification is presented in the study [4], which also describes the algorithm of optimal feature selection applied to the whole dataset to discover which features show the best performance. The final set was represented by {f5, f17, f65, f3, f14, f48, f66, f67, f68, f69, f77}, which includes both local and global features.

The best performance was achieved by Random Forest approach in the main classification performance measures: accuracy, precision, recall and specificity (TNR).

Comparing our results with those studies in which DPChallenge dataset was used for experiments we can conclude that our study obtained the best recall value of 0.81 at the rate of precision 0.99. The approaches [3] and [6] achieved less than 0.01 recall and a method described in [9] showed a value of 0.16. We also calculated an inverse projection of the selected point for the previous approaches and found the following precision values when recall is 0.81: [6] 0.65; [3] 0.81; [9] 0.86. Our result of 0.99 at the same recall level outperforms these methods and seems promising.

6. CONCLUSIONS AND FUTURE WORK

We reviewed the most commonly used datasets for image aesthetics evaluation. The existing datasets possess potential gaps in the methodology of their design that can lead to the losses in the quality of evaluation. One way to tackle this problem is including in the collection a new "gold standard" established by experts. Also, the approach organizing the datasets can be improved to correspond closer to the tasks from the real-life, where we deal not just with very good or bad photos. To do this, the data should include the photos with intermediate quality scores.

We listed those image features that showed good performance in our experiments and in the outcome of the previous studies. The variety of the features show the task of selecting the most useful ones is quite hard. The discrimination of selecting different features according to genres of photos is vital for improving the quality of classification.

In the experimental part we evaluated the quality of classification on the DPChallenge dataset using SVM, AdaBoost and Random Forest approaches. The latter we suspect showed particularly good results. However, the classification results can be improved better by more precise work with various genres of photos and using more complex types of high-level or low-level features.

7. REFERENCES

[1] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection", in *Proc. of CVPR, pp. 886-893*, 2005.

[2] R. Datta, D. Joshi, J. Li, and J. Wan, "Studying Aesthetics in Photographic Images Using a Computational Approach", in *Proc. of ECCV, pp. 288-301*, 2006.

[3] M. Desnoyer, and D. Wettergreen, "Aesthetic Image Classification for Autonomous Agents," in *Proc. of ICPR, pp. 3452 – 3455*, 2010.

[4] J. Faria, "What makes a good picture?", *Master of Sciences Thesis, Cranfield University*, 2012.

[5] P. Isola, D. Parikh, A. Torralba, and A. Oliva, "What makes an image memorable?", in *CVPR, pp. 145-152*, 2011.

[6] Y. Ke, X. Tang, and F. Jing, "The Design of High-Level Features for Photo Quality Assessment," in *Proc. of CVPR, pp. 419-426*, 2006.

[7] C. Li, T. Chen, "Aesthetic visual quality assessment of paintings". In *J-STSP, Vol. 3, No. 2, pp. 236-252*, 2009.

[8] Y. Luo, and X. Tang, "Photo and Video Quality Evaluation: Focusing on the Subject," in *Proc. of ECCV, pp. 386-399*, 2008.

[9] W. Luo, X. Wang, and X. Tang. "Content-based photo quality assessment", in *Proc. of ICCV, pp. 2206-2213*, 2011.

[10] N. Murray, L. Marchesotti, and F. Perronnin. "AVA: A Large-Scale Database for Aesthetic Visual Analysis", in *Proc. of CVPR, pp. 2408-2415*, 2012.

[11] M. J. Huiskes, B. Thomee, M. S. Lew. "New Trends and Ideas in Visual Concept Detection", in *Proc. of ACM MIR, pp. 527-536*, 2010.

[12] A. Oliva and A. Torralba. "Modeling the shape of the scene: a holistic representation of the spatial envelope", in *IJCV, pp. 145-1*, 2001.

[13] L. Breiman. "Random Forests", in *MLJ 45(1), pp. 5-32*, 2001.