IMAGE AESTHETIC PREDICTORS BASED ON WEIGHTED CNNS

Bin Jin[1]*, Maria V. Ortiz Segovia*[2] *and Sabine Süsstrunk*[1]

[1] EPFL, Lausanne, Switzerland; [2] Océ Print Logic Technologies, Creteil, France

ABSTRACT

Convolutional Neural Networks (CNNs) have been widely adopted for many imaging applications. For image aesthetics prediction, state-of-the-art algorithms train CNNs on a recently-published large-scale dataset, AVA. However, the distribution of the aesthetic scores on this dataset is extremely unbalanced, which limits the prediction capability of existing methods. We overcome such limitation by using weighted CNNs. We train a regression model that improves the prediction accuracy of the aesthetic scores over state-of-the-art algorithms. In addition, we propose a novel histogram prediction model that not only predicts the aesthetic score, but also estimates the difficulty of performing aesthetics assessment for an input image. We further show an image enhancement application where we obtain an aesthetically pleasing crop of an input image using our regression model.

***Index Terms*—** Aesthetics, sample weights, CNN

1. INTRODUCTION

Automatically assessing image aesthetics is useful for many applications. To name a few, aesthetics can be adopted as one of the ranking criteria for image retrieval systems or one of the objectives for image enhancement systems. Moreover, users can manage their images collections based on aesthetics. Hence, various algorithms [1–10] have been proposed in the recent years to perform image aesthetics assessment.

In this paper, we train convolutional neural networks (CNNs) for aesthetics assessment. Our model is trained on the recently-published AVA dataset [6], which contains more than 250,000 images collected from a digital photography challenge[1]. Each image has around 200 user ratings about its aesthetic quality, with each rating being an integer between 1 and 10 (1 implies the lowest quality and 10 means the highest quality). We show two sample images and their corresponding histograms of user ratings in Fig. 1. The average of user ratings is taken as the aesthetic score for each image.

The distribution of the aesthetic scores in the AVA dataset is extremely unbalanced, as shown in Fig. 2 (a), which introduces bias into all the previous CNN models that are trained on this dataset [8, 10]. To reduce such bias, we propose to use sample weights during training. The sample weights are first computed according to the occurrences of the aesthetic scores and later incorporated into a weighted loss function for training. This loss function is balanced over images with different aesthetic scores, thus enabling the trained CNNs to work for images of different aesthetic quality. Using sample weights, we train a regression model which can achieve a larger prediction range and better accuracy than previous methods.

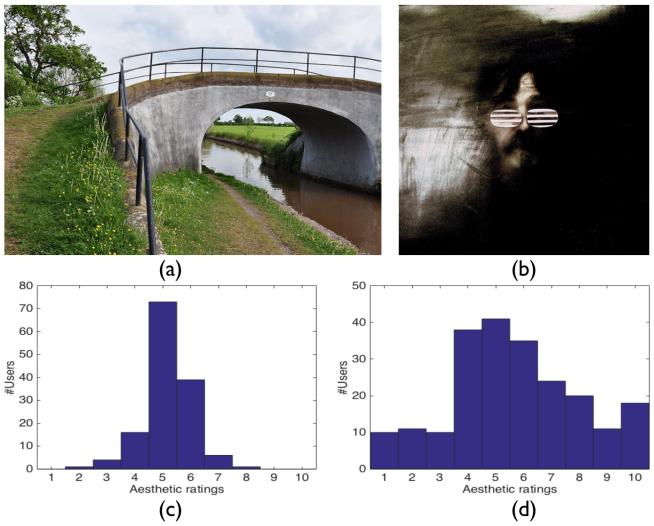

Fig. 1. (a) and (b) are two images of the AVA dataset, (c) and (d) are their corresponding histograms of user ratings.

All previous methods [6, 8–10] directly use the aesthetic scores for training while discarding the information of user ratings. As a matter of fact, the distribution of the ratings reveals not only the aesthetic score, but also how much users agree with each other when aesthetically assessing the image. Therefore, the distribution is an indicator of the difficulty of performing aesthetics assessment for a given image. Using difficulty estimation has been shown to give reliable aesthetic scores for images with user labels [11, 12]. For instance, the two histograms in Fig. 1 clearly indicate that Fig. 1(a) is agreed by the majority to be of average quality, thus being easy to judge, while Fig. 1(b) is less conclusive and more difficult to assess. To estimate the level of difficulty, we train a histogram prediction CNN model that can predict the normalized histogram of user ratings. Our experiments show that this model produces accurate aesthetic scores and reliable estimations of user ratings variety.

[1] http://www.dpchallenge.com/

To summarize, our contributions are: 1) the usage of sample weights during training, which helps to overcome the bias in the training set of the AVA dataset and extend the prediction capability of the trained CNN models; 2) a trained regression CNN model that achieves a larger prediction range and better accuracy than the state-of-the-art methods; 3) a trained histogram prediction model that reliably estimates the aesthetic scores as well as the difficulty of aesthetics assessment; 4) an image enhancement application that outputs an aesthetically pleasing crop of an input image by using the results of the trained CNN model.

2. STATE-OF-THE-ART

State-of-the-art aesthetics prediction methods can be characterized into three categories. The first category [1–4, 9] links aesthetics with handcrafted low-level image features, e.g., color distribution, edge distribution, hue channel, etc. Another category [5–7] uses generic image features such as SIFT [13] or Fisher Vector [14, 15], which have been shown to outperform the handcrafted low-level features. However, as aesthetics is a complex, subjective, and high-level concept, these methods often result in inferior performance.

Since CNNs have demonstrated their effectiveness in many imaging and computer vision tasks [16–19], the latest methods [8, 10] adopted CNNs for predicting aesthetics. For instance, Lu et al. [8] formulate the aesthetics assessment as a classification problem. They split the AVA dataset into two classes (high quality and low quality) and train a CNN model to predict the class labels. Such a classification model can only predict binary class labels while discarding the differences within a class. The applications of their model are thus limited: their model are not suitable for an image retrieval system or an image enhancement application. Kao et al. [10] propose a CNN regression model which provides continuous aesthetic scores. However, they ignore the unbalanced distribution of the aesthetic scores in the AVA dataset, as shown in Fig. 2(a). Their regression model is thus biased towards the scores between 4.5 to 6 and has limited prediction range. Consequently, it is less suitable for real world applications in which we encounter images of a variety of aesthetic quality.

3. METHODS

In this section we first explain how we derive the sample weights for the training set, followed by the two CNN models that we propose to predict aesthetics. We explain the regression model in Sec. 3.2 and the histogram prediction model in Sec. 3.3.

3.1. Sample weights

Assume the histogram of the aesthetic scores in the training set is $\{b_i, i = 1, 2...B\}$. B is the number of bins that evenly cover the range of the aesthetic scores. We set B to 90 for the aesthetic scores' range of 1 to 10. b_i is the occurrence number of the ith bin, namely the number of images assigned the aesthetic scores within the ith bin's range. The sample weight w_i for the ith bin is computed as:

$$b'_i = \frac{b_i}{\sum_{i=1}^{B} b_i}; \qquad w_i = \frac{1}{b'_i} \tag{1}$$

Images within the same bin share the same sample weights. The sample weight is inversely proportional to the normalized occurrence number. Consequently, images with rare scores are assigned larger sample weights than images with more frequent scores. Note that sample weights are only computed for the training set and only used during training, not during testing.

3.2. Regression model

The architecture of our regression CNN model is the same as the VGG16 network [19], which has shown superior performance on image classification. The last layer of the network is modified to have only one output neuron for predicting a single aesthetic score. We remove the last *softmax* activation function since the output is only one value.

The training of this model is done by minimizing the following Weighted Mean Squared Error (WMSE) loss function:

$$WMSE = \frac{1}{\sum_{i=1}^{N} w_i} \sum_{i=1}^{N} w_i \cdot (y_i - \hat{y_i})^2 \tag{2}$$

Here w_i is the sample weight computed according to Eqn. 1. y_i is the predicted aesthetic score and $\hat{y_i}$ is the groundtruth aesthetic score. N is the number of images in the training set.

Note that images with large sample weights do not occur very often, thus the overall contribution to the loss function is balanced across images with varying aesthetic scores. In this way, the sample weights help to reduce the bias in the training set.

3.3. Histogram prediction model

The histogram prediction model aims at predicting the normalized histogram of user ratings for an input image. The output of the model is a vector with 10 bins as user ratings are integers between 1 and 10. We adjust the last layer of VGG16 network [19] to have 10 output neurons. The loss function for training is the Weighted Mean χ^2 Error (WMCE):

$$WMCE = \frac{1}{\sum_{i=1}^{N} w_i} \sum_{i=1}^{N} w_i \cdot \chi^2(\mathbf{h_i}, \hat{\mathbf{h_i}}) \tag{3}$$

where w_i is the sample weight for image i. $\mathbf{h_i}$ is the output histogram from the network and $\hat{\mathbf{h_i}}$ is the groundtruth normalized histogram. χ^2 represents the chi-square distance.

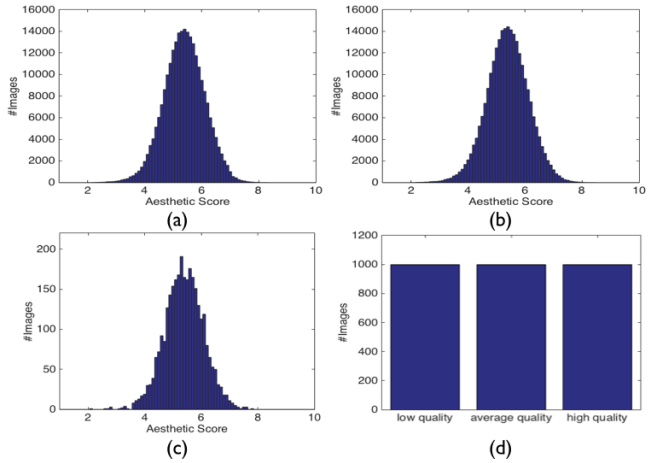

Fig. 2. The distribution of the average aesthetic scores for (a) the whole AVA dataset (b) the training set, (c) the *RS-test*, (d) the *ED-test*, which has an equal number of images from three categories: low, average, and high quality.

Based on the output histogram, two values are derived: the aesthetic score, which is the average of user ratings, and the standard deviation (std) of user ratings. This std value represents the difficulty of aesthetics assessment. A small std means consensus and simplicity of aesthetics assessment as user ratings concentrate around the average score, while a large std represents difficulty. By comparing the std values, we can evaluate whether one image is more difficult to aesthetically assess than another. For example, Fig. 1(c) has the std value of 0.8775 and Fig. 1(d) is 2.3228. The image in Fig. 1(b) is clearly more difficult to assess.

4. EXPERIMENTS

4.1. Training and test sets

We split the AVA dataset into three parts: training set, test set 1 (*RS-test*) and test set 2 (*ED-test*). The distributions of the aesthetic scores in these three sets are shown in Fig. 2(b)-(d). *RS-test* contains 3000 *Random Sampled* images, which is similar to the test set in [10] that contains 5000 random sampled images. *ED-test* is built to have 3000 images *Evenly Distributed* among three categories: the low quality images (aesthetic score < 4), the average quality images ($4 \leq$ aesthetic score ≤ 7) and the high quality images (aesthetic score > 7), as shown in Fig. 2 (d). The other 249530 images of the AVA dataset are used for the training set.

4.2. Processing

Since many aspects of the images can affect the aesthetics, such as composition and saturation, it is not recommended to apply data augmentation methods. We directly resize the whole image to 224×224, which is then fed into the network. Although this operation may change the aspect ratio of the image, we have experimentally found that it produces the best results as opposed to cropping the images, which is corroborated in [8]. The CNNs are initialized with the pre-trained ImageNet weights [16] and then fine-tuned for 20 epochs on the whole training set. Learning rate is set to 0.00001, and divided by 10 when the training loss stops decreasing. It takes around 4 days for each model to finish 20 epochs on a single NVIDIA TITAN X GPU.

4.3. Regression model results

For the regression task, we use the Mean Squared Error (MSE) as the evaluation metric, which is the same as in [10]:

$$MSE = \frac{1}{M}\sum_{i=1}^{M}(y_i - \hat{y}_i)^2 \quad (4)$$

Here, y_i and $\hat{y}_i$ are the predicted and the groundtruth aesthetic scores, respectively, for the ith image. M is the number of images in the test set. Note that sample weights are not applied in the evaluation metric.

Two regression CNN models with the same architecture are trained: a *Regression* model with *Sample Weights* (*SWR*) and a *Regression* model with *No Sample Weights* (*NSWR*). The performance is shown in Table 1.

Table 1. MSE of different models, results of the top 5 methods are taken from [10].

	RS-test	*ED-test*
GIST linear-SVR	0.5222	NA
GIST rbf-SVR	0.5307	NA
BoVW SIFT linear-SVR	0.5401	NA
BoVW SIFT rbf-SVR	0.5513	NA
Kao et al. [10]	0.4510	NA
No SW regression (*NSWR*)	**0.3373**	1.3951
SW regression (*SWR*)	0.4847	**0.9754**

The top four methods in Table 1 combine the generic image descriptors, GIST [20], SIFT [13] and Bag-of-Visual-Words (BoVW) [21], together with the Support Vector Regression (SVR) with linear or rbf kernel [22]. Refer to [4, 10] for details of these methods. Note that none of the previous methods was evaluated on a test set with balanced distribution, namely the *ED-test* we created.

Our regression model without sample weights (*NSWR*) outperforms all the state-of-the-art methods on the *RS-test*, while the model with sample weights (*SWR*) further outperforms *NSWR* on the *ED-test*, demonstrating the effectiveness of our regression model to predict aesthetics for images of a variety of aesthetic quality. Note that *SWR* produces larger MSE than *NSWR* and the method in [10] on the *RS-test*. This is because the *RS-test* and training set have similar unbalanced distribution. Hence, the bias introduced by the training set ac-

tually benefits these two models with better performance on the *RS-test*.

However, such bias in fact limits the prediction range of the models. The minimum and maximum values of the aesthetic scores predicted by the *NSWR* model on both test sets are 3.54 and 6.46. For the *SWR* model, these two values are 2.06 and 7.53. We further illustrate this effect in Fig. 3, which shows the mean MSE for different aesthetic scores. Using sample weights clearly contributes to reducing the MSE for images with aesthetic scores larger than 6 or smaller than 4.

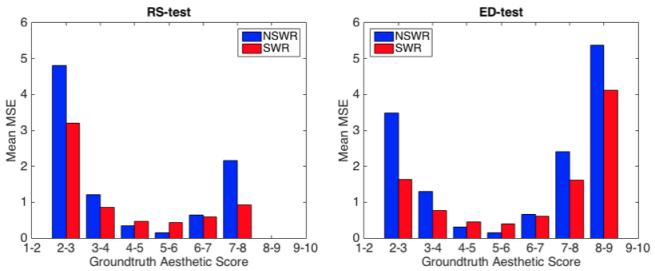

Fig. 3. Mean MSE for different aesthetic scores on the (a) *RS-test*, (b) *ED-test*.

We further evaluate our regression models on a classification task, following the same scheme as in [10]. We observe similar trends of the results as the regression task.

4.4. Histogram prediction model results

Two values can be extracted from the output of the histogram prediction model, the aesthetic score and the standard deviation (std) of the predicted user ratings. MSE in Eqn. 4 is used to evaluate the aesthetic score and the Root Mean Square Error Ratio (RMSER) is used for evaluating the std:

$$RMSER = \frac{\sqrt{\frac{1}{M}\sum_{i=1}^{M}(std_i - \hat{std}_i)^2}}{\frac{1}{M}\sum_{i=1}^{M}\hat{std}_i} \quad (5)$$

where std_i is the std of the predicted user ratings for image i and $\hat{std}_i$ is the std of the groundtruth histogram.

We train a *Histogram* prediction model with *Sample Weights* (*SWH*). Table 2 shows the results. *SWH* achieves comparable performance as the *SWR* for predicting the aesthetic scores on the *ED-test*, while producing less than 20% RMSER. Hence, the difficulty of aesthetics assessment for an image is also reliably estimated.

Table 2. MSE and RMSER for the histogram prediction model with sample weights (*SWH*).

	MSE	RMSER
RS-test	0.6358	26.75%
ED-test	1.0109	19.57%

5. APPLICATION

Our aesthetics prediction model can be used in many applications. We propose a simple application where our regression model *SWR* is used to automatically choose an aesthetically pleasing crop from the input image to fit into a target window, as users are often required to fit an image into a fixed-sized window. For an input image, we randomly take 1000 fixed-sized crops [2] and feed them into *SWR*. The one with the highest score is chosen as the output. Two examples are shown in Fig. 4. To prove the effectiveness of this application, we conducted a crowd-sourcing experiment on 50 images where we ask users to compare the crops chosen by our model with the random crops. In total, 40 users participated in the experiment. The results show that for 31 out of 50 images, users prefer the crops chosen by our system over the random crops.

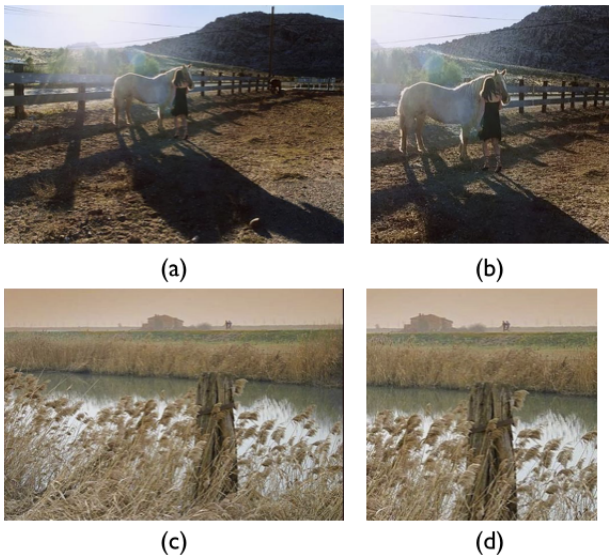

Fig. 4. Outputs from our image enhancement system. (a), (c) are original images and (b), (d) are the square crops that have the highest aesthetic scores.

6. CONCLUSION

In this paper, we propose to use sample weights while training CNN models on the AVA dataset for aesthetics assessment. Our experiments demonstrate the effectiveness of the sample weights for reducing the bias in the training set. We train two CNN models with sample weights, a regression model and a histogram prediction model. Our CNN models can output not only accurate aesthetic scores, but also reliable estimation of the difficulty of aesthetics assessment. Based on the results of our aesthetics prediction model, we further show an image enhancement system that crops the input image for better aesthetic quality. Further exploration of applications using our aesthetics prediction models will be conducted in the future.

[2] we use square crops in this experiment.

7. REFERENCES

[1] Yiwen Luo and Xiaoou Tang, "Photo and video quality evaluation: Focusing on the subject," in *Computer Vision–ECCV 2008*. 2008, pp. 386–399, Springer.

[2] Subhabrata Bhattacharya, Rahul Sukthankar, and Mubarak Shah, "A framework for photo-quality assessment and enhancement based on visual aesthetics," in *Proceedings of the 18th ACM International Conference on Multimedia*. 2010, pp. 271–280, ACM.

[3] Wei Luo, Xiaogang Wang, and Xiaoou Tang, "Content-based photo quality assessment," in *Computer Vision (ICCV), 2011 IEEE International Conference on*. 2011, pp. 2206–2213, IEEE.

[4] Luca Marchesotti, Florent Perronnin, Diane Larlus, and Gabriela Csurka, "Assessing the aesthetic quality of photographs using generic image descriptors," in *Computer Vision (ICCV), 2011 IEEE International Conference on*. 2011, pp. 1784–1791, IEEE.

[5] Luca Marchesotti and Florent Perronnin, "Learning beautiful (and ugly) attributes," in *Proceedings of the British Machine Vision Conference*, 2013.

[6] Naila Murray, Luca Marchesotti, and Florent Perronnin, "AVA: A large-scale database for aesthetic visual analysis," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. 2012, pp. 2408–2415, IEEE.

[7] Luca Marchesotti, Naila Murray, and Florent Perronnin, "Discovering beautiful attributes for aesthetic image analysis," *International Journal of Computer Vision*, vol. 113, no. 3, pp. 246–266, 2014.

[8] Xin Lu, Zhe Lin, Hailin Jin, Xin Yang, Lu, Zhe Lin, Hailin Jin, Jianchao Yang, and James Z. Wang, "Rapid: Rating pictorial aesthetics using deep learning," in *Proceedings of the 22nd ACM International Conference on Multimedia*. 2014, pp. 457–466, ACM.

[9] Florian Simond, Nikolaos Arvanitopoulos Darginis, and Sabine Süsstrunk, "Image aesthetics depends on context," in *Image Processing (ICIP), 2015 IEEE International Conference on*. 2015, pp. 3788–3792, IEEE.

[10] Yueying Kao, Chong Wang, and Kaiqi Huang, "Visual aesthetic quality assessment with a regression model," in *Image Processing (ICIP), 2015 IEEE International Conference on*. 2015, pp. 1583–1587, IEEE.

[11] Weibao Wang, Jan Allebach, and Yandong Guo, "Image quality evaluation using image quality ruler and graphical model," in *Image Processing (ICIP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 2256–2259.

[12] Yandong Guo Jianyu Wang and Jan Allebach, "A bayesian approach to infer ground truth photo aesthetic quality score from psychophysical experiment," in *IS&T/SPIE Electronic Imaging*, 2016.

[13] David G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[14] Gabriela Csurka and Florent Perronnin, "Fisher vectors: Beyond bag-of-visual-words image representations," in *Computer Vision, Imaging and Computer Graphics. Theory and Applications*. 2011, pp. 28–42, Springer.

[15] Florent Perronnin, Jorge Sánchez, and Thomas Mensink, "Improving the Fisher kernel for large-scale image classification," in *Computer Vision – ECCV 2010*. 2010, pp. 143–156, Springer.

[16] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25 (NIPS 2012)*. 2012, pp. 1097–1105, Curran Associates, Inc.

[17] Ken Chatfield, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman, "Return of the devil in the details: Delving deep into convolutional nets," in *Proceedings of the British Machine Vision Conference*, 2014.

[18] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich, "Going deeper with convolutions," in *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*, 2015, pp. 1–9.

[19] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2014.

[20] Aude Oliva and Antonio Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *International Journal of Computer Vision*, vol. 42, no. 3, pp. 145–175, 2001.

[21] Gabriella Csurka, Christopher Dance, Lixin Fan, Jutta Willamowski, and Cédric Bray, "Visual categorization with bags of keypoints," in *Workshop on Statistical Learning in Computer Vision, ECCV*, 2004, pp. 1–22.

[22] Alex J Smola and Bernhard Schölkopf, "A tutorial on support vector regression," *Statistics and Computing*, vol. 14, no. 3, pp. 199–222, 2004.