Predicting Aesthetic Score Distribution through Cumulative Jensen-Shannon Divergence

Xin Jin[1], Le Wu[1], Xiaodong Li[1], Siyu Chen[1], Siwei Peng[3], Jingying Chi[3], Shiming Ge[4,*], Chenggen Song[2], Geng Zhao[1]

[1]Department of Computer Sci. and Tech., Beijing Electronic Science and Technology Institute, Beijing, 100070, China
[2]Department of Info. Sec., Beijing Electronic Science and Technology Institute, Beijing, 100070, China
[3]College of Info. Sci. and Tech., Beijing University of Chemical Technology, Beijing 100029, China
[4]Institute of Information Engineering, Chinese Academy of Sciences, Beijing, 100093, China
∗Corresponding author email: geshiming@iie.ac.cn

Abstract

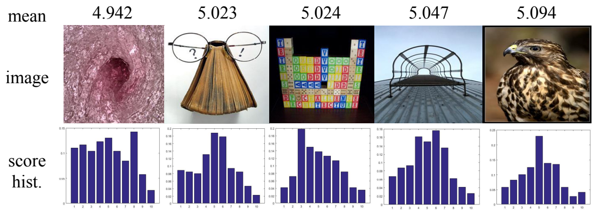
Aesthetic quality prediction is a challenging task in the computer vision community because of the complex interplay with semantic contents and photographic technologies. Recent studies on the powerful deep learning based aesthetic quality assessment usually use a binary high-low label or a numerical score to represent the aesthetic quality. However the scalar representation cannot describe well the underlying varieties of the human perception of aesthetics. In this work, we propose to predict the aesthetic score distribution (i.e., a score distribution vector of the ordinal basic human ratings) using Deep Convolutional Neural Network (DCNN). Conventional DCNNs which aim to minimize the difference between the predicted scalar numbers or vectors and the ground truth cannot be directly used for the ordinal basic rating distribution. Thus, a novel CNN based on the Cumulative distribution with Jensen-Shannon divergence (CJS-CNN) is presented to predict the aesthetic score distribution of human ratings, with a new reliability-sensitive learning method based on the kurtosis of the score distribution, which eliminates the requirement of the original full data of human ratings (without normalization). Experimental results on large scale aesthetic dataset demonstrate the effectiveness of our introduced CJS-CNN in this task.

Introduction

Recently, the ability of recognizing the semantic meaning of the objects in an image by computers is greatly increasing through deep convolutional neural networks. However, recognizing or assessing the aesthetic quality of an image by computers has not reached the practical precision people need.

Subjective Image Aesthetic Quality Assessment (IAQA) is still challenging (Mai, Jin, and Liu 2016) since the large intra class difference of images with high or low aesthetic quality, the large amount of low or high level aesthetic features, and the subjective evaluation of human rating. IAQA has been a hot topic in the communities of Computer Vision (CV), Computational Aesthetics (CA) and Computational Photography (CP).

Related work. As summarized by (Deng, Loy, and Tang 2017), in early work, various hand-crafted aesthetic features (i.e. aesthetic rule based features) are designed and connected with a machine classification or regression method. Another line is to use generic image description features. After that, the powerful deep feature representation learned from large amount of data has shown an ever-increased performance on this task, surpassing the capability of conventional hand-crafted features. (Karayev et al. 2014; Lu et al. 2014; Kao, Wang, and Huang 2015; Lu et al. 2015a; 2015b; Dong and Tian 2015; Kao, Huang, and Maybank 2016; Wang et al. 2016; Mai, Jin, and Liu 2016; Kong et al. 2016; Jin et al. 2016; Kao, He, and Huang 2017; Ma, Liu, and Chen 2017).



var.	6.665	5.291	5.161	4.987	4.911
median	4.850	5.012	4.783	5.003	5.076
skew.	0.094	0.003	0.320	0.048	0.134
kur.	1.890	2.368	2.305	2.414	2.613

Figure 1: Images with similar mean scores (i.e., around 5). The rating distributions are approximated by the score histograms (1-10). The hist., var., skew. and kur. are short for histogram, variance, skewness and kurtosis. The mean scores of the histogram are nearly the same. However, the histograms themselves with their statistics differ from each other. Images are from the AVA dataset (Murray, Marchesotti, and Perronnin 2012), which contains a list of photo IDs from www.dpchallenge.com.

The training data of aesthetic quality assessment are often collected from the online photo sharing communities such as photo.net and dpchallenge.com, in which people rate an image by selecting one of the predefined ordinal basic integer ratings (i.e., 1-7 or 1-10). Higher values indicate better rating (Wu, Hu, and Gao 2011). Most of the above studies use the following strategies to encode the aesthetic quality, namely, 1D numerical encoding and binary encoding.

- **1D numerical encoding**: the 1-dimension numerical encoding use the weighted mean scores of human ratings. A regression model can be learned to predict the numerical aesthetic quality.

- **binary encoding**: the binary encoding is used to classify the images into high or low aesthetic quality, which is determined by a threshold of the weighted mean scores of human ratings. A classifier can be learned to predict the high-low classification results.

However, although there exits consensus of the assessment of image aesthetic quality, it is still a subjective task in nature. The rated scores of multiple persons may differ greatly from each other. People tend to assign inconsistent scores to the same image (Wu, Bauckhage, and Thurau 2010). There is ambiguity in the image aesthetic quality assessment (Ke, Tang, and Jing 2006). A scalar value is insufficient to capture the true nature of the subjectivity of image aesthetic quality (Wu, Hu, and Gao 2011). The main limitation of the above representations is that they do not provide an indicator of the degree of consensus or diversity of opinion among annotators (Murray, Marchesotti, and Perronnin 2012).

Figure 1 shows some images from the AVA dataset (Murray, Marchesotti, and Perronnin 2012). Images with nearly the same mean scores (i.e., around 5) are listed. However, the distributions (approximated by the score histogram) are not that similar. Other statistics such as the variance, the median, the skewness, and the kurtosis differ greatly from each other. The human ratings are quite subjective. The mean score is greatly influenced by the low and high extremes of the rating scale, which makes it inappropriate to be a robust estimation of the whole distribution, especially when the distribution is skewed. For skewed distributions, the median value appears to be more appropriate to describe the distributions than the mean value (Wu, Hu, and Gao 2011). The Gaussian distribution is the best-performing model for only 62% of images in AVA (Murray, Marchesotti, and Perronnin 2012). The others are the skewed ones and can be best fitted by the Gamma distribution (Murray, Marchesotti, and Perronnin 2012).

Most recently, some methods are proposed to use modified or generated score distributions for binary classification and numerical assessment on aesthetics (Jin, Segovia, and Süsstrunk 2016; Wang et al. 2017; Hou, Yu, and Samaras 2016). Wu et al. (Wu, Hu, and Gao 2011) propose a modified support vector regression algorithm to predict the score distribution in two small aesthetic datasets, before the large scale AVA dataset released and the popularity of deep CNNs.

Jin et al. (Jin, Segovia, and Süsstrunk 2016) use the weighted Chi-square distance as the loss function to predict the mean score and the standard deviation from the score distribution. Wang et al. (Wang et al. 2017) explicitly modify the score distribution of the AVA dataset as Gaussian and jointly predict its mean and standard deviation. They use the asymmetrical Kullback-Leibler (KL) divergence as the loss function for their DBN network. Hou et al. (Hou, Yu, and Samaras 2016) generate score distribution by mapping the real number labels to 10 aesthetic bins of the AADB dataset (Kong et al. 2016). They propose to use squared Earth mover's distance (EMD) as the loss function, which can be equivalent to the Euclidean distance of the two cumulative distribution functions for the ordinal basic human ratings prediction. Thus, the loss functions of (Hou, Yu, and Samaras 2016) and (Wu, Hu, and Gao 2011) are the same. Note that, all these methods use modified or generated score distributions for binary classification and numerical assessment on aesthetics. While our work is to directly predict the score distribution itself. Murray et al. (Murray and Gordo 2017) use the Huber loss combined with ResNet and SPP-Net to predict the aesthetic score distribution of an image. Cui et al. (Cui et al. 2017) propose to use the traditional LDL (Label Distribution Learning) technology to predict the aesthetic score distribution of an image.

Our Approach. In this work, we learn from the large aesthetic dataset to predict the aesthetic score distribution of an image, which is represented as a score vector (histogram) using the deep convolutional neural network (DCNN), so as to better capture the subjectivity of aesthetic quality assessment. Conventional CNN which aims to minimize the difference between the predicted scalar numbers or 0-1 classification vectors and the ground truth cannot be directly used for the ordinal basic rating distribution. Inspired by recent work on non-parametric Jensen-Shannon Divergence by Nguyen et al. (Nguyen and Vreeken 2015), a Cumulative distribution with Jensen-Shannon divergence based CNN (CJS-CNN) is presented to predict the aesthetic score distribution of human ratings. In addition, to alleviate the problem of unreliable human ratings, we propose a new reliability-sensitive learning method based on the kurtosis of the score distribution. The proposed kurtosis can be directly computed using the normalized score histogram. While the rating number is additional information of the normalized score histogram and is not always available in the training set. We compare the recently proposed loss functions designed for score distribution and LDL method with our CJS loss and RS-CJS loss in the experiments. Experimental results on large scale aesthetic dataset demonstrate the effectiveness of our introduced CJS-CNN in this task. The main contributions of our work can be summarized as follows:

- The first work that predicts a score distribution vector of the ordinal basic human ratings under the deep convolutional neural network framework on the large scale AVA dataset, which is designed to capture the subjectiveness of the human aesthetic quality assessment.
- A novel CNN called the CJS-CNN (Cumulative distribution function with Jensen-Shannon Divergence) is introduced. Extensive comparisons with probability distribution function with Euclidean distance, cross entropy distance, Jensen-Shannon Divergence and cumulative distribution function with Euclidean distance are presented.
- A new reliability-sensitive learning method is proposed based on the kurtosis of the score distribution.

Besides the overall aesthetic quality of an image, there are other targets related to aesthetics. Our score distribution prediction can be used for aesthetic image retrieval, guide for shooting good photos, automatic selector for the most aesthetic or attractive cover of a video, etc. From the score distribution, rich information can be outputted, such as mean, median, variance, skewness and kurtosis. For skewed distributions, the median value appears to be more appropriate to describe the distributions than the mean value. The variance, skewness and kurtosis can be jointly used to measure the controversy of an image. The controversy is a measure of the degree of consensus and diversity of the aesthetic assessment of an image. Some artworks may not be accepted today, but may yield potential fashion or masterpieces in the future.

Subjectiveness Analysis of the AVA Dataset

The assessment of image aesthetic quality is subjective in nature. The perception of aesthetics is affected by the nationality, ethnicity, era, age, education, emotion and many other factors of human beings. In this section we make a statistical analysis of subjectiveness or diversity of the opinion among annotators in a large-scale database for aesthetic visual analysis (AVA) (Murray, Marchesotti, and Perronnin 2012). This dataset is specifically constructed for the purpose of learning more about image aesthetics. All those images are directly downloaded from dpchallenge.com. For each image in AVA, there is an associated distribution of scores (1-10) voted by different viewers. The number of votes that per image gets is ranged in 78-549, with an average of 210, which enables us to have a deeper understanding of such distributions and deduce more information from them.

The Standard Deviation or Variance. As described above, a mean score or a binary high-low label reveals only part of the information deduced from a score distribution. We make a statistical analysis on the number of images according to mean and standard deviation of the human ratings. The standard deviation represents the degree of consensus or diversity of human ratings for the same image, with a higher value meaning higher diversity. The number of images located in each mean and standard deviation interval is shown as a 2D histogram in Figure 2. Most images' mean values are located in $[4, 7]$. Images in this interval are not easy to be classified to a high-low label. Most images' standard deviation values are larger than 1.25, which shows the diversity of the human ratings for the same image. In addition, as described in (Murray, Marchesotti, and Perronnin 2012), the variance or standard deviation tends to increase with the distance between the mean score and the mid-point of the rating scale.

The Skewness. The skewness (Joanes and Gill 1998; Brown) is a measure of the asymmetry of the probability distribution of a real-valued random variable about its mean. If the bulk of the data is at the left and the right tail is longer, we say that the distribution is skewed right or positively skewed; if the peak is toward the right and the left tail is longer, we say that the distribution is skewed left or negatively skewed. Boxplots of the skewness of score distributions for images with mean scores within a specified range are shown in the left side of Figure 3. The skewness is a function of mean score in the AVA dataset. Images with mean score values from 4 to 7 tend to have a low absolute value of the skewness and can be considered as those with symmetrical score distributions. Images with mean score values lower than 4 and greater than 7 can be considered as those with positively and negatively skewed score distributions, respectively. This is likely due to the non-Gaussian nature of score distributions at the extremes of the rating scale (Murray, Marchesotti, and Perronnin 2012).

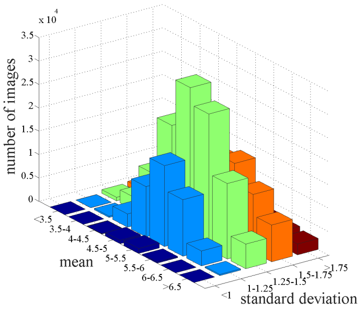


Figure 2: The histogram of numbers of images located in different intervals of the mean and standard deviation of the AVA dataset (Murray, Marchesotti, and Perronnin 2012).

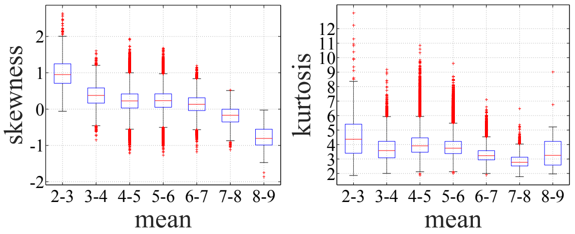


Figure 3: Left: Distributions of skewness of score distributions, for images with different mean scores. The red crosses are the outliers. The skewness tends to decrease from positive to negative with the mean score increasing. Right: Distributions of kurtosis of score distributions, for images with different mean scores.

Most representative distributions in the AVA dataset are slightly skewed or heavily skewed. For skewed distributions, the median value appears to be more appropriate to describe the distributions than the mean value (Wu, Hu, and Gao 2011). The mean and the median values of score distributions for images with skewness within a specified range are shown in Figure 4. Images with low and high absolute values of the skewness can use the mean and the median to describe their score distributions, respectively.

The Kurtosis. The other common measure of shape is called the kurtosis (Joanes and Gill 1998; Brown). As skewness is the third moment of the distribution, kurtosis is the fourth moment. The kurtosis of a normal distribuation is 3. A distribution with kurtosis < 3 and kurtosis > 3 are called platykurtic and leptokurtic, receptively. Compared with a normal distribution, the platykurtic has shorter and thinner tails and its central peak is lower and broader and vice versa. Score distributions with larger absolute values of the kurtosis (after normalized by minus 3, i.e., normalizing the kurtosis of the normal distribution to 0) have larger divergences from the normal distribution. Boxplots of the kurtosis of score distributions for images with mean scores within a specified range are shown in the right side of Figure 3.

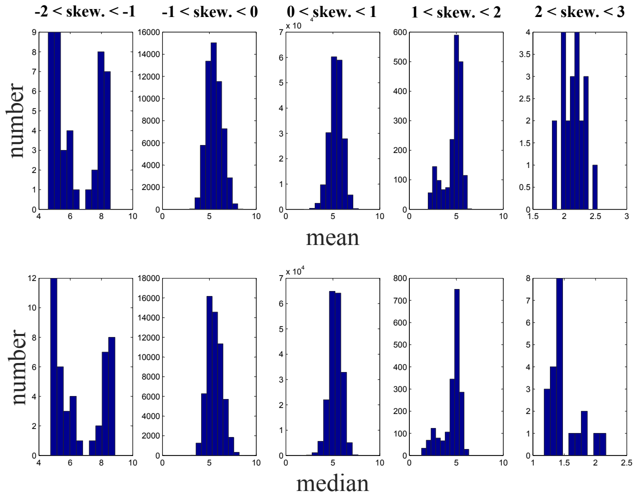


Figure 4: Distributions of mean and median of score distributions, for images with different skewness scores. The divergences between the mean and the median distributions tends to increase with the distance between the skewness values and 0, which is the skewness of the symmetrical normal distribution.

Within each range of the mean scores, there exist some images with high absolute values of kurtosis values (after normalized by minus 3), which are considered as those with unreliable score distributions.

CJS based CNN for Score Hist. Prediction

In this section, we introduce the proposed CJS-CNN (Cumulative distribution function with Jensen-Shannon divergence) and the reliability-sensitive learning method based on the kurtosis of the score distribution.

The Score Distribution Representation

With empirical data of the ordinal basic human ratings of an image from the AVA dataset, we use the score histogram to approximate the score distribution. We follow the definition in Wu et al. (Wu, Hu, and Gao 2011).

Assuming that there are Z ordinal basic ratings $R = \{R_1, ...R_Z\}$. In the AVA dataset, $Z = 10, R = \{R_1, ...R_{10}\}$. The human ratings for an image can be represented as $S = \{S(1), ..., S(L)\}$, where $S(i) \in R$ is given by the i^{th} person and L is the number of persons who have rated this image. (In the AVA dataset, $L \in [78, 549]$, with an average of 210). Then the score histogram or score vector of an image in the AVA dataset can be defined as:

$$\begin{aligned} y &= \{h(1), ..., h(i), ..., h(Z)\} \\ h(i) &= \frac{\sum_j^L \delta(S(j) = R_i)}{L}, \end{aligned} \tag{1}$$

where $\delta()$ is the indication function. With this representation, we can calculate the mean, median, variance, skewness, kurtosis using textbook methods.

The CJS-CNN

We use the first $1/3$ part of the GoogLeNet (layers before the first softmax layer) as our DCNN for fast training and extensive comparisons. We replace the full connected layer before the first softmax layer of the GoogLeNet with a output layer of $Z = 10$ dimensions. After each element of the output layer, we add a sigmoid layer to normalize each element to $[0, 1]$. The layers after the first softmax layer of the GoogLeNet are removed for fast training and comparisons.

The score vector defined by Eq. 1 can be considered as a vector. A straightforward way to calculate the loss is using the Euclidean distance. However, the score vector is an approximate of the underline probability distribution function (pdf). In addition, the score vector is built on the pre-defined ordinal basic ratings. Thus, a divergence between two cumulative distribution functions (cdf) is more appropriate for the loss function. Recently, Nguyen et al. (Nguyen and Vreeken 2015) propose a non-parametric Jensen-Shannon divergence, which performs well in detecting differences between distributions, outperforming the state-of-the-art methods in both statistical power and efficiency for a wide range of tasks. As verified by (Nguyen and Vreeken 2015), the CJS is quite suit for non-parametric computation on empirical data without estimating the underline distribution, such as the ordinal basic rating data of the AVA dataset. They define the asymmetrical continuous cumulative Jensen-Shannon divergence ($ACCJS(p(X)||q(X))$) of two continuous probability distribution functions $p(X)$ and $q(X)$ as follows.

$$\int P(x)log\,\frac{P(x)}{\frac{1}{2}P(x)+\frac{1}{2}Q(x)}dx+\frac{1}{ln2}\int(Q(x)-P(x))dx \tag{2}$$

The cumulative distribution function Y of the probability distribution function y defined by Eq. 1 is defined as follows.

$$Y(i) = \sum_{j=1}^{i} y(j) \tag{3}$$

CJS. Thus, we define the symmetrical discrete cumulative Jensen-Shannon divergence ($CJS(y_1||y_2)$) of two score histograms y_1 and y_2 defined by Eq. 1 as follows, derived from ($ACCJS(p(X)||q(X)) + ACCJS(q(X)||p(X))$).

$$\frac{1}{2}[\sum_{i=1}^{Z} Y_1(i)log\,\frac{Y_1(i)}{Y^s} + \sum_{i=1}^{Z} Y_2(i)log\,\frac{Y_2(i)}{Y^s}], \tag{4}$$

where Y_1 and Y_2 are defined by Eq. 3, $Y^s = \frac{1}{2}Y_1(i) + \frac{1}{2}Y_2(i)$. After that, we define our **CJS** loss function for the CJS-CNN as:

$$l^{CJS}(y, \hat{y}) = CJS(y||\hat{y}), \tag{5}$$

where y is the ground truth score histogram, and $\hat{y}$ is the predicted score histogram by our CJS-CNN.

The Reliability-sensitive Learning

In Eq. 1, the larger the rating number L is, the more reliable the distribution is. Wu et al. (Wu, Hu, and Gao 2011) use the rating numbers to model the reliability of the score distribution. In the AVA dataset, the number of votes that per image gets is ranged in 78-549 with an average of 210, which limits the performance of the rating number based reliability learning. Besides, one cannot obtain the rating numbers from normalized score histograms. If another dataset has only normalized score histograms, one cannot use the rating number for the reliability learning.

RS-CJS. We propose to use the kurtosis to measure the reliability of a score distribution y defined by Eq. 1. The kurtosis of a normal distribution is 3. Score distributions with kurtosis closer to 3 have smaller divergence from the normal distribution. Thus, inspired by Wu et al. (Wu, Hu, and Gao 2011), we define the reliability factor $r^{kurtosis}$ as follows.

$$r^{kurtosis}(y) = \mu(T(y)), T(y) = \frac{1}{|kus(y) - 3|}$$
$$\mu(T(y)) = \begin{cases} \frac{ln(T(y)+1)}{ln(T(y)+1)+1}, & T(y) < Th \\ 1, & \text{otherwise} \end{cases}, \tag{6}$$

where $r^{kurtosis}(y)$ equals to 1 if the kurtosis $kus(y)$ is sufficiently close to 3 and tends to 0 if $|kus(y) - 3|$ is very large. In practice, we add a small number ϵ to $|kus(y) - 3|$ to avoid the division by zero. We choose the threshold Th using cross validation. In practice, we use the percentage of the number of images above Th against the total number of the training images to determine Th. When the percentage equals 10%, we obtain the best performance of our score distribution prediction task (the other candidate percentages are: $5\%, 20\%, 30\%$).

Thus, the reliability-sensitive CJS loss is defined as:

$$l^{RS-CJS}(y, \hat{y}) = r^{kurtosis}(y) CJS(y, \hat{y}), \tag{7}$$

where y is the ground truth score histogram in the AVA dataset, and $\hat{y}$ is the predicted score histogram by our CJS-CNN. The more reliable the training image is, the more penalty it should obtain when the prediction is not correct.

Experiments

In this section, we present the experimental results in the AVA dataset. We follow the standard partition method of the AVA dataset in previous work (Murray, Marchesotti, and Perronnin 2012; Wang et al. 2016; Kong et al. 2016; Lu et al. 2015b; 2015a; Mai, Jin, and Liu 2016) . The training and test sets contain 235,599 and 19,930 images respectively. In all the experiments, for fair comparisons of various loss functions, we use the first $1/3$ part of the GoogLeNet as the DCNN. We show the predicted score histograms by our proposed CJS-CNN and other compared loss functions on the test set of AVA in Figure 5. Our CJS-CNN achieves the most similar results to the ground truth human rating distributions.

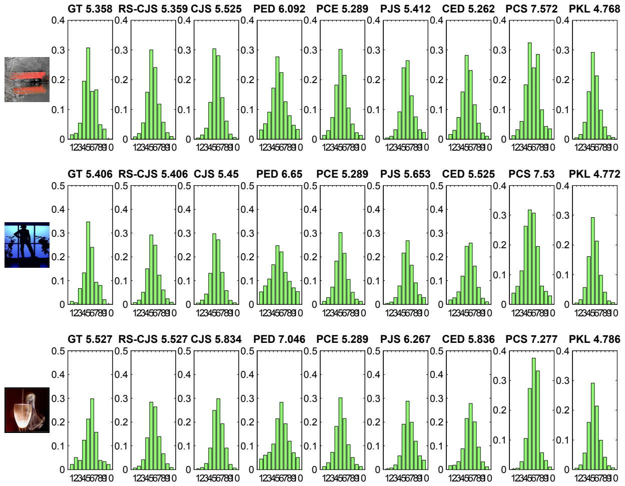


Figure 5: Predicted score histograms by the above loss functions. The numbers above each histograms are their mean scores. The first column is the images. The 2nd column is the human rating distributions (GT: Ground Truth). The 3rd and the 4th columns are the results predicted by our proposed RS-CJS and CJS based CNN, respectively. The other columns are the predicted results of other loss functions. Our results are more similar to the ground truth of human ratings than others. Images are from the AVA dataset (Murray, Marchesotti, and Perronnin 2012), which contains a list of photo IDs from www.dpchallenge.com.

Implementation Details

We fix the parameters of the layers before the first full connected layer of a pre-trained GoogLeNet model [1] on the ImageNet (Deng et al. 2009) and fine tune the 2 full connected layers on the training set of the AVA dataset. We use the Caffe framework (Jia et al. 2014) to train and test our models. The learning policy is set to *step*. Stochastic gradient descent is used to train our model with a mini-batch size of 48 images, a momentum of 0.9, a gamma of 0.5 and a weight decay of 0.0005. The max number of iterations is 480000. The training time is about 3 days using GTX980-Ti GPU and about 2 days using Titan X Pascal GPU.

Score Histogram Prediction and Comparisons

Baseline Loss Functions Besides the CJS loss function we proposed, we also evaluate other distribution divergences based loss functions as the baseline methods, parts of which are described below. These divergences or distances are often used in computer vision and pattern recognition tasks to compute the difference between two distributions or feature vectors. All the probability or cumulative distribution functions in our paper refer to discrete histograms. The DCNN cooperated with each divergence or distance based loss function is the first $1/3$ part of the GoogLeNet for fair comparisons.

[1] http://vision.princeton.edu/pvt/GoogLeNet/ImageNet/

Table 1: The mean divergences (MD, Eq. 14) between the predicted score histogram and the ground truth of various loss functions. The dataset is AVA. In all the methods listed below, the DCNN is the first 1/3 part of the GoogLeNet. The LDL method proposed by (Cui et al. 2017) does not use DCNN. Except the PED divergences, the performances of the other divergences we use are not reported in their work (Cui et al. 2017).

MD loss	PED	PCE	PJS	PCS	PKL	CED	CJS
PED	0.197	2.830	0.059	0.105	0.728	0.323	0.068
RS-PED	0.189	2.733	0.055	0.094	0.657	0.324	0.067
PCE	0.167	2.773	0.041	0.075	0.442	0.279	0.049
RS-PCE	0.169	2.771	0.046	0.071	0.438	0.279	0.047
PJS	0.185	2.828	0.051	0.093	0.527	0.326	0.053
RS-PJS	0.183	2.776	0.049	0.091	0.523	0.327	0.049
PCS (Jin, Segovia, and Süsstrunk 2016)	0.182	2.807	0.045	0.082	0.450	0.287	0.045
RS-PCS	0.175	2.783	0.045	0.079	0.423	0.277	0.044
PKL (Wang et al. 2017)	0.163	2.779	0.039	0.073	0.389	0.270	0.044
RS-PKL	0.164	2.778	0.037	0.071	0.386	0.268	0.043
MMD (Borgwardt et al. 2006)	0.201	2.831	0.064	0.112	0.710	0.339	0.068
RS-MMD	0.196	2.824	0.063	0.097	0.710	0.322	0.054
Huber (Murray and Gordo 2017)	0.184	2.775	0.044	0.078	0.409	0.279	0.053
RS-Huber	0.183	2.774	0.045	0.074	0.402	0.271	0.048
CED (Wu, Hu, and Gao 2011; Hou, Yu, and Samaras 2016)	0.182	2.799	0.047	0.085	0.502	0.294	0.049
RS-CED	0.180	2.792	0.048	0.082	0.502	0.283	0.047
Our CJS	0.163	2.779	0.039	0.072	0.382	0.266	0.041
Our RS-CJS	**0.158**	**2.760**	**0.037**	**0.068**	**0.381**	**0.260**	**0.040**
LDL Method (Cui et al. 2017)	0.303	-	-	-	-	-	-

Table 2: The ablation study of λ in Eq. 15. The dataset is AVA. The DCNN is the first 1/3 part of the GoogLeNet.

MD loss	PED	PCE	PJS	PCS	PKL	CED	CJS
$\lambda = 0$	0.159	2.760	0.037	0.068	0.387	0.260	0.040
$\lambda = 0.1$	0.159	2.764	0.038	0.069	0.384	0.262	0.040
$\lambda = 0.3$	0.159	2.762	0.038	0.069	0.386	0.262	0.040
$\lambda = 0.5$	0.160	2.766	0.038	0.070	0.386	0.264	0.040
$\lambda = 0.7$	0.158	2.761	0.037	0.068	0.385	0.261	0.041
$\lambda = 0.9$	0.159	2.763	0.038	0.069	0.384	0.262	0.040
Our RS-CJS ($\lambda = 1$)	**0.158**	**2.760**	**0.037**	**0.068**	**0.381**	**0.260**	**0.040**

PED. The loss function using the Euclidean distance of the two probability distribution functions is defined as:

$$l^{PED}(y, \hat{y}) = \sum_{i=1}^{Z}(y(i) - \hat{y}(i))^2 \tag{8}$$

PCE. The loss function using the cross entropy of the two probability distribution functions is defined as:

$$l^{PCE}(y, \hat{y}) = -\sum_{i=1}^{Z}[(y(i) log\, \hat{y}(i) + (1 - y(i)) log\, (1 - \hat{y}(i))] \tag{9}$$

This is the standard and widely used loss function for image classification problems and can be used as histogram difference for our task.

PJS. The loss function using the symmetrical version of the Jensen-Shannon divergence of the two probability distribution functions is defined as:

$$l^{PJS}(y, \hat{y}) = \frac{1}{2}[\sum_{i=1}^{Z} y(i) log\, \frac{y(i)}{m(y, \hat{y})} + \sum_{i=1}^{Z} \hat{y}(i) log\, \frac{\hat{y}(i)}{m(y, \hat{y})}], \tag{10}$$

where $m(y, \hat{y}) = \frac{1}{2}y(i) + \frac{1}{2}\hat{y}(i)$.

PCS (Jin, Segovia, and Süsstrunk 2016). The loss function using the Chi-square distance of the two probability distribution functions is defined as:

$$l^{PCS}(y, \hat{y}) = \frac{1}{2}\sum_{i=1}^{Z}\frac{(y(i) - \hat{y}(i))^2}{y(i) + \hat{y}(i)} \tag{11}$$

This loss function is proposed by Jin et al. (Jin, Segovia, and Süsstrunk 2016) to predict the mean score and standard deviation from the score distribution.

PKL (Wang et al. 2017). The loss function using the symmetrical version of the KullbackLeibler divergence of the two probability distribution functions is defined as:

$$l^{PKL}(y,\hat{y}) = \frac{1}{2}[\sum_{i=1}^{Z} y(i)log\,\frac{y(i)}{\hat{y}(i)} + \sum_{i=1}^{Z} \hat{y}(i)log\,\frac{\hat{y}(i)}{y(i)}] \quad (12)$$

The asymmetrical version of the KLD loss is used by Wang et al. (Wang et al. 2017), who explicitly modify the score distribution of the AVA dataset as Gaussian and jointly predict its mean and standard deviation.

CED (Wu, Hu, and Gao 2011; Hou, Yu, and Samaras 2016). The loss function using the Euclidean distance of the two cumulative distribution functions is defined as:

$$l^{CED}(y,\hat{y}) = \sum_{i=1}^{Z}(Y(i) - \hat{Y}(i))^2, \quad (13)$$

where Y and $\hat{Y}$ are the cumulative distribution functions of the original probability distribution functions y and $\hat{y}$, as defined in Eq. 3. This loss function is also used in Wu et al. (Wu, Hu, and Gao 2011) and can be derived from the squared Earth mover's distance (EMD) by Hou et al. (Hou, Yu, and Samaras 2016) for the ordinal basic human ratings prediction.

We show the predicted score histograms by our proposed CJS-CNN and other compared loss functions on the test set of AVA in the supplementary material. Our CJS-CNN achieves the most similar results to the ground truth human rating distributions.

Numerical Evaluation Results In Table 1, we summarize the evaluation results of the loss functions over the divergences. We use the Mean Divergences (MD) to evaluate various divergences between the predicted score histogram and the ground truth on the test set of AVA. The MD is defined as:

$$\text{MD} = \frac{1}{n}\sum_{i=1}^{n} l(y,\hat{y}), \quad (14)$$

where $l = \{l^{PED}, l^{PCE}, l^{PJS}, l^{PCS}, l^{PKL}, l^{CED}, l^{CJS}\}$ defined above. n is size of the test set.

The results in Table 1 reveal that, our proposed RS-CJS and CJS based CNN outperform other methods. Among all the odd lines with white background, our CJS achieves the best performance. All the mean divergences of our RS-CJS on the test sets are the smallest. Typically, in a learning setting, optimizing directly a certain criterion should lead to higher performance than optimizing a related one. However, although our methods are optimizing the CJS loss, the learned model can achieve best performance in other related loss. This is mainly because that, as verified by (Nguyen and Vreeken 2015), the CJS is quite suitable for non-parametric computation on empirical data, such as the ordinal basic rating data of the AVA dataset. The line with header 'RS-' means adding our reliability-sensitive learning strategy. Almost all the RS version methods (the even lines) perform better than the corresponding ones (the odd lines). The reliability sensitive learning based on the kurtosis reduces the impacts of the unreliable training samples.

The Ablation Study of the Reliability Factor Wu et al. (Wu, Hu, and Gao 2011) propose to use the number of ratings of each image for the reliability factor $r^{ratnum}(y)$. The larger the rating number is, the larger the reliability of rating is. To compare with our kurtosis based reliability factor $r^{kurtosis}(y)$ in Eq. 6 and Eq. 7, we use an balance factor λ as follow to make ablation study.

$$r(y) = \lambda r^{kurtosis}(y) + (1-\lambda)r^{ratnum}(y) \quad (15)$$

For a fair comparison, we use $r(y)$ on the CJS loss: $r(y)CJS(y,\hat{y})$

The comparison results are shown in Table 2. The results reveal that the performance of $r^{kurtosis}(y)$ are slightly better than that of $r^{ratnum}(y)$. The combination of these two reliability factors does not produce better performance. Note that, the kurtosis can be directly computed using the normalized score histogram. While the rating number is additional information of the normalized score histogram and is not always available in the training set.

Conclusions and Discussions

In this paper, we propose the CJS-CNN to predict the aesthetic score distribution of images. Unlike the object recognition, which definitely has right answers in most cases, the image aesthetic assessment is a subjective task in nature. Thus, only using a scalar to represent the aesthetics may not be the right direction.

Instead of only predicting the binary high-low label or the numerical score, we can output the aesthetic score distribution with rich statistics for various applications such as aesthetic image retrieval, aesthetic image enhancement. The overall aesthetic quality can be represented by the mean or median. The controversy or subjectiveness can be measured by the variance. The popularity of an image can be measured by the rating number of human. However, the rating number cannot be derived from the predicted score histogram. As shown in the experiments, we can use the kurtosis to approximate the popularity instead of the rating number.

The aesthetic quality assessment is a subjective task in nature. It has been a long time that people focused on the scalar representation (1D numerical or binary coding) of aesthetics. Wu et al. (Wu, Hu, and Gao 2011) pointed out this problem and made an attempt to predict the score distribution. However, it was submerged in rich literatures which aim to rise the classification or regression accuracy of the scalar representation. With the powerful deep representation learning technologies, we think it is the right time to let the aesthetic quality assessment return to it's subjective nature. This paper is a restart of this direction. We hope it can inspire more work in the future, such as (1) mapping more statistics to the subjective evaluation of vast amount of images, (2) designing new large scale aesthetic datasets with unbiased data and specially for subjective assessment of aesthetics, (3) using more powerful and larger DCNNs or other machine learning technologies to make the assessment by computer better match that of human.

Acknowledgments

We thank all the reviewers and ACs. This work is partially supported by the National Natural Science Foundation of China (Grant Nos. 61402021, 61401228, 61402463, 61772513), the Science and Technology Project of the State Archives Administrator (Grant No. 2015-B-10), the open funding project of State Key Laboratory of Virtual Reality Technology and Systems, Beihang University (Grant No. BUAA-VR-16KF-09), the Fundamental Research Funds for the Central Universities (Grant No. 3122014C017), the China Postdoctoral Science Foundation (Grant No. 2015M581841), and the Postdoctoral Science Foundation of Jiangsu Province (Grant No. 1501019A).

References

Borgwardt, K. M.; Gretton, A.; Rasch, M. J.; Kriegel, H.; Schölkopf, B.; and Smola, A. J. 2006. Integrating structured biological data by kernel maximum mean discrepancy. In *Proceedings 14th International Conference on Intelligent Systems for Molecular Biology 2006, Fortaleza, Brazil, August 6-10, 2006*, 49–57.

Brown, S. Measures of shape: Skewness and kurtosis. `http://brownmath.com/stat/shape.htm`.

Cui, C.; Fang, H.; Deng, X.; Nie, X.; Dai, H.; and Yin, Y. 2017. Distribution-oriented aesthetics assessment for image search. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, Shinjuku, Tokyo, Japan, August 7-11, 2017*, 1013–1016.

Deng, J.; Dong, W.; Socher, R.; Li, L.; Li, K.; and Li, F. 2009. Imagenet: A large-scale hierarchical image database. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), 20-25 June 2009, Miami, Florida, USA*, 248–255.

Deng, Y.; Loy, C. C.; and Tang, X. 2017. Image aesthetic assessment: An experimental survey. *IEEE Signal Process. Mag.* 34(4):80–106.

Dong, Z., and Tian, X. 2015. Multi-level photo quality assessment with multi-view features. *Neurocomputing* 168:308–319.

Hou, L.; Yu, C.; and Samaras, D. 2016. Squared earth mover's distance-based loss for training deep neural networks. *CoRR* abs/1611.05916.

Jia, Y.; Shelhamer, E.; Donahue, J.; Karayev, S.; Long, J.; Girshick, R. B.; Guadarrama, S.; and Darrell, T. 2014. Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the ACM International Conference on Multimedia, MM'14, Orlando, FL, USA, November 03 - 07, 2014*, 675–678.

Jin, X.; Chi, J.; Peng, S.; Tian, Y.; Ye, C.; and Li, X. 2016. Deep Image Aesthetics Classification using Inception Modules and Fine-tuning Connected Layer. In *The 8th International Conference on Wireless Communications and Signal Processing (WCSP)*, 1–6.

Jin, B.; Segovia, M. V. O.; and Süsstrunk, S. 2016. Image aesthetic predictors based on weighted cnns. In *2016 IEEE International Conference on Image Processing, ICIP 2016, Phoenix, AZ, USA, September 25-28, 2016*, 2291–2295.

Joanes, D. N., and Gill, C. A. 1998. Comparing measures of sample skewness and kurtosis. *Journal of the Royal Statistical Society: Series D (The Statistician)* 47(1):183–189.

Kao, Y.; He, R.; and Huang, K. 2017. Deep aesthetic quality assessment with semantic information. *IEEE Trans. Image Processing* 26(3):1482–1495.

Kao, Y.; Huang, K.; and Maybank, S. J. 2016. Hierarchical aesthetic quality assessment using deep convolutional neural networks. *Sig. Proc.: Image Comm.* 47:500–510.

Kao, Y.; Wang, C.; and Huang, K. 2015. Visual aesthetic quality assessment with a regression model. In *2015 IEEE International Conference on Image Processing, ICIP 2015, Quebec City, QC, Canada, September 27-30, 2015*, 1583–1587.

Karayev, S.; Trentacoste, M.; Han, H.; Agarwala, A.; Darrell, T.; Hertzmann, A.; and Winnemoeller, H. 2014. Recognizing image style. In *British Machine Vision Conference, BMVC 2014, Nottingham, UK, September 1-5, 2014*.

Ke, Y.; Tang, X.; and Jing, F. 2006. The design of high-level features for photo quality assessment. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), 17-22 June 2006, New York, NY, USA*, 419–426.

Kong, S.; Shen, X.; Lin, Z.; Mech, R.; and Fowlkes, C. 2016. Photo aesthetics ranking network with attributes and content adaptation. In *European Conference on Computer Vision (ECCV)*.

Lu, X.; Lin, Z.; Jin, H.; Yang, J.; and Wang, J. Z. 2014. RAPID: rating pictorial aesthetics using deep learning. In *Proceedings of the ACM International Conference on Multimedia, MM'14, Orlando, FL, USA, November 03 - 07, 2014*, 457–466.

Lu, X.; Lin, Z.; Shen, X.; Mech, R.; and Wang, J. Z. 2015a. Deep multi-patch aggregation network for image style, aesthetics, and quality estimation. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, 990–998.

Lu, X.; Lin, Z. L.; Jin, H.; Yang, J.; and Wang, J. Z. 2015b. Rating image aesthetics using deep learning. *IEEE Trans. Multimedia* 17(11):2021–2034.

Ma, S.; Liu, J.; and Chen, C. W. 2017. A-lamp: Adaptive layout-aware multi-patch deep convolutional neural network for photo aesthetic assessment. In *CVPR*, 722–731. IEEE Computer Society.

Mai, L.; Jin, H.; and Liu, F. 2016. Composition-preserving deep photo aesthetics assessment. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Murray, N., and Gordo, A. 2017. A deep architecture for unified aesthetic prediction. *ArXiv pre-print* abs/1708.04890.

Murray, N.; Marchesotti, L.; and Perronnin, F. 2012. AVA: A large-scale database for aesthetic visual analysis. In *IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, June 16-21, 2012*, 2408–2415.

Nguyen, H. V., and Vreeken, J. 2015. Non-parametric jensen-shannon divergence. In *Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2015, Porto, Portugal, September 7-11, 2015, Proceedings, Part II*, 173–189.

Wang, W.; Zhao, M.; Wang, L.; Huang, J.; Cai, C.; and Xu, X. 2016. A multi-scene deep learning model for image aesthetic evaluation. *Sig. Proc.: Image Comm.* 47:511–518.

Wang, Z.; Liu, D.; Chang, S.; Dolcos, F.; Beck, D.; and Huang, T. S. 2017. Image aesthetics assessment using deep chatterjee's machine. In *IJCNN*, 941–948. IEEE.

Wu, Y.; Bauckhage, C.; and Thurau, C. 2010. The good, the bad, and the ugly: Predicting aesthetic image labels. In *20th International Conference on Pattern Recognition, ICPR 2010, Istanbul, Turkey, 23-26 August 2010*, 1586–1589.

Wu, O.; Hu, W.; and Gao, J. 2011. Learning to predict the perceived visual quality of photos. In *IEEE International Conference on Computer Vision, ICCV 2011, Barcelona,Spain, November 6-13, 2011*, 225–232.