AVA: A Large-Scale Database for Aesthetic Visual Analysis

Naila Murray
Computer Vision Center
Universitat Autònoma de Barcelona, Spain
nmurray@cvc.uab.es

Luca Marchesotti, Florent Perronnin
Xerox Research Centre Europe
Meylan, France
firstname.lastname@xrce.xerox.com

Abstract

With the ever-expanding volume of visual content available, the ability to organize and navigate such content by aesthetic preference is becoming increasingly important. While still in its nascent stage, research into computational models of aesthetic preference already shows great potential. However, to advance research, realistic, diverse and challenging databases are needed. To this end, we introduce a new large-scale database for conducting Aesthetic Visual Analysis: AVA. It contains over 250,000 images along with a rich variety of meta-data including a large number of aesthetic scores for each image, semantic labels for over 60 categories as well as labels related to photographic style. We show the advantages of AVA with respect to existing databases in terms of scale, diversity, and heterogeneity of annotations. We then describe several key insights into aesthetic preference afforded by AVA. Finally, we demonstrate, through three applications, how the large scale of AVA can be leveraged to improve performance on existing preference tasks.

1. Introduction

Judging the aesthetic value of an image is a challenging task. It is becoming increasingly important in web-scale image search for retrieving high quality images, but also for recommending images in photofinishing and for on-line photo suggestion. This is a problem that is currently receiving increasing attention in the computer vision and multimedia retrieval communities.

Most of the research on aesthetic analysis has focused on feature design. Typically, features capturing the "aesthetic" properties of an image are proposed with the aim of mimicking photographic rules and practices such as the golden ratio, the rule of thirds and color harmonies [3, 11, 16, 6, 15]. More recently, Marchesotti et al. [17] showed that generic image descriptors can outperform traditional aesthetic features.

Despite all the work on designing image descriptors for aesthetics, little attention so far has been dedicated to the collection, annotation and distribution of ground truth data.

Figure 1. A sample challenge entitled "Skyscape" from the social network www.dpchallenge.com. Images are ranked according to average score and the top three are awarded ribbons.

We believe that *novel datasets shared by the community will greatly advance the research around this problem*. This has been the case for semantic categorization, where successful datasets such as Caltech 101 [13] and 256 [8], PASCAL VOC [7] and Imagenet[5] have contributed significantly to the advancement of research. Such databases are typically composed of images obtained by web-crawling and annotated by crowd-sourcing. In the specific case of aesthetic analysis, having rich and large-scale annotations is a key factor.

However, a major complication of aesthetic analysis in comparison to semantic categorization is the highly subjective nature of aesthetics. To our knowledge, all the image datasets used for aesthetic analysis were obtained from on-line communities of photography amateurs such as www.dpchallenge.com or www.photo.net. These datasets contain images as well as aesthetic judgments they received from members of the community. Collecting ground truth data in this manner is advantageous primarily because it is an inexpensive and expedient way to obtain aesthetic judgments from multiple individuals who are generally "prosumers" of data: they produce images and they

also score them on dedicated social networks. The interpretation of these aesthetic judgments, expressed under the form of numeric scores, has always been taken for granted. Yet a deeper analysis of the context in which these judgments are given is essential. The result of this lack of context is that it is difficult to understand what the aesthetic classifiers *really* model when trained with such datasets.

Additional limitations and biases of current datasets may be mitigated by performing analysis on a much larger scale than is presently done. To date, at most 20,000 images have been used to train aesthetic models used for classification and regression. In this work, we present a new database called AVA (Aesthetic Visual Analysis), which contains more than 250,000 images, along with a rich variety of annotations. We then investigate how this wealth of data can be used to tackle the problem of understanding and assessing visual aesthetics. The database is publicly available at `www.lucamarchesotti.com/ava`.

Below are the principal contributions of our work:

- We introduce a novel large-scale database for image aesthetics and we show how it can be used to advance research in the field using three sample applications.
- Through AVA we explore the factors that make aesthetic analysis such a challenging and intriguing research problem.
- We show in our experiments that not only does the *scale* of training data matter for increasing performances, but also the *aesthetic quality* of the images used for training.

The rest of the paper is organized as follows. In section 2 we present AVA and its components. We compare the database to existing image aesthetics databases in section 2.1. In section 3 we describe several important factors which should be addressed when modeling aesthetic preference but are currently ignored in the literature. In section 4 we provide three concrete examples of applications that can benefit from AVA. In section 5 we discuss possible future avenues of research that could be opened with the database.

2. Creating AVA

AVA is a collection of images and meta-data derived from `www.dpchallenge.com`. To our knowledge, it represents the first attempt to create a large database containing a unique combination of heterogeneous annotations. The peculiarity of this database is that it is derived from a community where images are uploaded and scored in response to photographic challenges. Each challenge is defined by a title and a short description (see Figure 1 for a sample challenge). Using this interesting characteristic, we associated each image in AVA with the information of its corresponding challenge. This information can be exploited in combination with aesthetic scores or semantic tags to gain an understanding of the context in which such annotations were provided. We created AVA by collecting approximately 255,000 images covering a wide variety of subjects on 1,447 challenges. We combined the challenges with identical titles and descriptions and we reduced them to 963. Each image is associated with a single challenge.

In AVA we provide three types of annotations:

Aesthetic annotations: Each image is associated with a distribution of scores which correspond to individual votes. The number of votes per image ranges from 78 to 549, with an average of 210 votes. Such score distributions represent a gold mine of aesthetic judgments generated by hundreds of amateur and professional photographers with a practiced eye. We believe that such annotations have a high intrinsic value because they capture the way hobbyists and professionals understand visual aesthetics.

Semantic annotations: We provide 66 textual tags describing the semantics of the images. Approximately 200,000 images contain at least one tag, and 150,000 images contain 2 tags. The frequency of the most common tags in the database can be observed in Figure 2.

Photographic style annotations: Despite the lack of a formal definition, we understand photographic style as a consistent manner of shooting photographs achieved by manipulating camera configurations (such as shutter speed, exposure, or ISO level). We manually selected 72 Challenges corresponding to photographic styles and we identified three broad categories according to a popular photography manual [12]: *Light*, *Colour*, *Composition*. We then merged similar challenges (e.g. "Duotones" and "Black & White") and we associated each style with one category. The 14 resulting photographic styles along with the number of associated images are: Complementary Colors (949), Duotones (1,301), High Dynamic Range (396), Image Grain (840), Light on White (1,199), Long Exposure (845), Macro (1,698), Motion Blur (609), Negative Image (959), Rule of Thirds (1,031), Shallow DOF (710), Silhouettes (1,389), Soft Focus (1,479), Vanishing Point (674).

2.1. AVA and Related Databases

In Table 1 we compare AVA to currently-used public databases containing aesthetic annotations. Below we also discuss the features that differentiate AVA from such datasets.

Photo.net (PN) [3]: PN contains 3,581 images gathered from the social network Photo.net. In this online community, members are instructed to give two scores from 1 to 7 for an image. One score corresponds to the image's aesthetics and the other to the image's originality. The dataset

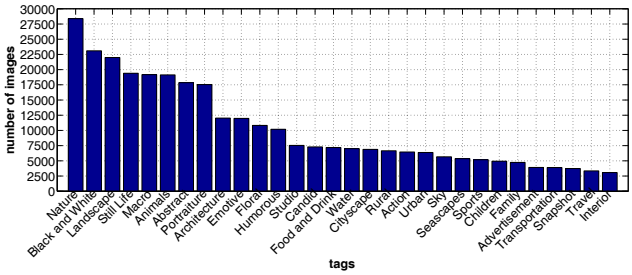


Figure 2. Frequency of the 30 most common semantic tags in AVA.

includes the mean aesthetic score and the mean originality score for each image. Each image in PN received two or more scores. AVA offers scores evaluated with much richer distributions and approximately 70× the number of images. PN is also affected by an important bias discovered in [17]. Images receiving high scores have frames manually created by the owners to enhance the visual appearance.

	AVA	Photo.net	CUHK	CUHKPQ	ImageCLEF
Large Scale	Y	N	N	N	N
Scores distr.	Y	Y	N	N	N
Rich annotations	Y	N	Y	Y	Y
Semantic Labels	Y	N	N	Y	Y
Style Labels	Y	N	N	N	Y

Table 1. Comparison of the properties of current databases containing aesthetic annotations. AVA is large-scale and contains score distributions, rich annotations, and semantic and style labels.

CUHK [11]: CUHK contains 12,000 images, half of which are considered high quality and the rest labeled as low quality. The images were derived from the same social network from which we derived AVA. Unfortunately, the images were obtained by retaining the top and bottom 10% (in terms of mean scores) of 60,000 images randomly crawled from `www.dpchallenge.com`. Our dataset differs from CUHK in several ways. CUHK only contains images with a very clear consensus on their score, while AVA also considers more ambiguous images. A consequence is that CUHK does not offer such a difficult challenge any more: recent methods achieved classification accuracies superior to 90% on this dataset [17]. Finally, CUHK provides only binary labels (1=high quality images, 0=low quality images) whereas AVA provides an entire distribution of scores for each image.

CUHKPQ [15]: CUHKPQ consists of 17,613 images obtained from a variety of on-line communities and divided into 7 semantic categories. Each image was labeled as either high or low quality by at least 8 out of 10 independent viewers. Therefore this dataset consists of binary labels of very high consensus images. Like CUHK, it is does not offer a very difficult challenge: the classification method of [15] obtained AROC values between 0.89 and 0.95 for all semantic categories. In addition, despite the fact that AVA shares similar semantic annotations, it differs in terms of scale and also in terms of consistency. In fact, CUHKPQ was created by mixing high quality images derived from photographic communities and low quality images provided by university students. For this reason, the dataset does not correspond to a real case scenario.

MIRFLICKR/Image CLEF: Visual Concept Detection and Annotation Task 2011 [9]: MIRFLICKR is a large dataset introduced in the community of multimedia retrieval. It contains 1 million images crawled by Flickr, along with textual tags, aesthetic annotations (Flickr's interestingness flag) and EXIF meta-data. A sub-part of MIRFLICKR was used by CLEF (the Cross-Language Evaluation Forum) to organize two challenges on "Visual Concept Detection". For these challenges, the basic annotations were enriched with emotional annotations and with some tags related to photographic style. It is probably the dataset closest to AVA but it lacks rich aesthetic preference annotations. In fact, only the "interestingness" flag is available to describe aesthetic preference. Some of the 44 visual concepts available might be related to AVA photographic styles but they focus on two very specific aspects: exposure and blur. Only the following categories are available: neutral illumination, over-exposed, under-exposed, motion blur, no blur, out of focus, partially blurred. In addition, the number of images with such style annotations is limited.

3. Analysis of AVA

We describe the main features of AVA by focusing on two particular aspects that we believe are very important for this kind of database: the aesthetic annotations and their relation to semantic annotations.

3.1. Aesthetic preference in AVA

Visual aesthetic preference can be described either as a single (real or binary) score or as a distribution of scores. In the first case, the single value is obtained by averaging all the available scores and by eventually binarizing the average with an appropriate threshold value. The main limitation of this representation is that it does not provide an indication of the degree of consensus or diversity of opinion among annotators. The recent work of Wu *et al.* [21] proposed a solution to this drawback by learning a model capable of predicting score distributions through structured-SVMs. However, they use a dataset composed of 1,224 images annotated with a limited amount of votes (on average 28 votes per image). We believe that such methods can greatly benefit from AVA where much richer scores distributions (consisting on average of approximately 200 votes)

are available. AVA also enables us to have a deeper understanding of such distributions and of what kind of information can be deduced from them.

Score distributions are largely Gaussian. Table 2 shows a comparison of Goodness-of-Fit (GoF), as measured by RMSE, between top performing distributions we used to model the score distributions of AVA. One sees that Gaussian functions perform adequately for images with mean scores between 2 and 8, which constitute 99.77% of all the images in the dataset. In fact, the RMSEs for Gaussian models are rarely higher than 0.06. This is illustrated in Figure 3. Each plot shows 8 density functions obtained by clustering the score distributions of images whose mean score lies within a specified range. Clustering was performed using k-means. The clusters of score distributions are usually well approximated by Gaussian functions (see Figures 3(b) and 3(c)). We also fitted Gaussian Mixture Models with three Gaussians to the distributions but we only found minor improvement with respect to one Gaussian. Beta, Weibull and Generalized Extreme Value distributions were also fitted to the score distributions, but gave poor RMSE results.

Non-Gaussian distributions tend to be highly-skewed. This skew can be attributed to a floor and ceiling effect [2], occurring at the low and high extremes of the rating scale. This can be observed in Figures 3(a) and 3(d). Images with positively-skewed distributions are better modeled by a Gamma distribution $\Gamma(s)$, which may also model negatively-skewed distributions using the transformation $\Gamma'(s) = \Gamma((s_{min} + s_{max}) - s)$, where s_{min} and s_{max} are the minimum and maximum scores of the rating scale.

Mean score	**Average RMSE**		
	Gaussian	Γ	Γ'
1-2	0.1138	**0.0717**	0.1249
2-3	0.0579	**0.0460**	0.0633
3-4	**0.0279**	0.0444	0.0325
4-5	**0.0291**	0.0412	0.0389
5-6	**0.0288**	0.0321	0.0445
6-7	0.0260	**0.0250**	0.0455
7-8	**0.0268**	0.0273	0.0424
8-9	0.0532	0.0591	**0.0403**
Average RMSE	**0.0284**	0.0335	0.0429

Table 2. Goodness-of-Fit per distribution with respect to mean score: The last row shows the average RMSE for all images in the dataset. The Gaussian distribution was the best-performing model for 62% of images in AVA.

Standard Deviation is a function of mean score. Boxplots of the variance of scores for images with mean scores within a specified range are shown in Figure 4. It can be seen that images with "average" scores (scores around 4, 5 and 6) tend to have a lower variance than images with scores greater than 6.6 or less than 4.5. Indeed, the closer the mean score gets to the extreme scores of 1 or 10, the higher the probability of a greater variance in the scores. This is likely due to the non-Gaussian nature of score distributions at the extremes of the rating scale.

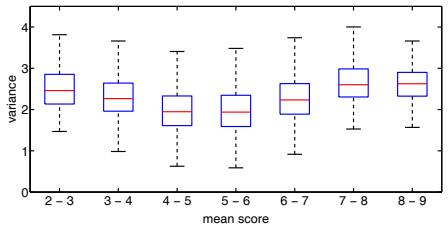


Figure 4. Distributions of variances of score distributions, for images with different mean scores. The variance tends to increase with the distance between the mean score and the mid-point of the rating scale.

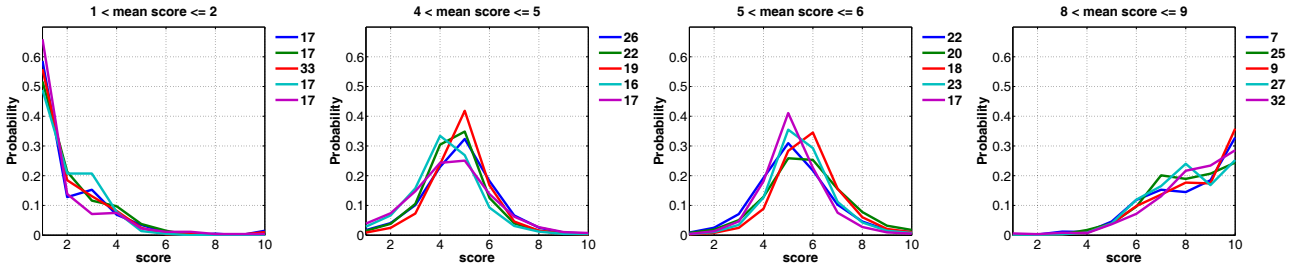
Images with high variance are often non-conventional. To gain an understanding of the additional information a distribution of scores may provide, we performed a qualitative evaluation of images with low and high variance. Table 3 displays our findings. The quality of execution of the styles and techniques used for an image seem to correlate with the mean score it receives. For a given mean value however, images with a high variance seem more likely to be edgy or subject to interpretation, while images with a low variance tend to use conventional styles or depict conventional subject matter. This is consistent with our intuition that an innovative application of photographic techniques and/or a creative interpretation of a challenge description is more likely to result in a divergence of opinion among voters. Examples of images with low and high score variances are shown in Figure 5. The bottom-left photo in particular, submitted to the challenge "Faceless", had an average score of 5.46 but a very high variance of 5.27. The comments it received indicate that while many voters found the photo humorous, others may have found it rude.

		variance	
		low	high
mean	low	poor, conventional technique and/or subject matter	poor, non-conventional technique and/or subject matter
	high	good, conventional technique and/or subject matter	good, non-conventional technique and/or subject matter

Table 3. Mean-variance matrix. Images can be roughly divided into 4 quadrants according to conventionality and quality.

3.2. Semantic content and aesthetic preference

AVA contains annotations for more than 900 photographic challenges, covering a vast range of different sub-



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

Figure 3. Clusters of distributions for images with different mean scores. The legend of each plot shows the percentage of these images associated with each cluster. Distributions with mean scores close to the mid-point of the rating scale tend to be Gaussian, with highly-skewed distributions appearing at the end-points of the scale.

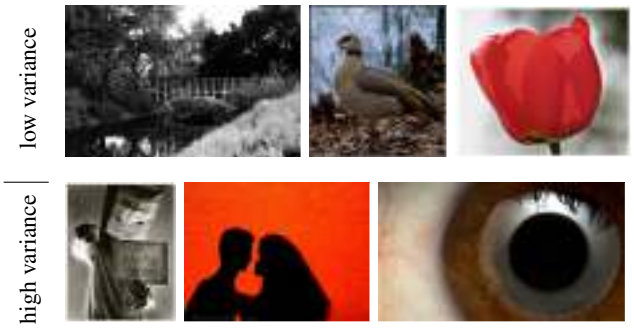


Figure 5. Examples of images with mean scores around 5 but with different score variances. High-variance images have non-conventional styles or subjects.

jects. We evaluated aggregated statistics for each challenge using the score distributions of the images that were submitted. Figure 6 shows a histogram of the mean score of all challenges. As expected, the mean scores are approximately normally distributed around the mid-point of the rating scale. We inspected the titles and associated descriptions of the challenges at the two extremes of this distribution. We did not observe any semantic coherence between the challenges in the right-most part of the distribution. However, it is worth noticing that two "masters' studies" (where only members who have won awards in previous challenges are allowed to participate) were among the top 5 scoring challenges. We use the arousal-valence plane [20] to plot the challenges on the left of the distribution (the low-scoring tail). The dimension of valence ranges from highly positive to highly negative, whereas the dimension of arousal ranges from calming or soothing to exciting or agitating. In particular, among the lowest-scoring challenges we identified: #1 "At Rest" (av. vote= 4.747), #2 "Despair" (av. vote=4.786), #3 "Fear" (av.vote=4.801), #4 "Bored" (av. vote=4.8060), # 6 "Pain" (av. vote=4.818), #23 "Conflict" (av. vote= 4.934), #25 "Silence" (av. vote= 4.948), #30 "Shadows" (av. vote= 4.953), #32 "Waiting" (av. vote.=4.953), #39 "Obsolete" (av.vote= 4.9740). In each case, the photographers were instructed to depict or interpret the emotion or concept of the challenge's title. This suggests that themes in the left quadrants of the arousal-valence plane (see Figure 6) bias the aesthetic judgments towards smaller scores.

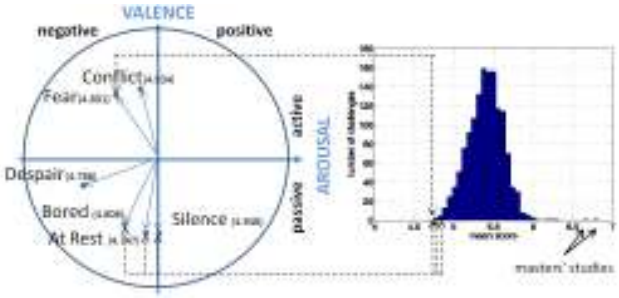


Figure 6. Challenges with a lower-than-normal average vote are often in the left quadrants of the arousal-valence plane. The two outliers on the right are masters' studies challenges.

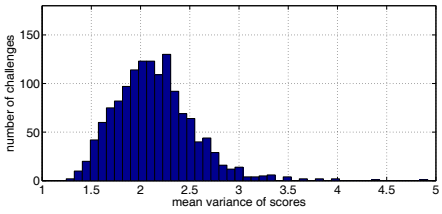


Figure 7. Histogram of the mean variance of score distributions over all challenges. Free studies tend to have low-variance score distributions.

Figure 7 shows a histogram of the mean variance of the score distributions for images in a given challenge, for all challenges. While we observed no trend among challenges with high-variance score distributions, we found that the majority of free study challenges were among the bottom 100 challenges by variance, with 11 free studies among the bottom 20 challenges. Free study challenges have no restrictions or requirements as to the subject matter of the submitted photographs. The low variance of these types of

challenges indicates that challenges with specific requirements tend to lead to a greater variance of opinion, probably with respect to how well entries adhere to these requirements. The number of votes per image is related to the challenge's theme. Of the top 5 challenges by number of votes per image, 4 involved capturing nude subjects or lingerie.

4. Applications of AVA

In this section, we present three applications, each related to aesthetic visual analysis. These applications illustrate the advantages of the AVA dataset not only for classic aesthetic categorization, but also to gain a deeper understanding of what makes an image appealing, *e.g.* what are the respective roles of the semantic content and the photographic technique. The first experiment shows the classification performance gains we achieve using a large amount of training data. The second application presents the results of content-dependent models trained by exploiting semantic annotations. Finally, we present a third scenario where AVA can be used to classify the photographic style of an image.

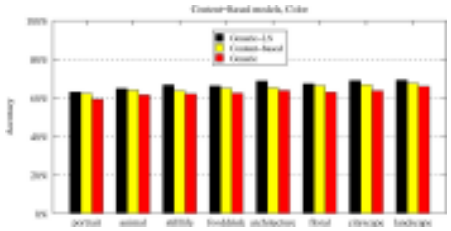
4.1. Large-Scale aesthetic quality categorization

Most approaches to the problem of aesthetic categorization involve fully-supervised learning. Typically, a classification model is trained to assign "high quality" or "low quality" labels to images [11, 16, 17, 15, 6, 10]. This framework is particularly interesting because preference information is currently collected at a web-scale through binary ratings (such as Facebook's "Like" button or Google's "+1" button). However, recent works [21] have interpreted this problem as a regression problem, which is possible only if appropriate annotations are available.

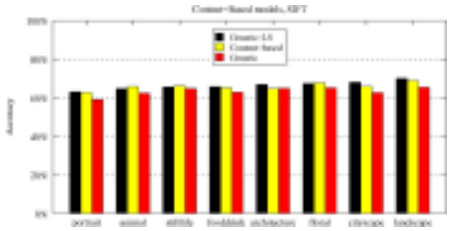
In our experiments, we followed [17]: we trained linear SVMs with Stochastic Gradient Descent (SGD) [1] on Fisher Vector (FV) signatures [18, 19] computed from color [19] and SIFT [14] descriptors . Three observations can be made with respect to scale, training data, and testing data.

The scale matters. Figures 8(a) and 8(b) show the learning curves with color and SIFT features respectively for a variable number of training samples and for more or less complex models. The model complexity is set by the number of Gaussians, $ngauss$, used to compute the FV as the FV dimensionality is directly proportional to $ngauss$. As expected, for both types of features, we consistently increase the performance with more training images but with diminishing returns. Also, more Gaussians lead to better results although the difference between $ngauss = 64$ and 512 remains limited (on the order of 1%).

The type of training images matters. We introduce a parameter δ to discard ambiguous images from the training set. More precisely, we discard from the training set all those images with an average score between $5-\delta$ and $5+\delta$.



(a)



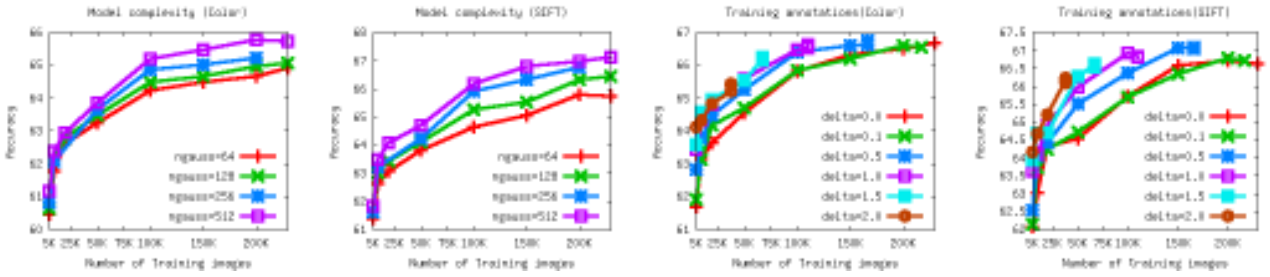
(b)

Figure 9. Results of content-based aesthetic quality categorization. Generic models trained on large-scale data out-perform small-scale content-based models.

As δ increases, we are left with increasingly unambiguous images. On the other hand, when $\delta = 0$, we use the full training set. This is somewhat similar to the protocol of [4, 17]. However, there is a major difference: in those works, δ was used to discard ambiguous images from the training *and the test set*, thus making the problem easier with larger values of δ. In our case, the test set is left unchanged, *i.e.* it includes both ambiguous and unambiguous images. Figures 8(c) and 8(d) show the classification results for color and SIFT descriptors respectively, as δ increases. There are two points to note. First, for the same number of training images, the accuracy increases with δ. Second, the same level of accuracy that is achieved by increasing the number of training samples can also be achieved by increasing δ. In this way, accuracy is preserved and computational cost is reduced by selecting the "right" training images.

4.2. Content-based aesthetic categorization

We experimented with the semantic tags of AVA. We selected 8 semantic categories equivalent to the ones picked by [15]. These categories are also the 8 most popular semantic tags in AVA, and they contain on average 14,368 images. With this dataset, we first trained 8 independent SVMs, one for each semantic category. Then we trained a single, generic classifier with an equivalent number of images randomly sampled among all categories. Finally, we trained a generic classifier using a large-scale training set composed of 150,000 images randomly sampled from AVA. The results in Figures 9(a) and 9(b) show that content-based models perform better than the generic model for the same



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

Figure 8. Results for large-scale aesthetic quality categorization for increasing model complexity ((a) and (b)) and increasing values of δ ((c) and (d)).

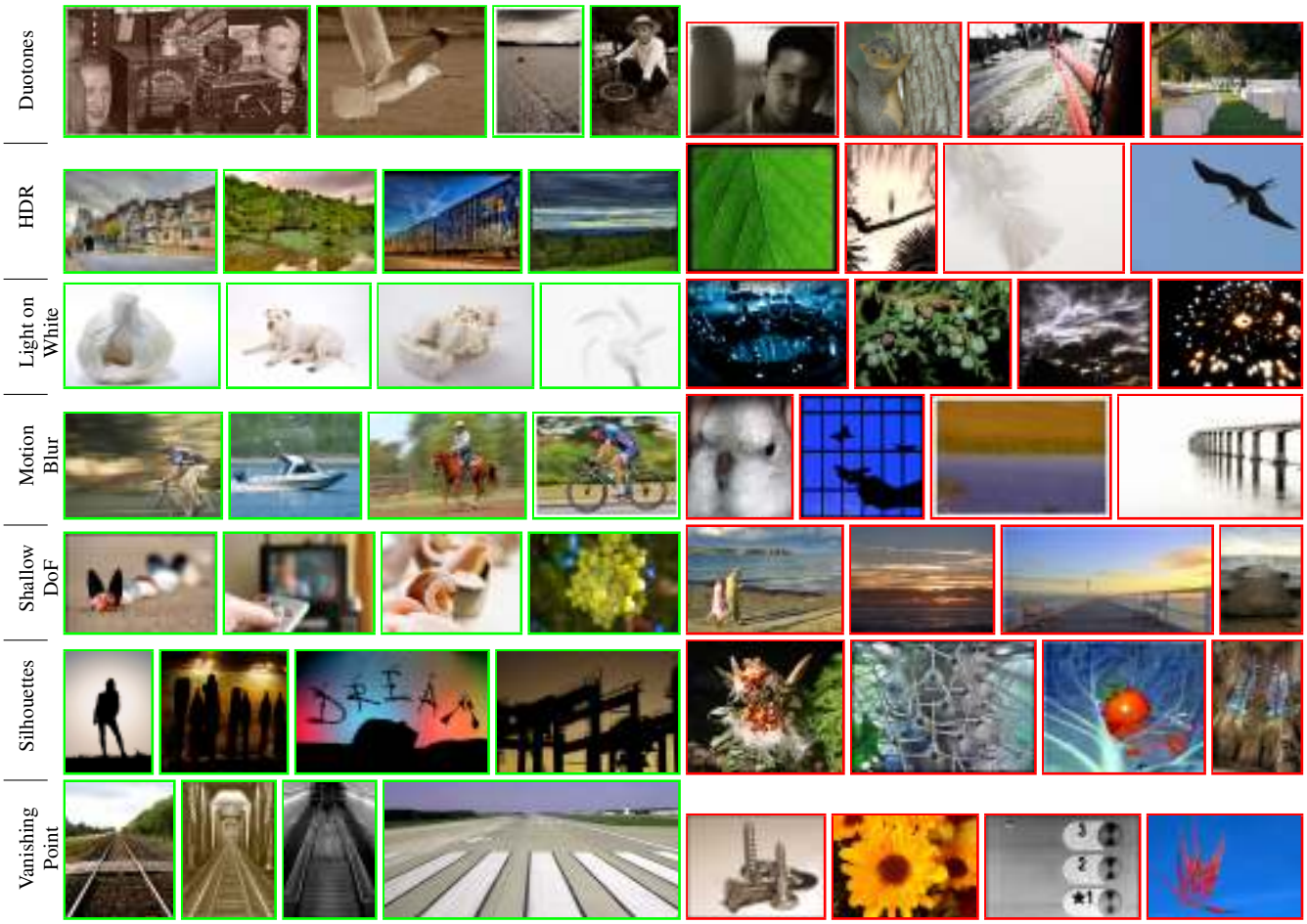


Figure 10. Qualitative results for style categorization. Each row shows the top 4 (green) and bottom 4 (red) ranked images for a category. Images with very different semantic content are correctly labeled.

number of training images. Similar trends were also noticed by Luo [15] and Dhar [6]. However, the generic large-scale model out-performs the content-based models for all categories using color features, and for 5 out of 8 categories using SIFT features.

4.3. Style Categorization

When asked for a critique, experienced photographers not only say *how* much they like an image. In general, they also explain *why* they like or dislike it. This is the behavior that we observed in social networks such as `www.dpchallenge.com`. Ideally, we would like to replicate this qualitative assessment of the aesthetic proper-

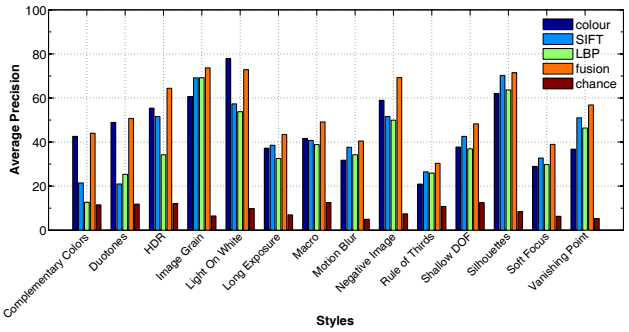


Figure 11. Mean average precision (mAP) for challenges. Late fusion results in a mAP of 53.85%.

ties of the image. This represents a novel goal that can be tackled using the style annotations of AVA.

To verify this possibility, we trained 14 classification models using the 14 photographic style annotations of AVA and their associated images (totaling 14,079). We trained 14 one-versus-all linear SVMs. Again, we learned all classifiers using Stochastic Gradient Descent (SGD). We computed separate FV signatures using SIFT, color histogram and LBP (Local Binary Patterns) features and combined them by late fusion.

Results are summarized in Figure 11. Not surprisingly, the color histogram feature is the best performer for the "duotones", "complementary colors", "light on white" and "negative image" challenges. SIFT and LBP perform better for the "shallow depth of field" and "vanishing point" challenges. Late fusion significantly increases the mean average precision (mAP) of the classification model, leading to a mAP of 53.85%. The qualitative results shown in Figure 10 illustrate that top-scored images are quite consistent with their respective styles, even while their semantic content differed.

5. Discussion and Future Work

In designing AVA, we aimed at three main objectives. The first objective was to provide a large-scale benchmark and training resource which would overcome the limitations of existing databases (*c.f.* section 2). The second one was to gain a deeper insight into aesthetic preference (*c.f.* section 3). The third one was to show how richer – and especially larger – datasets could help to improve existing applications and enable new ones (*c.f.* section 4).

In future work, we intend to use the annotations contained in AVA to explore the interplay between semantic content, photographic style and aesthetic quality. We would also like to further explore and leverage the relationship between an image's score distribution and its semantic and stylistic attributes.

References

[1] L. Bottou. Sgd. http://leon.bottou.org/ projects/sgd. 6

[2] D. Cramer and D. Howitt. *The SAGE dictionary of statistics*. SAGE, 1st edition, 2004. p. 21 (entry "ceiling effect"), p. 67 (entry "floor effect"). 4

[3] R. Datta, D. Joshi, J. Li, and J. Z. Wang. Studying aesthetics in photographic images using a computational approach. In *ECCV*, pages 7–13, 2006. 1, 2

[4] R. Datta, J. Li, and J. Z. Wang. Learning the consensus on visual quality for next-generation image management. In *ACM-MM*, 2007. 6

[5] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009. 1

[6] S. Dhar, V. Ordonez, and T. Berg. High level describable attributes for predicting aesthetics and interestingness. In *CVPR*, pages 1657–1664. IEEE, 2011. 1, 6, 7

[7] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *IJCV*, 2010. 1

[8] G. Griffin, A. Holub, and P. Perona. Caltech-256 object category dataset. Technical report, California Institute of Technology, 2007. 1

[9] T. D. Henning Muller, Paul Clough and B. Caput. Experimental evaluation in visual information retrieval. the information retrieval series. *Springer*, 2010. 3

[10] D. Joshi, R. Datta, E. Fedorovskaya, Q. Luong, J. Wang, J. Li, and J. Luo. Aesthetics and emotions in images. *Signal Processing Magazine, IEEE*, 28(5):94–115, 2011. 6

[11] Y. Ke, X. Tang, and F. Jing. The design of high-level features for photo quality assessment. In *CVPR*, 2006. 1, 3, 6

[12] Kodak. *How to take good pictures : a photo guide*. Random House Inc, 1982. 2

[13] R. F. L. Fei-Fei and P. Perona. Learning generative visual models from few training examples: an incremental bayesian approach tested on 101 object categories. In *CVPR*, 2004. 1

[14] D. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2), 2004. 6

[15] W. Luo, X. Wang, and X. Tang. Content-based photo quality assessment. In *ICCV*, 2011. 1, 3, 6, 7

[16] Y. Luo and X. Tang. Photo and video quality evaluation: Focusing on the subject. In *ECCV*, 2008. 1, 6

[17] L. Marchesotti, F. Perronnin, D. Larlus, and G. Csurka. Assessing the aesthetic quality of photographs using generic image descriptors. In *ICCV*, 2011. 1, 3, 6

[18] F. Perronnin and C. Dance. Fisher kernels on visual vocabularies for image categorization. In *CVPR*, 2007. 6

[19] F. Perronnin, J. Sánchez, and T. Mensink. Improving the Fisher kernel for large-scale image classification. In *ECCV*, 2010. 6

[20] J. Russell. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161, 1980. 5

[21] O. Wu, W. Hu, and J. Gao. Learning to predict the perceived visual quality of photos. In *ICCV*, 2011. 3, 6