

A Geometry-Sensitive Approach for Photographic Style Classification

Koustav Ghosal¹, Mukta Prasad^{1,2}, and Aljosa Smolic¹

¹*V-SENSE, School of Computer Science and Statistics, Trinity College Dublin*

²*Daedalean, Zurich*

Abstract

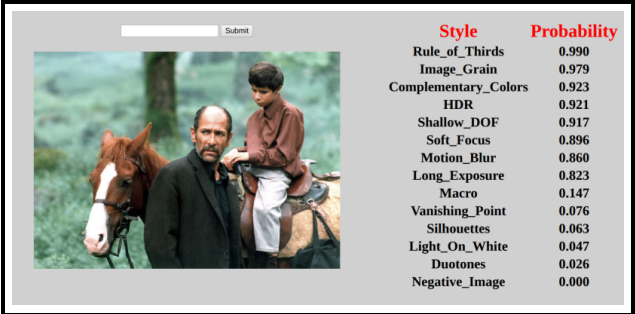
Photographs are characterized by different compositional attributes like the Rule of Thirds, depth of field, vanishing-lines *etc.* The presence or absence of one or more of these attributes contributes to the overall artistic value of an image. In this work, we analyze the ability of deep learning based methods to learn such photographic style attributes. We observe that although a standard CNN learns the texture and appearance based features reasonably well, its understanding of global and geometric features is limited by two factors. First, the data-augmentation strategies (cropping, warping, *etc.*) distort the composition of a photograph and affect the performance. Secondly, the CNN features, in principle, are translation-invariant and appearance-dependent. But some geometric properties important for aesthetics, *e.g. the Rule of Thirds* (RoT), are position-dependent and appearance-invariant. Therefore, we propose a novel input representation which is geometry-sensitive, position-cognizant and appearance-invariant. We further introduce a two-column CNN architecture that performs better than the state-of-the-art (SoA) in photographic style classification. From our results, we observe that the proposed network learns both the geometric and appearance-based attributes better than the SoA.

Keywords: Deep Learning, Convolutional Neural Networks, Computational Aesthetics

1 Introduction

Analyzing compositional attributes or styles is crucial for understanding the aesthetic value of photographs. At first, the computer vision community focused on modelling the physical properties of generic images for the more tangible, but very hard problems of object detection, localization, segmentation, tracking, *etc.* Popular datasets like Caltech, Pascal and ImageNet were created for training and evaluating such techniques effectively. The maturation of recognition and scene understanding has resulted in greater interest in the analysis of the subtler, aesthetic based aspects of image understanding. Furthermore, curated datasets such as AVA and Flickr-Style [Karayev et al., 2014, Murray et al., 2012] are now available and it is observed that learning from the matured areas of recognition/classification/detection transfers effectively to aesthetics and style analysis as well.

The aesthetic quality of a photograph is greatly influenced by its composition, that is a set of styles or attributes which guide the viewer towards the essence of the picture. Analyzing objectively, these



Style	Probability
Rule_of_Thirds	0.990
Image_Grain	0.979
Complementary_Colors	0.923
HDR	0.921
Shallow_DOF	0.917
Soft_Focus	0.896
Motion_Blur	0.860
Long_Exposure	0.823
Macro	0.147
Vanishing_Point	0.076
Silhouettes	0.063
Light_On_White	0.047
Duotones	0.026
Negative_Image	0.000

Figure 1: **Output from our network** : A screenshot from our web-based application. Attributes are shown with their probability values, ordered in descending order. This is a shot from Majid Majidi’s film ‘The Colours of Paradise’. We see that rule of thirds (for child’s position), shallow depth of field, complementary colours (green background and reddish foreground), image grain (because of the poor video quality) are all well identified.

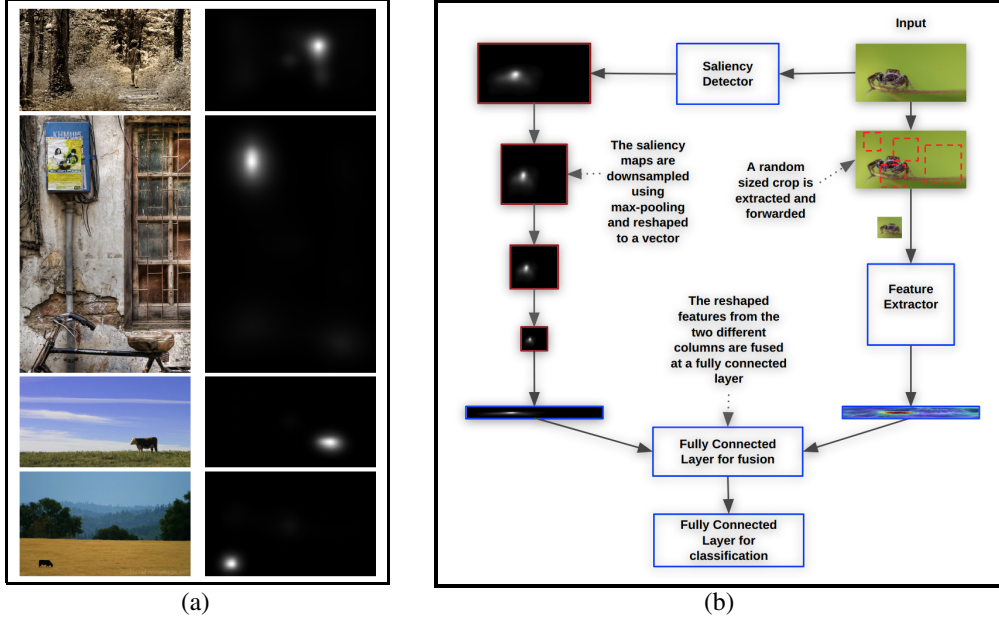


Figure 2: Our Contributions : **(a) Input (col 1), saliency maps (col 2)** : Saliency maps are generated using the method proposed in [Cornia et al., 2016]. The position of the main subjects can be obtained from the saliency maps. **(b) Our double-column CNN architecture:** One column accepts the regular RGB features and the other column accepts saliency maps. The features from RGB channel are computed using a pre-trained Densenet161 [Huang et al., 2016], fine-tuned on our datasets. They are fused using a fully-connected layer and finally passed to another final fully-connected layer for classification.

styles can be broadly categorized into local or appearance-based (focus, image-grain, *etc.*) and global or geometry-based (aspect ratio, RoT, framing, *etc.*). Figure 3 illustrates some popular styles adopted by photographers for a good composition.

In this work, we explore the ability of convolutional neural networks (CNN) to capture the aesthetic properties of photographic images. Specifically, can CNN based architectures learn both the local or appearance-based (such as colour) and global or geometry-based (such as RoT) aspects of photographs and how can we help such architectures capture location specific properties in images? Motivated by the recent developments in CNNs, our system takes a photograph as an input and predicts its style attributes (ordered by probabilities), as illustrated in Figure 1. There are several applications of automatic photographic style classification. For example, post-processing images and videos, tagging, organizing and mining large collections of photos for artistic, cultural and historical purposes, scene understanding, building assistive-technologies, content creation, cinematography, *etc.*

The traditional approach of using CNNs for natural image classification is to forward a *transformed* version of the input through a series of convolutional, pooling and fully connected layers and obtain a classification score. The transformation is applied to create a uniform sized input for the network (crop, warp, *etc.*) or to increase variance of the input distribution (flip, change contrast, *etc.*) for better generalization on the test data [Krizhevsky et al., 2012]. Clearly, such traditional transformations fail to preserve the aesthetic attributes of photographs. For example, a random fixed-sized crop cannot capture the arrangement of subjects within the picture. On the other hand, although warping the input photograph to a fixed size preserves the global context of the subjects better than crop, it distorts the aspect ratio and also smoothens appearance-based attributes like depth of field or image-grain.

This calls for a representation which preserves both the appearance-based and geometry-based properties of a photograph and which generalizes well over test data. Multiple solutions to these problems have been proposed. In [Lu et al., 2014], authors propose a double column CNN architecture, where the first column accepts

a cropped patch and the second column accepts a warped version of the entire input. In subsequent work [Lu et al., 2015], multiple patches are cropped from an input and forwarded through the network. The features from multiple patches are aggregated before the final fully-connected layer for classification. The authors argue that sending multiple patches from the same image encodes more global context than a single random crop. More recently, [Ma et al., 2017] follow a similar multiple-patch extraction approach, but the patches are selectively extracted based on saliency, pattern-diversity and overlap between the subjects. Essentially, these techniques attempt to incorporate global context into the features during a forward pass either by warping the whole input and sending it through an additional column or by providing multiple patches from the input at the same time.

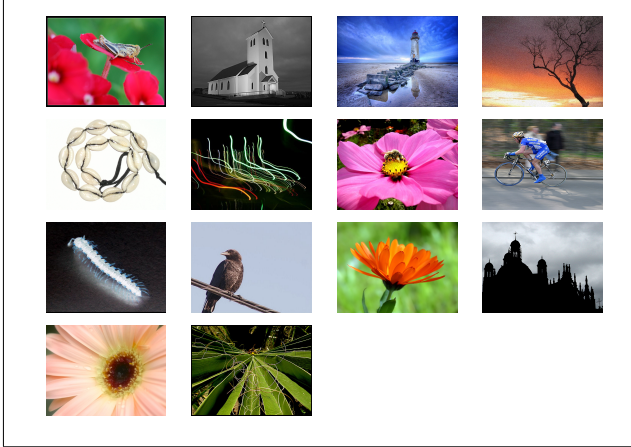


Figure 3: **Example images from the AVA dataset corresponding to 14 different styles: (L-R) Row 1 :** Complementary Colors, Duotones, HDR, Image Grain. **Row 2 :** Light On White, Long Exposure, Macro, Motion Blur. **Row 3 :** Negative Image, Rule of Thirds, Shallow DOF, Silhouettes. **Row 4 :** Soft Focus, Vanishing Point

are obtained from the saliency maps and then fused with the appearance features coming from a traditional CNN and finally passed to a classifier to identify the overall style of composition of the photograph. By definition, saliency maps are appearance-invariant. On the other hand, by avoiding convolution and fusing them directly with the CNN features we achieve location-cognizance. In Section 5, we show that our approach performs better than the SoA in photographic style classification especially for those styles which are geometry-sensitive.

Our second contribution is a comparative analysis of the traditional approaches for aesthetic categorization of images. Motivated both from the SoA and recent breakthroughs in deep learning, we implement multiple baselines, by trying different architectures and try to understand and identify the factors that are crucial for encoding the local and global aspects of photographic composition.

The rest of the paper is organized as follows. In Section 2, we summarize the relevant literature in image aesthetic quality prediction. In Section 3, we describe the double column CNN architecture we adopt. In Section 4, we provide a detailed description of the datasets used. In Section 5, we provide details of the experiments conducted and analyze the results.

2 Related Work

Image and video classification has always been a fundamental problem in computer vision. Understanding quantifiable visual semantics like the class, position and number of objects in an image were challenging enough and took the majority of focus. However, understanding the subtle, qualitative aspects especially from a creative perspective, due to its even more challenging nature, has only recently started being attacked.

Although these traditional double column or multi-patch strategies improve the overall performance, we argue that these networks cannot properly learn the geometry of a photograph. It is because CNNs, in principle, are designed to be translation invariant [Sabour et al., 2017]. While they can learn how the subjects look like, they cannot capture whether the subjects are rightly positioned. Since the convolutional filters corresponding to a feature map share weights, they become translation-invariant and appearance-dependent. In other words, they are activated for an object irrespective of its location in the image. As a result, they fail to understand photographic attributes like RoT. One option to tackle this could be training a fully-connected network on the full images, but they have too many parameters and are hard to train.

Our first contribution in this work is introducing a saliency-based representations (see Figure 2(a)) which we call **Sal-RGB** features. The position or relative geometry of the different subjects in the image

The initial works in photograph aesthetic assessment relied on explicitly modelling popular attributes like RoT, colour harmony, exposure, *etc.* [Datta et al., 2006, Ke et al., 2006, Luo and Tang, 2008]. Some recent works address the problem similarly, *i.e* explicitly defining the features but with improved performance [Obrador et al., 2012, Dhar et al., 2011, Joshi et al., 2011, San Pedro et al., 2012, Karayev et al., 2014]. [Aydin et al., 2015] propose a system which predicts the contribution of some photographic attributes towards the overall aesthetic quality of a picture. After estimating the extent of certain compositional attributes, they aggregate the scores for different attributes to predict the overall aesthetic score of a photograph by using a novel calibration technique. [Murray et al., 2012] published the Aesthetic Visual Analysis (AVA) dataset. Improved evaluation due to such datasets and parallel advances in deep learning resulted in a surge of research in this area in the last few years.

In recent years, deep learning has performed remarkably well in many computer vision tasks like classification [Krizhevsky et al., 2012], detection [Girshick, 2015], segmentation [Noh et al., 2015] and scene understanding [Karpthy and Fei-Fei, 2015, Xu et al., 2015]. Recent works like [Huang et al., 2016, He et al., 2016] have performed well in multi-tasking frameworks for detection and classification. In [He et al., 2016], the authors use a residual framework for tackling training error upon addition of new layers. In [Huang et al., 2016] the authors use dense connections by connecting outputs from all the previous layers as input to the next layer. As for many computer vision problems, deep learning has begun to be explored in the domain of image aesthetic assesment as well. Apart from [Lu et al., 2015, Lu et al., 2014, Ma et al., 2017] (discussed in Section 1), in [Kong et al., 2016], the authors learn styles and ratings jointly on a new dataset. Their algorithm is based on comparing and ranking a pair of images instead of directly predicting their coarse aesthetic scores. [Mai et al., 2016] propose a network that uses a composition-preserving input mechanism. They introduce an aspect-ratio aware pooling strategy that reshapes each image differently. In [Malu et al., 2017], the authors propose a network that predicts the overall aesthetic score and eight style attributes, jointly. Additionally, they use gradient-based feature visualization techniques to understand the correlation of different attributes with image locations.

In principle, our pipeline is similar to [Lu et al., 2015, Lu et al., 2014, Karayev et al., 2014] in the sense that we also perform a neural style prediction on the AVA dataset. However, our work differs in two important aspects. First, in the overall style prediction, our Sal-RGB features perform better than the strategies that use generic features [Karayev et al., 2014], the double column [Lu et al., 2014] or multi-patch aggregation [Lu et al., 2015]. Second, unlike [Lu et al., 2014, Lu et al., 2015] we analyze individual attributes and evaluate our strategy on multiple datasets.

3 Network Architecture

In this section, we describe our architecture, as illustrated in Figure 2(b). Our architecture consists of three main blocks — the saliency detector, the double-column feature-extractor and the classifier.

3.1 Saliency Detector

We compute the saliency maps using the method proposed in [Cornia et al., 2016]. Motivated from recent attention based models [Xu et al., 2015] that processes some regions of the input more attentively than others, the authors propose a CNN-LSTM (long and short term memory network) framework for saliency detection. LSTMs are applied to sequential inputs where output from previous states are combined with inputs to the next state using dot products. In this work, the authors modify the standard LSTM such that they accept a sequence of spatial data (patches extracted from different locations in the image) and combine them using convolutions instead of dot products. Additionally, they introduce a center-prior component, that handles the tendency of humans to fix attention at the center region of an image. Some outputs from the system can be found in Figure 2(a), second column.

3.2 Feature Extractor

The feature extractor consists of two parallel and independent columns, one for the saliency map and the other for raw RGB input.

Saliency Column : The saliency column consists of two max-pooling layers that downsample the input from 224×224 to 56×56 as shown in 2(b). Instead of max-pooling, we tried strided convolutions as they are known to capture low level details better than pooling [Johnson et al., 2016]. But pooling gave better results in our case which perhaps indicates that the salient position was more important than the level of detail captured.

RGB Column : We choose the DenseNet161 [Huang et al., 2016] network for its superior performance in the ImageNet challenge. Very deep networks suffer from the *vanishing-gradient* problem *i.e.* gradual loss of information as the input passes through several intermediate layers. Recent works like [He et al., 2016, Srivastava et al., 2015] address this problem by explicitly passing information between layers or by dropping random layers while training. The DenseNet is different from the traditional CNNs in the manner in which each layer receives input from the previous layers. The l^{th} layer in DenseNet receives as input, the concatenated output from all previous $l - 1$ layers.

We replace the last fully-connected layer from DenseNet with our classifier described in Section 3.3 and use the remaining as a feature extractor. Since we have less training images, we fine-tune a model pre-trained on ImageNet on our dataset instead of training from scratch. This works since the lower level features like edges and corners are generic image features and can be used for aesthetic tasks too.

3.3 Classifier

Feature-maps from the two columns are concatenated and fused together using a fully-connected layer. A second and final fully-connected layer is used as a classifier. During training, we use the standard cross-entropy loss function and the gradient is back-propagated to the two columns.

4 Datasets

We use two standard datasets for evaluation — AVA Style and Flickr Style. AVA [Murray et al., 2012] is a dataset containing 250,000 photographs, selected from www.dpchallenge.com. Dpchallenge is a forum for photographers. Users rate each photograph during the challenge on a scale of 10 and post feedback during and after the challenge.

Of these 250,000 photographs, the authors manually select 72 challenges, corresponding to 14 different photographic styles as illustrated in Figure 3 and create a subset called AVA Style containing about 14,000 images. While training images in the subset are annotated with a single label, the test images have multiple labels associated with them making them unsuitable for popular evaluation frameworks used for single-label multi-class classifiers.

Flickr Style [Karayev et al., 2014] is a collection of 80,000 images of 20 visual styles. The styles span across multiple concepts such as optical techniques (Macro, Bokeh, *etc.*), atmosphere (Hazy, Sunny, *etc.*), mood (Serene, Melancholy, *etc.*), composition styles (Minimal, Geometric, *etc.*), colour (Pastel, Bright, *etc.*) and genre (Noir, Romantic, *etc.*). Flickr Style is a more complex dataset than AVA not only because it has more classes, but because some of the classes like Horror, Romantic and Serene are subjective concepts and difficult to encode objectively.

5 Experiments

We investigate two different aspects of the problem. First, in Section 5.1 we report the overall performance of our features using mean average precision (MAP). Second, in Section 5.2 we observe the per-class precision (PCP) scores to understand how our features affect individual photographic attributes. For comparison, we use MAP reported in [Karayev et al., 2014, Lu et al., 2014, Lu et al., 2015]. PCP is compared only with [Karayev et al., 2014] since the implementations were unavailable for [Lu et al., 2014, Lu et al., 2015]. Additionally, we implement the following two benchmarks to evaluate our approach.

- **DenseNet161, ResNet152 :** These are off-the-shelf implementations [Huang et al., 2016, He et al., 2016] finetuned on our dataset and takes only RGB representation as input. These were chosen since they achieve the least error rates for ImageNet classification.

- **RAPID++** : Following [Lu et al., 2014], we implemented a two-column network. Each column takes as input, random crops and the whole image, as local and global representations, respectively. But, we used DenseNet161 architecture for the two columns whereas in the original work the authors use a shallower architecture with only three layers. We choose this as a benchmark in order to observe how their algorithm performs with a deeper architecture.

We train style classifiers on the AVA Style and Flickr Style datasets. The train-test partitions are followed from the original papers [Murray et al., 2012, Karayev et al., 2014].

For AVA, We use 11270 images for training and validation and 2573 images for testing. For Flickr Style we use 64000 images for training and 16000 images for testing. For testing, we follow the approach adopted by [Lu et al., 2014, Lu et al., 2015]. 50 patches are extracted from the test-image and each patch is passed through the network. The results are averaged to achieve the final scores.

5.1 Style Classification

The scores are reported in terms of Mean Average Precision (MAP). MAP refers to the average of per-class precision. The results are reported in Table 1. We observe that our method outperforms the SoA [Karayev et al., 2014, Lu et al., 2014, Lu et al., 2015] significantly. But, our own baselines perform more or less

Table 1: **Style Classification : Comparison with the SoA** : The results are reported in terms of Mean Average Precision(average of per class precision). We observe that for both the datasets, our method performs better than the state of the art. Flickr Style was not used in [Lu et al., 2014, Lu et al., 2015].

Network	Augmentation	AVA	Flickr Style
Fusion [Karayev et al., 2014]	centre crop	58.10	36.80
RAPID [Lu et al., 2014]	random crop, warp	56.81	-
Multi-Patch [Lu et al., 2015]	random crop	64.07	-
DenseNet161 [Huang et al., 2016]	random crop	71.68	43.83
ResNet152 [He et al., 2016]	random crop	70.57	43.65
RAPID++	random crop, warp	70.48	41.93
Sal-RGB	random crop	71.82	43.45

equally well. We deduce that for the improvement of MAP, the maximum impact is made by a more sophisticated CNN, followed by the location specific saliency. Both ResNet [He et al., 2016] and DenseNet [Huang et al., 2016] are residual networks and learn complex representations due to their very deep architectures. Such representations are crucial for learning photographic attributes, which have many overlapping properties (less inter-class variance).

From these results, one might argue that the improvement can be attributed largely to a better CNN, and so what does Sal-RGB bring to the representation ? We address this issue in Section 5.2.

5.2 Per-class Precision Scores

In [Karayev et al., 2014], the authors report per-class precision (PCP) scores on AVA Style and Flickr Style. We compare our algorithm with those results in Table 2.

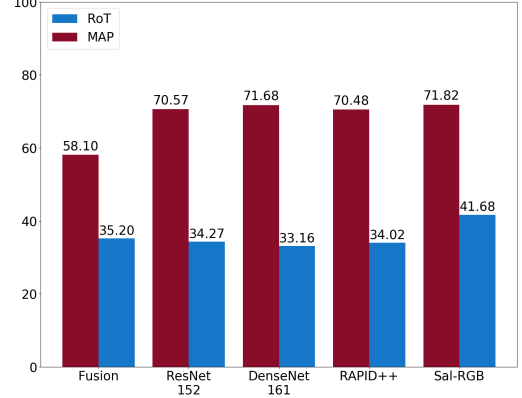
We observe that our method outperforms [Karayev et al., 2014] in almost all categories on both datasets. For the AVA Style dataset, a significant improvement is observed in the appearance-based categories like complementary colours, duotones, image grain, *etc.* Yet again, our own baselines DenseNet, ResNet and RAPID++ perform equally well in most categories except for RoT. For this category, Sal-RGB outperforms all others by a significant margin. This is an important result, since unlike others, RoT is a purely geometric attribute and important for image aesthetics and photography. A significant improvement in this category is a confirmation of our claim that the proposed approach efficiently encodes the geometry of a photograph. We highlight these observations in the bar plot beside Table 2.

5.3 Limitations

We tried to understand the limitations of our approach by plotting the confusion matrix for the different attributes of AVA.

Table 2: **PCP for AVA Style** : Sal-RGB outperforms the SoA [Karayev et al., 2014] by a significant margin in every category. Our own baselines DenseNet [Huang et al., 2016], ResNet [He et al., 2016], RAPID++ perform equally well for almost all categories except RoT, for which Sal-RGB performs much better. The bar plot on the right shows the relative improvements in overall MAP and RoT respectively.

Styles	Fusion(SoA)	Densenet161	ResNet152	RAPID++	Sal-RGB
Complementary_Colors	46.90	62.33	62.15	61.49	61.41
Duotones	67.60	86.58	84.82	84.77	87.58
HDR	66.90	74.95	70.08	71.51	72.86
Image_Grain	64.70	81.55	79.48	83.15	82.20
Light_On_White	90.80	84.69	83.41	85.64	82.99
Long_Exposure	45.30	64.16	65.38	63.94	61.94
Macro	47.80	64.89	65.52	64.90	66.58
Motion_Blur	47.80	63.93	62.12	61.21	61.98
Negative_Image	59.50	87.40	86.11	82.01	87.71
Rule_of_Thirds	35.20	33.16	34.27	34.02	41.68
Shallow_DOF	62.40	82.08	82.42	82.95	82.39
Silhouettes	79.10	93.73	92.49	91.14	93.05
Soft_Focus	31.20	49.89	44.91	44.57	46.41
Vanishing_Point	68.40	74.16	74.80	75.45	76.76



- The strongest classes are Light on White, Silhouettes, Vanishing Points. The weakest are Motion Blur and Soft Focus.
- Long Exposure and Motion Blur get confused with each other, which makes sense, since both attributes are captured using a slow shutter speed and mostly at night.
- Shallow DOF, Soft Focus and Macro are mutually confused classes, which is justified as all of them involve blur.
- The poorly performing classes have a high false-positive rate. We blame this on two factors. First, some classes such as Motion Blur and Soft-Focus have less samples as compared to others. Secondly, we observe that there is some ambiguity in the annotation of the training data of AVA. They are associated with a single label. But usually, most of the good photographs are captured with an interplay between multiple attributes. For example a macro image could very well conform to RoT or depth-of-field. Thus a single annotation incorporates undesired penalties to the loss during training the network

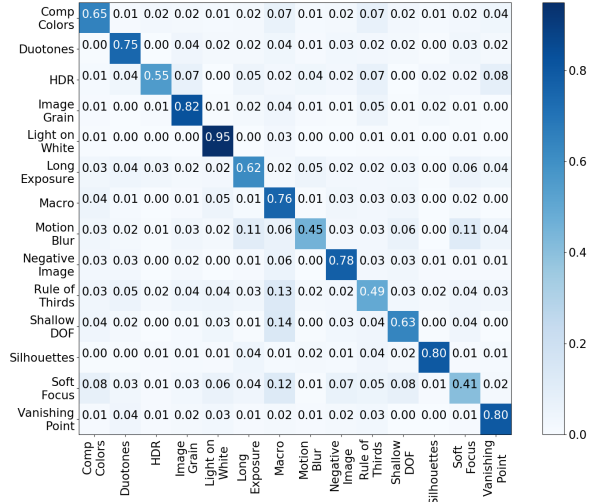


Figure 4: **Confusion matrix for AVA Style with our model:** For a test sample, the rows correspond to the real class and the columns correspond to the predicted class. The values are computed over 2573 test samples of AVA and then normalized. The diagonal elements represent correct classifications. The color scale ranges from 0.0 (light blue) to 0.8 (dark blue).

6 Conclusion

There are many potential applications of an automatic style and aesthetic quality estimator in the domain of digital photography such as interactive cameras, automated photo correction *etc.* Our system can be directly extended to video-processing for predicting shot-styles. For example, Figures 1 illustrates the aesthetic analysis of a shot taken from Majid Majidi’s movie *Colours of Paradise*. As future work, there are many possible directions. Generalizing the model to more style attributes could be one. Extending the system to the domain of video and 360 images would also be possible. A thorough mathematical analysis of seemingly intangible and subjective concepts in art and subsequently fixing ambiguities in the data-annotation could be another. We hope that this area will become more active in the future with its challenging and interesting set of problems.¹

¹This publication has emanated from research conducted with the financial support of Science Foundation Ireland (SFI) under the Grant Number 15/RP/2776

References

- [Aydın et al., 2015] Aydın, T. O., Smolic, A., and Gross, M. (2015). Automated aesthetic analysis of photographic images. *IEEE transactions on visualization and computer graphics*, 21(1):31–42.
- [Cornia et al., 2016] Cornia, M., Baraldi, L., Serra, G., and Cucchiara, R. (2016). Predicting human eye fixations via an lstm-based saliency attentive model. *arXiv preprint arXiv:1611.09571*.
- [Datta et al., 2006] Datta, R., Joshi, D., Li, J., and Wang, J. (2006). Studying aesthetics in photographic images using a computational approach. *Computer Vision—ECCV 2006*, pages 288–301.
- [Dhar et al., 2011] Dhar, S., Ordonez, V., and Berg, T. L. (2011). High level describable attributes for predicting aesthetics and interestingness. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 1657–1664. IEEE.
- [Girshick, 2015] Girshick, R. (2015). Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448.
- [He et al., 2016] He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.
- [Huang et al., 2016] Huang, G., Liu, Z., Weinberger, K. Q., and van der Maaten, L. (2016). Densely connected convolutional networks. *arXiv preprint arXiv:1608.06993*.
- [Johnson et al., 2016] Johnson, J., Alahi, A., and Fei-Fei, L. (2016). Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision*, pages 694–711. Springer.
- [Joshi et al., 2011] Joshi, D., Datta, R., Fedorovskaya, E., Luong, Q.-T., Wang, J. Z., Li, J., and Luo, J. (2011). Aesthetics and emotions in images. *IEEE Signal Processing Magazine*, 28(5):94–115.
- [Karayev et al., 2014] Karayev, S., Hertzmann, A., Winnemoeller, H., Agarwala, A., and Darrell, T. (2014). Recognizing image style. In *BMVC 2014*.
- [Karpathy and Fei-Fei, 2015] Karpathy, A. and Fei-Fei, L. (2015). Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3128–3137.
- [Ke et al., 2006] Ke, Y., Tang, X., and Jing, F. (2006). The design of high-level features for photo quality assessment. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 1, pages 419–426. IEEE.
- [Kong et al., 2016] Kong, S., Shen, X., Lin, Z., Mech, R., and Fowlkes, C. (2016). Photo aesthetics ranking network with attributes and content adaptation. In *European Conference on Computer Vision*, pages 662–679. Springer.
- [Krizhevsky et al., 2012] Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105.
- [Lu et al., 2014] Lu, X., Lin, Z., Jin, H., Yang, J., and Wang, J. Z. (2014). Rapid: Rating pictorial aesthetics using deep learning. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 457–466. ACM.
- [Lu et al., 2015] Lu, X., Lin, Z., Shen, X., Mech, R., and Wang, J. Z. (2015). Deep multi-patch aggregation network for image style, aesthetics, and quality estimation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 990–998.
- [Luo and Tang, 2008] Luo, Y. and Tang, X. (2008). Photo and video quality evaluation: Focusing on the subject. *Computer Vision—ECCV 2008*, pages 386–399.
- [Ma et al., 2017] Ma, S., Liu, J., and Wen Chen, C. (2017). A-lamp: Adaptive layout-aware multi-patch deep convolutional neural network for photo aesthetic assessment. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- [Mai et al., 2016] Mai, L., Jin, H., and Liu, F. (2016). Composition-preserving deep photo aesthetics assessment. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 497–506.
- [Malu et al., 2017] Malu, G., Bapi, R. S., and Indurkha, B. (2017). Learning photography aesthetics with deep cnns.
- [Murray et al., 2012] Murray, N., Marchesotti, L., and Perronnin, F. (2012). Ava: A large-scale database for aesthetic visual analysis. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2408–2415. IEEE.
- [Noh et al., 2015] Noh, H., Hong, S., and Han, B. (2015). Learning deconvolution network for semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1520–1528.
- [Obrador et al., 2012] Obrador, P., Saad, M. A., Suryanarayan, P., and Oliver, N. (2012). Towards category-based aesthetic models of photographs. In *International Conference on Multimedia Modeling*, pages 63–76. Springer.
- [Sabour et al., 2017] Sabour, S., Frosst, N., and Hinton, G. E. (2017). Dynamic routing between capsules. *arXiv preprint arXiv:1710.09829*.
- [San Pedro et al., 2012] San Pedro, J., Yeh, T., and Oliver, N. (2012). Leveraging user comments for aesthetic aware image search reranking. In *Proceedings of the 21st international conference on World Wide Web*, pages 439–448. ACM.
- [Srivastava et al., 2015] Srivastava, R. K., Greff, K., and Schmidhuber, J. (2015). Training very deep networks. In *Advances in neural information processing systems*, pages 2377–2385.
- [Xu et al., 2015] Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., and Bengio, Y. (2015). Show, attend and tell: Neural image caption generation with visual attention. In *International Conference on Machine Learning*, pages 2048–2057.