Recognizing Image Style

BMVC 2014 Submission # 468

Abstract

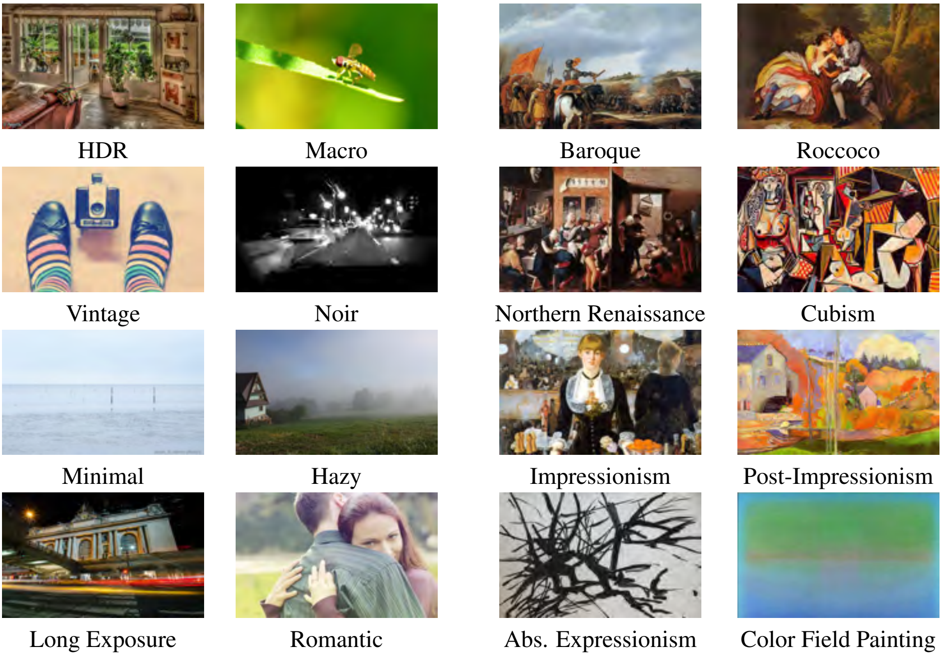
The style of an image plays a significant role in how it is viewed, but style has received little attention in computer vision research. We describe an approach to predicting style of images, and perform a thorough evaluation of different image features for these tasks. We find that features learned in a multi-layer network generally perform best – even when trained with object class (not style) labels. Our large-scale learning methods results in the best published performance on an existing dataset of aesthetic ratings and photographic style annotations. We present two novel datasets: 80K Flickr photographs annotated with 20 curated style labels, and 85K paintings annotated with 25 style/genre labels. Our approach shows excellent classification performance on both datasets. We use the learned classifiers to extend traditional tag-based image search to consider stylistic constraints, and demonstrate cross-dataset understanding of style.

1 Introduction

Deliberately-created images convey meaning, and *visual style* is often a significant component of image meaning. For example, a political candidate portrayed in the lush, springtime colors of a Renoir painting would tell a different story than if they were shown in the harsh, dark tones of a typical horror movie. Distinct visual styles are apparent in many everyday types of images, including art, design, cinematography, and advertising. Stylization has become extremely popular in amateur photography, spearheaded by the growth of mobile apps like Instagram. We argue that stylization has a significant impact on the viewer's response to an image and how it is interpreted, reflecting both aesthetics and meaning. Hence, understanding style is crucial to image understanding in many modern artistic and commercial contexts. Yet, very little research in computer vision has explored visual style.

Although is it very recognizable to human observers, visual style is a difficult concept to rigorously define. It depends on choices of colors, lighting, composition, scene objects, and optical techniques, and might suggest specific moods and genres. Most academic discussion of style has been in an art history context, but the distinctions between, say, Rococo versus pre-Rafaelite style are less relevant to modern photography and design. There has been some previous research in image style, but this has principally been limited to recognizing a few, well-defined optical properties, such as depth-of-field.

This paper studies the problem of photographic style recognition. We define several different *types* of image style, and gather a new, large-scale dataset of photographs annotated with style labels. This dataset embodies several different aspects of visual style, including photographic techniques ("Macro," "HDR"), composition styles ("Minimal," "Geometric"), moods ("Serene," "Melancholy"), genres ("Vintage," "Romantic," "Horror"), and types of



(a) Flickr Style: 80,000 images representing 20 different styles.

(b) Wikipaintings Style: 85,000 images labeled with 25 art historical styles.

Figure 1: Typical images in different style categories of our datasets.

scenes ("Hazy," "Sunny"). These styles are not mutually exclusive, and represent different attributes of style. We also gather a large dataset of visual art (mostly paintings) annotated with art historical style labels, ranging from Renaissance to modern art. Figure 1 shows some samples.

We test existing classification algorithms on these styles, evaluating several state-of-the-art image features. Most previous work in aesthetic style analysis has used hand-tuned features, such as color histograms. We find that deep convolutional neural network (CNN) features perform best for the task. This is surprising for several reasons: these features were trained on object class categories (ImageNet), and many styles appear to be primarily about color choices, yet the CNN features handily beat color histogram features. This leads to one conclusion of our work: mid-level features derived from object datasets are generic for style recognition, and superior to hand-tuned features.

We compare our predictors to human observers, using Amazon Mechanical Turk experiments, and find that our classifiers predict Group membership at essentially the same level of accuracy as Turkers. We also test on the AVA aesthetic prediction task [20], and show that using the "deep" object recogntion features improves over the state-of-the-art results.

Applications and code. Effective style predictors could be useful in several ways. First, we demonstrate an example of using our method to search for images by style. This could be useful for applications such as product search, storytelling, and creating slide presentations. In the same vein, visual similarity search results could be filtered by visual style, making possible queries such as "similar to this image, but more Film Noir" Second, style tags may provide valuable mid-level features for other image understanding tasks. For example, there has increasing recent effort in understanding image meaning, aesthetics, interestingness, popularity, and emotion (for example, [9, 11, 13, 15]), and style is an important part of meaning. Finally, learned predictors could be a useful component in modifying the style of an image.

All data, trained predictors, code, and a web interface for searching large image collections “with style” will be released upon publication.

Related Work. There has been growing interest in computer vision in predicting aesthetic and perceptual qualities of images, include beauty [4, 18, 20], memorability [11], “interestingness" [6, 9], sentiment based on object content [3], and similarity of compositions [24]. There has been some attention paid to predicting photographic style [20], but limited to a small number of optical techniques such as “HDR" and simple compositional qualities like “Duotones.” Several previous authors have developed systems to classify classic painting styles, including [14, 23]. These works consider only a handful of styles (less than ten apiece), with styles that are visually very distinct, e.g., Pollock vs. Dalí. These datasets comprise less than 60 images per style, for both testing and training. Mensink [19] provides a larger dataset of artworks, but does not consider style classification.

2 Data Sources

Building an effective model of photographic style requires annotated training data. To our knowledge, there is only one existing dataset annotated with visual style, and only a narrow range of photographic styles is represented [20]. We would like to study a broader range of styles, including different *types* of styles ranging from genres, compositional styles, and moods. Morever, large datasets are desirable in order to obtain effective results, and so we would like to obtain data from online communities, such as Flickr.

Flickr Style. Although Flickr users often provide free-form tags for their uploaded images, the tags tend to be quite unreliable. Instead, we turn to Flickr groups, which are community-curated collections of visual concepts. For example, the Flickr Group “Geometry Beauty” is described, in part, as “Circles, triangles, rectangles, symmetric objects, repeated patterns", and contains over 167K images at time of writing; the “Film Noir Mood” group is described as “Not just black and white photography, but a dark, gritty, moody feel..." and comprises over 7K images.

At the outset, we decided on a set of 20 visual styles, further categorized into types:

- **Optical techniques:** Macro, Bokeh, Depth-of-Field, Long Exposure, HDR
- **Atmosphere:** Hazy, Sunny
- **Mood:** Serene, Melancholy, Ethereal
- **Composition styles:** Minimal, Geometric, Detailed, Texture
- **Color:** Pastel, Bright
- **Genre:** Noir, Vintage, Romantic, Horror

For each of these stylistic concepts, we found at least one dedicated Flickr Group with clearly defined membership rules. From these groups, we collected 4,000 positive examples for each label, for a total of 80,000 images. Example images are shown in Figure 1a. The exact Flickr groups used are given in the Supplementary Materials.

The derived labels are considered clean in the positive examples, but may be noisy in the negative examples, in the same way as the ImageNet dataset [5]. That is, a picture labeled as *Sunny* is indeed *Sunny*, but it may also be *Romantic*, for which it is not labeled. We consider this an unfortunate but acceptable reality of working with a large-scale dataset. Following ImageNet, we still treat the absence of a label as indication that the image is a negative example for that label. Mechanical Turk experiments described in section 4.1 serve to allay our concerns.

Wikipaintings. We also provide a new dataset for classifying painting style. To our knowledge, no previous large-scale dataset exists for this task – although very recently a large dataset of artwork did appear for other tasks [19]. We collect a dataset of 100,000 high-art images – mostly paintings – labeled with artist, style, genre, date, and free-form tag information by a community of experts on the `Wikipaintings.org` website.

Analyzing style of non-photorealistic media is an interesting problem, as much of our present understanding of visual style arises out of thousands of years of developments in fine art, marked by distinct historical styles. Our dataset presents significant stylistic diversity, primarily spanning Renaissance styles to modern art movements (Supplementary Materials provides further breakdowns). We select 25 styles with more than 1,000 examples, for a total of 85,000 images. Example images are shown in Figure 1b.

3 Learning algorithm

We learn to classify novel images according to their style, using the labels assembled in the previous section. Because the datasets we deal with are quite large and some of the features are high-dimensional, we consider only linear classifiers, relying on sophisticated features to provide robustness.

We use an open-source implementation of Stochastic Gradient Descent with adaptive subgradient [1]. The learning process optimizes the function

$$\min_{w} \lambda_1 \|w\|_1 + \frac{\lambda_2}{2}\|w\|_2^2 + \sum_i \ell(x_i, y_i, w)$$

We set the L_1 and L_2 regularization parameters and the form of the loss function by validation on a held-out set. For the loss $\ell(x,y,w)$, we consider the hinge ($\max(0, 1 - y \cdot w^T x)$) and logistic ($\log(1 + \exp(-y \cdot w^T x))$) functions. We set the initial learning rate to 0.5, and use adaptive subgradient optimization [7]. Our setup is of multi-class classification; we use the One vs. All reduction to binary classifiers.

Features. To effectively classify diverse visual styles, we must choose appropriate image features. At the outset of this work, we hypothesized that image style may be related to many different types of features, including low-level statistics [17], color choices, composition, and content. In particular, color palette seems to be a distinctive feature of many styles: for example, *Noir* images are nearly all black-and-white, while most *Horror* images are very dark, and *Vintage* images use old photographic colors. However, that image content could be predictive of individual styles, e.g., *Macro* images include many images of insects and flowers. Hence, we test features that embody these different elements, including features from the object recognition literature.

Table 1: Mean APs on three datasets for the considered single-channel features and their second-stage combination. As some features were clearly worse than others on the AVA Style dataset, only the better features were evaluated on larger datasets.

	Fusion x Content	DeCAF_6	MC-bit	L*a*b* Hist	GIST	Saliency	random
AVA Style	0.604	0.577	0.529	0.291	0.220	0.149	0.127
Flickr	0.419	0.391	0.360	-	-	-	0.066
Wikipaintings	0.476	0.356	0.443	-	-	-	0.043

We test the following features; full details are given in the Supplemental Materials. **L*a*b color histogram,** using Palermo et al.'s representation [22]; **GIST descriptor [21]** ; **Graph-based visual saliency [10]**; **Meta-class (MC) binary object features [2]**; and **Deep convolutional neural networks (CNN)**, using the Caffe [12] implementation of Krizhevsky's ImageNet-trained classifier [16] (henceforth referred to as the **DeCAF** feature, with subscript denoting network layer). Notably, the last two of these are features designed and trained for object recognition.

For all features except binary ones, values are standardized: each column has its mean subtracted, and is divided by its standard deviation. For feature combinations, we use two-stage late fusion. First, single-feature classifiers are trained, then their scores are linearly combined with weights learned by a second classifier.

As we hypothesize that style features may be content dependent, we also trained **Content classifiers** (following Dhar et al. [6]) using the CNN features and an aggregated version of the PASCAL VOC [8] dataset. To enable our style classifiers to learn content-dependent style, we take the outer product of a feature channel with the aggregate content classifiers in doing feature combination.

4 Experiments

Details of our experiments follow, with a concluding discussion section. Due to space constraints, we are not able to provide as many figures as we would like here; however, the Supplementary Materials provides a full accounting.

4.1 Flickr Style

We learn and predict style labels on the 80,000 images labeled with 20 different visual styles of our new Flickr Style dataset, using 20% of the data for testing, and another 20% for parameter-tuning validation.

There are several performance metrics we consider. Average Precision evaluation (as reported in Table 1 and in detailed tables in the Supplementary Materials) is computed on a random class-balanced subset of the test data (each class has equal prevalence). We compute confusion matrices on the same data. Per-class accuracies are computed on subsets of the data balanced by the binary label, such that chance performance is 50%. We follow these decisions in all following experiments.

The best single-channel feature is DeCAF_6 with 0.391 mean AP; feature fusion obtains 0.419 mean AP. Per-class APs range from 0.11 [Bright] to 0.44 [Macro]. Per-class accuracies range from 68% [Romantic, Depth of Field] to 85% [Sunny, Noir, Macro]. The average per-class accuracy is 78%. We show the most confident style classifications on the test set of Flickr Style in Figure 3.

Upon inspection of the confusion matrices, we saw points of understandable confusion: Depth of Field vs. Macro, Romantic vs. Pastel, Vintage vs. Melancholy. There are also surprising sources of mistakes: Macro vs. Bright/Energetic, for example. To explain this particular confusion, we observed that lots of Macro photos contain bright flowers, insects, or birds, often against vibrant greenery. Here, at least, the content of the image dominates its style label.

To explore further content-style correlations, we plot the outputs of PASCAL object class classifiers (one of our features) on the Flickr dataset in Figure 2. We can observe that some styles have strong correlations to content (e.g., "Hazy" occurs with "vehicle", "HDR" doesn't occur with "cat").

We hypothesize that style is content-dependent: a Romantic portrait may have different low-level properties than a Romantic sunset. We form a new feature as an outer product of our content classifier features with the second-stage late fusion features ("Fusion × Content" in all results figures). These features gave the best results, thus supporting the hypothesis.

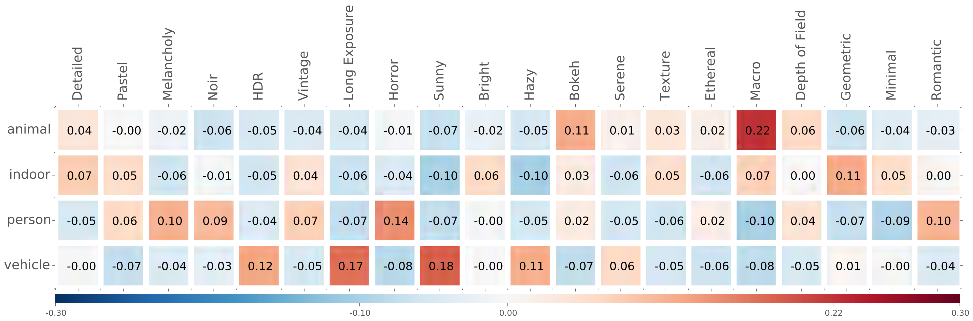


Figure 2: Correlation of PASCAL content classifier predictions (rows) against ground truth Flickr Style labels (columns). We see, for instance, that the Macro style is highly correlated with presence of animals, and that Long Exposure and Sunny style photographs often feature vehicles.

Mechanical Turk Evaluation. In order to provide a human baseline for evaluation, we performed a Mechanical Turk study. For each style, Turkers were shown positive and negative examples for each Flickr Group, and then they evaluated whether each image in the test set was part of the given style. We treat the Flickr group memberships as ground truth as before, and then evaluate Turkers' ability to accurately determine group membership. Measures were taken to remove spam workers; see the Supplemental Material details on the experimental setup. For efficiency, one quarter of the test set was used, and two redundant styles (Bokeh and Detailed) were removed. Each test image was evaluated by 3 Turkers, and the majority vote taken as the human result for this image.

In total, Turkers achieved 75% mean accuracy (ranging from 61% [Romantic] to 92% [Macro]) across styles, in comparison to 78% mean accuracy (ranging from 68% [Depth of Field] to 87% [Macro]) of our best method. Our algorithm did significantly worse than Turkers on Macro and Horror, and significantly better on Vintage, Romantic, Pastel, Detailed, HDR, and Long Exposure styles.

Some of this variance may be due to subtle difference from the Turk tasks that we provided, as compared to the definitions of the Flickr groups, but may also due to the Flickr groups' incorporating images that do not quite fit the common definition of the given style. For example, there may be a mismatch between different notions of "romantic" and "vintage," and how inclusively these terms are defined.

We additionally used the Turker opinion as ground truth for our method's predictions. In switching from the default Flickr to the MTurk ground truth, our method's accuracy hardly changed from 78% to 77%. However, we saw that the accuracy of our Vintage, Detailed, Long Exposure, Minimal, HDR, and Sunny style classifiers significantly decreased, indicating machine-human disagreement on those styles. Detailed tables are provided in Supplemental Results.

4.2 Wikipaintings

With the same setup and features as in the Flickr experiments, we evaluate 85,000 images labeled with 25 different art styles. The results are given in Table 1 here, and many more in Supplementary Materials. The best single-channel feature is MC-bit with 0.443 mean AP; feature fusion obtains 0.476 mean AP. Per-class accuracies range from 72% [Symbolism, Expressionism, Art Nouveau] to 94% [Ukiyo-e, Minimalism, Color Field Painting].

4.3 AVA Style

AVA [20] is a dataset of 250K images from `dpchallenge.net`. We evaluate classification of aesthetic rating and of 14 different photographic style labels on the 14,000 images of the AVA dataset that have such labels. For the style labels, the publishers of the dataset provide a train/test split, where training images have only one label, but test images may have more than one label [20]. Although the provided test split has an uneven class distribution, we found that to compare with the reported results, a class-balanced set is needed. Consequently, we adhere to the provided split but compute evaluation metrics on a random class-balanced subset of the test data.

For style classification, the best single feature is the DeCAF$_6$ convolution network feature, obtaining 0.577 mean AP. Feature fusion improves the result to 0.604 mean AP; both results beat the previous state-of-the-art of 0.538 mean AP [20].

In all metrics, the DeCAF and MC-bit features significantly outperformed more low-level features on this dataset. Accordingly, we do not evaluate the low-level features on the larger Flickr and Wikipaintings datasets.

4.4 Application: Style-Based Image Search

Style classifiers learned on our datasets can be used toward novel goals. For example, sources of stock photography or design inspiration may be better navigated with a vocabulary of style. Currently, companies expend labor to manually annotate stock photography with such labels. With our approach, any image collection can be searchable and rankable by style.

To demonstrate, we apply our Flickr-learned style classifiers to a new dataset of 80K images gathered on Pinterest (this data will also be made available with this paper's code release); some results are shown in Figure 5. Interestingly, styles learned from photographs can be used to order paintings, and styles learned from paintings can be used to order photographs, as illustrated in Figure 4.

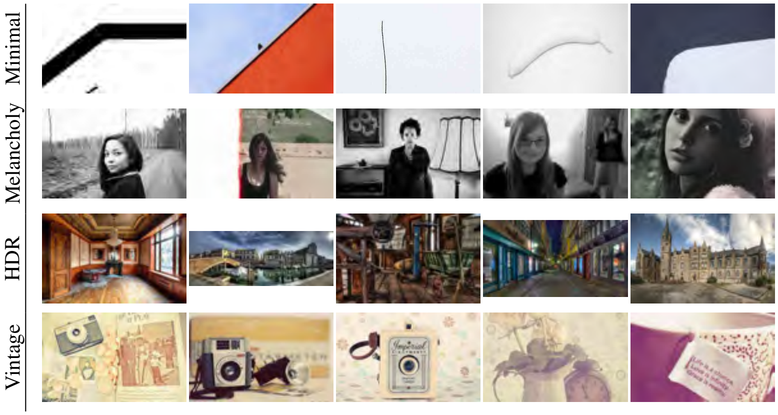


Figure 3: Top five most-confident positive predictions on the Flickr Style test set, for a few different styles. **See Figures 1-3 of the Supplemental Material for more results.**

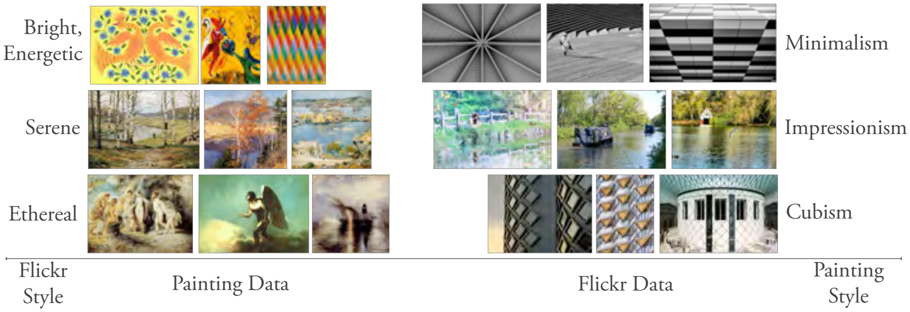


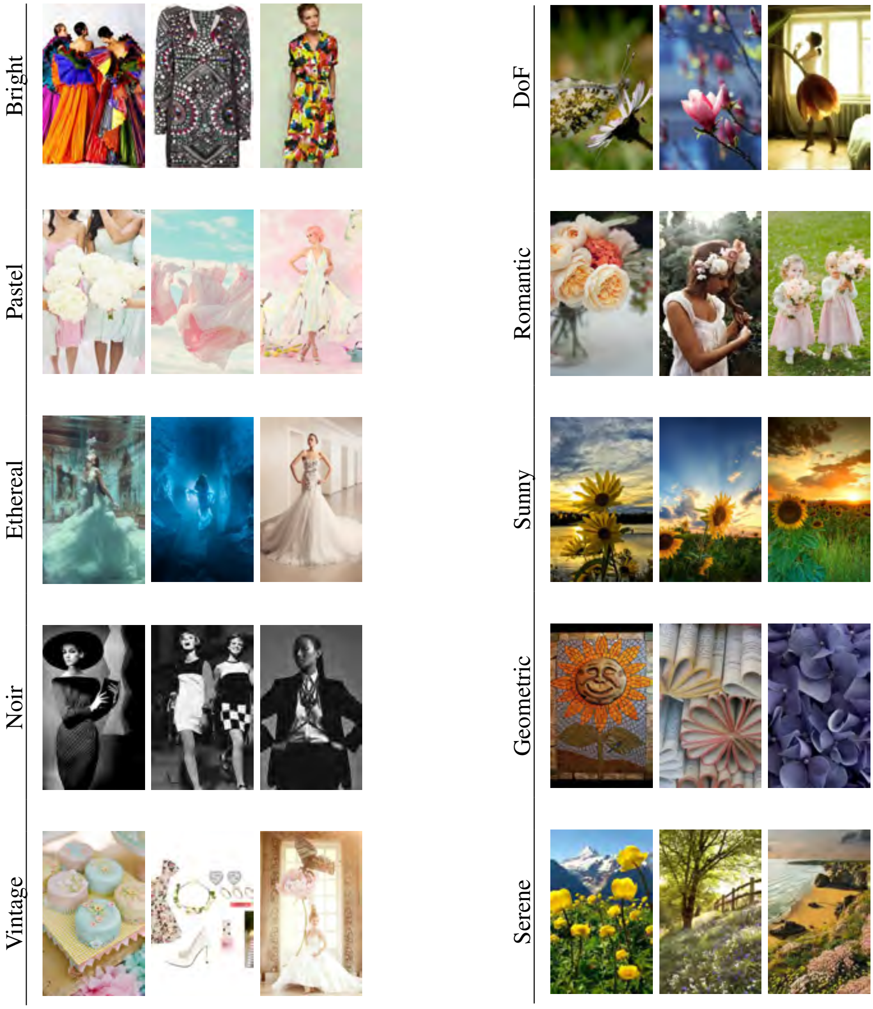
Figure 4: Cross-dataset style. On the left are shown top scorers from the Wikipaintings set, for styles learned on the Flickr set. On the right, Flickr photographs are accordingly sorted by Painting style. (Figure best viewed in color.)

4.5 Discussion

We have made significant progress in defining the problem of understanding photographic style. We provide a novel dataset that exhibits several types of styles not previously considered in the literature, and we demonstrate state-of-the-art results in prediction of both style and aesthetic quality. These results are comparable to human performance. We also show that style is highly content-dependent.

Style plays a significant role in much of the manmade imagery we experience daily, and there is considering need for future work to further answer the question “What is style?"

One of the most interesting outcomes of this work is the success of features trained for object detection, particularly MC-bit for aesthetic prediction, and CNNs for style prediction. We propose several possible hypotheses to explain these results. Perhaps the network layers that we use as features are extremely good as general visual features for image representation in general. Another explanation is that object recogntion depends on object appearance, e.g., distinguishing red from white wine, or different kinds of terriers, and that the model learns to repurpose these features for image style. Understanding and improving on these results is fertile ground for future work.



(a) Query: “dress”.

(b) Query: “flower”.

Figure 5: Example of filtering image search results by style. Our Flickr Style classifiers are applied to images found on Pinterest. The images are searched by the text contents of their captions, then filtered by the response of the style classifiers. Here we show three out of top five results for different query/style combinations.

References

[1] Alekh Agarwal, Olivier Chapelle, Miroslav Dudik, and John Langford. A Reliable Effective Terascale Linear Learning System. *Journal of Machine Learning Research*, 2012.

[2] A. Bergamo and L. Torresani. Meta-class features for large-scale object categorization on a budget. In *CVPR*, 2012.

[3] Damian Borth, Rongrong Ji, Tao Chen, and Thomas M Breuel. Large-scale Visual Sentiment Ontology and Detectors Using Adjective Noun Pairs. In *ACM MM*, 2013.

[4] Ritendra Datta, Dhiraj Joshi, Jia Li, and James Z Wang. Studying Aesthetics in Photographic Images Using a Computational Approach. In *ECCV*, 2006.

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

[6] Sagnik Dhar, Vicente Ordonez, and Tamara L Berg. High Level Describable Attributes for Predicting Aesthetics and Interestingness. In *CVPR*, 2011.

[7] John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *The Journal of Machine Learning Research*, 2011.

[8] M Everingham, L Van Gool, C K I Williams, J Winn, and A Zisserman. The PASCAL VOC Challenge Results, 2010.

[9] Michael Gygli, Fabian Nater, and Luc Van Gool. The Interestingness of Images. In *ICCV*, 2013.

[10] Jonathan Harel, Christof Koch, and Pietro Perona. Graph-Based Visual Saliency. In *NIPS*, 2006.

[11] Phillip Isola, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. What makes an image memorable? In *CVPR*, June 2011.

[12] Yangqing Jia. Caffe: An open source convolutional architecture for fast feature embedding. `http://caffe.berkeleyvision.org/`, 2013.

[13] Jungseock Joo, Weixin Li, Francis Steen, and Song-Chun Zhu. Visual persuasion: Inferring communicative intents of images. In *Proc. CVPR*, 2014.

[14] Daniel Keren. Painter identification using local features and naive bayes. In *Proc. ICPR*, 2002.

[15] A. Khosla, A. Das Sarma, and R. Hamid. What makes an image popular? In *Proc. WWW*, 2014.

[16] Alex Krizhevsky, Ilya Sutskever, and Geoff E. Hinton. Imagenet classification with deep convolutional neural networks. In *Proc. NIPS*, 2012.

[17] Siwei Lyu, Daniel Rockmore, and Hany Farid. A digital technique for art authentication. *PNAS*, 101(49), 2004.

[18] Luca Marchesotti and Florent Perronnin. Learning beautiful (and ugly) attributes. In *BMVC*, 2013.

[19] Thomas Mensink and Jan van Gemert. The rijksmuseum challenge: Museum-centered visual recognition. In *Proc. ICMR*, 2014.

[20] Naila Murray, De Barcelona, Luca Marchesotti, and Florent Perronnin. AVA: A Large-Scale Database for Aesthetic Visual Analysis. In *CVPR*, 2012.

[21] Aude Oliva and Antonio Torralba. Modeling the Shape of the Scene: A Holistic Representation of the Spatial Envelope. *IJCV*, 42(3):145–175, 2001.

[22] Frank Palermo, James Hays, and Alexei A Efros. Dating Historical Color Images. In *ECCV*, 2012.

[23] Lior Shamir, Tomasz Macura, Nikita Orlov, D. Mark Eckley, and Ilya G. Goldberg. Impressionism, expressionism, surrealism: Automated recognition of painters and schools of art. *ACM Trans. Applied Perc.*, 7(2), 2010.

[24] Jan C. van Gemert. Exploiting photographic style for category-level image classification by generalizing the spatial pyramid. In *Proc. ICMR*, 2011.