Predicting Image Aesthetics with Deep Learning

Simone Bianco, Luigi Celona, Paolo Napoletano(✉), and Raimondo Schettini

Department of Informatics, Systems and Communication,
University of Milano-Bicocca, viale Sarca, 336, 20126 Milan, Italy
{bianco,luigi.celona,napoletano,schettini}@disco.unimib.it

Abstract. In this paper we investigate the use of a deep Convolutional Neural Network (CNN) to predict image aesthetics. To this end we fine-tune a canonical CNN architecture, originally trained to classify objects and scenes, by casting the image aesthetic prediction as a regression problem. We also investigate whether image aesthetic is a global or local attribute, and the role played by bottom-up and top-down salient regions to the prediction of the global image aesthetic. Experimental results on the canonical Aesthetic Visual Analysis (AVA) dataset show the robustness of the solution proposed, which outperforms the best solution in the state of the art by almost 17 % in terms of Mean Residual Sum of Squares Error (MRSSE).

1 Introduction

The automatic assessment of image aesthetic is a novel challenge for the computer vision community that has wide applications, e.g. image retrieval, photo management, photo enhancement, image cropping, etc. [13,20]. Because of the subjectivity of humans' aesthetic evaluation, in recent years, many research efforts have been made and various approaches have been proposed [5,19,24,26,29]. According to the way the problem is formulated, computational approaches can be divided into two groups: aesthetic classification and aesthetic regression. The first group of methods treats aesthetic quality assessment as a binary classification problem, i.e. distinguish between aesthetic and unaesthetic images. Most of these methods have focused on designing features able to replicate the way people perceive the aesthetic quality of images. For example, Datta et al. [12] design special visual features (colorfulness, the rule of thirds, low depth of field indicators, etc. [3,4,7]) and use the Support Vector Machine (SVM) and Decision Tree (DT) to discriminate between aesthetic and unaesthetic images. Nishiyama et al. [28] propose an approach based on color harmony and bags of color patterns to characterize color variations in local regions. Marchesotti et al. [25] demonstrate that generic image descriptors, such as GIST, Bag-of-Visual-words (BOV) encoded from Scale-Invariant Feature Transform (SIFT) information, and Fisher Vector (FV) encoded from SIFT information, are able to capture a wealth of statistics useful for aesthetic evaluation of photographs. Simon et al. [29] show that aesthetic quality depends on context since they obtain more accurate predictions by selecting features for specific image categories. Methods able to learn effective aesthetic features directly from images

© Springer International Publishing AG 2016
J. Blanc-Talon et al. (Eds.): ACIVS 2016, LNCS 10016, pp. 117–125, 2016.
DOI: 10.1007/978-3-319-48680-2_11

have been proposed. Lu et al. [24] present the RAting PIctorical aesthetics using Deep learning (RAPID) system, which adopts a Convolutional Neural Network (CNN) approach to automatically learn features for aesthetic quality categorization. Kao et al. [19] train a linear SVM using the features extracted from a CNN pre-trained on ImageNet classification task.

The second group of approaches considers aesthetic quality assessment as a regression problem, i.e. they predict an aesthetic rating or score of the images. Datta et al. [12] propose the use of Linear Regression (LR) with polynomial terms of the features to predict the aesthetic score. Bhattacharya et al. [2] propose to use a saliency map and a high-level semantic segmentation technique for extracting aesthetic features subsequently used for training a Support Vector Regression (SVR) machine. Wu et al. [30] design a new algorithm called Support Vector Distribution Regression (SVDR) in order to use a distribution of user ratings instead of a scalar for model learning. More recently, Kao et al. [19] propose a regression model based on CNNs, which achieves the state-of-the-art results on aesthetic quality assessment.

In this paper we investigate the use of a deep CNN to predict image aesthetic scores. To this end we fine-tune [1,31] a canonical CNN architecture, originally trained to classify both objects and scenes, by casting the image aesthetic prediction as a regression problem. We also investigate whether image aesthetic is a global or local attribute, and the role played by bottom-up and top-down salient regions [16,18] to the prediction of the global image aesthetic. For the evaluation we use the AVA dataset [26], because it is actually the largest dataset available and the only one providing aesthetic ratings instead of binary classification of aesthetic quality (e.g. "high" or "low"). Experimental results show the robustness of the solution proposed, which outperforms the best solution in the state of the art by almost 17 % in terms of Mean Residual Sum of Squares Error (MRSSE).

The rest of the paper is organized as follows: Sect. 2 describes the data and the evaluation metric; Sect. 3 describes the proposed approach; Sect. 4 analyzes the experimental results; finally, Sect. 5 presents our final considerations.

2 Database and Evaluation Criterions

In this work we use the Aesthetic Visual Analysis (AVA) dataset [26], that is a large-scale collection of images and meta-data obtained from the on-line community of photography amateurs and covering a wide variety of subjects on almost 1,000 challenges derived from www.dbchallenge.com. Figure 1 shows some samples from the AVA dataset. It contains over 255,000 images, both in RGB and grayscale with three types of annotations: aesthetic ratings ranging from 1 to 10; semantic annotations consisting in 66 textual tags describing the semantics of the images; photographic style annotations corresponding to 14 photographic techniques.

For the experiments we follow the same experimental procedure as in [19]. We discard the images whose longest dimension is three times more than the smallest

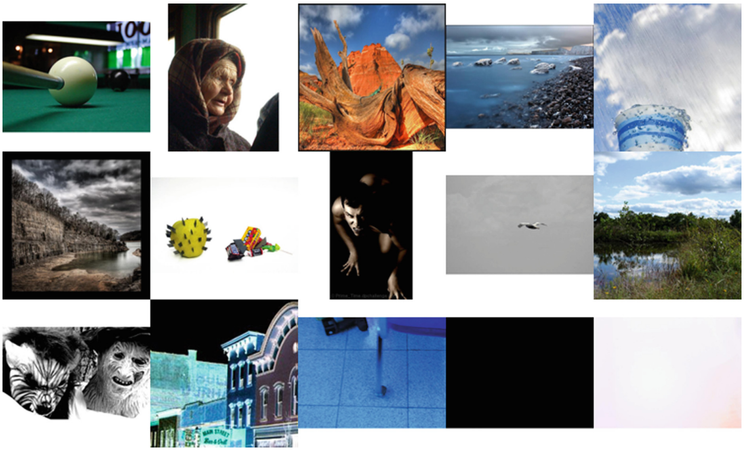


Fig. 1. Sample images from the Aesthetic Visual Analysis (AVA) database sorted by their aesthetic score (decreasing from left to right).

dimension, resulting in a total of 255,099 images. Among them, 250,129 images are selected for train and 4,970 for test. The average score of user ratings is taken as the images aesthetic quality ground truth. For performance evaluation, we use the Mean Residual Sum of Squares Error (MRSSE), that is defined as follows:

$$MRSSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

where $\hat{y}_i$ is the predicted aesthetic score and y_i is the ground truth of image i.

3 Proposed Approach for Image Aesthetic Assessment

Deep CNNs have demonstrated to be very effective in many image domains [22]. CNNs consist in a stack of layers involving linear, non-linear and spatial operators and are usually trained using back-propagation [23]. Most of the methods, due to the lack of very large datasets, take a CNN that is pre-trained for a different task (e.g. ImageNet competition [15]) and then use it as an initialization for a transfer learning process, known as fine-tuning [1,31]. In this work, we modify and fine-tune the Caffe network architecture [17] (inspired by the AlexNet architecture [21]) to model image aesthetic. We replace the last fully connected layer with a single-neuron layer in order to produce, given an input image, a predicted aesthetic score as a real number ranging between 1 and 10. We evaluate the effect of several design choices for pre-processing including the use of cropping and visual attention models for salient regions masking [6,8,27].

3.1 DeepIA: A CNN for Image Aesthetic Assessment

In this paper, we treat the aesthetic quality assessment as a regression problem, because it is closer to the human photo rating process [14]. Thus, the output of our CNN is a single-real value indicating the predicted aesthetic score.

Image aesthetic may depend on both the scenes and objects depicted. To this end we have chosen to fine-tune a pre-trained CNN as generic as possible to predict the aesthetic of an unseen image. The network used is the Hybrid-CNN [32], originally trained by merging the scene categories from Places dataset [32] and the object categories from ImageNet [15] for a total of 1,183 different classes. The proposed CNN is obtained by fine-tuning the Hybrid-CNN after replacing the last fully connected with a single-neuron layer and using the Euclidean loss layer instead of the Softmax loss layer:

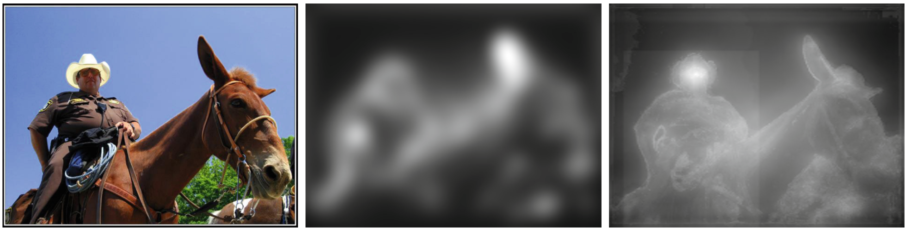
$$\min \sum_{i=1}^{n} \|y_i - \hat{y_i}\|_2^2$$

where y_i is the ground truth of image i, $\hat{y_i}$ is the predicted aesthetic score and n is the number of images. We call our final CNN *DeepIA*.

We fine-tune our CNN using Stochastic Gradient Descent (SGD) by chopping and retraining the last fully connected and by slightly updating the weights for the other layers. We use a batch size of 256, momentum set to 0.9, and a weight decay parameter of 0.0005. Then, we initialize the learning rate to a value of 0.001, and drop it every 20,000 iterations. We fine-tune for a total of 50,000 iterations. In all the experiments we use the Caffe open-source framework [17] for both the CNN training and prediction processes. During the training process, the original images are resized to 256 × 256 pixels without preserving the aspect ratio and then a random region of 227 × 227 pixels is extracted from the resized image. This approach increases the training set size in order to avoid overfitting. The mean-pixel values calculated across the training set images is the subtracted from the resized images.

At test time, we resize the original images to fixed dimensions and then we evaluate different design choices:

- we resize the images to 256 × 256 pixels and we use the 227 × 227 pixels central crop for image aesthetic prediction.
- we resize the images to 256 × 256 pixels and we average the prediction of multiple 227 × 227 pixels sub- regions (i.e. crops) of the input the image. We consider 10 crops corresponding to the four corners, the center region and their horizontal reflections.
- we resize the input image to 314 × 314 pixels and extract 10 crops with size 227 × 227 pixels.
- we weight the image pixels on the basis of their saliency using both a top-down and a bottom-up saliency models. To this end, the saliency map values have been scaled to fit the range $[0, 1]$. We use two different algorithms for estimating salient regions: the Itti et al. [16], which is built upon a biologically plausible computational model of focal bottom-up attention, and the Judd et al. [18], integrating a set of low, mid and high-level image features. In Fig. 2, we show the saliency maps predicted by the two considered algorithms.



Original image Itti saliency map Judd saliency map

Fig. 2. Saliency maps predicted using the Itti et al. [16] and the Judd et al. [18] algorithms on an image of the Aesthetic Visual Analysis (AVA) dataset [26].

Table 1. Performances of aesthetic quality assessment on the AVA dataset.

Method	Image size	#crops	MRSSE
DeepIA+Itti saliency map	256	1	0.5822
DeepIA+Itti saliency map	256	10	0.5766
DeepIA+Judd saliency map	256	1	0.4900
DeepIA+Judd saliency map	256	10	0.4829
DeepIA	314	10	0.4034
DeepIA	256	1	0.3866
DeepIA	256	10	**0.3727**

4 Experimental Results

The MRSSE obtained on the AVA dataset by our DeepIA for the different design choices outlined in Sect. 3, is reported in Table 1. The best results are obtained using the average prediction over 10 crops of size 227 × 227 extracted from the 256 × 256 image. The second best result is obtained by considering only the central 227 × 227 crop extracted from the image of size 256 × 256. The use of relatively smaller crops (i.e. 227 × 227 from 314 × 314 images) is not able to improve the results, giving a hint that image aesthetic is a global rather than a local attribute. The use of both top-down and bottom-up saliency models to filter out not-salient image content does not help to improve the accuracy of the prediction. In Table 2 we compare our best solution with different methods in the state of the art. As a reference, we also report the performance that could be achieved by always predicting an average score of 5. From the results it is possible to see that our DeepIA outperforms all the methods considered, with a reduction of MRSSE by almost 17 % with respect to the best method in the state of the art.

We report in Fig. 3 the five test images with the smallest MRSSE between ground truth and predicted aestetic scores. Figure 4 reports the ten test images

Table 2. Performance comparison of aesthetic quality assessment on the AVA dataset.

Method	MRSSE
Always predicting 5 as aesthetic score	0.5700
BOV-SIFT+rbfSVR ([25] adapted in [19])	0.5513
BOV-SIFT+linSVR ([25] adapted in [19])	0.5401
GIST+rbfSVR ([25] adapted in [19])	0.5307
GIST+linSVR ([25] adapted in [19])	0.5222
Aest-CNN [19]	0.4501
DeepIA	**0.3727**

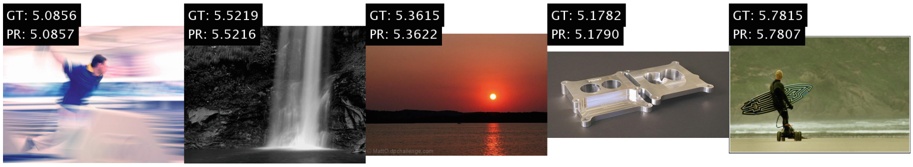


Fig. 3. Top 5 test images with the lowest error between ground truth (GT) and predicted (PR) aesthetic score.

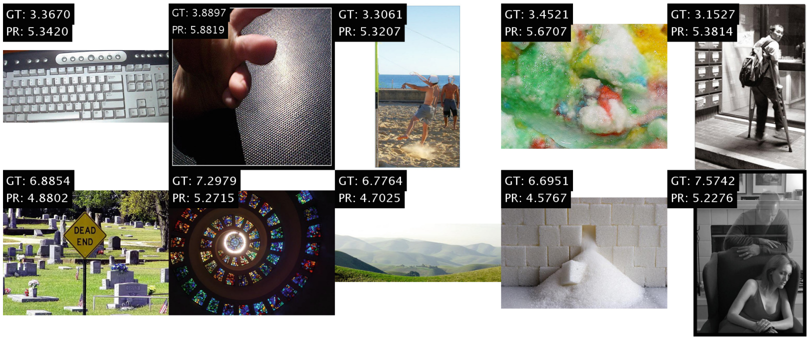


Fig. 4. Top 10 test images with the highest error between ground truth (GT) and predicted (PR) aesthetic score. Test images for which the predicted aesthetic score is overestimated (first row), and images whose predictions are underestimated (second row).

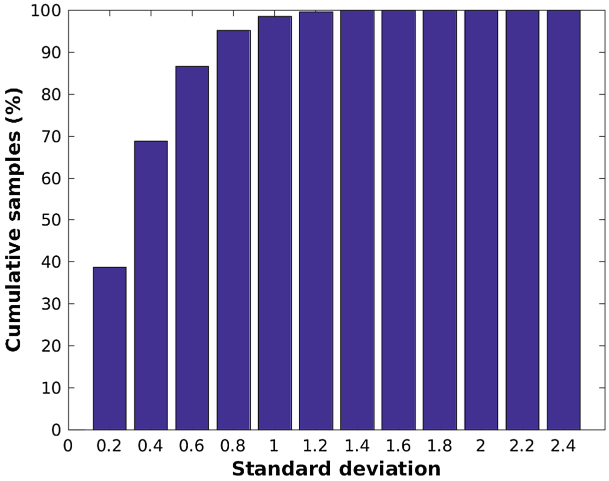


Fig. 5. Number of samples (%) with respect to the ratio between absolute estimation error and standard deviations (σ) of human scores.

with the largest errors: in the first row we report the top five overestimation errors, while in the second row the top five underestimation errors. The highest errors reported in Fig. 4 show that sometimes bad predictions reflect a lack of information consisting in the already defined *aesthetic gap* [14], defined as follows: *The aesthetics gap is the lack of coincidence between the information that one can extract from low-level visual data (i.e., pixels in digital images) and the interpretation of emotions that the visual data may arouse in a particular user in a given situation.*

Finally, since human aesthetic scores are noisy, we investigate how close is the score predicted by our DeepIA with the whole distribution of scores given by the humans to each image. To this end, for each image, we measure the ratio between our estimation error and the standard deviation of human scores. The cumulative histogram is reported in Fig. 5. From the plot it is possible to see that almost 99 % of the predictions have an error smaller or equal to a standard deviation value of 1.

5 Conclusions

We have investigated the use of a deep Convolutional Neural Network (CNN) to predict image aesthetics. Our approach consists in fine-tuning a canonical CNN architecture, originally trained to classify both objects and scenes, by casting the image aesthetic prediction as a regression problem. Experimental results on the canonical Aesthetic Visual Analysis (AVA) dataset show the robustness of the solution proposed, which outperforms the best solution in the state of the art by almost 17 % in terms of Mean Residual Sum of Squares Error (MRSSE). We also investigated whether image aesthetic is a global or local attribute, and the role played by bottom-up and top-down salient regions to the prediction of the global image aesthetic. Experimental results indicate that image aesthetic is a global attribute, and that the use of a saliency map to filter out not salient regions in the prediction stage does not help to achieve more accurate aesthetic score predictions. As a future work we plan to further investigate how can we exploit additional textual information, such as user comments or tagging, to predict image aesthetics [9–11].

References

1. Bengio, Y.: Deep learning of representations for unsupervised and transfer learning. Unsupervised Transf. Learn. Challenges Mach. Learn. **7**, 19 (2012)
2. Bhattacharya, S., Sukthankar, R., Shah, M.: A framework for photo-quality assessment and enhancement based on visual aesthetics. In: Proceedings of the 18th ACM International Conference on Multimedia, pp. 271–280. ACM (2010)
3. Bianco, S.: Reflectance spectra recovery from tristimulus values by adaptive estimation with metameric shape correction. JOSA A **27**(8), 1868–1877 (2010)
4. Bianco, S., Bruna, A.R., Naccari, F., Schettini, R.: Color correction pipeline optimization for digital cameras. J. Electron. Imaging **22**(2), 023014–023014 (2013)
5. Bianco, S., Ciocca, G., Marini, F., Schettini, R.: Image quality assessment by preprocessing and full reference model combination. In: IS&T/SPIE Electronic Imaging, p. 72420O. International Society for Optics and Photonics (2009)
6. Bianco, S., Ciocca, G., Napoletano, P., Schettini, R.: An interactive tool for manual, semi-automatic and automatic video annotation. Comput. Vis. Image Underst. **131**, 88–99 (2015)
7. Bianco, S., Schettini, R.: Adaptive color constancy using faces. IEEE Trans. Pattern Anal. Mach. Intell. **36**(8), 1505–1518 (2014)
8. Cagli, R.C., Coraggio, P., Napoletano, P., Boccignone, G.: What the draughtsman's hand tells the draughtsman's eye: a sensorimotor account of drawing. Int. J. Pattern Recogn. Artif. Intell. **22**(05), 1015–1029 (2008)
9. Colace, F., De Santo, M., Greco, L., Napoletano, P.: A query expansion method based on a weighted word pairs approach. In: Proceedings of the 3rd Italian Information Retrieval (IIR) vol. 964, pp. 17–28 (2013)
10. Colace, F., De Santo, M., Greco, L., Napoletano, P.: Weighted word pairs for query expansion. Inf. Process. Manag. **51**(1), 179–193 (2015)
11. Cusano, C., Napoletano, P., Schettini, R.: Evaluating color texture descriptors under large variations of controlled lighting conditions. JOSA A **33**(1), 17–30 (2016)
12. Datta, R., Joshi, D., Li, J., Wang, J.Z.: Studying aesthetics in photographic images using a computational approach. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006. LNCS, vol. 3953, pp. 288–301. Springer, Heidelberg (2006). doi:10.1007/11744078_23
13. Datta, R., Li, J., Wang, J.Z.: Learning the consensus on visual quality for next-generation image management. In: Proceedings of the 15th International Conference on Multimedia, pp. 533–536. ACM (2007)
14. Datta, R., Li, J., Wang, J.Z.: Algorithmic inferencing of aesthetics and emotion in natural images: an exposition. In: 15th IEEE International Conference on Image Processing, ICIP 2008, pp. 105–108. IEEE (2008)

15. Deng, J., Berg, A., Satheesh, S., Su, H., Khosla, A., Fei-Fei, L.: Imagenet large Scale Visual Recognition Competition (ILSVRC 2012) (2012)
16. Itti, L., Koch, C.: Computational modelling of visual attention. Nat. Rev. Neurosci. **2**(3), 194–203 (2001)
17. Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., Darrell, T.: Caffe: convolutional architecture for fast feature embedding. In: Proceedings of the ACM International Conference on Multimedia, pp. 675–678. ACM (2014)
18. Judd, T., Ehinger, K., Durand, F., Torralba, A.: Learning to predict where humans look. In: IEEE International Conference on Computer Vision (ICCV) (2009)
19. Kao, Y., Wang, C., Huang, K.: Visual aesthetic quality assessment with a regression model. In: 2015 IEEE International Conference on Image Processing (ICIP), pp. 1583–1587. IEEE (2015)
20. Ke, Y., Tang, X., Jing, F.: The design of high-level features for photo quality assessment. In: Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on. vol. 1, pp. 419–426. IEEE (2006)
21. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Proceedings of Advances in Neural Information Processing Systems, pp. 1097–1105 (2012)
22. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. Nature **521**(7553), 436–444 (2015)
23. LeCun, Y., Bottou, L., Orr, G.B., Müller, K.-R.: Efficient backprop. In: Orr, G.B., Müller, K.-R. (eds.) Neural Networks: Tricks of the Trade. LNCS, vol. 1524, pp. 9–50. Springer, Heidelberg (1998). doi:10.1007/3-540-49430-8_2
24. Lu, X., Lin, Z., Jin, H., Yang, J., Wang, J.Z.: Rapid: rating pictorial aesthetics using deep learning. In: Proceedings of the ACM International Conference on Multimedia, pp. 457–466. ACM (2014)
25. Marchesotti, L., Perronnin, F., Larlus, D., Csurka, G.: Assessing the aesthetic quality of photographs using generic image descriptors. In: 2011 IEEE International Conference on Computer Vision (ICCV), pp. 1784–1791. IEEE (2011)
26. Murray, N., Marchesotti, L., Perronnin, F.: Ava: a large-scale database for aesthetic visual analysis. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2408–2415. IEEE (2012)
27. Napoletano, P., Boccignone, G., Tisato, F.: Attentive monitoring of multiple video streams driven by a Bayesian foraging strategy. IEEE Trans. Image Process. **24**(11), 3266–3281 (2015)
28. Nishiyama, M., Okabe, T., Sato, I., Sato, Y.: Aesthetic quality classification of photographs based on color harmony. In: 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 33–40. IEEE (2011)
29. Simond, F., Arvanitopoulos Darginis, N., Süsstrunk, S.: Image aesthetics depends on context. In: International Conference on Image Processing, vol. 1 (2015)
30. Wu, O., Hu, W., Gao, J.: Learning to predict the perceived visual quality of photos. In: 2011 IEEE International Conference on Computer Vision (ICCV), pp. 225–232. IEEE (2011)
31. Yosinski, J., Clune, J., Bengio, Y., Lipson, H.: How transferable are features in deep neural networks? In: Proceedings of Advances in Neural Information Processing Systems, pp. 3320–3328 (2014)
32. Zhou, B., Lapedriza, A., Xiao, J., Torralba, A., Oliva, A.: Learning deep features for scene recognition using places database. In: Proceedings of Advances in Neural Information Processing Systems, pp. 487–495 (2014)