Bilinear-CNN for Non-reference Image Quality Prediction

by Kehan Zhang

B.S. in Electronical Engineering, June 2016, Beijing University of Posts and Telecommunications

A Thesis submitted to

The Faculty of
The School of Engineering and Applied Science
of the George Washington University
in partial fulfillment of the requirements
for the degree of Master of Science

May 19, 2019

Thesis directed by

Miloš Doroslovački
Associate Professor of Electrical and Computer Engineering
Acknowledgements

I would like to acknowledge and thank my thesis advisor, Professor Miloš Doroslovački, for his patient and encouragement throughout the entire thesis research course. I also thank the members of my thesis committee, Professor Kie-Bum Eom and Professor Hermann J. Helgert for their guidance and suggestions.
Abstract of Thesis

Bilinear-CNN for Non-reference Image Quality Prediction

Image quality assessment (IQA) is widely applied in many areas such as camera technology, digital television, stream video service. With the increase of computer networks transmission bit rate, the greater requirements for non-references image quality assessment in practice prompts the researchers to find an accurate algorithm.

Image recognition based on convolution neural network (CNN) has proved its superiority in computer vision and digital signal processing tasks. However, applying deep CNN in non-reference image quality assessment (NR-IQA) remains challenges due to the lack of sample database, superabundant sample categories, and other problems.

In this thesis, first, we provide an overview of several successful approaches to NR-IQA such as BRESQUE, FRIQUEE, and DIQA. And we obtain the conclusion and comparison of those algorithms’ advantages and disadvantages.

In addition, we introduce the Bilinear CNN structures which are well used in image recognition area. The approaches and flow chart of Bilinear CNN are given in detail.

Furthermore, we applied the structure of Bilinear CNN to patch-wise and image-wise methods in non-reference image quality prediction. The detailed training progress and CNN networks specifications are given. The performance of prediction and challenge problems in NR-IQA are discussed.
Table of Contents

Acknowledgements ........................................................................................................ ii

Abstract ........................................................................................................................... iii

Table of Contents ........................................................................................................... iv

List of Figures ................................................................................................................. vi

List of Tables .................................................................................................................. ix

Chapter 1 Introduction ................................................................................................. 1

1.1 Purpose of image quality assessment ................................................................. 1

1.2 IQA Method ............................................................................................................. 2

1.2.1 Subjective evaluation method ..................................................................... 2

1.2.2 Objective evaluation method ..................................................................... 5

1.3 Types of IQA ......................................................................................................... 5

1.3.1 Full reference method .............................................................................. 5

1.3.2 Reduced reference method ..................................................................... 9

1.3.3 Non-reference method .......................................................................... 10

1.4 CNNs for NR-IQA ............................................................................................. 11

1.4.1 Image-wise ............................................................................................... 14

1.4.2 Patch-wise ............................................................................................... 15

Chapter 2 Related Works in NR-IQA .......................................................................... 16

2.1 BRISQUE ............................................................................................................. 16

2.1.1 MSCN ....................................................................................................... 16

2.1.2 GGD ........................................................................................................ 19

2.2 FRIQUEE ........................................................................................................... 20
5.1 Performance on databases.................................................................................. 51
5.2 Performance on distortion types......................................................................... 51
5.3 Conclusion and discussion.................................................................................. 52

References ................................................................................................................ 55
List of Figures

Figure 1-1 SSIM generation ........................................................................................................ 8
Figure 1-2 SSIM example with test and reference images ....................................................... 9
Figure 1-3 Flow chart of image wise training ........................................................................... 14
Figure 1-4 Flow chart of patch wise training ....................................................................... 15
Figure 2-1 Comparison of original and MSCN results ......................................................... 18
Figure 2-2 Original and MSCN images .................................................................................. 18
Figure 2-3 Flow chart of GGD ................................................................................................. 20
Figure 2-4 Original and DoG feature images ......................................................................... 22
Figure 2-5 Original and Yellow channel images ................................................................... 23
Figure 2-6 Original, LAB space, and Chroma feature images ............................................ 24
Figure 2-7 $\hat{l}$, $\hat{a}$, and $\hat{b}$ images .................................................................................. 26
Figure 2-8 Flow chart of DIQA ............................................................................................. 28
Figure 2-9 CNN network used in DIQA ................................................................................ 28
Figure 2-10 Reliable and error maps in DIQA ......................................................................... 30
Figure 3-1 Last convolution output ......................................................................................... 33
Figure 3-2 Original images and last conv. layer output ........................................................ 34
Figure 3-3 Flow chart of bilinear CNN .................................................................................. 35
Figure 3-4 Gradient of bilinear vector .................................................................................. 36
Figure 4-1 Example images in LIVE IQA ............................................................................ 37
Figure 4-2 Example images in LIVE Challenge .................................................................. 38
Figure 4-3 Flow chart of VGG nets ....................................................................................... 41
Figure 4-4 Examples of VGG last conv. layer output ........................................................... 42
Figure 4-5 Flow chart of bilinear with image wise training................................. 45
Figure 4-6 Flow chart of bilinear with patch wise training.................................. 47
Figure 4-7 Testing image with sixteenth layer output image.................................. 48
List of Tables

Table 1-1 Classification levels of all-optional scale ........................................................... 3
Table 1-2 Classification levels of cluster excellence scale ................................................... 3
Table 2-1 Comparison of algorithms.................................................................................. 32
Table 4-1 Comparison of different database....................................................................... 38
Table 4-2 Detail parameters of VGG nets.......................................................................... 40
Table 4-3 Detail data size of each layer............................................................................. 47
Table 4-4 Normalized patches scores.............................................................................. 48
Table 5-1 Prediction results on different database......................................................... 51
Table 5-2 Prediction results on different distortion types............................................... 52
Chapter 1 – Introduction

1.1 Purpose of image quality assessment

Image quality prediction, also referred as image quality assessment (IQA), is mainly based on analyzing characteristics of images as well as estimating the image quality (the degree of distortion). Since image quality score is an essential part of algorithm analysis and performance evaluation of signal processing system, image quality prediction, as the basic techniques in image processing, has been comprehensively applied in the field of camera imaging technology, digital television, streaming media services as well as social media.

Image quality prediction is trying to accurately measure the digital image quality after being transported from distance, decompressed from condensation, generated by camera and so forth. Attenuation, distortion, and noise frequently happen in signal processing of digital images. For instance, White Noise(WN), Fast Fading(FF), and Gaussian Blur(Blur) are common distortions in image processing. Since distortion is inevitable, the reliable IQA method can be helpful in estimating image quality and confirming necessity in regard to reproduction or retransmission without human’s observation. Image quality assessment can accurately predict human’s subjective judgment about the picture and thus can be utilized to extremely increase the human’s satisfaction of the servers. By automatically monitoring images or videos, the servers are able to affect the viewing of image quality with several image processing algorithms, like compression engine, denoising and other algorithms system.

In recent years, with increasingly higher requirements of high-quality digital images for human customers, the study of image quality assessments have drewed
increasing attention from researchers, and many significant indicators and approaches of IQA have been established and further improved.

1.2 IQA Method

The image quality evaluation method including subjective evaluation method and objective evaluation method. On the one hand, subjective evaluation takes people as observers, and makes subjective evaluation on the image, striving to reflect people's visual perception truly. On the other hand, the objective evaluation method tries to reflect the subjective perception of human’s eyes by some mathematical model, obtaining the results based on numerical calculation.

1.2.1 Subjective evaluation method

Subjective evaluation takes people as observers to conduct subjective qualitative evaluation on the quality of images. Those observers are generally made up of untrained observers or trained observers. Such method is based on statistics conclusion in general. It’s worth to mention that since the subjective evaluation of images requires statistically significant meaning, the sufficient number of observers should participate in the evaluation. Furthermore, subjective evaluation method includes absolute evaluation method and relative evaluation method.

Absolute evaluation is the evaluating method that scores the absolute image quality according to certain specific evaluation performance based on the observer's own knowledge and understanding. To evaluate the quality directly, Double Stimulus Continuous Quality Scale (DSCQS) [1] is applied to obtain the absolute evaluation of image quality. The specific approaches are as follows: First of all, the testing image and the original image should be shown alternately to the observer in a certain period of time
according to the rules. Subsequently, a certain time interval should be left after the image is shown to the observer to score. Finally, the average of all scores should be calculated as the evaluation quality value of the image. There are several international regulations as the evaluation scales. The image quality is graded and represented by levels, which is also known as the 5-point "all-optimal scale" of image evaluation (see table 1-1).

### Table 1-1

<table>
<thead>
<tr>
<th>Rating point</th>
<th>Label</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>Excellent</td>
<td>There is no sign of distortion in the quality of the images</td>
</tr>
<tr>
<td>4</td>
<td>Good</td>
<td>There are signs of distortion, but no hinder to view</td>
</tr>
<tr>
<td>3</td>
<td>Fair</td>
<td>There are signs of distortion, and hinders the viewing slightly</td>
</tr>
<tr>
<td>2</td>
<td>Pool</td>
<td>Hinder the viewing</td>
</tr>
<tr>
<td>1</td>
<td>Bad</td>
<td>Hinder the viewing seriously</td>
</tr>
</tbody>
</table>

In the relative evaluation method, there is no original image as reference. A batch of testing images are compared by the observer to estimate the order of each image and present the corresponding evaluation value. In general, Single Stimulus Continuous Quality Evaluation (SSCQE) is utilized for judgment. Specifically, a batch of images are shown in a certain sequence, at which the observer gives the corresponding evaluation score of each image while viewing the images. Similar to the subjective absolute evaluation, the subjective relative evaluation provides a corresponding scoring system which is known as the "cluster excellence scale" (see table 1-2).

### Table 1-2

<table>
<thead>
<tr>
<th>Rating point</th>
<th>Label</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>Excellent</td>
<td>The best of the bunch</td>
</tr>
<tr>
<td>4</td>
<td>Good</td>
<td>Slightly better than average</td>
</tr>
<tr>
<td>3</td>
<td>Fair</td>
<td>Average of the bunch</td>
</tr>
<tr>
<td>2</td>
<td>Pool</td>
<td>Slightly worse than average</td>
</tr>
<tr>
<td>1</td>
<td>Bad</td>
<td>The worst of the bunch</td>
</tr>
</tbody>
</table>
The research of IQA aims to accurately identify and measure the image quality value. Mean opinion score (MOS), as the arithmetic mean of all individual observer’s scores, is usually collected in subjective quality assessment tests and is applied to represent the overall image quality. Although MOS is commonly applied in audio and audio-visual quality assessment at first, it is not limited to these methods. Nowadays, image and video quality evaluation also use MOS for the labeling of the quality score. The function of MOS is here:

\[
\text{MOS} = \frac{\sum_{n=1}^{N} R_n}{N}
\]

Where \( R_n \) are the individual ratings of image given by \( N \) observers. Nonetheless, generating a MOS rating can be time-consuming and expensive since from the aspect of statistics, sufficient number of human evaluators is required. For various cases of image quality assessment, researchers usually utilize the MOS values of samples to train and develop prediction models. The closer artificial MOS value to real MOS, the more accurate the algorithm is. One of the questions raised by these models is whether the MOS difference is significant to the user. For example, when an image is rated at a 5-point MOS scale, the expected quality of an image with MOS which equals to 5 is significantly better than the one with MOS which equals to 1. By contrast, it is not clear that the quality of MOS which equals to 3.8 is significantly better than the quality of MOS which equals to 3.6. The studies on the determination of the minimum MOS difference perceived by users in digital photos show that 75% of users need approximately 0.46 MOS difference to detect higher quality images[2]. However, image quality expectation or MOS varies with users’ expectations. To accurately measure the image, researcher uses MOS for IQA with the formed scale from 1 to 100 responding to the rating point scale from 1 to 5. With the larger scale of MOS, the
differences between two images can be measured more accurately in the IQA image database.

1.2.2 Objective evaluation method

The objective evaluation method focuses on the scoring of image quality through mathematical models or algorithms. In this way, human observer’s evaluation of the test image can be approximately reflected. In other words, the purpose of the objective evaluation method is to simulate human’s image quality evaluation through mathematical model. When compared with subjective evaluation, objective evaluation is featured by the batch processing and reproducible characteristics without deviation incurred by human factors.

In fact, the current researches on image quality assessment mainly emphasize on objective evaluation methods. To simulate the results closing to subjective evaluation, algorithms and models are trained according to the database that contains subjective evaluation scores.

1.3 Types of IQA

The objective evaluation of image quality can be reduced to three categories: Full-reference image quality assessment (FR-IQA), Reduced-reference image quality assessment (RR-IQA) and Non-reference image quality assessment (NR-IQA).

1.3.1 Full-reference method

Full-reference method means test image are evaluated with the reference image. FF-IQA focus on analyzing the distortion degree of the image under the condition that the original image is selected as the reference image. The commonly-used objective evaluation
of all-reference image quality is mainly based on three aspects, including pixel statistics, information theory, and structure information.

For the FF-IQA, Peak Signal to Noise Ratio (PSNR) and Mean Square Error (MSE) are two frequently used quality estimation methods to measure image quality by calculating the differences between the gray value of the pixels in the test image and the reference image. Let testing image be F, the reference image be R, and their size be M*N. i and j are pixel coordinates in image. Then the calculation method to characterize the image quality by MSE is:

$$MSE = \frac{1}{MN} \sum_{i=1}^{M} \sum_{j=1}^{N} |R(i,j) - F(i,j)|^2$$

And relation function of MSE and PSNR is:

$$PSNR = 10\log_{10}\left(\frac{MAX_i^2}{MSE}\right)$$

Since 255 is maximum intensity value, the calculation method to characterize the image quality by PSNR (in dB) is as follows:

$$PSNR = 10\log_{10}\left(\frac{255^2}{\frac{1}{MN} \sum_{i=1}^{M} \sum_{j=1}^{N} |R(i,j) - F(i,j)|^2}\right)$$

Both PSNR and MSE measure the image quality by calculating the pixel differences between testing image and reference image. The larger the peak signal to noise ratio value is, the smaller the distortion between the two images is. Similarly, the smaller the MSE value, the better the image quality. MSE and PSNR are simplest and easiest to implement and are extensively used in image denoising. Nevertheless, such type of
algorithm focuses on the global statistics of image pixel values without considering the local vision factor of human’s eyes. Thus, it is impossible to truly grasp the image quality.

For the evaluation based on information entropy in the information theory, mutual information is commonly used in the estimation of the image quality value. Sheikh and Bovik et al [3] proposed two algorithms, including Information Fidelity Criterion (IFC) and Visual Information Fidelity (VIF). Image quality method are measured by calculating the related information between testing and reference images. These two methods have certain theoretical support and extend the connection between images and human’s eyes in terms of information fidelity. Nevertheless, these methods do not respond to the structural information of images.

For the evaluation based on structural information, Wang Zhou and Bovik et al. [4] introduced a novel concept about image structural information for the first time on the basis of the research works on image processing, image compression and image visual quality evaluation. They believe that the human vision system (HVS) is usually recognize the images by extracting structural information, and the human eyes can achieve this goal highly and adaptively. Thus, it should be significant for the objective evaluation to achieve the evaluation of the structural information in the image. On this basis, a standard evaluation method of image quality, Structural Similarity (SSIM), is provided.

SSIM constructs structural similarity according to the correlation between image pixels. Given two images A and B, X and Y are the small N*N windows in the same locations of A and B where the mean value, standard deviation, and the covariance of X and Y are represented by \( \mu_x, \mu_y, \sigma_x, \sigma_y, \) and \( \sigma_{xy} \), respectively. The comparison functions that define luminance, contrast and structure are as below:
\[ l(x, y) = \frac{2\mu_x\mu_y + c_1}{\mu_x^2 + \mu_y^2 + c_1} \]

\[ c(x, y) = \frac{2\sigma_x\sigma_y + c_2}{\sigma_x^2 + \sigma_y^2 + c_2} \]

\[ s(x, y) = \frac{\sigma_{xy} + c_3}{\sigma_x\sigma_y + c_3} \]

The normal constant numbers \( c_1, c_2, c_3 \) are adopted to adjust the instability of the denominator near zero. And it is worth mentioning that an image is typically calculated with the usage of a Gaussian window or a block window.

With the realization of the structural similarity theory, the mean value has been used as the estimation of luminance, variance as the estimation of contrast, and covariance as the measurement of structural similarity. The above-mentioned three components combined are SSIM index and Figure 1-1 shows the generation of SSIM.

\[
SSIM = [l(x, y)]^\alpha [c(x, y)]^\beta [s(x, y)]^\gamma
\]

Figure 1-1
As we can see, function shows that the mean, variance, and covariance of X and Y are combined to generate the structural similarity. The larger SSIM value is, the better the image quality is. And the example of SSIM image with the testing image quality score is shown in figure 1-2.

![Figure 1-2](image)

It is simple to implement the index algorithm in quality evaluation. Meanwhile, since many researchers have improved it by combining with the human visual system, it has been extensively applied in many different aspects of image processing.

1.3.2 Reduced reference method

Reduced reference, also called partial reference, means that takes partial feature of the reference image as a reference to compare and analyze the testing image and then obtain the evaluation results of testing image quality. Since the reference information is an extracted feature from the image, it should first extract partial features of testing image and the reference. Also, the reference methods are classified into three types, including the method using original image feature, the method based on digital watermarking and the method using the statistical algorithms in the Wavelet domain [5]. Since the RR-IQA
counts on multiple features of the reference image and comparison with the whole testing image, used data size has decreased a lot. Nowadays, the current application is concentrated in the system of image transmission.

1.3.3 Non-reference method

The non-reference IQA method is also called the blind IQA method. Since it is hard to get the reference image in real world evaluation, such type of image quality assessment method without any ideal reference image has been widely used. The non-reference methods are generally based on the mathematic algorithms to predict image quality.

In recent years, More and more researches attempt to apply the mathematical models and algorithms to truly predict the image quality of images by learning subjective scores given by human observers. Many previous algorithms are hitherto based on the theorem of FR-IQA and RR-IQA. Nevertheless, NR-IQA will be more and more required in practice since the source image references are not available.

With the rapid improvement of bit rate in computer networks transmission, greater requirements for non-reference image quality assessment in practical application prompt researchers to find accurate algorithm. Plenty researches based on traditional machine learning and feature extraction have been taken in non-reference IQA. And several of them provide successful works with high accuracy prediction results.

Non-reference IQA is also referred to as Blind IQA. It is an evaluation method that the image quality is estimated according to its own characteristics and value without reference image at all. In image prediction, there are varieties of specific distortion types that have been reliably predicted. In addition, there are different directions in non-reference image quality assessment. Some methods focus on accurately classifying the distortion
cause before quantitative evaluation, while some methods try to evaluate images with different distortion types at the same time. After all, the non-reference method is the most practical approach with amount of applications.

Significant researches are conducted to improve the prediction performance of the NR-IQA. Many NR-IQA algorithms proposed by the researchers used the application of machine learning combined with feature extraction. Traditional machine learning like support vector machines (SVM) and support vector machine regression (SVR) were involved in NR-IQA. Researches have shown that the performance of traditional machine learning largely counts on the design of the feature. And NR-IQA has made significant progress after some difficulties appeared in obtaining the reliable features was solved by Natural scene statistics (NSS). Nowadays, some NR-IQA studiers are trying to apply deep learning for NR-IQA. Although traditional machine learning has good performance in several classic databases and deep learning is still confronted with great challenges in NR-IQA, lots of current studies attempt to use hand-crafted features with deep learning models rather than traditional machine learning.

1.4 CNNs for NR-IQA

Due to strong representation capability and impressive performance, convolutional neural networks (CNNs) show significant performances for CNN application to varieties of image recognition tasks. There's no denying that CNNs have become the most popular nowadays in the field of deep learning.

Although CNNs have been shown to be effective in performing outstanding applications on different works of visual information, such rapidly developing technique has not been systematically applied to image quality prediction until recently, mainly due
to the limitation of subjective data size as well as superabundant sample categories. As data collection methods and novel approaches are evolved based on deep learning for the perception and prediction of image quality, an increasing number of researchers have turned their attention to the application of CNN in image quality prediction.

In recent years, researchers try to utilize deep learning technology to NR-IQA problem. And several useful deep learning models have proven its significance in IQA. Li et al. [7] deduced the characteristics associated with NSS by Shearlet-transform generation. Subsequently, the NSS features are normalized to the subjective score with the use of an encoder. Similarly, Hou et al. utilized DBN [20] in NSS features extraction in the wavelet domain and then used them as input for deep learning model. Ghadiyaram and Bovik [8] tried to combine multiple NSS features with DBN training to predict subjective scores accurately. Even though most studies use deep learning models instead of traditional regression machines, algorithm involves the design of small hand-crafted functional neural networks that are not deep enough to make good use of deep learning application. Kang et al. [21] firstly used convolution neuron network deep learning to the non-reference image quality prediction with the purpose of performing better performance. They abandoned the extraction of features. To solve lack of database, they decided to divided testing image into small patches. Then they proposed a method which called patch-wise training. The mean opinion score (MOS) of the image is assigned for all the patches training. However, patch-wise training cannot response to the attributes of human vision system (HVS) which are perceived fine changes reflecting the differences of the testing and reference image. Bosse et al. [22] applied a 12 layers CNN network and proposed an additional model that can understand and express the significance of each patch. Several studies were trying to solve
the issue of the lack of sample sets, the method of generating recognizable images has also been tried. Jongyoo Kim et al. [13] proposed DIQA to solve the problem by using the reference image to generate error maps and reliable maps in training. In the mathematic aspect, DIQA have no relationship with complex matrix from full-reference image quality assessment. Furthermore, DIQA uses the convolution layer in the pre-training stage and the second training stage, hence the model is deeper than the previous research.

Some CNNs algorithms in this rapidly growing field are analyzed in detail in chapter 2 of this thesis. The performance of CNN model highly related with the number of the training samples. Nowadays, the available image quality databases, such as LIVE IQA, TID2013, LIVE Challenge and so on, are insufficient for the training. For instance, the LIVE IQA database (nearly 400 images) includes nearly 30 original images and five different types of distortion. Those distortion images are composite images generated by the database designer. Several common distortions, such as Gaussian blur (GB), average displacement, and white noise are included in the databases. Nevertheless, it is different from image damage actually appeared in the real digital photos. Even the JPEG/JP2000-encoded type distortion images are compressed by a more free-form way when compared with practice. As for the recent LIVE challenge database [3], the size of the available resources (nearly 1200 different photos, unknown combinations estimated by nearly half of million individual human observer’s subjective scores) is greatly expanded across most dimensions. Although the database presents a good challenge for researchers, there is no reference image for further study. By comparison, the image recognition database is far less than the database containing tens of millions of tagged images, such as ImageNet[14].
Nevertheless, creating a larger set of subjective image quality data is still a tricky problem as the study of subjective data consumes a lot of human resources.

### 1.4.1 Image-wise

So as to overcome the overfitting problem, most researchers have used the method of horizontal flip and divide the image into small patches. Those methods are trying to enlarge the sample numbers. The processing of patches before the full connection layer divide CNN into two branches. One is called the image-wise method, and another is called the patch-wise method.

In general, the difference between the image-wise method and the patch-wise method lies on the image processing method before training samples. To be specific, the difference lies in the sample processing after the CNN convolution layer and before the fully connection layer. The image-wise approach is that all patches of the paper image after the CNN convolutional layer shall be connected together again with the usage of aggression and pooling, and then further learning and training will be conducted for the full connection layer and the soft-max layer. The image-wise training process is shown in Figure 1-3.

![Image-wise training diagram](image.png)
The main steps of image-wise training are: At first, an input image would be divided into several patches; In addition, those patches are passing through same CNN net; In image-wise training, the convolution layer outputs of each patch are combined together with the method of aggression and polling. And then the combination with whole image quality score will continue to further training.

1.4.2 Patch-wise

The patch-wise method adopted to train the prediction models of the patch-based picture quality is similar to image-wise training. Firstly, divided patches are also passing through one CNN net. Then each patch will be assigned with its patch score which is generated by global subjective score or local score. As the samples in training are actually the image patches, the number of samples is enlarged significantly. The flow chart of patch-wise training is shown in Figure 1-4.
2.1 BRISQUE

Anish Mittal et al.[6] proposed a pretty early NR-IQA model using traditional machine learning and feature extraction, which is called Blind/Referenceless Image Spatial Quality Evaluator (Brisque). Instead of calculating any particular distortion feature, the model calculates the mean subtracted contrast normalized (MSCN) coefficients of the image by locally normalized luminance coefficients. The basic features are combination of Gaussian distribution and local luminance normalization. When compared with the previous reference IQA method, the Brisque model does not use another coordination domain and contains the simple structure. In comparison with the Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity (SSIM) for image quality prediction algorithms, Brisque is ideal for real-time applications since it has low computational complexity.

2.1.1 MSCN

The general principle of the Brisque algorithm is to extract the MSCN coefficients from images, and then fit histogram of MSCN map with asymmetric generalized Gaussian distribution (AGGD) as well as extract the characteristics of the fitted Gaussian distribution. The evaluation results of the image quality are obtained by inputting them into SVR. The Brisque method of NR-IQA can be summarized as follows. For the testing images, the local normalized average brightness subtraction and division normalization are first computed. MSCN coefficients can be applied to testing image intensity $I(i,j)$ to generate the below formula:

$$MSCN = \frac{I(i,j) - \mu(i,j)}{\sigma(i,j) + C}$$
where \( i \in 1,2,\ldots M, j \in 1,2,\ldots N \) refer to coordinate index, \( M \) and \( N \) refer that image should be \( M \times N \) size, \( C \) is the constant value used to prevent instabilities. And the functions of \( \mu(i,j) \) and \( \sigma(i,j) \) are:

\[
\mu(i,j) = \sum_{k=-K}^{K} \sum_{l=-L}^{L} w_{k,l} I_{k,l}(i,j)
\]

\[
\sigma(i,j) = \sqrt{\sum_{k=-K}^{K} \sum_{l=-L}^{L} w_{k,l} (I_{k,l}(i,j) - \mu(i,j))^2}
\]

where \( w = \{w_{k,l}|k = -K, \ldots, K, l = -L, \ldots, L\} \) is a 2D Gaussian weighting window with size of \( 7 \times 7 \) and the its variance are equal to unit. And \( K, L \) equal to 3, which means a \( 7 \times 7 \) size window is employed. Graph blew shows the distribution of \( 7 \times 7 \) window used.

The advantage of MSCN coefficients is that the region correlation caused by texture is very small. The difference between the left (original pixels) and right (MSCN) results is shown in Figure 2-1 and the original colored image (a), intensity mean image (b), \((I - \mu)\) image (c), variance image (d) and MSCN images(e) are presented in Figure 2-2.
2.1.2 GGD

Those images show that MSCN has no strong dependence on the texture strength. Thus, the extracted features are more applicable for image quality prediction since the generalized Gaussian distribution (GGD) would be utilized to fitted with histogram of the MSCN coefficient map effectively. It usually shows the variation coefficient distribution in the tail behavior, where GGD with zero mean value is given by:

$$f(x; \alpha, \sigma^2) = \frac{\alpha}{2\beta \Gamma(1/\alpha)} \exp\left(-\frac{|x|}{\beta}\right)^\alpha$$

Where,

$$\beta = \sigma \sqrt{\frac{\Gamma(1/\alpha)}{\Gamma(3/\alpha)}}$$

And $\Gamma(\cdot)$ is the Gamma function:

$$\Gamma(a) = \int_0^{\infty} t^{a-1} e^{-t} dt \quad a > 0$$

Where $\alpha$ used to control the distribution shape whereas $\sigma$ refers to variance. After obtaining feature maps, the next step is applying the regression module from feature space to quality fraction in order to generate the model. The detailed progress is: GGD model was applied to fit with histogram of MSCN map, to obtain statistical features of MSCN coefficient $(\alpha, \sigma^2)$. A support vector regression (SVR), that usually used in image quality prediction, is applied in Brisque. The flow chart of BRISQUE is showed in figure 2-3. It’s worth to mention that Brisque also applied different empirical distribution of adjacent MSCN coefficients to get more features.
In general, Brisque is traditional machine learning with feature extraction of MSCN and AGGD. And the evaluation results show that Brisque is statistically superior to PSNR and SSIM according to the correlation with human’s perception.

**2.2 FRIQUEE**

Deepti Ghadiyaram et al. [8] proposed one of the most effective NR-IQA evaluation algorithms. When compared with the previous works that most non-reference IQA models are usually trained with the database of MOS of artificially distorted images, feature maps based on referenceless image quality evaluation engine (FRIQUEE)[23] trying to accurately predict quality of nature distortion images that always contain combination of multiple types of distortion. By studying the perception-related natural scene statistics of the real-distorted images, they are able to effectively predict human’s visual quality judgment as well as the general isolation (single) distortion. Using a large real distortion
image database with human’s perception (subjective score) of it and computing a large number of features, they train a regression for image quality prediction. As for feature extraction, FRIQUEE contains 3 types of feature extraction method for several detailed features.

2.2.1 Luminance feature maps

Luminance map: Given an image I which is intensity image, the image size is M*N*3. The general progress is that: First of all, its luminance component diagram L is extracted; Then L is processed to produce normalized luminance MSCN results; A generalized Gaussian distribution model is applied to simulate the MSCN coefficient images distribution. Similar as BRISQUE, the algorithm estimates the ($\alpha$, $\sigma$) feature parameters and (kurtosis, skewness) parameters. Therefore, nearly 8-dimensional features are taken into consideration.

Neighboring paired products: This algorithm simulates the statistical relationship between neighborhoods as well as calculates four neighborhood product graphs from the testing images. AGGD distribution fitting is computed and sample parameters (kurtosis and skewness) are extracted on those four graphs. The algorithm extracts these features on two scales. Thus, 48-dimensional features in total are taken into consideration.

Sigma DoG map: Since the sigma coefficients of the gray-scale images show that regular structure and distortion can interfere the sigma coefficients of the image structure. Therefore, the algorithm can extract the skewness, kurtosis, and mean of sigma on two scales. A DoG model is generating the two-dimensional difference of the individual Gaussian filters:
\[ DoG = \frac{1}{\sqrt{2\pi}} \left( \frac{1}{\sigma_1} e^{-\frac{(x^2+y^2)}{2\sigma_1^2}} - \frac{1}{\sigma_2} e^{-\frac{(x^2+y^2)}{2\sigma_2^2}} \right) \]

Where \( \sigma_2 = 1.5\sigma_1 \) and the value of \( \sigma_1 \) in their implementation is 1.16. x and y are pixel coordinate in the N*N size window. Figure 2-4 is an example of DoG feature extraction of distortion image. The extraction of DoG also makes use of windows. In my example of feature extraction, the 7*7 block window is used.

Figure 2-4

<table>
<thead>
<tr>
<th>Distortion Image</th>
<th>DoG Feature Image</th>
</tr>
</thead>
</table>

Standardization coefficient of DoG for sigma parameter (expressed as \( DoG_{\sigma} \)) also conforms to the GGD distribution. At the same time, distortion will also affect the characteristics of Gaussian distribution. The sigma field of \( DoG_{\sigma} \) is extracted, and its average subtraction and division normalization coefficients are expressed as \( DoG'_{\sigma} \). Thereby, the shape of \( DoG_{\sigma} \) distribution, standard deviation parameters, skewness and kurtosis characteristics of samples can be extracted.

Laplacian of the luminance map: As the gray image of the original image has a similar distribution with the Laplacian function, the AGGD model can be used to fit and match. Since distortion will infect the attributes of image, the estimation model \((\alpha, \sigma)\)
parameters and sample features (skewness and kurtosis) of the image are calculated and then used as image features to evaluate the image quality.

Yellow color channel map: a yellow color channel map is constructed and defined as follows:

\[ Y = \frac{R + G}{2} - \frac{|R - G|}{2} - B \]

Where the G, B, and R are the green, blue, and red channels. Figure 2-5 indicates the visualized example of generated yellow channel map.

The normalized coefficients of Y calculated on the original image show the Gaussian-like behavior on the high-quality image. In addition, the normalized coefficients of Y sigma graph show the Gaussian-like behavior on the original image. Such behavior usually cannot be estimated by the distorted images. Therefore, the GGD of the X and Y normalized coefficients on testing image can be extracted and added as features.

The extraction of image features in wavelet domain: For a given image, C-DIIYINE can be used to extract the features of three scales and six directions from its luminance image, and a total of 82-dimensional statistical features can be extracted.
2.2.2 Chroma feature maps

A luminance (L*) and two chroma features (a* and b*) components are defined in the perception-related CIELAB colors. The coordinates L* in the CIELAB space denote color luminance, a* is related to red and green, and b* relates to yellow and blue. In addition, on the one hand, the non-linear relationship between L*, a* and b* simulates the cones in the retina of the non-linear response to the perceived color difference uniformly. And chroma feature captures the perceived yellow color channel map which is defined as below:

\[ C_{ab}^* = \sqrt{a^{*2} + b^{*2}} \]

where LAB color space have three components l*, a* and b*. And the Figure 2-6 indicates the example image which is transferred to the LAB space and the chroma feature image.

Figure 2-6

Test Image

Test Image In LAB Space

Chroma Feature Image
For the image to be generally measured, firstly, color transformation is carried out into the CIELAB space, various features are then obtained from the A and B chrominance components.

A and B chromaticity maps: The A and B normalized coefficients components of the original image as similar as Gaussian distribution, and distortion affects this distribution. Therefore, the GGD model can be utilized to capture these statistical biases. The method extracts two model parameters (α, σ) on the two scales and two sample statistic (kurtosis and skewness) features in the image. The sigma parameter of the chromaticity diagram: The feature processing method is the same as the luminance component.

2.2.3 LMS feature maps

The LMS color space is the closest color space to human’s vision perception. Similar as chroma feature map, image transferred from RGB color space to LMS color space. Assuming that the LMS color space constituting L, M and S three components, the symmetry axis are generated as follows:

\[
\hat{l} = \frac{1}{\sqrt{3}} (\hat{L} + \hat{M} + \hat{S})
\]

\[
\hat{\alpha} = \frac{1}{\sqrt{6}} (\hat{L} + \hat{M} - 2\hat{S})
\]

\[
\hat{\beta} = \frac{1}{\sqrt{2}} (\hat{L} - \hat{M})
\]
Where $\hat{L}, \hat{M}, \hat{S}$ refer to the logarithmic parameters of the L, M, and S components, respectively. $\hat{l}$ refers to luminance, $\hat{\alpha}$ refers to blue and yellow color, $\hat{\beta}$ refers to red and green color. And the function of $\hat{L}(i,j)$ is:

$$\hat{L}(i,j) = \frac{\log L(i,j) - \mu_L(i,j)}{\sigma_L(i,j) + 1}$$

where $\mu_L(i,j)$ refers to the mean, $\sigma_L(i,j)$ denotes the standard deviation of log L, $\hat{M}(i,j)$ and $\hat{S}(i,j)$ are defined from log $M(i,j)$ and log $S(i,j)$, respectively. In addition, several other feature maps are employed accordingly. The visualized images of $\hat{l}, \hat{\alpha}, \hat{\beta}$ are showed in figure 2-7.

![Figure 2-7](image)

BY and RG color-opponent maps: The algorithm quantifies the image along the opposite axis and the projection of two colors, then uses the AGGD model to fit and match, and finally captures the model parameters as well as the kurtosis and skewness of the color pair.

M and S channel maps: After the image is transformed to the LMS color space, the M and S chromaticity features are obtained, and their normalization coefficients and corresponding sigma parameters are modeled. Such method also contains the extractions
of Laplacian, DoG sigma statistical features for two chroma components. Meanwhile, the method also applies C-DIIVINE to extract the characteristics of three scales and six directions.

Overall, FREQUEE applies many high-quality features within the support vector regression to complete non-reference image prediction. Such algorithm has achieved excellent prediction results in the real image library, such as the LIVE challenge database. Nevertheless, the algorithm is still a combination of traditional machine learning and feature extraction.

2.3 DIQA

Jongyoo Kim et al. [13] put forward a novel means of image quality assessment by CNN deep learning network. The proposed approach, also referred to as deep image quality assessment (DIQA), using patch wise method with 2 stages training.

DIQA separates CNN training of the NR-IQA into two steps. In the first step, error map are trained in CNN to generate pre-trained model which is an objective distortion score training model. Subsequently, the pre-trained model training with MOS is applied to complete the prediction model in the second stage. A reliable map to predict the uniform region map is proposed to supplement the imprecision of the error graph. Besides, there are several handmade features utilized in the predictions.

When compared with the previous deep learning method, two stages of training are applied in DIQA to avoid overfitting in the learning process as indicated in figure 2-8:
From the flow chart of DIQA, it can be seen that such approach divides NR-IQA into two aspects, including objective distortion score training and subjective distortion score training based on the HVS part. In the objective distortion score training, an error map would be training through the CNN model. And in subjective score prediction, this pre-trained CNN model is used for further training for the HVS part.

For prediction, the usage of the CNN network is shown in figure 2-9.

The CNN architecture is proposed as in figure 2-9. For the prediction part of the error graph, the convolution layer with zero padding is used on the boundary before each convolution since the output can avoid the loss of relative position information. Except for the last layer, each layer has a 3×3 pooling layer with rectified linear element (ReLU). The result of the 8th convolution layer output is referred to the feature map and further training is continued in the second phase. In the first stage of training, the results are simplified into
a single channel target error mapping with the mapping for a 1*1. A simple linear combination of channels is applied to generate a meaningful feature map. The Conv9 (9th convolution layer) output is actually 1/4 size of the input image. This corresponds to a 1/4 reduction in the basic truth map of the objective error. For the down sampling operation, the convolution with step size two is utilized. After the reliable map training, two full connected layers are involved after the extracted data is inputted into the pooling layer.

Another specialty of DIQA is that the low-pass filter is applied for the input images to obtain low-frequency images. The low-frequency images are generated by reducing to 1/4 size of original image and rescaled back to the normal size. In addition, the Gaussian low-pass filter and the subsampling method are applied in this operation.

To complement the error map in potential defects, a reliable map is proposed in DIQA. Let $I_d$ be the corresponding distorted and image $I_r$ be the reference image and. $\hat{I}_r, \hat{I}_d$ refers to the normalized reference image and distortion image, respectively. The function of the reliable map generation is shown as below:

$$r = \frac{2}{1 + \exp(-\alpha(|\hat{I}_d|))} - 1$$

where $\alpha$ is empirical constant value. The reliable map will be normalized as follows in order to make the reliable map more accurate,:;

$$\hat{r} = \frac{1}{H_r \cdot W_r \sum(i, j) r(i, j)} r$$

where $H_r$ and $W_r$ means that $r$ is $H_r \times W_r$ size image, and i and j refer to the coordinate index. And the samples of reliable maps are illustrated in figure 2-10.
And the MSE between the prediction error graph and the ground truth error graph are applied in generating loss function:

$$
\mathcal{L}_1(\hat{I}_d; \theta_f, \theta_g) = \| g(f(\hat{I}_d; \theta_f) - e_{gt}; \theta_g) \odot \hat{r} \|_2
$$

where $\theta$ represents the parameters of CNN, $f(\cdot)$ and $g(\cdot)$ are identified in figure 2-9, and error $e_{gt}$ is defined by

$$
e_{gt} = |\hat{I}_r - \hat{I}_d|^P
$$

For the training detail, DIQA uses the patch-wise training method and the patch size is 112*112. On the other hand, for subjective points learning, the trained CNN model will be continue training in the full connected layer and the soft-max layer. Subsequently, the mean value of the feature graph is taken in the spatial domain within 128 feature vectors. Let $\sigma_{I_d}$ be variance of the low frequency distorted image and $\mu_r$ be the mean of the
nonnormalized reliability map. To compensate the lost information in the feature map, the loss function is applied as:

\[ \mathcal{L}_2(I_d; \theta_f, \theta_h) = \| (h(v, \mu_r, \sigma_{I_d}^{\text{low}}; \theta_h) - S) \|^2_2 \]

Where S denotes the subjective score of testing image, \( f(\cdot) \) represents the nonlinear regression function, and definition of v is shown as follow, where GAP is the kind of average pooling operation.

\[ v = GAP(f(\hat{I}_d; \theta_f)) \]

2.4 Summary of the NR-IQA method

Traditional machine learning for NR-IQA, such as BRISQUE, shows a sound learning effect in some classical databases, such as LIVE IQA which contains artificially distorted images. Nonetheless, the performance of some real distorted image databases (such as LIVE challenge) is relatively not well enough. In fact, some researchers are trying to improve the accuracy of machine learning by the combination of multiple fine feature extraction algorithms, such as the FRIQUEE algorithm. Others have turned to deep learning application for NR-IQA. The main factor which limits the development and progress of IQA in deep learning is the insufficiency of database capacity. In this thesis, one approach that may solve this problem based on the Bilinear CNN architecture will be introduced. Table 2-1[12] shows the prediction accuracy of several NR-IQA databases:
<table>
<thead>
<tr>
<th>Type</th>
<th>Methods</th>
<th>LIVE IQA</th>
<th>CSIQ</th>
<th>TID2013</th>
<th>LIVE MD</th>
<th>LIVE challenge</th>
<th>Weighted Average</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>SRCC</td>
<td>PLCC</td>
<td>SRCC</td>
<td>PLCC</td>
<td>SRCC</td>
<td>PLCC</td>
</tr>
<tr>
<td>FR</td>
<td>PSNR</td>
<td>0.876</td>
<td>0.872</td>
<td>0.806</td>
<td>0.800</td>
<td>0.636</td>
<td>0.706</td>
</tr>
<tr>
<td></td>
<td>SSIM</td>
<td>0.948</td>
<td>0.945</td>
<td>0.876</td>
<td>0.861</td>
<td>0.775</td>
<td>0.691</td>
</tr>
<tr>
<td></td>
<td>FSIIMe</td>
<td>0.963</td>
<td>0.960</td>
<td>0.931</td>
<td>0.919</td>
<td>0.851</td>
<td>0.877</td>
</tr>
<tr>
<td></td>
<td>DeepQA</td>
<td>0.981</td>
<td>0.982</td>
<td>0.961</td>
<td>0.956</td>
<td>0.939</td>
<td>0.947</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NR</td>
<td>BLIINDSII</td>
<td>0.912</td>
<td>0.916</td>
<td>0.780</td>
<td>0.832</td>
<td>0.536</td>
<td>0.628</td>
</tr>
<tr>
<td></td>
<td>BRISQUE</td>
<td>0.939</td>
<td>0.942</td>
<td>0.775</td>
<td>0.817</td>
<td>0.572</td>
<td>0.651</td>
</tr>
<tr>
<td></td>
<td>CORNIA</td>
<td>0.942</td>
<td>0.943</td>
<td>0.714</td>
<td>0.781</td>
<td>0.549</td>
<td>0.613</td>
</tr>
<tr>
<td></td>
<td>ILNIQE</td>
<td>0.902</td>
<td>0.908</td>
<td>0.821</td>
<td>0.865</td>
<td>0.521</td>
<td>0.648</td>
</tr>
<tr>
<td></td>
<td>GMLOG</td>
<td>0.950</td>
<td>0.954</td>
<td>0.803</td>
<td>0.812</td>
<td>0.675</td>
<td>0.683</td>
</tr>
<tr>
<td></td>
<td>HOSA</td>
<td>0.948</td>
<td>0.949</td>
<td>0.781</td>
<td>0.841</td>
<td>0.688</td>
<td>0.764</td>
</tr>
<tr>
<td></td>
<td>NRSL</td>
<td>0.952</td>
<td>0.956</td>
<td>0.851</td>
<td>–</td>
<td>0.661</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>SESANIA</td>
<td>0.934</td>
<td>0.948</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>CNN</td>
<td>0.956</td>
<td>0.953</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>MDNN</td>
<td>0.951</td>
<td>0.949</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>deepIQA</td>
<td>0.960</td>
<td>0.972</td>
<td>0.825</td>
<td>0.838</td>
<td>0.835</td>
<td>0.855</td>
</tr>
<tr>
<td></td>
<td>BIECON</td>
<td>0.958</td>
<td>0.962</td>
<td>0.825</td>
<td>0.838</td>
<td>0.721</td>
<td>0.765</td>
</tr>
<tr>
<td></td>
<td>DIQA-BASE</td>
<td>0.963</td>
<td>0.964</td>
<td>0.812</td>
<td>0.791</td>
<td>0.800</td>
<td>0.803</td>
</tr>
<tr>
<td></td>
<td>DIQA</td>
<td>0.975</td>
<td>0.977</td>
<td>0.884</td>
<td>0.915</td>
<td>0.825</td>
<td>0.850</td>
</tr>
</tbody>
</table>

|      |         |         |      |         |        |                |                  |
|      |         | SRCC    | PLCC | SRCC    | PLCC   | SRCC           | PLCC             |
|      |         |         |      |         |        |                |                  |
|      |         | 0.725   | 0.815 | 0.643   | 0.507  | 0.664          | 0.729            |
|      |         | 0.845   | 0.882 | 0.607   | 0.645  | 0.689          | 0.746            |
|      |         | 0.900   | 0.915 | 0.618   | 0.662  | 0.666          | 0.717            |
|      |         | 0.902   | 0.914 | 0.594   | 0.589  | 0.662          | 0.747            |
|      |         | 0.824   | 0.863 | 0.543   | 0.571  | 0.751          | 0.761            |
|      |         | 0.902   | 0.926 | 0.659   | 0.678  | 0.761          | 0.819            |
|      |         | 0.932   | 0.946 | 0.631   | 0.654  | 0.760          | –                |
|      |         | 0.933   | 0.948 | –       | –      | –              | –                |
|      |         | 0.956   | 0.953 | –       | –      | –              | –                |
|      |         | 0.951   | 0.949 | –       | –      | –              | –                |
|      |         | 0.960   | 0.972 | 0.835   | 0.855  | –              | –                |
|      |         | 0.912   | 0.928 | 0.595   | 0.613  | 0.790          | 0.821            |
|      |         | 0.900   | 0.803 | 0.663   | 0.705  | 0.836          | 0.836            |
|      |         | 0.939   | 0.942 | 0.703   | 0.704  | 0.867          | 0.888            |

32
Chapter 3 – Bilinear CNN

3.1 Principal of the convolution layer

Marcel Simon [10] has found that the last layer of the convolution layer contains valid image information. Their research also shows that the latest layer is sensitive to more abstract image patterns which can even correspond to the entire object or part. According to the study from this convolution information, he found that the parts of the image that are highlighted or are deep color areas are the key parts of the images.

The purpose of CNNs is to extract the key eigenvalues of images through the convolution layer. There are examples of those two layers. The highlight part corresponds to high activation whereas the deeper part corresponds to lower activation. The figure 3-1 shows that the last convolution layer contains the 13*13*N features. After simple calculation of the mean of data, the image of cats and dogs last convolution layer 13*13 outputs shows that the heads and tails of cats and dogs contain key information in figure 3-2.

\[
egin{bmatrix}
  f_{1,1} & \cdots & f_{1,13} \\
  \cdots & \cdots & \cdots \\
  f_{13,1} & \cdots & f_{13,13}
\end{bmatrix}
\]

Figure 3-1
3.2 Bilinear CNN

3.2.1 Structure of Bilinear CNN

Bilinear CNN is the new method of CNN structure created by Tsung-Yu Lin [9]. Bilinear CNN works quite well in the fine-grained classification with the usage of the principal of the convolution layer performance. The main approach of Bilinear CNN is amplifying the key region information of the convolution layer through two CNNs so as to improve the prediction performance. And the bilinear CNN flow chart is shown in figure 3-3.
The basic steps of Bilinear CNN are: a training sample image will pass through two different CNNs or the same CNN with different weights, and the outer product is used to complete the product of the last convolution layer of the two CNNs. Subsequently, the obtained bilinear vector will be further trained and learned through the full connection layer. It should be noted that both CNNs need to have the same or similar input format. Furthermore, the convolutional layer format of the two CNNs needs to be consistent.

The Bilinear CNN has the advantage of effectively utilizing the particularity of the CNN feature extraction using last convolution layer. Even though two CNNs contain different layers and weights, the last convolution layer maps have similar structure. And through outer product, the feature value of the essential region of the image is magnified. Therefore, bilinear CNN has better prediction accuracy than a single CNN structure.

3.2.2 Bilinear models

The bilinear model B can be shown as $B = (f_A, f_B, P, C)$ where $f_A$ and $f_B$ refer to feature extraction functions, $C$ denotes classification algorithm and $P$ refers to pooling operation. And $f: f(L \ast I) \rightarrow R^{c \times D}$ for image $I$ and location $L$ and output is $R^{c \times D}$. $L$ called
location which generally contains position and scale. Those feature results are aggressed together by the outer product matrix, such as \((L, l, f_A, f_B) = f_A(L, l)^T f_B(L, l)\).

It is noted that Both \(f_A\) and \(f_B\) should contain the same or similar feature dimension \(c\) to be fitted. Since the overall architecture is kind of end-to-end training, the DIQA are more simply to complete training. If the two networks outputs are \(L^*M\) size data \(A\) and \(L^*N\) size data \(B\) respectively, the obtained bilinear vector will be \(x = A^T B\) of size \(M^*N\).

Let gradient function of loss \(l\) is \(dl/dx\), then we have:

\[
\frac{dl}{dA} = B \left( \frac{dl}{dx} \right)^T, \quad \frac{dl}{dB} = A \left( \frac{dl}{dx} \right)
\]

Subsequently, the gradients computing of the whole operation of outer product in bilinear CNN are shown in the figure 3-4.

Figure 3-4

\[
\begin{align*}
& x = A^T B \\
& y = x^{\text{sqrt}} \\
& z = y\ell_2
\end{align*}
\]

Bilinear CNN streamlines overall CNN training by simplifying gradient calculations and completing end-to-end training of both networks only using the image labels. Nonetheless, instead of utilizing the original end-to-end bilinear CNN structure, the idea of bilinear structure is used in thesis. To be more specific, two CNNs are utilized to complete training which is similar to Bilinear CNN.

36
4.1 Database

In this thesis, the image quality estimation databases, including LIVE IQA and LIVE Challenge, are applied in this thesis to estimate the prediction results of the algorithm.

4.1.1 LIVE IQA

The LIVE IQA [14] phase I which contains symmetric distortion is applied. Phase I dataset including JPEG compression and JPEG2000 standard compression, Gaussian white noise, Gaussian blur, and Fast fading images based on the Rayleigh fading channels. There are 20 original stereo images with 365 distortion images. The image’s format is 640*360. The image scores can be obtained by the mean of the observer subjective score and each image has a differential mean opinion score (DMOS) between 1 to 100 value within 4 decimals. For image quality, the higher DMOS, the worse image quality is. Figure 4-1 shows examples of images from the database.

![Figure 4-1](image-url)
4.1.2 LIVE Challenge

LIVE challenge [15] is the database with a larger number of challenging images in the IQA database which contains 1164 images. The image format is 500*500. Each image has mean opinion score (MOS) from 1 to 100 with 4 decimals. The higher the MOS value, the better the image quality. Here is an example image from the database.

![Example image from the database](image_url)

4.1.3 Database comparison

In addition to the LIVE IQA and LIVE Challenge databases, there are others. Table 4-1 shows the detail comparison of databases. In this thesis, the LIVE IQA and LIVE challenge databases will be selected.

<table>
<thead>
<tr>
<th>Database</th>
<th>reference images</th>
<th>Distortion images</th>
<th>Distortion types</th>
</tr>
</thead>
<tbody>
<tr>
<td>Live IQA</td>
<td>29</td>
<td>779</td>
<td>5</td>
</tr>
<tr>
<td>CSIQ</td>
<td>30</td>
<td>866</td>
<td>6</td>
</tr>
<tr>
<td>TID2013</td>
<td>25</td>
<td>3000</td>
<td>24</td>
</tr>
<tr>
<td>LIVE MD</td>
<td>15</td>
<td>405</td>
<td>2</td>
</tr>
<tr>
<td>LIVE Challenge</td>
<td>N/A</td>
<td>1162</td>
<td>N/A</td>
</tr>
</tbody>
</table>
4.2 CNN networks applied

In this research, two VGG nets are used to implement CNN training. One is the vgg-m-net which contains 16 layers structure in general (13 convolution layers and 3 fully-connected layers). The convolution layer of VGG net has a remarkable feature [17]: the spatial resolution of the features size decreases monotonously, and the number of the feature map dimensions increases monotonously. For the structure of bilinear CNN, as one outer product layer is added, it should be treated as 17 layers. It is applied for image-wise training with the standard input of 224*224. The other one is the VGG-s-net with 11 layers (8 convolution layers and 3 fully-connected layers) and is used for training images patches with the standard input of 32*32.

As shown in table 4-2, VGG uses several convolution layers with smaller convolution kernels (3x3) instead of a larger convolution layer with smaller convolution kernels. On the one hand, VGG can reduce parameters. On the other hand, more non-linear mappings can be carried out to increase the fitting ability of the network.

Small convolution kernel is an essential feature of VGG [18]. Although VGG imitates the Alex-net network structure, it does not use Alex-net's larger convolution core size (such as 7x7) yet achieves the same performance by reducing the size of convolution core (3x3) as well as increasing the number of convolution sub-layers (VGG: from 1 to 4 convolution sub-layers, Alex-net: 1 sub-layer).
Table 4-2

<table>
<thead>
<tr>
<th>ConvNet Configuration</th>
<th>A</th>
<th>A-LRN</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>11 weight layers</td>
<td>11 weight layers</td>
<td>13 weight layers</td>
<td>16 weight layers</td>
<td>16 weight layers</td>
<td>19 weight layers</td>
</tr>
<tr>
<td>Input (224 × 224 RGB image)</td>
<td>conv3-64</td>
<td>conv3-64</td>
<td>conv3-64</td>
<td>conv3-64</td>
<td>conv3-64</td>
<td>conv3-64</td>
</tr>
<tr>
<td>conv3-128</td>
<td>conv3-128</td>
<td>conv3-128</td>
<td>conv3-128</td>
<td>conv3-128</td>
<td>conv3-128</td>
<td>conv3-128</td>
</tr>
<tr>
<td>maxpool</td>
<td>conv3-256</td>
<td>conv3-256</td>
<td>conv3-256</td>
<td>conv3-256</td>
<td>conv3-256</td>
<td>conv3-256</td>
</tr>
<tr>
<td>maxpool</td>
<td>conv3-512</td>
<td>conv3-512</td>
<td>conv3-512</td>
<td>conv3-512</td>
<td>conv3-512</td>
<td>conv3-512</td>
</tr>
<tr>
<td>maxpool</td>
<td>conv3-512</td>
<td>conv3-512</td>
<td>conv3-512</td>
<td>conv3-512</td>
<td>conv3-512</td>
<td>conv3-512</td>
</tr>
<tr>
<td>maxpool</td>
<td>conv3-512</td>
<td>conv3-512</td>
<td>conv3-512</td>
<td>conv3-512</td>
<td>conv3-512</td>
<td>conv3-512</td>
</tr>
<tr>
<td>maxpool</td>
<td>conv3-512</td>
<td>conv3-512</td>
<td>conv3-512</td>
<td>conv3-512</td>
<td>conv3-512</td>
<td>conv3-512</td>
</tr>
<tr>
<td>FC-4096</td>
<td>FC-4096</td>
<td>FC-1000</td>
<td>soft-max</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

It is considered that the size of the receptive field obtained by two 3x3 convolution stacks is equivalent to that of a 5x5 convolution, while the size of the receptive field obtained by three 3x3 convolution stacks is equivalent to that of a 7x7 convolution. This can enlarge non-linear mappings and decrease the parameter numbers.
The commonly overall data size of each layer of VGG CNN is shown in the graph. The detailed steps of the standard VGG CNN are shown as below

1. Input $224 \times 224 \times 3$ pictures and make two convolutions + ReLU through $64 \times 3 \times 3$ convolution cores. After convolution, the size becomes $224 \times 224 \times 64$.

2. Max pooling (maximizing pooling), pooling cell size $2 \times 2$ (halving image size), and pooling output size $112 \times 112 \times 64$.

3. After two convolutions of $128 \times 3 \times 3$ convolution cores + ReLU, the size becomes $112 \times 112 \times 128$.

4. Max pooling with the output size of $56 \times 56 \times 128$.

5. After $256 \times 3 \times 3$ convolution cores for three convolutions + ReLU, the size is changed to $56 \times 56 \times 256$.

6. Max pooling with the output size of $28 \times 28 \times 256$.

7. After $512$ convolution kernels of $3 \times 3$ are convoluted three times + ReLU, the size is changed to $28 \times 28 \times 512$.

8. Max pooling with the output size of $14 \times 14 \times 512$. 
9. After 512 convolution cores of 3*3 are convoluted three times + ReLU, the size is changed to 14*14*512.

10. Max pooling with the output size of 7*7*512.

11. Full connection with Layer 1*1*4096 and ReLU (three layers).

12. Output results through the soft-max layer.

VGG-m-net is used to obtain the last convolution layer output for 3 types of images in figure 4-4.

As shown in figure 4-4, the VGG-m-net model is utilized to display the last convolution layer output of three different distorted images. The three types of distorted images are blur, fast fading, and JP2K. Even without the operation of visualization highlighting, the resulting graph of the convolution layer of three images still contains similar structure, which may imply that the convolution properties of an image may be
similar. This structure shows that the center of the image is the key area of high activity, which makes sense when human sights focus on the detail of the building in example image.

4.3 Image quality assessment

In this thesis, the CNN structure is constructed based on the basic idea of bilinear CNN and the CNN algorithm of image-wise and patch-wise is completed in a bilinear way. This is the basic idea of the overall research on training and the testing methods.

4.3.1 Image preprocessing

For this experiment, as two ways of CNN learning are required, the image preprocessing can be separated as two parts. For the image-wise method:

Firstly, all the images are randomly classified according to the ratio of training samples to testing samples (3:1). Second of all, the image samples with 640*360 resolution in the LIVEIQA database and the image samples with 500*500 resolution in LIVE challenge database are divided into 6 portions in the format of 180*180 and 4 portions in the format of 250*250, respectively. Thirdly, those images are reshaped into format of 224*224 due to the standard input of VGG-m-net.

There are two reasons for those processing steps. First of all, the sample size of two databases are too small to avoid overfitting, particularly for LIVEIQA. The method of dividing images is a way to extend the data size. Secondly, the images are required to fit the standard input of the VGG net.

For the patch-wise method, the preprocessing steps are similar to the image-wise method. The image samples in the LIVEIQA and LIVE challenge databases are divided into 5*5 portions in the format of 128*72 and the format of 100*100, respectively. Those
images are resized into the 112*112 format. The reason why one image is divided into 5*5 part will be explained later.

### 4.3.2 Image-wise training

The performance and accuracy of the CNN models highly depend on the size of the training samples and labels. Nonetheless, due to the difficulty of collecting human’s subjected evaluation, the process of obtaining the image labels will request complex physical experiments and information gathering. Thereby, the IQA database is much smaller than the normal computer vision and image recognition database. Furthermore, the image from the IQA database is much larger than the standard input of CNN. Dividing image in patches is an efficient way to increase the training sample number for the reason that the small number of images may not be complicated for highly accurate CNN training. For those reasons, the training sample can be divided into several patches in the preprocessing step.

For each patch in the training, the patch image score is set as same as the original image score. And each patch counts a score and makes average to generate the final image quality in testing.

After the preprocessing steps, the image-wise training basically follows the bilinear structure similar to the bilinear CNN.

Take Live IQA database as an example. First of all, 224*224 input images pass through the first 13 convolution layers of two VGG CNN. The data obtained from the last convolution layer is called the weight map. The other one obtained through the last convolution layer of another CNN is referred to the feature map. Subsequently, the bilinear
vector is constructed by computing the outer product of two maps. In the end, bilinear vector will continue to be trained in next fully-connected layers and soft-max layers. The flow chart of image-wise training is in below figure 4-5. In general, the training images are trained in two CNN networks at the same time and are combined with the convolution layer result of each other for the training of the soft-max layer and fully-connected layer.

4.3.3 Patch-wise training

As mentioned previously, the training sample images are divided into patches, and the patches quality score is set equivalent to the image score in several previous image-wise trainings.

However, this is not an ideal way for patches to get trained. Some pitches may have different qualities that do not equal to the image quality for some reasons. Therefore, the training does not involve every correct pitch quality, which will decrease the currency of the predicted result.
The research shows that the last convolution layer contains key information. And the bilinear CNN using two CNNs to enlarge the essential part in training gets better results than normal CNN. The proposed idea to solve this problem is using data from the last several convolution layers to redefine the scores of the image patches.

The flow chart of the patch-wise training is in figure 4-6. The steps of patch-wise training are also bilinear ways: one is extracting the patch score map from the original images by the VGG-s-net convolution layer output and making normalized patch scores. The other is applying patch with normalized patch scores in VGG-m-net for training.

To redefine the image patches through the CNN convolution layer, the output size of the CNN convolution layer and image patches need to be determined. The advantage of VGG net is that its convolution kernel is 3*3. Thus, the output of the convolution layer changes little. In order to get better data size, I modified the convolution kernel size in the first convolution layer and the convolution layers’ output data size are shown in table 4-3. There are a variety of CNN convolution layer output formats to choose. In this experiment,
13*13 and 6*6 can be selected for the convolution layer output to observe the experimental results, since the image size of 640*360 and 500*500 can be basically divided by 12 and 5, respectively.

The location of the two convolution layers are the ninth and sixteenth layers of VGG net, respectively. Here is the data size output for the training results.

<table>
<thead>
<tr>
<th>Layer</th>
<th>Data size</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>224x224x3 single</td>
</tr>
<tr>
<td>2</td>
<td>109x109x96 single</td>
</tr>
<tr>
<td>3</td>
<td>109x109x96 single</td>
</tr>
<tr>
<td>4</td>
<td>109x109x96 single</td>
</tr>
<tr>
<td>5</td>
<td>54x54x96 single</td>
</tr>
<tr>
<td>6</td>
<td>26x26x256 single</td>
</tr>
<tr>
<td>7</td>
<td>26x26x256 single</td>
</tr>
<tr>
<td>8</td>
<td>26x26x256 single</td>
</tr>
<tr>
<td>9</td>
<td>13x13x256 single</td>
</tr>
<tr>
<td>10</td>
<td>13x13x512 single</td>
</tr>
<tr>
<td>11</td>
<td>13x13x512 single</td>
</tr>
<tr>
<td>12</td>
<td>13x13x512 single</td>
</tr>
<tr>
<td>13</td>
<td>13x13x512 single</td>
</tr>
<tr>
<td>14</td>
<td>13x13x512 single</td>
</tr>
<tr>
<td>15</td>
<td>13x13x512 single</td>
</tr>
<tr>
<td>16</td>
<td>6x6x512 single</td>
</tr>
<tr>
<td>17</td>
<td>1x1x4096 single</td>
</tr>
<tr>
<td>18</td>
<td>1x1x4096 single</td>
</tr>
<tr>
<td>19</td>
<td>1x1x2048 single</td>
</tr>
<tr>
<td>20</td>
<td>1x1x2048 single</td>
</tr>
<tr>
<td>21</td>
<td>1x1x1000 single</td>
</tr>
<tr>
<td>22</td>
<td>1x1x1000 single</td>
</tr>
</tbody>
</table>

For the 12th convolution layer, the output result will be 6*6*512. After average operation, 6*6 data, also called the feature map, can be obtained. \( n^{th} \) patch score \( S_p \) is defined by normalization function:

\[
S_p(n) = a \left[ \left( S_i/v_m \right) v_p(n) \right] + b
\]
Where \( n \in \{(i, j) | i = 1, 2, ..., 6, j = 1, 2, ..., 6\} \), \( v_m \) is the mean value of the 6*6 feature data, \( v_p(n) \) refers to the value of \( n^{th} \) patch in 6*6 data, and \( S_i \) denotes the whole image quality score. \( a \) and \( b \) are the constant values to limit the maximum and minimum values of \( S_p(n) \).

Figure 4-7 indicates the 6*6 output data from the convolution layer and table 4-4 shows the calculated quality scores for each patch when image quality score is 28.61328125, \( a = 5/S_i \), \( b = S_i - 5 \). The values of \( a \) and \( b \) are empirical selections for training and prediction.

![Figure 4-7](image)

<table>
<thead>
<tr>
<th>Table 4-4</th>
</tr>
</thead>
<tbody>
<tr>
<td>31.82775</td>
</tr>
<tr>
<td>28.19564</td>
</tr>
</tbody>
</table>

This function is a simple function to simplify the generation of the patch score. Actually, there can be more complicated function to normalize the data.

After incising the data to 5*5 image for the convenience of sample training, each patch score can be obtained from the function and the image can be divided by the format of 5*5. For the LIVEIQA and LIVE challenge databases, the image patch is 128*72 and 100*100, respectively. Here, the 5*5*sample size can be trained in the VGG-s-net CNN, which is enough for CNN to complete accurate training and learning. It is worth mentioning...
that throughout the training process, the cores of all the patches are converted into integers with 1 decimal.

**4.3.4 Image prediction**

The preprocessing of the testing image is completed in the preprocessing step and the testing progress will be similar to the general CNN training step. The predicted performance can be generated by comparison of truth-ground image scores with prediction scores.

The results of the predicted image are the average values of all the predicted image patches not only from image-wise testing but from patch-wise testing as well.

**4.3.5 Evaluation method**

To evaluate the performances of the CNN predicted result, standard measurement, Spearman’s rank-order correlation coefficient (SRCC), is utilized. And SRCC is defined by:

\[
    \text{SRCC} = 1 - \frac{6 \sum_i d_i^2}{n(n^2 - 1)}
\]

Where \(d_i\) refers to difference between the truth score and the predicted score of the number \(i\) image, and \(n\) denotes the total number of the test images. And Pearson’s linear correlation coefficient (PLCC) is defined by

\[
    \text{PLCC} = \frac{\sum_i (\hat{S}_i - \mu_{\hat{S}})(S_i - \mu_S)}{\sqrt{\sum_i (\hat{S}_i - \mu_{\hat{S}})^2} \sqrt{\sum_i (S_i - \mu_S)^2}}
\]
Where $S$ and $\hat{S}_i$ are the ground-truth and prediction scores and $\mu_S$ and $\mu_{\hat{S}}$ indicate the mean of ground-truth score and the predicted score. The reason for choosing two methods to evaluate is to obtain more reliable results of prediction through comparison of the groups.
Chapter 5 – Prediction result conclusion

5.1 Performance on databases

As mentioned in the training step, all the distortion images are randomly divided into two subsets as the rate of 3:1. Upon the completion of the testing images prediction, the prediction result of the bilinear CNN is evaluated with the image-wise and patch-wise approaches with the usage of the two standard measures, including SRCC and PLCC.

Table 5-1 shows the prediction performance of two methods of bilinear CNN in all the databases. According to the results in table 5-1, bilinear CNN with the image-wise method have better results compared with patch-wise method in LIVEIQA database. Meanwhile, patch-wise method works well in LIVE challenge database.

<table>
<thead>
<tr>
<th></th>
<th>LIVE IQA</th>
<th>LIVE challenge</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SRCC</td>
<td>PLCC</td>
</tr>
<tr>
<td>bilinear CNN+image wise</td>
<td>0.931</td>
<td>0.939</td>
</tr>
<tr>
<td>bilinear CNN+patch wise</td>
<td>0.917</td>
<td>0.922</td>
</tr>
</tbody>
</table>

5.2 Performance on distortion types

To discuss the prediction results of the algorithm in 5 different types distortion images in detail, the predicted image results are classified according to the individual image types. As the LIVE Challenge database does not label the image types for challenge, only the accuracy performance of LIVEIQA database is summarized as results. To show a more detailed comparison of different evaluated methods, I set SRCC and PLCC types as well.

From the accuracy result of five types of distortion images, the image-wise method and the patch-wise method show similar tendency in the prediction results of five types of
distortion. White Gaussian distortion image has the best prediction performance, while Fast Fading image has the opposite performance. The prediction performance of the other three types of images is similar to the overall prediction value, which shows the stable prediction results.

<table>
<thead>
<tr>
<th>Table 5-2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>LIVE IQA</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>

5.3 Conclusion and discussion

In this thesis, an overview of the NR-IQA is made and several algorithms are listed. Many previous successful NR-IQA algorithms applied with the machine learning and feature extraction to complete training and learning. The advantage of fine feature extraction within SVR is that it can provide sound prediction performance in the case of the small size of samples (hundreds of samples). Nevertheless, finer feature extraction may lead to over-fitting and poor performance in real-time sample databases since the real distortion images contain multiple types of distortion. In view of the excellent performance of CNN in image processing and similar learning structure compared with the human visual system, lots of researchers are trying to apply deep learning networks to the training and learning in NR-IQA. Nonetheless, the disadvantage of CNN is that it needs sufficient samples to accomplish accurately deep learning. In this case, researchers expand the number of training samples by dividing the sample image into patches. The differences of
dealing patches before soft-max classified algorithms into the image-wise method or the patch-wise method.

The image-wise approach is to recombine all the patches back into image structures and classify them according to the score of the test image. Regarding the advantages of this method, it enables further training even the score of each patch remains unknown, and the training samples can enlarged by training and combining each patch separately. However, the combined data still counts one category when it comes to soft-max layer. Furthermore, there will be a great amount of computation in combining patches into the original image and proceeding with the SoftMax layer.

The patch-wise approach completes the entire training and learning based on each patch by scoring each patch with their testing image scores. The advantage of this method is that there will be enough samples for each score category to learn since patches are actually training in CNN. The size of the database is extended thoroughly, which is more appropriate for the requirements of CNN learning. However, for each patch, directly setting the score of the patches equal to the sample image score will inevitably lead to the error of the patches learning.

With no doubt, there are some other methods to complete image training in deep learning. For instance, the DIQA algorithm is a combination of subjective image scoring training and objective image scoring training.

Subsequently, the bilinear CNN which has been well-used in image recognition is introduced. In addition, the bilinear CNN structure is applied in the image-wise method and the patch-wise method. Although the application of CNN in IQA is very challenging for many reasons, bilinear methods are utilized to solve those problems.
For overall predictive performance in comparison with other algorithms, bilinear CNN with image-wise method is not much far behind the latest and best algorithms. For the case that the patch-wise method does not show many advantages in comparison to the image-wise in LIVEIQA, two explanations can be figured out. First of all, it is highly possibility that the patch-wise method suffers the overfitting in training process at LIVEIQA database. Secondly, different score types between LIVE IQA and LIVE Challenge cause different performances. LIVEIQA uses DMOS score which is the measure of differences between the testing images and reference images. LIVE Challenge uses MOS score which is the subjective score without references. The function that produces patches scores is basically the normalized function which may lead to poor representation in the DMOS score type. The fact that could support this explain is that, in the LIVE challenging databases, the patch-wise method has better performance than the image-wise method.
References


[16] L. Fei-Fei, R. Fergus, and P. Perona. One-shot learning of object categories. PAMI, 28(4), 2006. 1, 2


