Research on Image Quality Evaluation Method of Depth Learning Model Based on Core Concept

Wen Huang*, Juan Wang
School of Cyberspace Security, Chengdu University of Information Technology, Chengdu 610225, China
*Corresponding author e-mail: willfreelyhw@gmail.com

Abstract. In view of the fact that the existing image quality evaluation methods are generally artificial design features, it is difficult to automatically and effectively extract image features that conform to the human visual system. Inspired by human visual characteristics, a new full reference image quality evaluation method based on depth learning model based on core concepts is proposed. Firstly, depth learning algorithm is used to extract multi-layer features from reference images and distorted images respectively. Then, the local similarity of the feature map of the reference image and the distorted image in each layer is calculated as the local quality description of the corresponding depth. Finally, the local quality of all layers is synthesized to obtain the overall quality score of the image. On the basis of the pre-training model, the depth model network is fine-tuned by using the image visual evaluation data set to obtain a depth model for evaluation. The standard experiment shows that fine-tuning training of each pre-training model on the standard data set achieves good classification results. Experiments show that the designed depth learning model based on core concepts is superior to the existing full reference image quality evaluation methods, and its prediction results have good accuracy and consistency with subjective quality evaluation.

1. Introduction
Images are subject to different degrees of distortion in the process of acquisition, transmission and compression, which leads to the decline of image quality. The quality of image will directly affect the ability of people and computers to obtain information [1]. In addition, due to the popularity of various intelligent mobile terminals, it has become more convenient for us to acquire images. We can record the people and things around us with our mobile phones or cameras at any time [2]. Images are the medium of information expression for status publishers, and often attract more attention than mere text status. In the news media layout, high-quality images are an essential way of presenting news [3]. On the other hand, more and more people are keen on photography. Photographic works are an important part of the world image database. Under this background, the research on objective image quality evaluation has become a hot spot in the field of image processing. Objective quality evaluation refers to the establishment of mathematical models through computer algorithms to accurately perceive the image quality, so as to finally achieve the use of computers to replace the human visual system to watch and recognize images. To solve this problem, a convolution neural network is used to generate a visual
difference map, which is weighted according to the local spatial characteristics of distorted images to evaluate the image quality. Therefore, the contributions of this paper are as follows: A new depth learning full reference image quality evaluation model based on core concepts is proposed. On the basis of a large number of image samples, this method can capture the characteristic information affecting the image quality very effectively and sensitively.

2. Depth Image Quality Evaluation Model

2.1. Network structure.
The deep learning model network framework of the core concept proposed in this paper includes 8 convolution layers to extract feature maps, 3 maximum pools to reduce the size of feature maps while extracting local strong features, and a nonlinear regression function composed of two full connection layers. Compared with the quality of the original reference image and other distorted images, people can easily estimate the relative quality of the test image [4]. In order to reduce the influence of image scale on features and fully consider the overall composition of the image, we use the idea of image pyramid to segment the image. Different local areas of distorted images suffer from different degrees of degradation, image gradients are sensitive to image distortion, and human eyes are sensitive to image gradients [5]. In the fields of image restoration and fusion, image quality evaluation technology can optimize the parameters of the objective function and guide the image fusion algorithm to obtain high-quality images. The purpose of this operation is to simulate the synthesis of images by human vision, and then adjust the coefficients of different frequency bands. Finally, the quality of distorted images is determined by comparing the coefficients of original images and distorted images. Then the extracted feature vector is input into DBN containing 3 hidden layers; Finally, a regression model between the features and DMOS values is established. In the testing phase, the corresponding objective quality evaluation values are predicted according to the established regression model.

2.2. Image preprocessing.
Before the image input, the normalization pretreatment of the image is helpful to reach the optimal value faster. Reference [6] mentioned that image normalization can improve the performance of the algorithm to some extent. The phase consistency and gradient series are used to calculate the local similarity mapping. The method has also obtained the best results at that time on many image databases. This method also proves the importance of HVS characteristics for image quality evaluation methods. The model is established by simulating various stages in the actual human visual system pathway [7]. The response of 2D Gabor filter in different scales and directions is similar to the mammalian visual processing process, which can accurately reflect the texture characteristics of the image. Each training instance displays a label distribution instead of a label or a related label set in the multi-label learning problem.

The description intensity corresponding to all labels of an example constitutes the label distribution. The learning process based on label distribution description is called label distribution learning algorithm. Figure 1 introduces the overall flow of LDL algorithm.

![Figure 1: The framework of label distribution learning algorithm](image-url)
Although no one has been able to define the image structure features accurately, the methods proposed according to the image structure features have verified their effectiveness, and such methods can obtain quality prediction results highly correlated with subjective quality scores. First, the reference image and the distorted image are converted into gray scale images. Then, the reference image and the distorted image are normalized to obtain image results [8]. Calculating Euclidean distances of singular values of blocks corresponding to the distorted image and the original reference image, wherein Euclidean distances of all blocks form a local distortion map, and taking an average value of absolute values of differences between Euclidean distances of each block and average Euclidean distances of all blocks as a standard for measuring image quality.

### 2.3. The definition of gradient difference graph based on the core concept of deep learning model.

Image gradient is highly sensitive to human visual system (HVS) because it can effectively capture the local structure of the image and is often applied to image quality evaluation algorithms. Since we can only extract a small number of features of the original reference image, it is required that the features are highly correlated with the image quality. Therefore, the key of some reference image quality evaluation methods lies in the selection of features. The label (single label or multiple labels) indication values used in the multi-label learning algorithm are manually constructed from the original data, and these labels are used as subsequent decisions. The distribution of labels originates from the real problem itself. Since the human eye is the final receiver of the signal, it is of great significance to establish an objective method that conforms to the characteristics of human vision. Scanning the image with blocks with scanning step size of one unit; Finally, the feature vectors of all blocks are combined to obtain the image features. Gradient difference map is defined by enhancing the gradient difference in the image, so that the depth learning model network of core concepts can better learn the edge information in the image. Gradient features are extracted by convolution of images with linear filters. The method extracts the features of the original reference image and the test image, which include the types and degrees of image distortion, as well as the definition of the image, and combines these features to obtain the quality score of the image. Although the description strength is not equal to probability, fortunately, the label distribution and the probability distribution have the same constraint conditions. Therefore, many theoretical methods of probability distribution are also applicable to label distribution learning problems.

### 3. Image Quality Evaluation Method Based on Depth Feature Similarity of Core Concept

#### 3.1. Calculation of local similarity.

We measure the local similarity between the features of the test image and the corresponding features of the original reference image through a sliding window with a size of \(d \times d\). The local similarity can be regarded as a quality index and is expressed in the form of a matrix, and we can regard it as a quality feature. Before that, there were many indicators to measure the similarity or distance between probability distributions, based on the fact that probability distributions and label distributions have the same form. Due to simple calculation, it is widely used, but the result is not very good because only the difference between pixels is calculated, which is inconsistent with human visual characteristics.

We use \(M_{i,k}\) to represent the mass characteristics of the \(k\)-th channel of the \(l\)-th layer of VGGNet, \(k = 1, 2, \ldots, k_l, l = 1, 2, \ldots, L\). \(Y = \{y_i | i = 1, 2, \ldots, n\}\) is used to represent the elements in the \(d \times d\) region centered on \((i, j)\) in the corresponding original reference image features, \(n = d \times d\), and the quality of the image at points \((i, j)\) can be calculated by the following formula [9]:

\[
M_{i,k}(i,j) = \frac{2\mu_{x,y} + c_1(2\sigma_{x,y} + c_2)}{(\mu_{x}^2 + \mu_{y}^2 + c_1)(\sigma_{x}^2 + \sigma_{y}^2 + c_2)}
\]  

(1)

We use the same settings as in [5], we use a Gaussian kernel \(w = \{w_i | i = 1, 2, \ldots, n\}\) with a size of \(d \times d\) and a standard deviation of 1.5 about center symmetry to correct the local statistics in the following way:
In order to abstract simple features, unsupervised pre-training model is used to adjust and update the weights of DBN layers. When the input reference image is the same as the distorted image, that is, the image has no distortion, and each point on the gradient difference map reaches the maximum value of 1. It is worth noting that when the input image is preprocessed, the output features of layers 32 to 35 are 4096-dimensional vectors, and the output features of layers 36 and 37 are 1000-dimensional vectors. In this case, the local similarity is calculated through a window with a size of $1 \times d$. Based on the obtained score distribution, we can obtain the average quality score as the image quality score by calculating the expectation of its distribution.

3.2. Quality integration.

In order to obtain the global quantification of image quality, we will pool the local quality features step by step in three steps to obtain the final integer, that is, the image quality score. Then, for the final image quality score, we can classify the quality by setting the appropriate score threshold. The local quality map is calculated by similarity measure, and then the quality score is obtained by pooling. Because the sensitivity of human visual system to different features is different, the importance of different features to the final quality score is different. The weights of the generated model are obtained through pre-training. Bayesian belief network is used in the part near the visual layer. In order to simulate the human visual system and strengthen the proportion of gradient in the image evaluation process, the gradient difference map is used to weight each pixel in the visual difference map to obtain the visual perception map [10]. First, the local similarity in each quality feature graph is pooled to obtain a scalar score. In this way, we obtain the quality scores of the L-level feature graph $K_l$ in the network. We express these scores by the following formula:

$$
q_{AVG} = \frac{1}{N} \sum_{i=1}^{N} q_i \quad s = \{s_l | l = 1, 2, \cdots, L\}
$$

Since each feature map corresponds to a convolution kernel, we call the obtained integer mass fraction convolution kernel mass. Then, the convolution kernel quality of each layer of the network is pooled to obtain a layer quality index, which is expressed by the following formula:

$$
m_l = \{m_{l,k} | k = 1, 2, \cdots, K_l, l = 1, 2, \cdots, L\}
$$

Finally, pool the quality indexes of L layer to obtain the final global quality score.

Before extracting depth features, we do a simple pretreatment on the image. Convolution neural network has strict size requirements on the input image, resulting in the loss of the image as a whole. In order to reduce the error caused by zero filling in the convolution process, the image boundary in the experimental process is cut. We use the idea of image pyramid to cut the image in order to capture the local and global information of the image at the same time. Similarly, other frequency domain features such as discrete cosine domain are also used as extraction features. Local features describe local block information of an image, such as entropy difference model based on information theory, singular value decomposition model, local fractal and other methods. Unsupervised learning is carried out from bottom to top. Each layer is regarded as an RBM. The weight of each layer is trained by greedy learning method, and training is carried out layer by layer from bottom to top.

In this method, we studied five common and simple pooling strategies: average pooling, standard deviation pooling, average absolute deviation pooling, full deviation pooling and percentage pooling.

Average pooling: The average value of the quality index list $Q$ is used as the global quality score of the image for average pooling, and the calculation formula is as follows:

$$
q_{AVG} = \frac{1}{N} \sum_{i=1}^{N} q_i \quad s = \{s_l | l = 1, 2, \cdots, L\}
$$

Percentage pooling: According to our usual subjective observation, the image quality depends more on the poor quality areas in the image, just like buckets effect. Therefore, we reasonably assume that we
can estimate the global quality score of the image by using the quality score of the region with the worst quality in the image. The corresponding operation formula for percentage pooling is as follows:

\[ q_{p} = \frac{1}{M} \sum_{i=1}^{M} q^{(worst)}_{i} \]  

(6)

Standard Deviation: As the name indicates, when using standard deviation, we use the standard deviation of the mass fraction to estimate the final mass fraction. The formula for standard deviation is as follows:

\[ q_{SD} = \frac{1}{N-1} \sum_{i=1}^{N} (q_{i} - q^{AVG})^2 \]  

(7)

Average absolute deviation pooling: average absolute deviation pooling takes the average absolute difference between a given local mass fraction and its average as the result, and the formula is as follows:

\[ q_{MAD} = \frac{1}{N} \sum_{i=1}^{N} |q_{i} - q^{AVG}| \]  

(8)

Total deviation pooling: the operation of total deviation pooling is to combine standard deviation pooling with average absolute deviation, and the formula is as follows:

\[ q_{FD} = \alpha q_{SD} + (1 - \alpha)q_{MAD} \]  

(9)

3.3. Training method.

Due to the different image sizes of LIVE databases, it is necessary to input the images into fixed-size blocks, and ensure that all image blocks of the same distorted image should be in the same batch. Note that the image blocks of the input distorted image correspond to the image blocks of the gradient difference map one by one. Considering the different number of images in the four databases, we take the number of images in each database as the weight of the three indexes calculated using the corresponding database. On the basis of the predicted score distribution, we can further excavate the information contained in the distribution and make a multi-dimensional evaluation of the image properties. We choose the expected value of the distribution as the quality score of the image and the distribution variance value as the degree of dispute over the image quality. In this process, attention should be paid to avoid cross overlapping of feature image blocks. Therefore, when acquiring distorted image blocks from distorted pictures, it is necessary to set an appropriate sliding step size, which is set to 80 in this paper. The average image of the training set is adjusted to the size of the input image, and then the adjusted average image is subtracted from the input image to perform the experiment in this way. If the quality scores of multiple images are the same, the image quality with smaller distribution variance is relatively high. Based on the predicted scores, we can classify the image quality by setting appropriate thresholds as classification criteria.

### Table 1 Specific parameters of in-depth learning model for core concepts

| Fusion  | Activation | Input channel | Output channel | Filter sizes | Stride |
|---------|------------|---------------|----------------|--------------|--------|
| Conv1-1 | LReLU      | 2             | 33             | (2,3)        | (1,1)  |
| Conv2-1 | LReLU      | 33            | 33             | (3,3)        | (1,1)  |
| Maxpool1| -          | 2             | 31             | (3,4)        | (2,2)  |
| Conv1-2 | LReLU      | 31            | 32             | (3,3)        | (1,1)  |

The specific parameters of the network architecture of the deep learning model model of the core concept are shown in Table 1. At the initial input, the gradient difference map and distorted image are input to Conv1-1 and Conv2-1 respectively. After MaxPool is performed respectively, the images are directly bonded together by Fusion and then input to Conv3. The specific transformation process of the network is shown in Table 1. Changing the size of the input image before input can improve the accuracy of the image quality score predicted by the depth learning model of the core concept in all pooling
operations. Moreover, changing the size of the input image makes the final performance of the depth learning model of the core concept less affected by the type of pooling operations. Through a series of experiments, the effectiveness of our proposed quality evaluation framework based on label distribution learning is verified from three angles of image score distribution prediction, image quality score prediction and quality classification. Under the condition of not using subjective quality scores, a model is established to predict the quality scores by using a depth learning method. The quality evaluation method is a plane image quality evaluation method and cannot be directly applied to stereo images, so the left and right image features are weighted to obtain input features, the network is trained end to end, the features are extracted by convolution layer, and nonlinear regression is performed by two completely connected layers to obtain the final quality score.

4. Summary

According to the sensitivity of human eye vision to gradient, a full reference image quality evaluation method based on depth learning model of core concepts is designed by using gradient difference graph. The model has been trained in image classification tasks. The trained model is used to extract the features of the image, and the quality estimation is obtained by comparing the features of the test image and the original reference image on each network layer. In our proposed method, we accept image features as input and output image quality score distribution. Compared with a single category label or score value, the image quality score distribution proposed by us can describe the quality score more comprehensively. Experimental results of perceptual feature analysis show that the results obtained by this feature extraction method are in good agreement with subjective evaluation values of human eyes, indicating that the proposed feature extraction method is of perceivable quality. The experimental results show that the proposed method can accurately predict the quality of stereo images, which is superior to the existing evaluation methods. To further improve the performance, conventional methods can be considered for preprocessing, followed by in-depth learning, so as to reduce training parameters and improve the operation speed of the algorithm.

References

[1] Yang S J, Berndl M, Ando D M, et al. Assessing microscope image focus quality with deep learning[J]. BMC Bioinformatics, 2018, 19(1):77.
[2] Hong X, Zan Y, Weng F, et al. Enhancing the image quality via transferred deep residual learning of coarse PET sinograms[J]. IEEE Transactions on Medical Imaging, 2018, PP(99):1-1.
[3] Yu F L, Sun J, Li A, et al. Image quality classification for DR screening using deep learning[J]. Conf Proc IEEE Eng Med Biol Soc, 2017:664-667.
[4] Yang J, Jiang B, Song H, et al. No-Reference Stereomage Quality Assessment for Multimedia Analysis towards Internet-of-Things[J]. IEEE Access, 2018, PP(99):1-1.
[5] Yang H F, Dillon T S, Chen Y P P. Optimized Structure of the Traffic Flow Forecasting Model With a Deep Learning Approach[J]. IEEE Transactions on Neural Networks and Learning Systems, 2017, 28(10):2371-2381.
[6] Ravi, Daniele, Szczoatka A B, Shakir D I, et al. Effective deep learning training for single-image super-resolution in endomicroscopy exploiting video-registration-based reconstruction[J]. International Journal of Computer Assisted Radiology & Surgery, 2018, 13(6):1-8.
[7] Gharbi, Michaël, Chen J, Barron J T, et al. Deep Bilateral Learning for Real-Time Image Enhancement[J]. Acm Transactions on Graphics, 2017, 36(4):118.
[8] Maltezos E, Doulamis N, Doulamis A, et al. Deep convolutional neural networks for building extraction from orthoimages and dense image matching point clouds[J]. Journal of Applied Remote Sensing, 2017, 11(4):042620-1-042620-22.
[9] Kwon K C, Kwon K H, Erdenebat M U, et al. Resolution-Enhancement for an Integral Imaging Microscopy Using Deep Learning[J]. IEEE Photonics Journal, 2019, PP(99):1-1.
[10] Chen J, Shen L, Zheng L, et al. Naturalization Module in Neural Networks for Screen Content Image Quality Assessment[J]. IEEE Signal Processing Letters, 2018, PP(99):1-1.