USRRM: Pairwise Ranking and Scoring Images using its Aesthetic Quality

Binay Dahal and Justin Zhan, *University of Nevada Las Vegas*

Abstract—**Image Aesthetics Analysis is a challenging research problem as aesthetics of an image is a subjective quality and it is quite difficult to formulate it into a mathematical or algorithmic problem. On the other hand, its applications are numerous, ranging from aesthetic based image retrieval to image editing. Earlier works on image aesthetics analysis relied upon hand picking the standard features from the image, based upon which, its aesthetics was quantified. This method was applied, in belief that sufficient aesthetic features have been taken into consideration and no more features impacts its aesthetic quality. This is not always the case, since, subjective quality as this, is defined individually and depends upon personal perspective. With the advent of deep learning, automatic feature learning became prevalent. Classification works on image aesthetics using deep learning have had some success. However, we are concerned with giving a tentative score to images and perform some sort of relative ranking among them. We formulate a unified loss objective accounting both of these factors and also devise a model named USRRM which learns from the global view and pixel level finer details from an image. The external style and semantic information also aids in model learning. The scoring result of USRRM has the best correlation with the ground truth scores among the most similar previous models and comparable results with less similar models but evaluated with same metrics. We quantify our results using different correlation coefficients and accuracy metric.**

Keywords*—*Image Aesthetics, Convolutional Neural Network, Image ranking

I. Introduction

Image Aesthetics Analysis is a research problem of automatically assessing the aesthetic quality of an image. This research problem is finding its application in wide areas ranging from image retrieval to image editing. Analyzing image aesthetics by computer is a challenging problem as aesthetics is a highly subjective attribute. It is related to visual perception. It is very hard to define a metric that differentiates aesthetically sound images from low aesthetic images. Even for humans, judging an image on its aesthetic quality is not a straightforward task. It may vary from person to person. Various factors play role in its assessment. Hence, representing the aesthetics of image quantitatively is not obvious. Initial attempts of analyzing image aesthetics mainly comprised of hand picking the features from image and representing it in terms of those features[1]. Then the problem is to classify or rate the image according to the result of some weighted combination of these features. The defining features are often inspired based on photography rules or psychological intuition that describes the aesthetics. Some of those photography rules are color, the rule of thirds, composition, sharpness, clarity etc. These image attributes are approximated as features and aesthetics is assessed based on it.

This method of manually picking the features can yield satisfactory result but misrepresentation of any image attribute or failure to take into consideration any one of those can affect the result negatively. Also, emergence of deep neural network has increasingly substituted the feature extraction process to the model itself. Motivated by this, we use deep convolutional neural network(CNN) to the problem of analysing image aesthetics. We have seen the success of CNNs in image processing and pattern analysis tasks such as image retrieval, semantic classification of image etc. CNNs can be effectively applied to image tasks due to its features like local receptive fields and weight sharing. Prior works have noted that the performance of image aesthetics model is enhanced when the model is fed with both the global view and local information from the image[2]. They have also injected the external information related to style of images and it's semantic separately in a model and shown each of these information is beneficial to the classification. However, most of the works concentrate on binary classification of aesthetic categories. Application wise and for more rigorous analysis, ranking of images or giving them some absolute score seem more appropriate. The scoring task is more challenging too, given that the inter-class variation in binary classification can be relatively bigger which is not the case in scoring. In addition, giving true rank of images necessitates the model to be able to draw line between images of comparable aesthetic score. Hence, we work to come up with such scoring and ranking model.

Specifically, in this paper we develop a CNN model incorporating global view, local view, style and semantic information. We use content or semantic of an image interchangeably. The model is guided towards an objective of scoring the image based on it's aesthetic quality and relatively ranking images in pair at the same time. For this objective, we propose a unified loss function which guides the model training. We call this model Unified Scoring and Relative Ranking Model. Section II of the paper discusses some of the prior works done in this area. First we summarize some methods based on manual features extraction and then present the approaches based on automatic features learning. We propose our approach and discuss it in detail in section III. Next, we talk about the experimental setup and results obtained from the experiments. We also provide a brief discussion of our results with some analysis. Finally, we conclude our paper with some conclusion

B. Dahal(binay.dahal@unlv.edu) is a Ph.D. student of Computer Science at University of Nevada, Las Vegas

J. Zhan(justin.zhan@unlv.edu) is a professor of Computer Science and Director of Big Data Hub at University of Nevada, Las Vegas

in section V.

II. Related Works

Automatic aesthetics analysis of images can be broadly studied under two major approaches. First there were studies that focused on common visual cues. Color[3, 4, 5], texture[3, 6], and composition[7, 8, 9] are some of such visual cues. These were manually regarded as the features based on which aesthetics analysis of images are to be performed. Then, after the successful applications of deep learning and especially Convolutional Neural Nets on several image related problems, this area of research has also commenced applying CNNs to automatically extract related features. The later approaches has achieved state of the art performance compared to the former manual approaches.

A. Approaches based on Manual Features Extraction

One of the earliest work in image aesthetics analysis was done by Datta et al.[3]. They manually enlisted various features to form classification and regression model to classify the images into aesthetic category and to provide aesthetic score. Some of the features they used are Light and Colorfulness, Saturation and Hue, Rule of Thirds, Familiarity of images, Size and Aspect ratio etc. [6] did a more thorough study of factors that differentiates high quality professional photographs to low quality snapshot and came of with bunch of features for classification. They develop models based on high level features such as Spatial Distribution of Edges, Color Distribution, Hue Count, and Blur and two low level features namely: contrast and brightness. Bhattacharya et al.[10] proposed an interactive framework for photo quality enhancement by learning support vector regression model to capture image aesthetics. The framework provides compositional recommendation to user based on the learned model.

Content based assessment of images was done in [9]. They extract regional features from the subject areas that draw most attentions to human eye and combine it with global features to perform the analysis. [11] used general image descriptors to assess the aesthetic quality. Particularly, they used two family of image descriptors: Bag-of-Words(BOV)[12] and Fisher Vector (FV)[13]. Other work employing Fisher vector is [11]. Works have been done representing content of image using generic image features such as SIFT[14] and GIST[15]. One slightly related work is done by Tong et al. [16] grouping images based on whether they are taken by professional photographers or home users. They investigate some low-level features related to the tasks and feed it to the classifier. [17] extracts the salient region from an image use the region to classify it using some visual descriptors and [18] generates various candidates by cropping the image and applies quality score methods to come up with the most agreeable cropped region. All of these works are related to image aesthetics in one way or another.

B. Approaches based on automatic learning of features

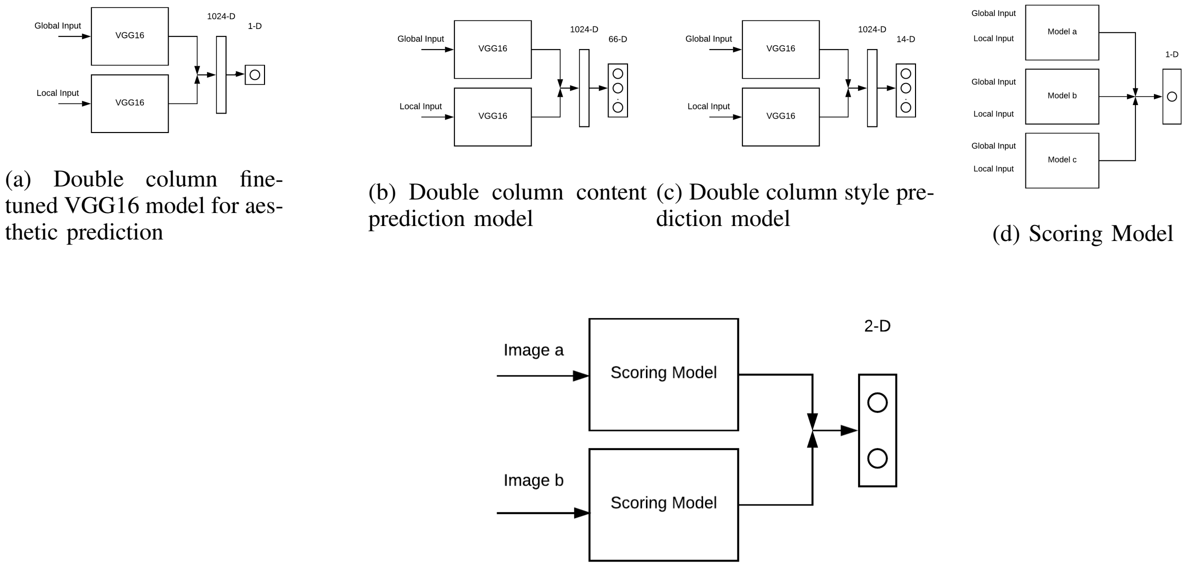
After deep learning and in particular Convolutional Neural Nets took off following its successful application on image classification task by [19], most of the computer vision research begin applying CNNs to their problems. Similarly, aesthetic analysis of images also saw works that employed CNNs to automatically extract features from the images and perform the analysis. This was made easier by the introduction of AVA[20]: a large scale image dataset containing the ratings of images from users.

Lu et al.[2] proposed Single Column and Double Column CNN model to classify images as high or low based on its aesthetics. Single Column network took as input one of the transformed image to perform classification. The transformation they applied were central cropping, warping image to a fixed size, and padding. Double column network is given two input; one a random cropping of image of capture local features and second a whole image warped to a certain size to capture global features. Later, they also used the semantic and style attributes of an image to form RDCNN[2]. They showed, RDCNN gives the best performance. This work transforms the images to a fixed size so that it can be fed to a CNN. However, [21] argues, this hampers the image composition by reducing the resolution or introducing the distortion in images. Hence, they proposed Composition-preserving photo aesthetics assessment. Using adaptive spatial pooling strategy[22], they remove the fixed-size constraint in images. Brain-inspired Deep Networks has been proposed in [23]. They introduced parallel supervised pathway which helps learn various attributes on different dimensions. Neural Image Assessment(NIMA) is proposed in [24] where they introduce a CNN model to predict the distribution of human opinion scores for any task and experimented with aesthetic score as well using AVA dataset. Using NIMA, image enhancement pipeline is proposed in [25]. Another more recent work is done by Hosu et al.[26] where they propose a method to support full resolution image without downsizing them preserving most of the information. Wang et al.[27] uses deep network to perform the photo cropping retaining the important parts of an image while being aesthetically pleasing. EnhanceGAN is proposed in [28] which performs image enhancement based on adversial learning.

III. Our Approach

Convolutional Neural Networks(CNN) are one of the widely used approaches in deep learning. It has found much success in computer vision tasks such as object identification. CNNs work by hierarchically extracting most primitive features of input and building those features to construct more abstract features. We construct a deep convolutional neural network which takes two images as an input and attempts to output the difference of aesthetic score between them. This way we can find the relative ranking of the images while correctly predicting their difference in aesthetic score. Aesthetics of an image is impacted by both of its global view and local view. Global view gives meaning to the image while local view takes care of technical aspect like resolution.

Our first model is a Double Column VGG16[29] network trained and fine-tuned for binary classification of image aesthetics. Then we extend this model for a pair of image and train it to predict the difference in aesthetic score between then.



(a) Double column fine-tuned VGG16 model for aesthetic prediction

(b) Double column content prediction model

(c) Double column style prediction model

(d) Scoring Model

(e) Unified Model

Fig. 1: Overview of our models

A. Double column Fine Tuned VGG16

VGG16 network was trained on ImageNet[30] dataset for large scale image classification. Since, analysis of image aesthetics entails working out various features from the image and combining them to yield a aesthetic prediction, we argue using the model that was pre-trained on large image database is beneficial than training a model from scratch. Two columns of our model takes the global view and randomized local view of an image to properly capture the overall representation and patterns of the image to fine resolution level information. Each of the columns is a VGG16 network which are then merged later and fine-tuned together.

B. Style and Content Prediction Model

Fine-tuned VGG16 Model is trained solely on the image pixel data. It constructs the features based on the spatial patterns and colors of the images. The model lacks any external information such as what category does the image belong to or what style of photography does this image represent. Any external relevant information apart from the image itself provided to the model will definitely help the model performance. Our dataset contains the style and content information of some of the images while majority of images lack this information. Hence, we create a double column style and double column content prediction model which is trained on the data which has style and content labels. Style and Content Model try to predict the style and semantic of an image respectively. Some of the styles of images include complementary Colors, Duotones, Long Exposure, Macro etc. Model is trained for 14 total styles. The semantics of an image can be any of the 66 categories including architecture, animal, food and drinks or landscape. The fine-tuned VGG16 model, style model and the content model can be combined together to perform aesthetic related analysis including aesthetic category prediction or score prediction. The base model for all of our prediction model is VGG16.

C. Scoring Model

We combine fine-tuned VGG16 aesthetic prediction model with Style and Content model to create a Scoring model. These three models are merged at a later stage and the output is fed to a sigmoid layer whose output represents the score of an image. Our dataset contains score from 1 to 10 for each images from number of users. We normalize the images score in range 0 to 1 so that it will be within the range of sigmoid layer output. We don't train the scoring model as a separate unit rather it is taken as a functional unit of our unified scoring and relative ranking model.

D. Unified Scoring and Relative Ranking Model

Unified Scoring and Relative Ranking Model(USRRM) is a all in one model which is trained with pairwise image samples. We develop USSR with a combined objective of reducing the loss in scores prediction of a pair of image as well as ranking them in correct order. The fundamental unit of Unified Scoring and Relative Ranking Model is the Scoring Model we developed in the previous section. Each members of a pair of image is fed to a scoring model and it's output is combined as the last layer of USRRM. Training the model with just the scoring loss may work well when the scores are far apart between the images in a pair. However, for images with close scores, we have to make sure the ranking is correct too.

1) Scoring Loss: The first part of the combined loss is the the scoring loss which is defined as follows:

$$loss_{scoring} = \frac{1}{n}\sum_{i=1}^{n}(y-\widehat{y})^2 \tag{1}$$

where y is the true labels and $\widehat{y}$ is the predicted output of our USRRM. Since, the model is 2 column scoring model merged at the output layer, it's output is 2 dimensional. The scoring loss minimizes the distance between true score and predicted score of 2 images in a pair.

2) Ranking loss: Next, it is imperative that we deal with the situation where the scoring loss brings the predicted score as close to the true score as possible but the true scores are themselves very close meaning the images have almost similar scores, the order of the image may end up being incorrect. Hence, the ranking loss is defined as:

$$loss_{ranking} = \frac{1}{n}\sum_{i=1}^{n}(min(0, f(y)(\widehat{y}[0]-\widehat{y}[1]))^2$$

where,

$$f(y) = \begin{cases} 1, & \text{if } y[0] > y[1]. \\ -1, & \text{if } y[0] < y[1]. \end{cases} \tag{2}$$

3) Unified loss: The Unified Scoring and Relative Ranking Model is trained with unified loss given as follows:

$$loss_{unified} = \alpha loss_{scoring} + \beta loss_{ranking}$$

, where α and β are the weight given to scoring loss and ranking loss.

So, our unified loss will be:

$$loss_{unified} = \alpha\left[\frac{1}{n}\sum_{i=1}^{n}(y-\widehat{y})^2\right] + \beta\left[\frac{1}{n}\sum_{i=1}^{n}(min(0, f(y)(\widehat{y}[0]-\widehat{y}[1]))^2\right] \tag{3}$$

IV. Experimental Setup and Results

A. Dataset

AVA is the standard dataset for image aesthetics analysis. It contains around 255,000 images with about 100 ratings for each images. A score on a scale of 10 is given by human users to each of these images. Most of the images have at most 2 semantic category associated with it. All together there are 66 categories of image in the dataset. About 14,000 images in the dataset are associated with style labels. There are 14 total styles.

For relative ranking of images, we create a new subset dataset from AVA by sampling the pair of images. We calculate the average score for each images in AVA dataset which is in the scale of 10 and normalize it within the range of 0 and 1 and sample 195,462 pairs of images out of which training size is 156,373 and test size is 39,089. While doing so, we proportionally sample out the pairs having the score difference of at least 0.5, at least 0.4 and so on till the score difference between the pairs is less than 0.1. By doing so, we ensure the training process is not skewed.

We also train and test our model with TID2013[31] dataset which contains 25 reference images and 3000 distorted images obtained from the reference images. Each of the 3000 distorted images has a mean opinion score which we used as a label. We randomly sampled pair of 5000 images from the dataset out of which 4000 pairs are used for training.

B. Initial Finetuning and training of models

To create a fine-tuned VGG16 model as shown in Fig 1a. we take two VGG16 model without it's top 3 layers and concatenate them. All the layers in VGG16 model are not trainable. Only the layers added after concatenating the models are trained. A Global Average Pooling layer is added on top followed by a 1024 dimensional dense layer. A batch normalization layer follows this layer. At last a sigmoid layer is added. This model is fine tuned for aesthetic category prediction. An image with average score more than 0.5 is considered aesthetic whereas an image less than 0.5 is non-aesthetic. This fine tuning is done with 230,000 training images and 25,507 test images from original AVA dataset. 5% of training data is allocated as validation data. The model is finetuned for 5 epochs with binary cross entropy as a loss function.

Next, we train the model shown in figure 1b and 1c to predict the content category and style category of an image. Each of the models is created as described above for fine-tuned VGG16 but instead of a 1 unit sigmoid layer at output, we place a 66 units sigmoid layer resembling the 66 content categories and 14 units sigmoid layer resembling the 14 style categories. The models are trained for 5 epochs for multi label classification using binary cross-entropy as a loss function. To create a scoring model as shown in figure 1d, we combine each of these 3 models without their output layer and add a final 1 unit sigmoid layer to predict the score of an image. All of our models are trained on a Tesla K80 GPU machine with 4 GPU cores each having around 12GB of RAM.

C. Pairwise Training of USRRM and Evaluation

USRRM is trained with the dataset we created from AVA and TID2013 with an objective unified loss we described in the previous section. The model is trained with different values of α and β. The two units in it's output layer represents the scores of a pair of image provided as an input. The model is trained for additional 5 epochs. We also experimented with different values of α and β in our unified loss function. Some of the values chosen was α=0.3 and β=0.7, α=0.5 and β=0.5 and α=0.7 and β=0.3.

The model is evaluated with three different metrics. The first one is the correlation coefficient(Spearman's and Pearson's). The correlation coeeficient measures if the two series of data are positively correlated or negatively correlated. It's value lies in the range of [-1,1]. The larger the coefficient, the more

Methods	ρ	Accuracy (%)	Ranking Accuracy(%)
Murray et al.[20]	-	68.0	-
SPP [22]	-	72.85	-
AlexNet_FT_Conf	0.48	71.52	-
DCNN[2]	-	73.25	-
RDCNN_style[2]	-	74.46	-
RDCNN_semantic[2]	-	75.42	-
DMA[32]	-	74.46	-
DMA_AlexNet$_FT$[32]	-	75.41	-
Reg[33]	0.499	72.04	-
Reg+Rank+Att[33]	0.544	75.48	-
Reg+Rank+Att+Cont[33]	0.558	77.33	-
NIMA(VGG16)[24]	0.592	80.6	-
Hosu et al.[26]	0.75	81.72	-
USRRM	**0.58**	**80.69**	**72.2**

TABLE I: Performance evaluation of USRRM on AVA with other baseline models. Our model outperforms almost all of the previous models both on correlation coefficient and on accuracy and has the comparable performance with the couple of most recent models. However, it should be noted that we trained our model with an objective of comparing a pair of images and provide a relative ranking which none of those models do. Hence, being trained with different objective, our model still yields comparable result.

Loss	ρ	r
Scoring	0.565	0.59
Scoring+ranking	0.58	0.61

TABLE II: Two correlation coefficients on AVA dataset

correlated the data is. The Spearman's coefficient is computed as:

$$\rho = 1 - \frac{6\sum_{i=1}^{N} {d_i}^2}{N(N^2-1)}$$

where ${d_i}^2$ is the square of difference between i^{th} element of two data series. Similarly, Pearson's correlation coefficient is given by:

$$r = \frac{N\sum_{i=1}^{N} x_i y_i - (\sum_{i=1}^{N} x_i)(\sum_{i=1}^{N} y_i)}{\sqrt{\left[N\sum_{i=1}^{N} {x_i}^2 - (\sum_{i=1}^{N} x_i)^2\right]\left[N\sum_{i=1}^{N} {y_i}^2 - (\sum_{i=1}^{N} y_i)^2\right]}}$$

where x and y are two series of data.

Next, we compute the accuracy of binary classification into aesthetic classes which is a primary evaluation metrics for most of the related works. For this, we perform a logistic regression using the predicted score of images as the only feature and try to linearly separate the two classes. Lastly, we also compute the relative ranking accuracy which is defined as the fraction of image pairs correctly ranked.

$$Ranking \quad Accuracy = \frac{\sum_{i=1}^{N} Cr(y_i)}{Total \quad pairs}$$

where,

$$Cr(y) = \begin{cases} 1, & \text{if } y[0] > y[1] \text{ and } \widehat{y}[0] > \widehat{y}[1]. \\ 1, & \text{if } y[0] < y[1] \text{ and } \widehat{y}[0] < \widehat{y}[1]. \\ 0, otherwise \end{cases} \tag{3}$$

Methods	$\rho(SRCC)$	r(LCC)	Ranking Acc.(%)
Kim et al.[34]	0.8	0.8	-
Moorthy et al.[35]	0.88	0.89	-
Mittal et al.[36]	0.89	0.92	-
Saad et al.[37]	0.88	0.91	-
Kottayil et al.[38]	0.88	0.89	-
Xu et al.[39]	0.95	0.96	-
Bianco et al.[40]	0.96	0.96	-
NIMA(MobileNet)[24]	0.69	0.78	-
NIMA(VGG16)[24]	0.94	0.94	-
USRRM	**0.937**	**0.939**	**87.97**

TABLE III: Performance evaluation of USRRM on TID2013. Our model was designed to use the external information like semantic labels and style labels while training whereas TID2013 images lacked such labels. Hence, we trained the model with just global and pixel level information and obtained very comparable results.

D. Results and Discussion

The Spearman's coefficient and the pearson's coefficient of correlation or Linear Correlation Coefficient(LCC) along with the accuracy is tabulated in table I,II and III. Other baseline models with similar metrics are also shown for comparison. USRRM has an Spearman's coefficient of 0.58 and accuracy of 80.69% on AVA dataset. Most of the previous works don't have the earson's coefficient computed. For our model, it is slightly higher at 0.61. On TID2013, Spearman's coefficient is 0.937 and LCC is 0.939. We trained our model for different values of α and β in the loss functions, the best result is obtained for $\alpha = 0.7$ and $\beta = 0.3$.

To be sure that both parts of our unified loss i.e. scoring loss and ranking loss have contributed to the obtained values of correlation, we separately trained the model using scoring

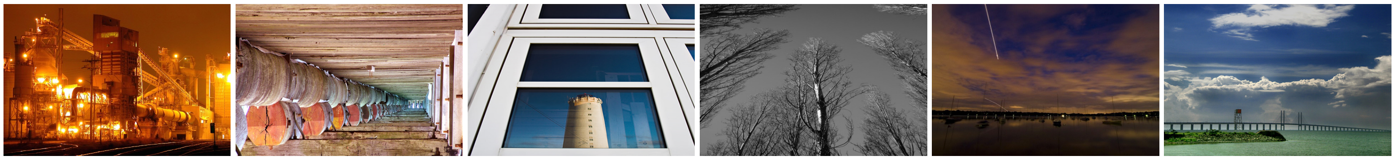


Fig. 2: Aesthetically high images incorrectly predicted by the scoring model trained with binary cross entropy and correctly predicted by the USRRM model trained with unified loss(AVA dataset)

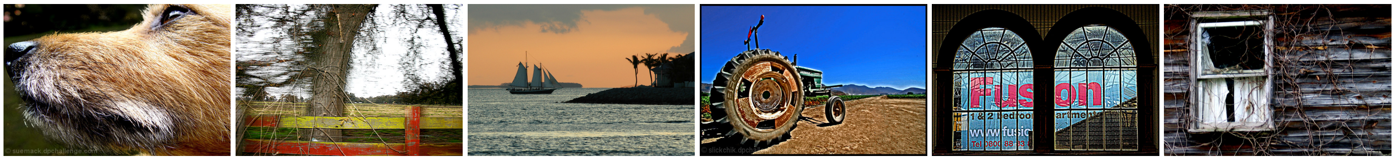
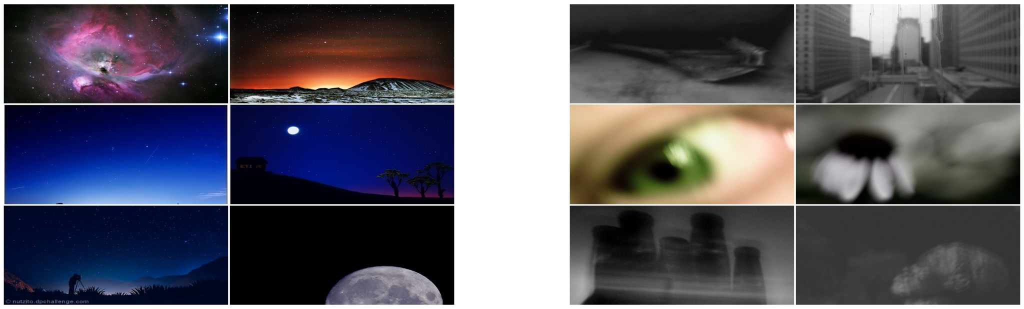


Fig. 3: Aesthetically low images incorrectly predicted by the scoring model trained with binary cross entropy and correctly predicted by the USRRM model trained with unified loss(AVA dataset)



(a) Images receiving highest score by USRRM

(b) Images receiving lowest score by USRRM

Fig. 4: Highest and Lowest scores given by USRRM(AVA dataset)

loss only. Table II shows that scoring loss only gives a slightly lesser value of correlation. To compare the improvements of the USRRM over the previous models, we trained the scoring model separately using binary cross-entropy loss and tested with aesthetic categorization. Some of the results are shown in figure 3 and 4. We can see the instances of images correctly predicted by the USRRM which the scoring model which was not trained with unified loss was not able to predict. We also did some analysis on the ground truth score of the images correctly predicted by the USRRM on AVA dataset. Some of the sampled scores are listed in table IV. We can see from the table, most of the images have the ground truth score on borderline of 5. It is noteworthy that the threshold value of the aesthetic categorization classes for our experiment is 5. This gives a strong indication that the USRRM model which takes both the scoring loss and ranking loss into consideration has more fine-grained view on the images and hence is able to distinguish the images that are in the borderline.

Next, the accuracy comparison of these two models over various semantic categories was performed. The USRRM model outperforms the scoring model in this regard too. Finally, in addition to accuracy, we wanted to make sure the high accuracy of the model is not too deceiving. We evaluated our model using F1 score as well on AVA dataset. The model achieved an F1 score of 0.68. As the F1 score of the model is impacted by both the precision and recall, the majority class of our dataset is non-aesthetic. Hence, this unbalanced nature of the dataset brings the recall down to some extent and influence the F1 measure. However, no other existing works have analysed the F1 score of the classification so we have no way to compare our result on this metric.

V. Conclusion

We provided a unified quantitative model of scoring and ranking images based on their aesthetic value. Our model takes into consideration the global view of the image, finer pixel level details of the image, and leverages the external information in the form of it's style and semantic. Furthermore,

Aesthetically high image	Aesthetically low image
5.83	5.44
5.52	5.01
5.558	4.76
5.552	5.44

TABLE IV: Sample ground truth score of images on AVA incorrectly predicted by scoring model and correctly predicted by USRRM

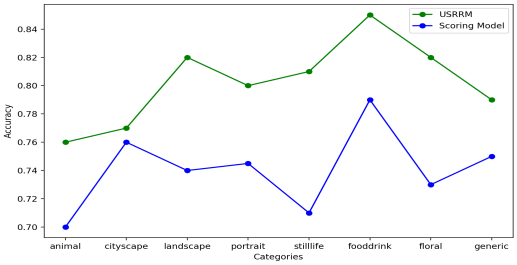


Fig. 5: Accuracy comparison between USRRM trained with unified loss and Scoring Model trained with binary crossentropy on AVA

our approach is optimized for ranking the image pairs in addition to giving scores to each images in a pair. This scoring and ranking combined in a single loss objective makes our model truly a unified solution to image aesthetic analysis. This unified approach gives us a state of the art correlation coefficient of 0.58 and an accuracy of 80.69%. With this result, our research may find application in enhancing automatic photo management tools or other image editing software. That being said, there is still plenty of rooms for improvement in image aesthetic analysis. Context aware image aesthetics analysis is one such area. Our human brain process the images putting it's context into perspective. Image semantic may be somewhat taken as a context however the big picture of the setting or the knowledge of what's going on in the picture is largely responsible for how we judge them in terms of aesthetics. Coming up with some representation of the context and using them as a metadata may be the next big challenge in this topic.

References

[1] Tunç Ozan Aydın, Aljoscha Smolic, and Markus Gross. "Automated aesthetic analysis of photographic images". In: *IEEE transactions on visualization and computer graphics* 21.1 (2015), pp. 31–42.

[2] Xin Lu et al. "Rating image aesthetics using deep learning". In: *IEEE Transactions on Multimedia* 17.11 (2015), pp. 2021–2034.

[3] Ritendra Datta et al. "Studying aesthetics in photographic images using a computational approach". In: *European conference on computer vision*. Springer. 2006, pp. 288–301.

[4] Masashi Nishiyama et al. "Aesthetic quality classification of photographs based on color harmony". In: *CVPR 2011*. IEEE. 2011, pp. 33–40.

[5] Peter O'Donovan, Aseem Agarwala, and Aaron Hertzmann. "Color compatibility from large datasets". In: *ACM Transactions on Graphics (TOG)* 30.4 (2011), p. 63.

[6] Yan Ke, Xiaoou Tang, and Feng Jing. "The design of high-level features for photo quality assessment". In: *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*. Vol. 1. IEEE. 2006, pp. 419–426.

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

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

[9] Wei Luo, Xiaogang Wang, and Xiaoou Tang. "Content-based photo quality assessment". In: *2011 International Conference on Computer Vision*. IEEE. 2011, pp. 2206–2213.

[10] 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*. ACM. 2010, pp. 271–280.

[11] Luca Marchesotti et al. "Assessing the aesthetic quality of photographs using generic image descriptors". In: *2011 International Conference on Computer Vision*. IEEE. 2011, pp. 1784–1791.

[12] Gabriella Csurka et al. "Visual categorization with bags of keypoints". In: *Workshop on statistical learning in computer vision, ECCV*. Vol. 1. 1-22. Prague. 2004, pp. 1–2.

[13] Tommi Jaakkola and David Haussler. "Exploiting generative models in discriminative classifiers". In: *Advances in neural information processing systems*. 1999, pp. 487–493.

[14] David G Lowe. "Distinctive image features from scale-invariant keypoints". In: *International journal of computer vision* 60.2 (2004), pp. 91–110.

[15] Aude Oliva and Antonio Torralba. "Modeling the shape of the scene: A holistic representation of the spatial envelope". In: *International journal of computer vision* 42.3 (2001), pp. 145–175.

[16] Hanghang Tong et al. "Classification of digital photos taken by photographers or home users". In: *Pacific-Rim Conference on Multimedia*. Springer. 2004, pp. 198–205.

[17] Lai-Kuan Wong and Kok-Lim Low. "Saliency-enhanced image aesthetics class prediction". In: *2009 16th IEEE International Conference on Image Processing (ICIP)*. IEEE. 2009, pp. 997–1000.

[18] Masashi Nishiyama et al. "Sensation-based photo cropping". In: *Proceedings of the 17th ACM international conference on Multimedia*. ACM. 2009, pp. 669–672.

[19] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. "Imagenet classification with deep convolutional neural networks". In: *Advances in neural information processing systems*. 2012, pp. 1097–1105.

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

[21] Long Mai, Hailin Jin, and Feng Liu. "Composition-preserving deep photo aesthetics assessment". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 497–506.

[22] Kaiming He et al. "Spatial pyramid pooling in deep convolutional networks for visual recognition". In: *IEEE transactions on pattern analysis and machine intelligence* 37.9 (2015), pp. 1904–1916.

[23] Zhangyang Wang et al. "Brain-inspired deep networks for image aesthetics assessment". In: *arXiv preprint arXiv:1601.04155* (2016).

[24] Hossein Talebi and Peyman Milanfar. "Nima: Neural image assessment". In: *IEEE Transactions on Image Processing* 27.8 (2018), pp. 3998–4011.

[25] Hossein Talebi and Peyman Milanfar. "Learned perceptual image enhancement". In: *2018 IEEE International Conference on Computational Photography (ICCP)*. IEEE. 2018, pp. 1–13.

[26] Vlad Hosu, Bastian Goldlucke, and Dietmar Saupe. "Effective Aesthetics Prediction with Multi-level Spatially Pooled Features". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2019, pp. 9375–9383.

[27] Wenguan Wang, Jianbing Shen, and Haibin Ling. "A deep network solution for attention and aesthetics aware photo cropping". In: *IEEE transactions on pattern analysis and machine intelligence* 41.7 (2018), pp. 1531–1544.

[28] Yubin Deng, Chen Change Loy, and Xiaoou Tang. "Aesthetic-driven image enhancement by adversarial learning". In: *2018 ACM Multimedia Conference on Multimedia Conference*. ACM. 2018, pp. 870–878.

[29] Karen Simonyan and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition". In: *arXiv preprint arXiv:1409.1556* (2014).

[30] Jia Deng et al. "Imagenet: A large-scale hierarchical image database". In: *2009 IEEE conference on computer vision and pattern recognition*. Ieee. 2009, pp. 248–255.

[31] Nikolay Ponomarenko et al. "Color image database TID2013: Peculiarities and preliminary results". In: *european workshop on visual information processing (EUVIP)*. IEEE. 2013, pp. 106–111.

[32] Xin Lu et al. "Deep multi-patch aggregation network for image style, aesthetics, and quality estimation". In: *Proceedings of the IEEE International Conference on Computer Vision*. 2015, pp. 990–998.

[33] Shu Kong et al. "Photo aesthetics ranking network with attributes and content adaptation". In: *European Conference on Computer Vision*. Springer. 2016, pp. 662–679.

[34] Jongyoo Kim et al. "Deep convolutional neural models for picture-quality prediction: Challenges and solutions to data-driven image quality assessment". In: *IEEE Signal Processing Magazine* 34.6 (2017), pp. 130–141.

[35] Anush Krishna Moorthy and Alan Conrad Bovik. "Blind image quality assessment: From natural scene statistics to perceptual quality". In: *IEEE transactions on Image Processing* 20.12 (2011), pp. 3350–3364.

[36] Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. "No-reference image quality assessment in the spatial domain". In: *IEEE Transactions on image processing* 21.12 (2012), pp. 4695–4708.

[37] Michele A Saad, Alan C Bovik, and Christophe Charrier. "Blind image quality assessment: A natural scene statistics approach in the DCT domain". In: *IEEE transactions on Image Processing* 21.8 (2012), pp. 3339–3352.

[38] Navaneeth K Kottayil et al. "A color intensity invariant low-level feature optimization framework for image quality assessment". In: *Signal, Image and Video Processing* 10.6 (2016), pp. 1169–1176.

[39] Jingtao Xu et al. "Blind image quality assessment based on high order statistics aggregation". In: *IEEE Transactions on Image Processing* 25.9 (2016), pp. 4444–4457.

[40] Simone Bianco et al. "On the use of deep learning for blind image quality assessment". In: *Signal, Image and Video Processing* 12.2 (2018), pp. 355–362.