https://doi.org/10.2352/issn.2694-118X.2021.LIM-11

Modeling image aesthetics through aesthetics-related attributes

Marco Leonardi[1]***, Paolo Napoletano***[1]***, Alessandro Rozza***[2]***, Raimondo Schettini***[1]
[1] ***Department of Informatics, Systems and Communication, University of Milano - Bicocca, viale Sarca, 336 Milano, Italy***
{m.leonardi6@campus.unimib.it, paolo.napoletano@unimib.it, raimondo.schettini@unimib.it}
[2] ***lastminute.com, Chiasso, Switzerland***
{alessandro.rozza@lastminute.com}

Abstract

Automatic assessment of image aesthetics is a challenging task for the computer vision community that has a wide range of applications. The most promising state-of-the-art approaches are based on deep learning methods that jointly predict aesthetics-related attributes and aesthetics score. In this article, we propose a method that learns the aesthetics score on the basis of the prediction of aesthetics-related attributes. To this end, we extract a multi-level spatially pooled (MLSP) features set from a pretrained ImageNet network and then these features are used to train a Multi Layer Perceptron (MLP) to predict image aesthetics-related attributes. A Support Vector Regression machine (SVR) is finally used to estimate the image aesthetics score starting from the aesthetics-related attributes. Experimental results on the "Aesthetics with Attributes Database" (AADB) demonstrate the effectiveness of our approach that outperforms the state of the art of about 5.5% in terms of Spearman's Rank-order Correlation Coefficient (SROCC).

Introduction

Easy access to a camera and the consequent nearly effortless task of taking photos has made shooting a picture similar to natural action. We took photos in every moment of our days, for example to remind something or to capture events. Despite cameras are becoming increasingly sophisticated and smart, it is not rare to shoot images that are not pleasing in terms of aesthetics. Given the exponential growth of the number of images taken and stored, selecting pleasing images has become a tedious and boring task. Being able to automatically distinguish good aesthetics images from bad ones can help various types of applications, such as automatic photo album creation, media storage techniques and so on.

Automatic aesthetics assessment of images is usually treated as a classification or regression task based on ratings provided by human annotators [2]. In recent years, many research efforts have been made and various approaches have been proposed.

Datta et al. [4] carefully selected 56 hand-crafted visual features based on standard photography and visual design rules to discriminate between aesthetically pleasing and displeasing images.

Dhar et al. [5] proposed a method for predicting image interestingness by exploiting high-level describable image attributes divided into three categories: compositional (image layout or configuration), content (objects or scene types depicted) and sky-illumination (natural lighting conditions).

With the availability of more labeled data the trend has been moved from methods based on hand crafted features to deep learning methods. Recent works have both been focused on sophisticated training loss [9, 15, 11, 3] and more powerful features [10, 13, 7].

Given the importance of photography rules and aesthetics attributes, Kong et al. have collected the "Aesthetics with Attributes Database", or AADB [9]. This collection includes images that have been rated by several human observers in terms of both global aesthetics and visual aesthetics-related attributes. They proposed a Convolutional Neural Network (CNN) architecture to jointly predict semantic photo content, global aesthetics and aesthetics-related attributes.

Malu et al. [10] proposed a multi-task network based on features extracted from a ResNet-50 [6]. To better encapsulate the information from the ResNet-50 they extracted 16 rectified convolution maps from the ReLU output of the 16 residual blocks of the ResNet-50. The proposed architecture is used to predict eight aesthetics attributes alongside the global aesthetics score.

In [15] authors explored the relationship between aesthetics score and aesthetics attributes introducing the PI-DCNN: a ResNet optimized over three different loss functions: regression, ranking and a privileged information loss which rely on some domain knowledge and additional information between attributes and aesthetics.

In [11] Pan et al. exploited the feature extracted from a ResNet-50, and proposed a multi-task neural network to predict both the aesthetics score and the attributes. Different from the other works, they proposed a framework in which the network is trained in an adversarial manner: the discriminator distinguishes the predictions given by the proposed multi-task network from the real labels.

Chen [3] proposed a different training framework based on data covariance learning to improve performance of baseline architectures: the method proved that training an architecture modeling the data uncertainty is more effective than training with the mean squared error.

In [13] authors combined low-resolution, semantically strong features with the high-resolution, semantically weak features from the EfficientNets B4 [16] to predict simultaneously 8 aesthetics tags and the global aesthetics score.

Taking inspiration from the work by Dhar et al. [5] we propose a method based on a MLP that, from features extracted by an ImageNet pretrained CNN, predicts eleven aesthetics-related attributes. Then we train a SVR [12] to predict image aesthetics on the basis of the aesthetics-related attributes computed in the previous stage.

The main contributions of the paper are the following:

- We propose an aesthetics quality estimation method that relies on the prediction of aesthetics-related attributes.
- We show how predicting the aesthetics of an image is more accurate through aesthetics-related attributes rather than modeling only the aesthetics or jointly the aesthetics-related attributes and the global aesthetics.
- We demonstrate on the "Aesthetics with Attributes Database" (AADB) the effectiveness of our approach that

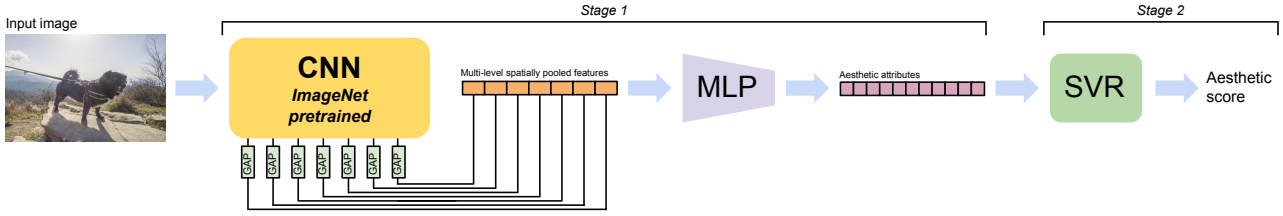

Figure 1. *The proposed method. Given an input image, a multi-level spatially pooled features set is extracted from a Convolutional Neural Network pretrained on ImageNet. This feature set is then fed to a Multi Layer Perceptron to predict image aesthetics-related attributes. Finally a Support Vector Regression machine is used to estimate the image aesthetics score starting from the aesthetics-related attributes.*

outperforms the state of the art of about 5.5% in terms of SROCC.

Proposed method

Figure 1 shows the pipeline of the proposed method. The first stage is a MLP that predicts the eleven aesthetics-related attributes of an input image on the basis of MLSP features extracted from an ImageNet pretrained network. The second stage is based on a SVR that takes the attributes predicted in the first stage as input and it estimates the global image aesthetics score. Our design choices have been driven by the following considerations.

MLSP features demonstrated to be very effective for image aesthetics prediction [7]. The main idea was to create a feature vector that encodes information from multiple levels of a CNN. This is achieved by concatenating Global Average Pooled (GAP) activations from fixed blocks of a given CNN trained for image classification.

The use of aesthetics-related attributes for the estimation of image aesthetics has been investigated in previous works [4][5]. Datta et al. [4] proposed several visual attributes and studied the correlation between those properties and the aesthetics score. Some examples of attributes are: light exposure, colorfulness, depth of field and rule of thirds. They also proposed a Support Vector Classification machine that uses 15 visual attributes for the classification of high and low rated photographs in terms of interestingness.

Dhar et al. [5] proposed an aesthetics estimation method that from low-level features (e.g. Color spatial distribution map, Spatial Pyramid of shape features etc.) predicts aesthetics-related attributes such as presence of a salient objects, opposing colors, presence of people and clear skies.

The proposed method takes inspiration from the paper by Hosu et al. [7] for what concerns the use of MLSP features and from the paper by Dhar et al. [5] for what concerns the use of aesthetics-related attributes to predict image aesthetics.

To assess the effectiveness of the proposed method we experimented different CNN architectures (ResNet-50, EfficientNets B4, Inception-v3, InceptionResNet-v2) pretrained on ImageNet [1, 14] from which we extract the MLSP features. We also compare the proposed method with two variants: a single-task MLP trained to predict solely the aesthetics score and a multi-task MLP trained jointly over the eleven aesthetics attributes and the aesthetics score.

Experiments

Dataset

We train and test the performance of the proposed method on the AADB [9], a database composed of 10,000 images. Each image of the database has the aesthetics rating and the assessment of eleven aesthetics-related attributes provided by five different subjects. The images are divided into training (8,500), validation (500) and testing sets (1,000), and they were collected from the Flickr website and curated manually. With the help of professional photographers the authors has selected eleven attributes that are closely related to image aesthetics judgements: *interesting_content, object_emphasis, good_lighting, color_harmony, vivid_color, shallow_depth_of_field, motion_blur, rule_of_thirds, balancing_element, repetition,* and *symmetry*.
To gather the data, authors ask qualified Amazon Mechanical Turk (AMT) workers to rate ”positive” if an attribute conveyed by the image can enhance the image aesthetics level, or ”negative” if the attribute degrades image aesthetics. According to the authors, the default value was ”null”, meaning that the attribute does not affect image aesthetics. The collected labels were then translated into real values encoding ”positive” as 1, ”negative” as -1 and ”null” as 0. For each image, the attribute score is the average over all the users judgements. Figure 2 reports, for each of the eleven aesthetics attributes, the distribution of the mean values to underlying imbalances of the ground truths: attributes like motion blur, symmetry or repetition contains many ”null” values.

For the aesthetics score, AMT workers were allowed to express their judgement on a scale from 1 to 5. For each image, the aesthetics score is the average over all the users judgements. The aesthetics score was further normalized in order to fit a range of [0, 1]. Figure 3 depicts the distribution of the aesthetics scores.

In order to highlight the intrinsic power of the aesthetics-related attributes, table 1 reports the SROCC between the values of each attribute and the overall aesthetics score. Color harmony, Content, Light, Object, Vivid color correlate more than others with the image aesthetics with an SROCC higher than 0.6. To better highlight the prediction power of the aesthetics-related

Table 1. *Correlation between aesthetics properties and the aesthetics scores.*

Property	srocc
Balacing elements	0.3830
Color harmony	0.6227
Content	0.7279
Depth of field	0.5098
Light	0.6221
Motion blur	0.2204
Object	0.6415
Repetition	0.1023
Rule of thirds	0.3892
Symmetry	0.1063
Vivid color	0.6161
All of the above (SVR)	0.9374

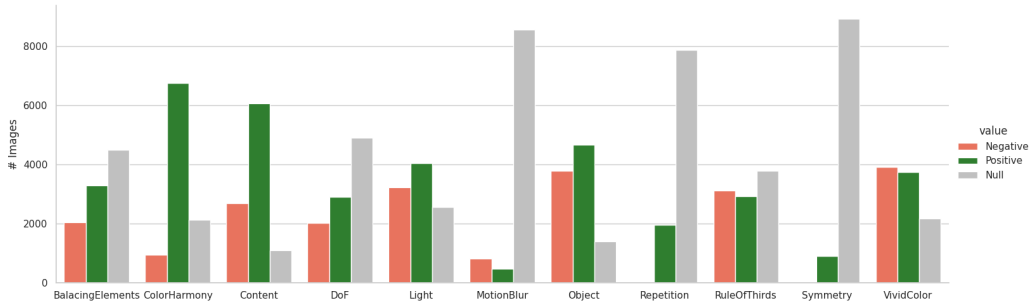

Figure 2. *Value distribution for each of the eleven aesthetics attributes. Null values are those which have a mean score of 0. Positive values are those images with on average more positive labels than negative, vice-versa for the Negative.*

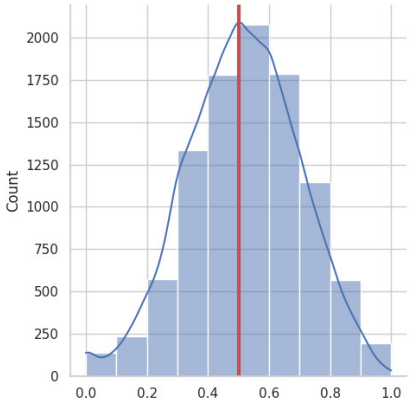

Figure 3. *Score distribution of the AADB database. The red line indicates the 0.5 value.*

attributes, we trained a SVR for image aesthetics score estimation based on the human ground truth. Overall all the attributes achieve an SROCC value higher than 0.9. We also study the correlation between the SROCC of each of these attributes and the percentage of null values in the ground truth and we found a SROCC of -0.9182. This suggests that the poor correlation of some attributes with the aesthetics score is more likely due to the null values rather than the expressiveness of the attribute itself.

Experimental Setup

The models have been developed in PyTorch and trained on an Nvidia GTX 1070 GPU. The MLP is composed of three stacked linear layers with ReLu activations. The training has been done with a batch size of 16 and for a maximum of 100 epochs adopting the early stop technique with patience of 6 epochs over the average of the SROCC with respect to the validation set. As optimizer was used Adam [8] with a learning rate of 1.5e-05. The first two layers of the MLP were trained with a dropout probability of 0.5.

For the feature extraction part, as in [7], we decide to extract and store from images having a different resolution, fixed sized narrow MLSP features of dimension $(1 \times 1 \times b)$ where b is the number of kernels from which features are computed.
To extract these featuress we adopt the Global Average Pooling layer (GAP) over selected activation blocks output: for the EfficientNets B4, Inception-v3 and InceptionResNet-v2 we decide to select the same block as done by the original works ([13, 7]). Note that for the EfficientNets B4 authors extract features from given blocks in a different way, while for the ResNet-50 we selected all the five convolutional blocks. There are 6 blocks in EfficientNets B4 (3,056 kernels), 11 blocks in Inception-v3 (10,048 kernels), 43 in InceptionResNet-v2 (49,248 kernels) and 5 blocks in ResNet-50 (3,904 kernels).

The most commonly used metric to evaluate the performance of automatic image aesthetics assessment is the SROCC. It is used to compare the scores predicted by the models and the subjective opinion scores provided by the dataset and it evaluates the monotonic relationship between two continuous or ordinal variables. The SROCC operates on the rank of the data points ignoring the relative distances between them. It varies in the interval $[+1, -1]$ and for n samples it is defined as follows:

$$SROCC = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n\left(n^2-1\right)}, \tag{1}$$

where $d_i = (rank(x_i) - rank(y_i))$ is the difference between the two ranks of each sample.

Results

As stated before, to better understand the contribution given by the proposed method, we compare our pipeline against two different alternatives which are common in the state of the art: a single-task Neural Network [2, 7] trained to predict the aesthetics score directly from the image and a multi-task Neural Network [9, 11] trained to predict at the same time the eleven aesthetics attributes and the aesthetics score. Table 2 reports the mean of SROCC achieved on 10 repetitions of the experiments. The table reports, for each of CNN archictecture experimented, the proposed approach and the two variants mentioned above. The table also reports results taken from the most recent state-of-the-art approaches which jointly predict aesthetics-related attributes and image aesthetics.

Overall, independently from the architectural choice, the proposed method outperforms the state of the art of about 5.5% in terms of SROCC. The improvement of the proposed method with respect to the other two alternatives is of about 2.4%. The enhancement given by the proposed method confirms that predicting image aesthetics through the estimation of aesthetics-related attributes is more effective than a multi task CNN. Moreover, it is more effective predicting the aesthetics score on the basis of aesthetics-related attributes rather than predicting the aesthetics score along with attributes or solely the aesthetics score. Figure 4 shows an example of the predicted images attributes with respect to the ground truth values.

Table 2. *Spearman's Rank-order Correlation Coefficient (SROCC) between the predicted image aesthetics quality and the ground truth. (* srocc are taken from the authors pubblication)*

Name (base architecture)	**Architecture type**	SROCC
Kong et al. (Alexnet) [9]	Multi-Task CNN	0.6782*
Malu et al. (Resnet-50) [10]	Multi-Task CNN	0.6890*
PI-DCNN (Resnet-50) [15]	Multi-Task CNN	0.7051*
Chen (Resnet-50) [3]	Multi-Task CNN	0.7080*
Pan et al.(ResNet-50) [11]	Multi-Task CNN	0.7041*
Reddy et al. (Efficientnet_b4) [13]	Multi-Task CNN	0.7059*
EfficientNets_B4	Single-Task MLP	0.7281 $\pm$ 0.0138
	Multi-Task MLP	0.7219 $\pm$ 0.0039
	SVR over MLP's tag prediction	**0.7454** $\pm$ 0.0033
ResNet-50	Single-Task MLP	0.7083 $\pm$ 0.0067
	Multi-Task MLP	0.7194 $\pm$ 0.0060
	SVR over MLP's tag prediction	**0.7384** $\pm$ 0.0023
Inception-v3	Single-Task MLP	0.7242 $\pm$ 0.0065
	Multi-Task MLP	0.7197 $\pm$ 0.0029
	SVR over MLP's tag prediction	**0.7354** $\pm$ 0.0025
InceptionResNet-v2	Single-Task MLP	0.7316 $\pm$ 0.0029
	Multi-Task MLP	0.7308 $\pm$ 0.0036
	SVR over MLP's tag prediction	**0.7429** $\pm$ 0.0015

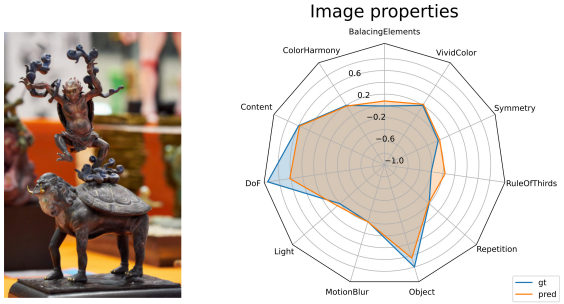

Figure 4. *Example of predicted aesthetics-related attributes (orange line) with respect to the ground truth (blue line).*

Conclusions

In this work, we have introduced a method for the automatic image aesthetics assessment based on the prediction of eleven attributes that are closely related to image aesthetics judgements. The designed model exploit MLSP features extracted from a CNN pretrained on ImageNet to predict these eleven attributes with a MLP. Then, a SVR is trained to infer the aesthetics score of the input images over the prediction of the aforementioned MLP. Experimental results with four different architectures demonstrated the effectiveness of the proposed approach: predicting the image aesthetics through related attributes lead to an improvement of 5.5% in terms of SROCC with respect to the state of the art. These promising results encourage us to continue working in this direction with a major focus on the improvement of the attributes prediction.

References

[1] Simone Bianco, Remi Cadene, Luigi Celona, and Paolo Napoletano. Benchmark analysis of representative deep neural network architectures. *IEEE Access*, 6:64270–64277, 2018.

[2] Simone Bianco, Luigi Celona, Paolo Napoletano, and Raimondo Schettini. Predicting image aesthetics with deep learning. In *International Conference on advanced concepts for intelligent vision systems*, pages 117–125. Springer, 2016.

[3] Zhihong Chen. Data covariance learning in aesthetic attributes assessment. *Journal of Applied Mathematics and Physics*, 8(12):2869–2879, 2020.

[4] Ritendra Datta, Dhiraj Joshi, Jia Li, and James Z Wang. Studying aesthetics in photographic images using a computational approach. In *European conference on computer vision*, pages 288–301. Springer, 2006.

[5] Sagnik Dhar, Vicente Ordonez, and Tamara L Berg. High level describable attributes for predicting aesthetics and interestingness. In *CVPR 2011*, pages 1657–1664. IEEE, 2011.

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[7] Vlad Hosu, Bastian Goldlucke, and Dietmar Saupe. Effective aesthetics prediction with multi-level spatially pooled features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9375–9383, 2019.

[8] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[9] Shu Kong, Xiaohui Shen, Zhe Lin, Radomir Mech, and Charless Fowlkes. Photo aesthetics ranking network with attributes and content adaptation. In *European Conference on Computer Vision*, pages 662–679. Springer, 2016.

[10] Gautam Malu, Raju S Bapi, and Bipin Indurkhya. Learning photography aesthetics with deep cnns. *arXiv preprint arXiv:1707.03981*, 2017.

[11] Bowen Pan, Shangfei Wang, and Qisheng Jiang. Image aesthetic assessment assisted by attributes through adversarial learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 679–686, 2019.

[12] John Platt et al. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3):61–74, 1999.

[13] Gajjala Viswanatha Reddy, Snehasis Mukherjee, and Mainak Thakur. Measuring photography aesthetics with deep cnns. *IET Image Processing*, 14(8):1561–1570, 2020.

[14] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya

Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

[15] Yangyang Shu, Qian Li, Shaowu Liu, and Guandong Xu. Learning with privileged information for photo aesthetic assessment. *Neurocomputing*, 404:304–316, 2020.

[16] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*, pages 6105–6114. PMLR, 2019.

Author Biography

Marco Leonardi received the B.Sc. degree and the M.Sc. degree in computer science from the Department of Informatics, Systems and Communication (DISCo), University of Milano-Bicocca, Italy, respectively in 2016 and 2018. He is currently a Ph.D. student at DISCo, University of Milano-Bicocca. His work has focused on the development of machine learning systems capable to predict perceptual properties of images.

Paolo Napoletano is associate professor at the University of Milano Bicocca (Italy). In 2007, he received a Doctor of Philosophy degree (PhD) in Information Engineering from the University of Salerno (Italy). In 2003, he received a Master's degree in Telecommunications Engineering from the University of Naples Federico II. His current research interests focus on signal, image and video analysis and understanding, multimedia information processing and management and machine learning for multi-modal data classification and understanding.

Alessandro Rozza is the Chief Scientist of lastminute.com group. In 2011, he received a Doctor of Philosophy degree (PhD) in Computer Science from the Department of Scienze dell'Informazione, Universitá degli Studi di Milano. From 2012 to 2014 he was Assistant Professor at Universitá degli Studi di Napoli-Parthenope. From 2015 to 2017, he was head of research at Waynaut. His research interests include machine learning and its applications.

Raimondo Schettini is a professor at the University of Milano Bicocca (Italy). He is head of the Imaging and Vision Lab. He has been associated with the Italian National Research Council since 1987, where he led the color imaging lab from 1990 to 2002. He has been a team leader in several research projects and published more than 300 refereed papers and six patents about color reproduction, and image processing, analysis, and classification. He is a fellow of the International Association of Pattern Recognition for his contributions to pattern recognition research and color image analysis.