VIETNAM NATIONAL UNIVERSITY, HANOI

UNIVERSITY OF ENGINEERING AND TECHNOLOGY

Nguyen Trung Nghia

MODELING HUMAN VISUAL SYSTEM IN PATCH-BASE IMAGE QUALITY ASSESSMENT USING DEEP LEARNING

Major: Computer Science

HA NOI - 2019

VIETNAM NATIONAL UNIVERSITY, HANOI
UNIVERSITY OF ENGINEERING AND TECHNOLOGY

Nguyen Trung Nghia

MODELING HUMAN VISUAL SYSTEM IN PATCH-BASE IMAGE QUALITY ASSESSMENT USING DEEP LEARNING

Major: Computer Science

Supervisor: Assoc. Prof., Ph.D. Le Thanh Ha
Co-Supervisor: Assoc. Prof., M.Sc. Pham Thanh Tung

HA NOI - 2019

AUTHORSHIP

"I hereby declare that the work contained in this thesis is of my own and has not been previously submitted for a degree or diploma at this or any other higher education institution. To the best of my knowledge and belief, the thesis contains no materials previously published or written by another person except where due reference or acknowledgement is made."

Signature ....

SUPERVISOR'S APPROVAL

"I hereby approve that the thesis in its current form is ready for committee examination as a requirement for the Bachelor of Computer Science degree at the University of Engineering and Technology."

Signature ....

ACKNOWLEDGEMENT

First and foremost, I would like to express my sincere thanks to my supervisor Ph.D. Le Thanh Ha and M.Sc. Pham Thanh Tung for their support and guidance throughout this research work.

I greatly appreciated the Department of Computational Science and Engineering, and HMI Lab both at the VNU UET for the generous support.

I would also like to thanks my teachers and friends for their support throughout my time in University of Engineering and Technology, Vietnam National University, Hanoi

This thesis was partly supported by the University of Fire Fighting and Prevention under the subjective image quality experiment and InfoRe Technology for providing working environment.

ABSTRACT

As humans are the ultimate receivers of the majority of visual signals being processed, the most accurate way of assessing image quality is to ask humans for their opinions of an image's quality, known as the subjective image quality assessment (IQA). The subjective image quality scores gathered from all subjects are processed to be the mean opinion score (MOS), which is regarded as the ground truth of image quality. Conventionally, a number of full-reference image quality assessment (FR-IQA) methods adopted various computational models of the human visual system (HVS) from psychological vision science research.

The image compression is one of the most prominent applications that require IQA metrics to be highly correlated with human vision. To explore IQA algorithms that are more consistent with human vision, several calibrated databases have been constructed. However, the distorted images in the existing databases are usually generated by corrupting the pristine images with various distortions in coarse levels, such that the IQA algorithms validated on them may be inefficient to optimize the image compression with fine-grained quality differences. In addition, HVS is differently sensitive to features of image patch, the 'ground truth' quality of patch is essential for training patch-based methods, but in practice it's easy to obtain the ground truth quality of images rather than patches.

So an experimental quality assessment to approach database for image patch has been developed. We propose Full-reference Deep Image-Patch Quality Assessment (DIPQA), a novel image-patch quality assessment that used deep neural network to estimate the 'ground truth' for patches with the developed database.

Seven well-know IQA algorithms are evaluated and analyzed on the proposed database to show that there is still large room for improvement regarding fine-grained patch-based method. In the following experiment, we train and evaluate the proposed DIPQA on the proposed database and show competitive performance to other FR methods. DIPQA is expected to improve the performance of many applications that require patch's "ground truth" especially in image compression.

Contents

List of Figures **vii**

List of Tables **viii**

Abbreviations **ix**

1 Introduction **1**
1.1 Motivation 4
1.2 Contributions 5
1.3 Thesis Outline 6

2 Background **7**
2.1 Image Quality Assessment 7
2.1.1 Subjective image quality assessment 7
2.1.2 Objective image quality assessment 8
2.1.2.1 Full-reference image quality assessment (FR-IQA) 9
2.1.2.2 Reduced-reference image quality assessment (RR-IQA) . . . 11
2.1.2.3 No-reference image quality assessment (NR-IQA) 11
2.1.3 IQA in Visual Data Compression 12
2.2 Neural Networks 14
2.2.1 Artificial Neural Network 14
2.2.2 Training Neural Network 14
2.2.3 Convolutional Neural Network 17

3 Methodology **21**
3.1 Database Construction 21
3.1.1 Testing image database 21
3.1.2 Database creation 22
3.1.3 Testing methodology 23
3.1.4 Benchmark Analyses 24
3.2 Deep Image-Patch Quality Assessment 25
3.2.1 Architecture 25
3.2.2 Feature Fusion 26
3.2.3 Training Method 26

4 Evaluation **27**
4.1 Evaluation Method 27
4.2 Experiment results 28
4.2.1 HMII Benchmark Analysis 28
4.2.2 Image-Patch Models 29

5 Conclusions **31**
5.1 Conclusions 31
5.2 Future work 31

References **33**

List of Figures

1.1 Digital images suffer from distortions on every step. 2

1.2 Example of JPEG distorted image. Different patches have different qualities. . 5

2.1 General Objective image quality assessment framework 9

2.2 Examples of quantization table and the corresponding compressed JPEG images. 12

2.4 The basic structure of Neural Network. 14

2.5 The function overfit the train samples. 17

2.6 CNN share the kernel on each layer. 18

2.7 Kernel that maps $m \times n \times d_1$ block in the previous layer to an $1 \times 1 \times d_2$ block in next layer. 18

2.8 Common nonlinear functions used in CNN: ReLU, Sigmoid and hyperbolic tangent. 19

2.9 A max pooling layer with a 2×2 filter size and a stride of 2 20

3.1 Selected image patch. 22

3.2 Testing software screenshot. 24

3.3 Deep Image-Patch Quality Assessment Network Architecture. 25

4.1 Objective Score by top 3 IQA on HMII 29

List of Tables

3.1 Example of testing image database. 23

4.1 PLCC and SRCC for different IQA algorithms 28

4.2 Comparing different Full-Reference Image-Patch approaches 30

Abbreviations

IQA	Image Quality Assessment
CNN	Convolutional Neural Network
DIPQA	Deep Image-Patch Quality Assessment
DMOS	Differential Mean Opinion Score
FR	Full-reference
MOS	Mean Opinion Score
NR	No-reference
RR	Reduced-reference

Chapter 1

Introduction

With the growing popularity of smartphones, compact cameras, and Internet services such as Facebook and Instagram, the production and sharing of digital images has grown tremendously over the past few years. The Fig.1.1 show the journey of a picture begins with it being obtained by a camera, which changes over it into a digital format and compresses it utilizing lossy compression algorithms to meet the onboard storage accessibility. This image is then transmitted over wired or wireless transmission channels and is altered in its resolution to meet the available bandwidth. Finally, the end user receives this image and watches it over devices ranging from smartphones to 4K displays, which require further alterations to its resolution.

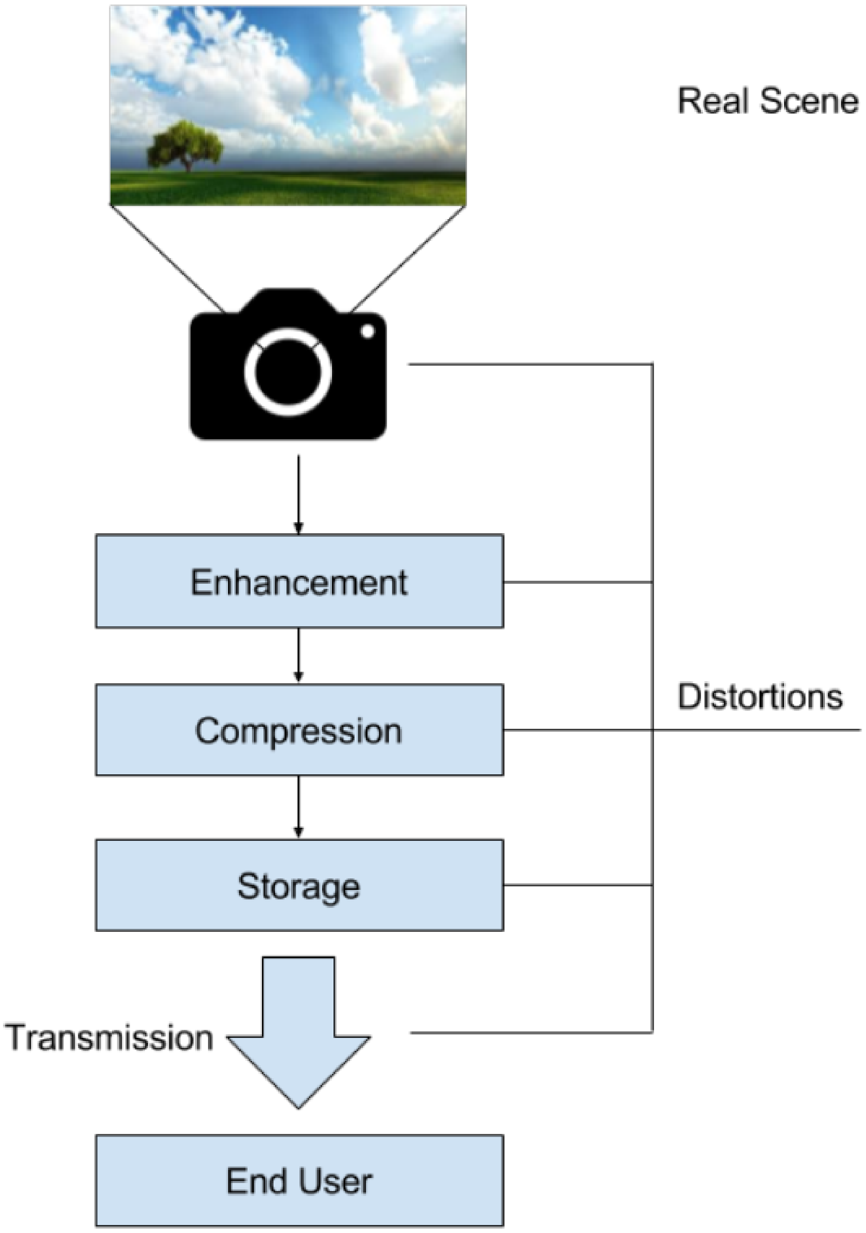


FIGURE 1.1: Digital images suffer from distortions on every step.

End users tend to be more inclined to select a content provider, a service provider, and a display device that can better meet their image quality expectations at delivery. Thus, optimizing these respective technologies to deliver perceptually good results becomes crucial for all content providers, service providers, and display providers, and in order to do so, perceptual image quality needs to be estimated. In addition, in order to determine the perceptual quality, this estimation process should be automated as much as possible to make it independent of the availability of human observers.

Image Quality Assessment (IQA) aims to measure the perceived quality of the visual signal based on its statistical characteristics and human perceptual mechanism, which is widely required in numerous applications for image processing. IQA plays a vital role in guiding, implementing, optimizing and verifying many visual processing algorithms and systems [1–4]. In particular, image compression is one of IQA's most representative applications, which can be used in the process of optimizing rate distortion to obtain compressed images with better visual quality at the same bit-rate level. [5–10]. The traditional methods of image compression mainly use the quality metrics based on signal-fidelity, which are less correlated with human perceptual quality , e.g., MAE (mean absolute error), MSE (mean square error), SNR (signal-to-noise ratio), PSNR (peak SNR) and their relatives. While these metrics have many favorable properties, e.g. clear physical meaning and high calculation efficiency, they severely impede the improvement in compression performance by further reducing image visual redundancies due to their poor consistency with human visual perception.

Many perceptual quality metrics have been proposed over the past few years to obtain more consistent measures with human visual perception. According to the availability of a reference image, these methods can be divided into three categories, i.e., full reference (FR) ones where the pristine reference image is available, reduced reference (RR) ones where partial information of the reference image is available and no reference (NR) ones where the reference image is unavailable. For image compression problem, the reference images are available at the encoder side such that the FR-IQA algorithms are applicable.

Many FR-IQA based algorithms have been proposed over time. One class of these algorithms including SSIM [7], FSIM [11], RFSIM [12] use handcrafted features (attributes (edge, color, etc.) in data (images) that are relevant to the modeling problem) that supposedly captures relevant factors affecting image quality. Although their performance is acceptable, there is still large room for improvement regarding the accuracy with which they reproduce human judgment of quality. Another set of algorithms, including convolutional neural network (CNN) based approaches [13, 14], employ automatic learning of features from the raw image pixels, which are superior and more efficient as they make feature selection automatic and embedded within the system itself.

1.1 Motivation

Most of the existing IQA databases usually contain limited distortion levels (5-6 levels) covering the whole quality range from “Bad” to “Excellent”, which make the images in adjacent distortion levels obviously different and easy to rank. To describe the obvious and subtle quality differences between two images, Zhang *et al* [15] use the terms “coarse-grained” and “fine-grained”. More specifically, the images with “coarse-grained” quality differences correspond to the compressed ones generated using the same codec at obvious different bitrates, while the images with “fine-grained” quality difference correspond to the compressed ones generated using different optimization methods at the same bitrate. Therefore, these databases with coarse-grained distortion variations for the same image may not be able to provide sufficient information to further improve the performance of IQA algorithms in evaluating fine-grained quality differences.

Another weakness for the existing IQA databases is that they only contain a few reference images with limited visual content. To solve this problem patch-based methods are gradually used in IQA, e.g. CNN-IQA [16], CORNIA [17] The patch-based learning methods requires the ‘ground truth’ of patch quality for training but there are only the ground truth quality of images instead of patches in IQA datasets. To deal with this problem, existing works usually assign the image quality score to all patches in this image as their ‘ground truth’, e.g. CNN-IQA [16]. This approach might introduce much noise in patches labels because in some distortion types the quality of patches in one image varies much and the patches quality score can’t be simply assigned as the image quality core.

Based on all these observations, this project promotes IQA in the new challenges of fine-grained quality assessment task by constructing a large-scale Image-Patch Quality Assessment database with fine-grained distortion differences. We also analyze 7 state-of-the-art IQA algorithms on the proposed database and show that there is still a large room to improve the IQA in the prediction of the fine-grained quality preference. Finally, we propose an FR Image-Patch model to help estimate the ‘ground truth’ quality of patches based on a state-of-the-art CNN architecture.

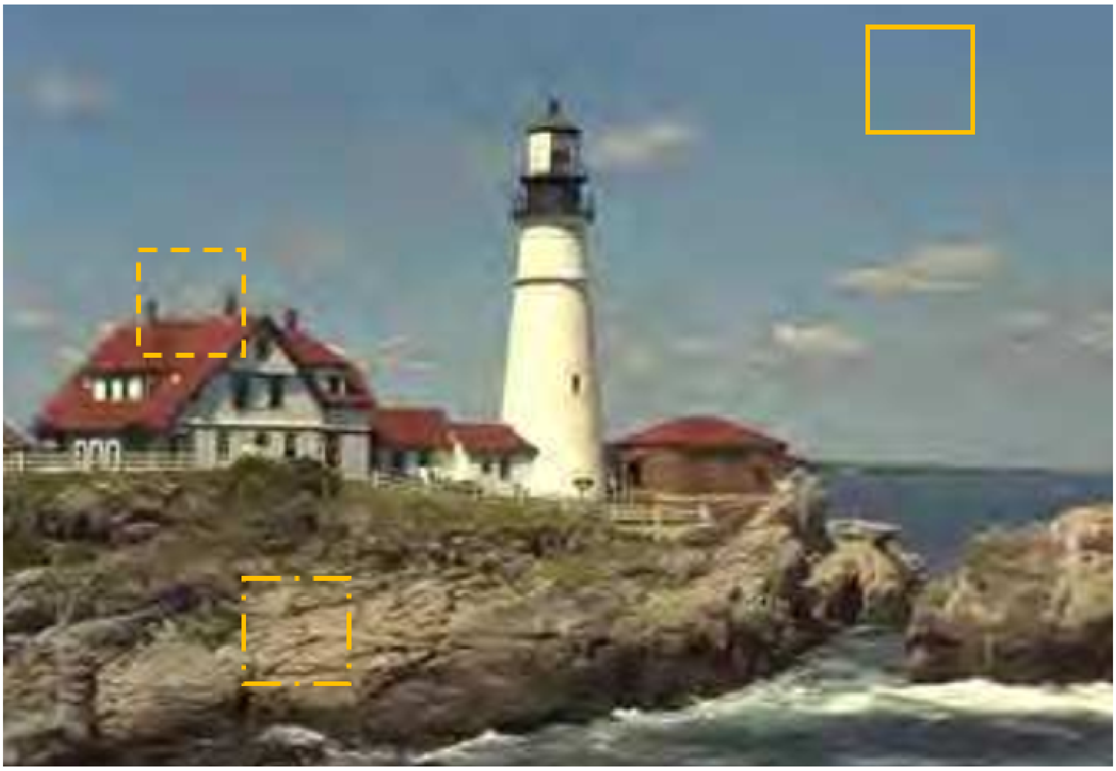


FIGURE 1.2: Example of JPEG distorted image. Different patches have different qualities.

1.2 Contributions

This thesis provides the following contributions:

1. **Image-Patch Quality Assessment dataset**

To our knowledge, this dataset is the first one constructing to provide benchmark for compressed image patch quality assessment, and also benefit for perceptual-based image compression. The existing databases with coarse-grained quality are inefficient to evaluate IQA algorithms especially patch-based methods on images with fine-grained quality differences. In perceptual-based image compression problem, for each coding block there are many coding modes to select a according their rate-distortion costs. Therefore, the proposed dataset can help researchers in image compression community to select the best IQA method to do the perceptual based image optimization. 7 well-know IQA algorithms are evaluated and analyzed on the proposed database to reveal some limitations of the existing algorithms.

2. **Deep Image-Patch Neural Network Design**

We also investigate different FR methods to model the relationship between the image patch and patch quality score. After multiple of experiments, Deep Image-Patch Quality Assessment (DIPQA) is proposed to address the problem in and end-to-end optimization. We adapt the concept of Siamese networks know from classification task [18, 19] that allow for a join regression of the features extracted from the reference and distorted patch using a deep convolution neural network.

1.3 Thesis Outline

The rest of this thesis is organized as follows. After this introduction, we present the literature review in Chapter 2 in which we introduce the fields of image quality assessment and deep learning. Next, our methodology is described in Chapter 3. Chapter 4 shows the evaluation of our database and proposed neural network based on experiment result. Finally, the conclusions and future directions are given in Chapter 5.

Chapter 2

Background

2.1 Image Quality Assessment

As portrayed in the introduction, image quality assessment (IQA) is vital for many applications, but on the other hand is hard to achieve as such because of its reliance on the quantification of human perception. The most solid approach to embrace IQA is through subjective assessments, but this is not practical in real-life applications since users can't always be dependent upon to comment on the perceived quality. On the other hand, objective image quality assessment focuses on implementing human perception models that can estimate the quality of an image as perceived by a person based solely on pixel analysis information.

In the following, we briefly review current subjective quality assessment methods to then go deeper in the state-of-the-art of methods for objective quality assessment.

2.1.1 Subjective image quality assessment

Subjective image quality assessment methods use human observers to express their personal opinion on the quality of images which are used to be assessed. Because humans are the end users in a large portion of the multimedia applications, subjective IQA methods are the most reliable and accurate for image quality assessment.

Several international standards have been proposed for performing subjective image quality assessment such as , ITU P913 [20], ITU P910 [21] and ITU BT 500 [22]. The main objective of subjective IQA methods for a given set of images is to assign a score to each of them that quantifies the perceived quality of the user. In most cases, a scaling process can be achieved, either explicitly or implicitly.

Subjective testing usually focuses on quantifying average observer's perceived quality. A group of subjects is requested to evaluate an image and give its perceived quality score. These scores are then accumulated and the final score is calculated to reflect the quality perceived by an average observer. For the calculation of this final score, different scales could be used, for example, direct scaling in which the perceived quality of an image is calculated as the mean of the scores assigned to that image by each subject. (Mean Opinion Score (MOS) or Differential Mean Opinion Score (DMOS)). The objective IQA methods (to be followed) are intended to use different models to predict these mean values.

Despite being the most accurate and reliable, subjective IQA methods are highly impractical for real-world applications as it is very expensive and time-consuming to gather an adequate number of observers to evaluate image quality. Consequently, more practical objective IQA methods are used for many applications.

2.1.2 Objective image quality assessment

Rather than using human observers, objective IQA methods are aimed at using relevant models that can predict image visual quality as perceived by humans. Because these algorithms require no human observer, they are fast and very practical for many applications in the real world, such as image enhancement, image restoration, etc.

To estimate the perceptual quality of the given image (called test image), either in the presence or absence of its reference image, most of the objective IQA methods share a common framework of three main phases as illustrated in Fig.2.1. These three phases are described in the following

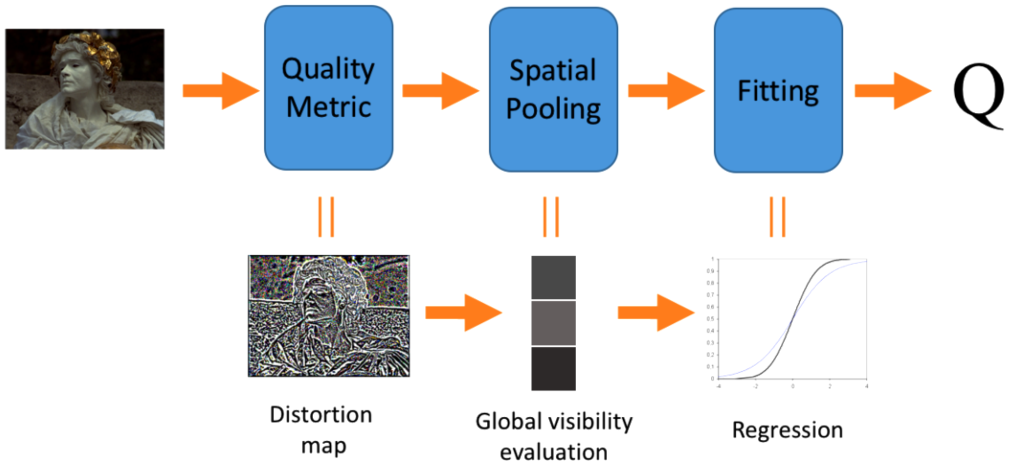


FIGURE 2.1: General Objective image quality assessment framework

1. The test image is processed pixel by pixel or region by region in accordance with the objective IQA method used to measure the amount of distortion present in it. This phase then outputs the distortion measured in the form of a distortion map containing the image quality local description. This step is equivalent to the feature extraction.

2. A multidimensional phase is produced in the first phase, but humans perceive image quality as a single global entity rather than the local properties of an image. A spatial pooling strategy is generally used to downsample the multidimensional distortion map to a single quality score in order to produce a global quality assessment. [23]

3. Since in the first two phases non-linearity, which characterizes perception, is not used, the output may not be sufficiently accurate. Thus, to increase the overall accuracy of the framework, an appropriate strategy could be applied. This requires a set of images along with their subjective quality scores (obtained through subjective testing), and a parametric model whose parameters are learned through the analysis of image model predictions and their actual subjective scores through regression. This learned model is then used to transform predicted scores into better estimates allegedly consistent with human perception.

Objective IQA strategies are classified into three large categories.

2.1.2.1 Full-reference image quality assessment (FR-IQA)

FR-IQA methods aim to achieve objective IQA goals while taking as input both reference and test images. Because these algorithms also require reference images to estimate visual quality, their scope is limited to a few applications where reference images are easily available, such as compression of images and watermarking.

Over time, a lot of FR-IQA algorithms were proposed. According to one of these methods, image quality can be computed as a peak-signal-to-noise ratio (PSNR), which is simply a ratio of a signal's maximum power and distortion power. The distortion power is generally calculated to calculate the pixel-wise difference between the reference and the distorted image in terms of mean-square-error (MSE). PSNR has the advantages of being simple and very inexpensive computationally, but it does not deliver very good performance because the essential physiological and psychophysical characteristics of the human visual system (HVS) are not included in this algorithm.

Another FR-IQA algorithm, the Structural Similarity Index (SSIM) [24], advances FR-IQA from raw pixels to structures. It is based on the assumption that HVS is highly adapted to extract structural information present in an image, and degradation of images is perceived as a change in this structural information. SSIM therefore aims to evaluate the quality of an image by measuring variations in the structural information of distorted images (in relation to their reference image). In evaluating the perceptual quality of images, SSIM has been shown to outperform PSNR

Another lately proposed FR-IQA [11] algorithm is the Function Similarity Index (FSIM). It is based on the fact that HVS uses low-level features to understand images (like edges and zero crossing). FSIM uses two features to estimate an image's quality: a primary feature called Phase Congruency, which is a contrast-invariant dimensionless measure of the local structure's significance, and an image gradient magnitude feature. FSIM shows superior performance on different datasets than PSNR and SSIM algorithms.

Recently, Bosse *et al.* [13] presents an IQA data-driven approach based on deep neural networks. The network consists of 10 convolution layers and 5 pooling layers for extraction of features, and 2 fully connected layers for regression, making it significantly deeper than related IQA models. Unique features of the proposed architecture are that I it can be used in a no-reference (NR) as well as in a FR-IQA setting with slight adaptations and (ii) it enables joint learning of local quality and local weights in a unified framework, i.e. the relative importance of local quality to the global quality estimate.

2.1.2.2 Reduced-reference image quality assessment (RR-IQA)

RR-IQA methods aim to achieve objective IQA goals by estimating the quality of the test image while using partial reference image information. Usually this partial information is in the form of features extracted from the images of the reference.

In communication networks, RR-IQA finds its application that is used to transmit images and videos. Using RR-IQA algorithms, partial reference image information transmitted through these communication networks can be used to track visual quality degradation of images and videos transmitted. In similar applications, therefore, RR-IQA algorithms are preferred over FR-IQA algorithms as presented in [25, 26].

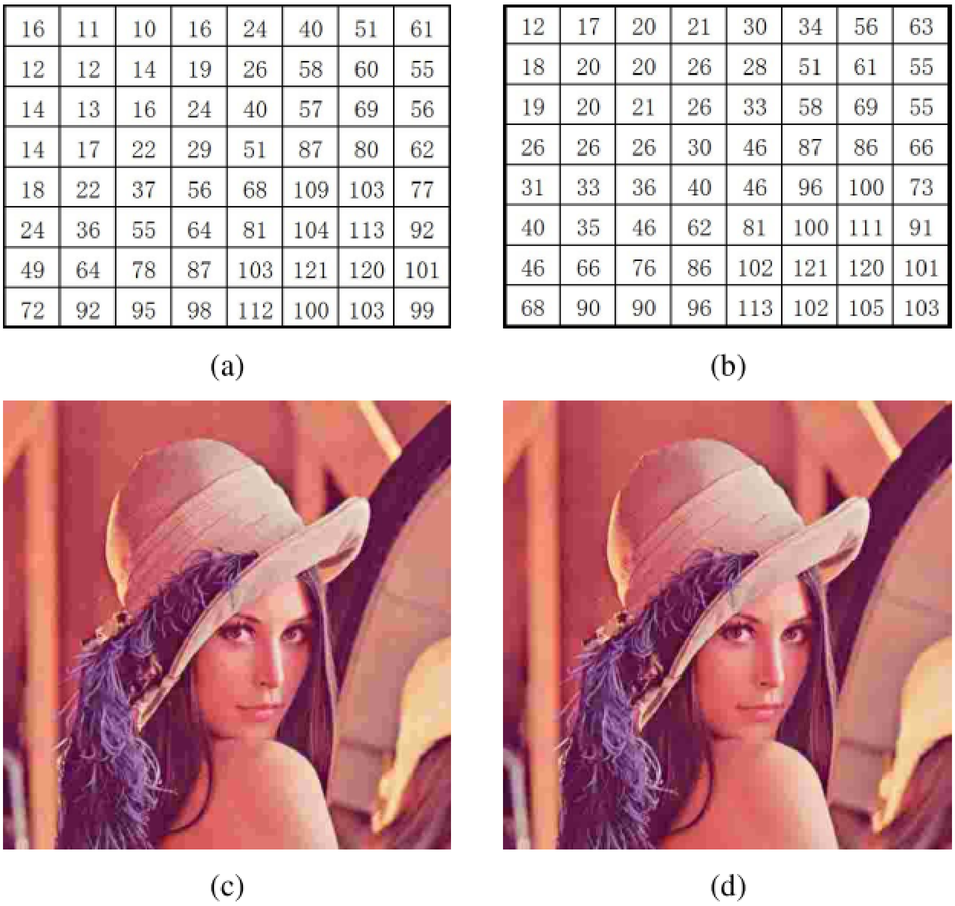
2.1.2.3 No-reference image quality assessment (NR-IQA)

NR-IQA methods are intended to achieve the objectives of objective IQA by using only test images to estimate the quality of the image. Due to the lack of information on reference images, these methods are considerably more challenging than FR-IQA and RR-IQA. But due to their application in the wide variety of fields, they are also more desirable, ranging from image processing to image enhancement, where reference images are usually not available. NR-IQA methods are also used in a wide range of online applications, such as communication systems, image acquisition systems, etc. [27], making it very important for them to be computationally cheap.

Some early NR-IQA attempts used distortion-specific methods that approach IQA tasks by using models very specific to a specific type of distortion. These methods are more specific to applications where there is prior knowledge of the type of distortion. For example, in an application to measure quality losses in compressed images, knowledge of the appearance of compression artifacts, such as blocking and ringing, could be used to design NR-IQA methods that can detect their visibility

It is more useful to have algorithms, regardless of the types of distortion, that can be applied for general purpose NR-IQA. Existing NR-IQA approaches for general purposes could be further divided into two broad categories: Natural scene statistic based approaches (NSS) and Feature learning based approaches

2.1.3 IQA in Visual Data Compression



16	11	10	16	24	40	51	61
12	12	14	19	26	58	60	55
14	13	16	24	40	57	69	56
14	17	22	29	51	87	80	62
18	22	37	56	68	109	103	77
24	36	55	64	81	104	113	92
49	64	78	87	103	121	120	101
72	92	95	98	112	100	103	99

(a)

12	17	20	21	30	34	56	63
18	20	20	26	28	51	61	55
19	20	21	26	33	58	69	55
26	26	26	30	46	87	86	66
31	33	36	40	46	96	100	73
40	35	46	62	81	100	111	91
46	66	76	86	102	121	120	101
68	90	90	96	113	102	105	103

(b)

(c)

(d)

FIGURE 2.2: Examples of quantization table and the corresponding compressed JPEG images.

(a) JPEG default quantization table at quality factor equal to 50;
(b) Optimized quantization table with the optimization goal of MS-SSIM;
(c) JPEG image using default quantization table at QF = 10, 0.234 bbp, PSNR = 30.45, SSIM = 0.819, MS-SSIM = 0.946;
(d) JPEG image using MS-SSIM optimized quantization table, 0.226 bpp, PSNR = 30.49, SSIM = 0.818, MS-SSIM = 0.953

While the bitstream has been normalized by image coding standards, different coding parameters or modes determined by different IQA metrics will obviously result in distinct compression performance. In JPEG, the custom quantization table is one of the optional coding parameters, and the default table is empirically determined based on human perception [28]. For example, in Fig.2.2(a), which is scaled to generate quantization tables for other quality factors, the quality factor (QF) quantization table of the luminance component equal to 50 is shown. In addition to the standard JPEG quantization table, the open source and well-optimized

JPEG codec, *libjpeg*, has adopted another 8 quantization tables, one of which is an optimized MS-SSIM-based quantization table as shown in Fig.2.2(b).

In [29], the researchers proposed optimization of the image-dependent quantization table based on the signal fidelity-based metric, MSE, which achieved substantial bit rate savings at the same quality as PSNR. However, because of the poor correlation between perceptual quality and PSNR, these optimization strategies can not ensure the same visual quality improvement. The researchers introduced SSIM and its variants into image and video coding in [5], [7] and [30] to optimize the process of rate distortion, but the performance improvement is not yet so satisfying. The upper and lower boundaries of the average SSIM index were derived for the first time by Channappayya *et al.*[5] as a function of the quantization rate for different source distributions, e.g. uniform, Gaussian and Laplacian. Wang *et al.*[7] used SSIM as the quality metric for optimizing rate distortion instead of MSE and achieved a bit rate saving of about 5%-10% compared to the original H.264/AVC. Ou *et al.* [30] applied SSIM to the problem of perceptual rate control with a gain of 0.008 SSIM (corresponding to a saving of 14 percent bitrate). We can see from this work that the improvements in quality are still small.

Essentially, with regard to the compression of the perceptual image, although different encoding optimization strategies can improve the image quality at the same bit rate level, the quality fluctuations are usually limited within a small range. However, most traditional IQA databases contain only coarse-grained compression distortion levels, and IQA algorithms on the fine-grained quality prediction for image compression problem can not be evaluated well. For example, the scaled quantization tables in Fig.2.2(a) and 2.2(b), respectively, compress the JPEG images in Fig.2.2(c) and 2.2(d) at similar bitrates. Although the image shows less blocking artifacts in Fig.2.2(d), it has a lower SSIM value but higher PSNR and MS-SSIM values than the image shown in Fig.2.2(c). These different IQA algorithms show opposite quality rankings on the level of fine-grained distortion, which motivates us to re-examine existing IQA algorithms and examine their suitability to distinguish fine-grained distortions

2.2 Neural Networks

2.2.1 Artificial Neural Network

Artificial Neural Network (ANN) is not a brand new idea. It was first introduced as a computational model of "nerve net" in the human brain by Warren McCulloch and Walter Pitts [31] in 1943. After that, the concept and architecture of neural networks are further developed by follow-up researchers. Neural networks have long been constrained by hardware performance. The advancement in GPU design and brain science led to a boom in the development of neural networks only in recent decades.

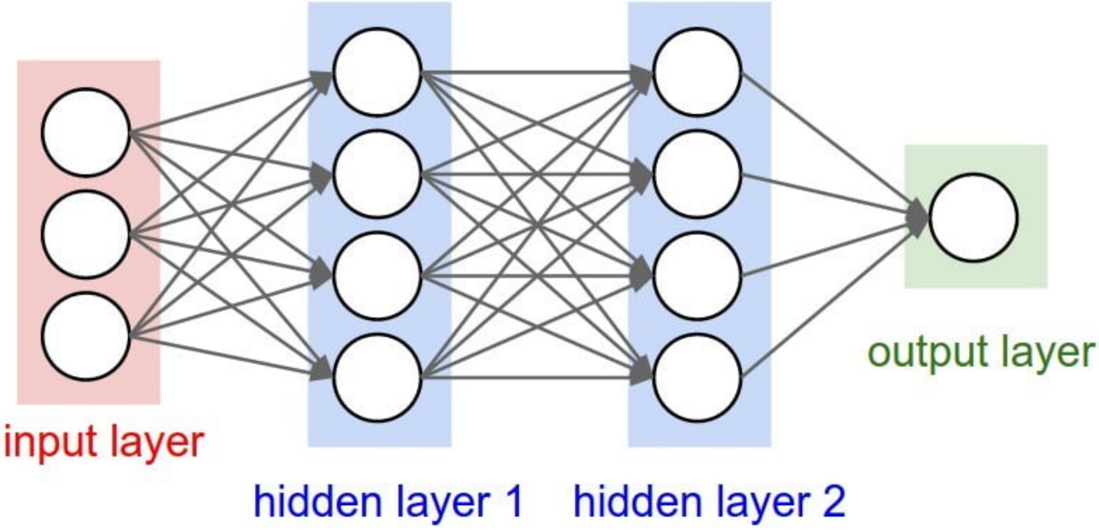


FIGURE 2.4: The basic structure of Neural Network.

A common modern neural networks consist of a large number of nodes called neurons. Each neuron does a simple calculation, usually $y = Wx + b$, where W is called Weight and b is called Bias. The neurons form multiple layers and the result value y of each neuron is then passed to the neurons in the next layer. The first layer is called input layer as shown in Fig.2.4. As its name implies, it takes features from outside the network as input. The last layer is output layer and its output value is the prediction given by the neural network.

2.2.2 Training Neural Network

First, Weight Metric and Bias Metric are initialized with random values (or pre-trained value obtained from other benchmark data). A training of the neural networks is required to adjust those parameters to fit into a particular task.

The most common and popular method of training neural network is back-propagation (BP) [32]. The goal of back-propagation is to compute the partial derivative, or gradient, $\frac{\partial E}{\partial w}$ of a loss function E with respect to any weight w in the network. The loss function E calculates the difference between prediction of neural network and its expected output, after one or a batch of sample data go through the network. A loss function is usually defined as:

$$E = \frac{1}{N}\sum_{i=1}^{N}|f(x_i) - y_i| \tag{2.1}$$

or

$$E = \frac{1}{N}\sum_{i=1}^{N}(f(x_i) - y_i)^2 \tag{2.2}$$

where $f(x)$ is the equivalent function fo the neural network. Equation 2.1 is called $L1$ loss while equation 2.2 is called $L2$ loss. In practice, $L2$ loss is the most popular one because it is more sensitive to examples that far away from expected output. Thus the trained neural network is hopefully more general. On the other hand, $L1$ loss is not that sensitive to a minority of output that far from the expectation and takes care of the average error of the majority. It is especially useful when training data is not very carefully collected and may contain incorrect samples.

Thus the progress of training by BP can be presented as:

1. Put one batch of training data through the neural network
2. Calculate the loss between output and ground truth
3. Go backward the network and calculate the partial derivative, or gradient, $\frac{\partial E}{\partial w}$ of loss function E with respect to each weight w in the network
4. Update the weights in the network according to loss, gradient and learning rate (LR)
5. Repeat step 1 to 4 until training ends, usually when a certain number of cycles set by researcher is reached or the loss value is smaller than a threshold

This method is called “back-propagation” partly because the partial derivative is calculated using chain rule:

$$\frac{\partial E}{\partial w_{ij}} = \frac{\partial E}{\partial o_j}\frac{\partial o_j}{\partial net_j}\frac{\partial net_j}{\partial w_{ij}} \tag{2.3}$$

where E is the loss function, w_{ij} is the weight from neuron i to neuron j, o_j is the output of the neuron j, net_j is the weighted sum of outputs to neuron j from the previous layer.

We can calculate them one by one from the output layer to input layer and use the result in later layers for calculating the former layers. Therefore, calculating partial derivative is actually quite cheap when doing backward.

A neural network with multiple layer structure has proved its power on image recognition [33]. However, it suffers from “the curse of dimensionality” heavily. It means that the number of parameters in the network goes up quickly when the dimension (resolution) of input image increases. Early neural networks work on low resolution images such as 20 and 32. Early benchmark datasets, MNIST [34] and CIFAR10 [35] for instance, are also collections of small images with 20 and 32 pixels. At that time, neural networks can take care of those images with hundreds or thousands of parameters. When the size of target image rise to around 200, an input layer with 40000 neurons is needed. Assuming the first hidden layer is fully connected and has the same number of neurons as the input layer, which is quite common in practice, at least 40000 individual parameters is needed in just 2 layers. The mass of parameters not only consumes computing resources, but also causes serious overfitting problems.

Overfitting means that a statistical model tries to describe each training sample rather than to find out regular patterns among the sample collection. Fig.2.5 shows a simple case of overfitting. The regression function tends to "remember" the distinguishing features of each sample individually but fails to figure out the trend of all samples. Although it passes every sample point and has 0 loss, it is not generalizable to unseen data. Even a linear function has more prediction power than it.

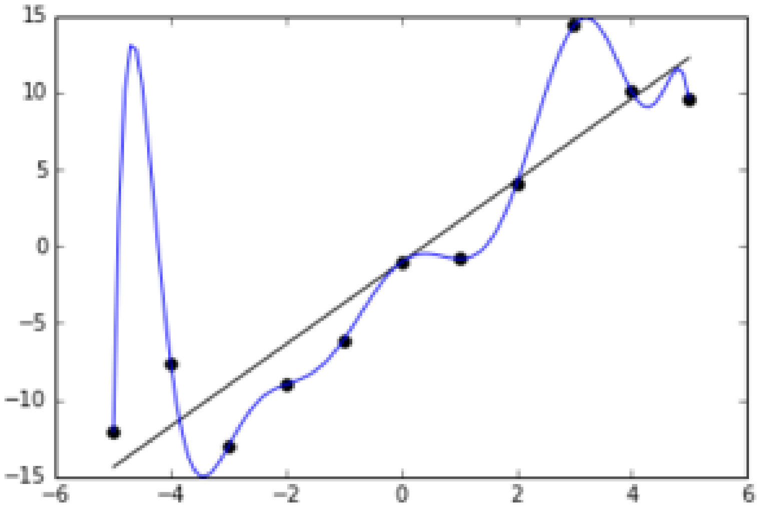


FIGURE 2.5: The function overfit the train samples.

Overfitting usually happens when neural network has too many parameters compared to the number of training samples, which enables the network easily remembering all samples. To limit the number of parameters in neural network, researchers find a way to reuse parameters in different parts of the image, which is a Convolutional Neural Network (CNN)

2.2.3 Convolutional Neural Network

A CNN is one type of neural network that specially designed for image recognizing. The architecture of CNN comes from the organization of animal visual cortex. Fig.2.6 shows the basic structure of a convolutional neural network. $X_{i,j}$ represents the pixels in the input image. A is the kernel of the first convolutional layer and is repeatedly used on each block of four pixels. The neurons on the second layer then takes the outputs of the first layer as their inputs and use the same kernel B. Fig.2.7 shows the mapping between 2 layers. One blocked in the front layer, which is a $m \times n \times d_1$ tensor ($m = n$ in most cases), is multiplied by a $m \times n \times d_1 \times d_2$ kernel and mapped to a $1 \times 1 \times d_2$ block in the next layer. The $m \times n \times d_1$ block in the front layer is called receptive field, which means all neurons in such block is connected to one neuron in the next layer, and their information is gathered together by a neuron in next layer.

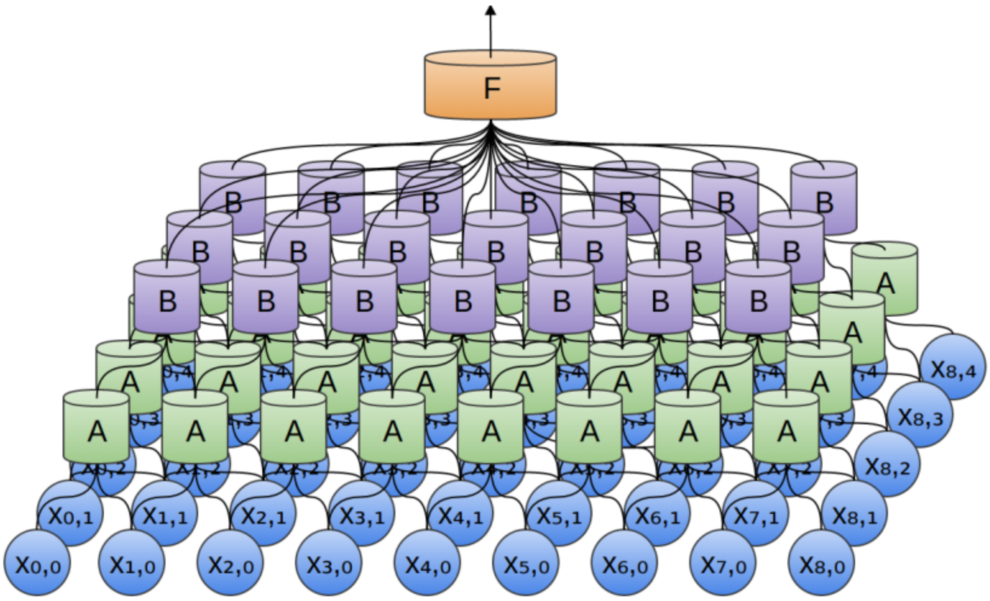


FIGURE 2.6: CNN share the kernel on each layer.

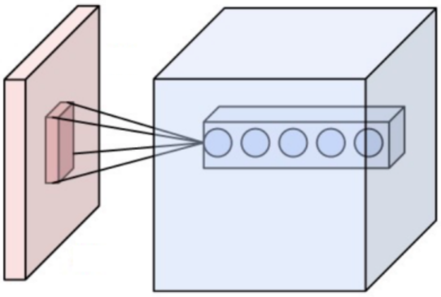


FIGURE 2.7: Kernel that maps $m \times n \times d_1$ block in the previous layer to an $1 \times 1 \times d_2$ block in next layer.

CNN is proved to be very efficient in pattern recognition and other image classification tasks. Its superior performance comes from some particular features. The most important feature is perhaps its spatial invariant. Since the same kernel is used repeatedly in the whole input space, it can detect its corresponding pattern no matter where the pattern shows up. This feature significantly reduces the number of patterns the network needs to learn.

Another important feature is its ability of abstracting and concentrating information. In Fig.2.6, each neuron A (instance of kernel) on first layer accesses information from 4 pixels. On the second layer, each neuron B connect to 4 neurons in the first layer, which means it can access information gathered from 9 pixels in the input image. As the network goes deeper, the neurons in later layers get access to larger area of the input image. At last, at the final layer, the network gets an overall abstract sense of the input image. All of these concentration and abstract procedure are learned automatically by back-propagation. It is still a mystery to researches that how those things exactly happen because the mid product of hidden layers are really difficult to understand by human beings.

ReLU Layers

ReLU layers usually stand between 2 convolutional layers. ReLU stands for Rectified Linear Units. ReLU layer applies the non-saturating activation function to the outputs of convolutional layers:

$$f(x) = max(0,x) \tag{2.4}$$

ReLU layers are very simple but they efficiently add nonlinear properties to the decision making function of the overall network as well as the sigmoid function:

$$f(x) = \frac{1}{1+e^{-x}} \tag{2.5}$$

and the hyperbolic tangent function:

$$f(x) = \tanh(x) \tag{2.6}$$

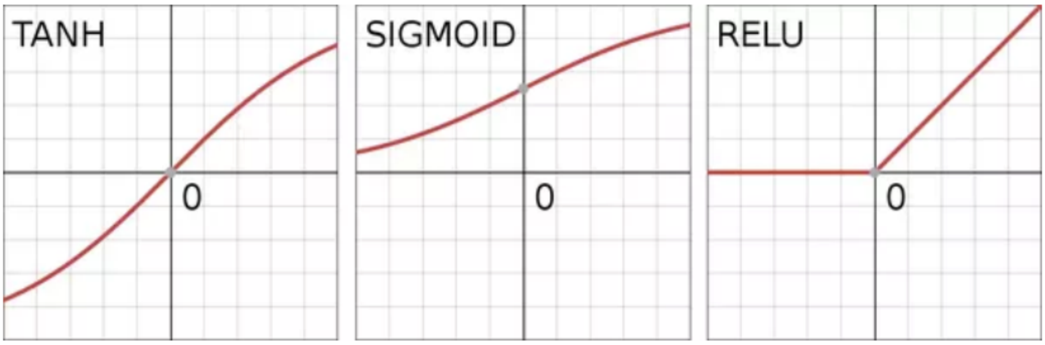


FIGURE 2.8: Common nonlinear functions used in CNN: ReLU, Sigmoid and hyperbolic tangent.

Fig.2.8 shows the response of 3 methods. The 3 methods share the same idea of inhibiting negative outputs and amplifying/keeping positive outputs of the early layers, which is a simulation of how human brain cells work. Hyperbolic tangent function and sigmoid function are widely used in old models but ReLU function becomes more preferable recently because it is proved to be much computational cheaper without making any significant differences in accuracy [36].

Pooling Layers

"Pooling" is a nonlinear down-sampling method widely used in CNNs. Fig.2.9 shows a common max pooling layer with a 2 × 2 filter size and a stride of 2. The filter move through the entries with a certain stride, pooling layer maps each block in former layer to a single value. Pooling layer concentrates the information in former layer and provides the later layers a larger "vision" in the original image. Also, pooling helps reducing the number of parameters in the network and hence has an effect of overfit control. The most popular pooling methods are max pooling and min pooling, where the filter takes the max or min value in each block as the output. Average pooling, which uses the average of all values in the block as output is also commonly used in old days. However, it has given its place to max pooling since the later one is proved to work better in practice [37].

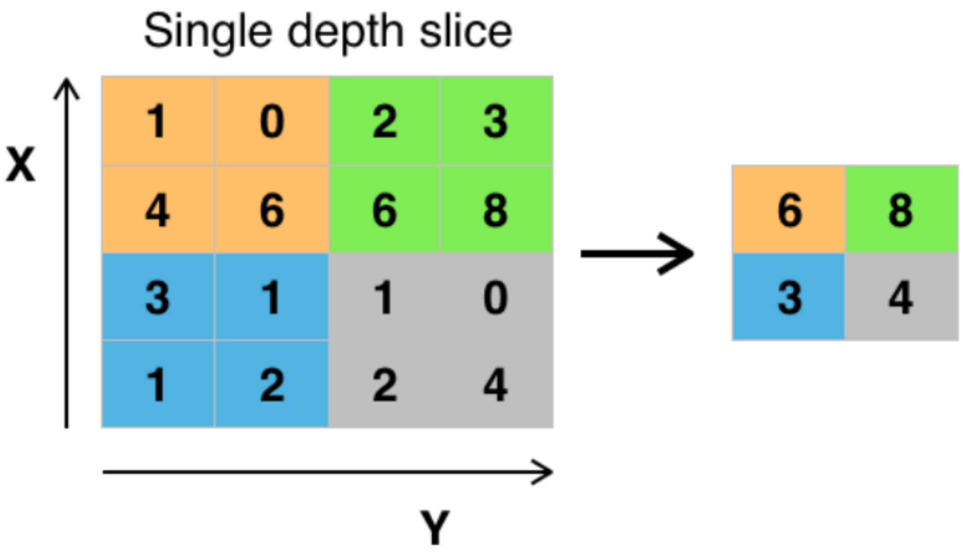


FIGURE 2.9: A max pooling layer with a 2 × 2 filter size and a stride of 2

Chapter 3

Methodology

3.1 Database Construction

All available image quality benchmark databases are only suitable for evaluating the quality of images as a whole and not able to investigate which parts of the testing image contribute to the testing results or the score for a particular patch of image. In this project, we set up an experimental database to evaluate the quality that human perceive for each image patch.

3.1.1 Testing image database

A good database for testing is critical to be the success of the research. Due to the research orientation for video encoding, testing images are cropped from extracted frames in the video test sequence and noise types are added to the original video by H265/HEVC compression before extracting. In this database, we randomly select several patches from each image so that the database includes at least three attributes: smooth texture, complex texture and edge texture.

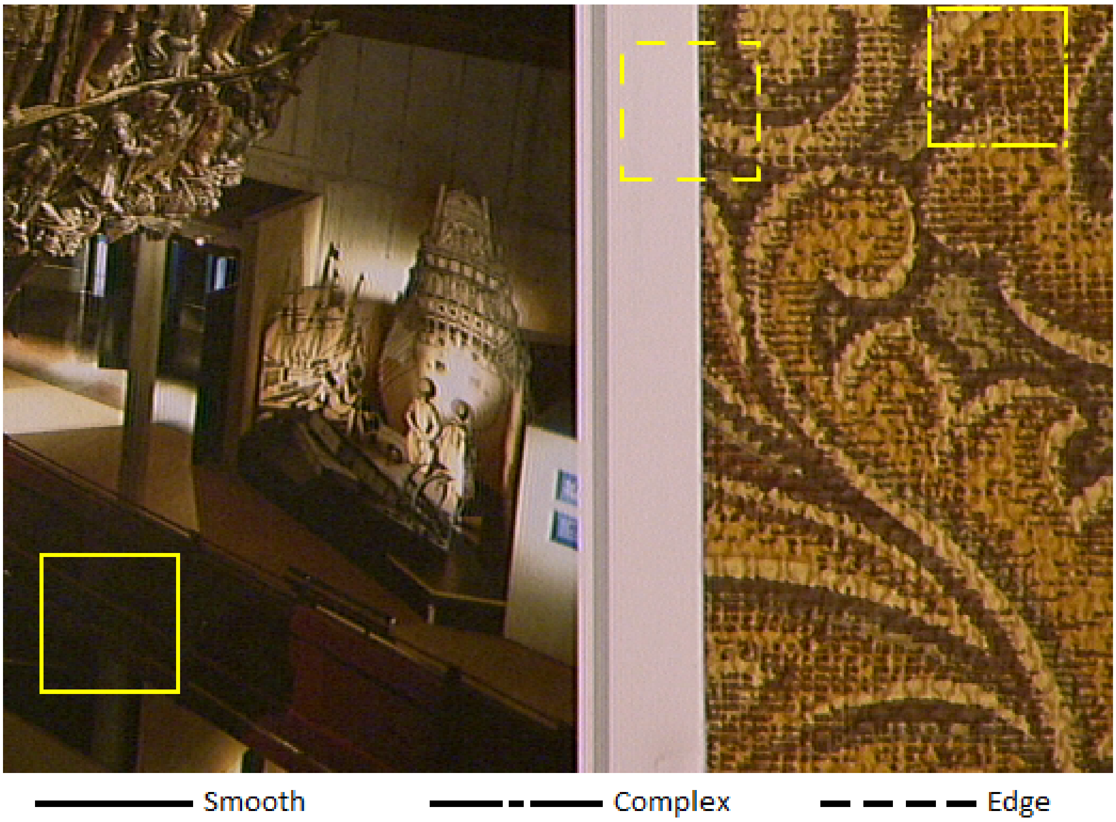


FIGURE 3.1: Selected image patch.

3.1.2 Database creation

There are 40 original source videos of high-definition (1280x720) and full high-definition (1920x1080) being compressed by H.265/HEVC with different Quantization parameters (QPs) with the range from 2 to 50. Testing images are extracted from those testing video sequences. For each video sequence, we select a different number of frames depend on the original video, this number drops in the values from 5 to 15. After that, we select random positions of the image to crop different 128x128 patches of each pair of image. We also crop the center 64x64 patches from the original pair 128x128 to evaluate in the experiments. Finally, we obtain 161,144 images: 40286 pairs of 64x64 patches and 40286 pairs of 128x128 patches.

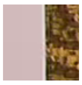
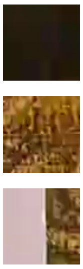
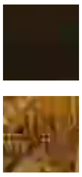
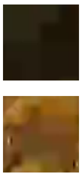
Orignal	QP						
	25	30	35	40	45	50	55
							

TABLE 3.1: Example of testing image database.

3.1.3 Testing methodology

Depending on the nature of the test, observers may be expert or non-expert. Studies have found that systematic differences can occur between different laboratories conducting similar tests [22]. One of the reasons for this is that expert observers have different view in compare with no-experts. Other explanations may include gender, age, and occupation. However, in reality, the majority of consumers should be non-expert observers are chosen for this experiment. Before final selection, all candidates have been checked to ensure that they possess normal visual acuity.

For the purpose of this experiment, 1200 subjects who are undergraduates, graduates, researchers, and lecturers of University of Fire Fighting and Prevention are employed. These subjects have been trained and practiced quality assessment of several sample images.

For the purpose of subject testing methodology, the International Telecommunication Union set the ITU-R BT.500-11 [22] standard. In such standard, there are several popular subjective methodologies for testing such as Single stimulus categorical rating, Double stimulus categorical rating, Ordering by force-choice pairwise comparison and Pairwise similarity judgments. Double stimulus categorical rating is chosen in this practical. In this method, both the test and reference images are displayed for a fixed amount of time. After that, the images disappear from the screen and observers are asked to rate the quality of the test image according to the abstract scale containing one of the five categories: excellent, good, fair, poor or bad. All those images are displayed randomly. At the beginning of each session, an explanation is given to the observers about the type of assessment, the grading scale, the sequence and timing (reference image, grey, test image, voting period).

The previous image quality assessment methods are only suitable for assess quality of image as a whole. It cannot be directly applied for our testing experiments. Therefore, we modify this image selection method in the standard so that the users can only concentrate and assess the local image patch instead of the whole image. Each pair quality is assessed with the following procedure: The subjects observe the original image within the time T1 at minimum 5s then click on the observing image patch to observe the compressed image within the time T2. After watching at least twice per image, the observers would score on scale of 5 as in Fig.3.2.

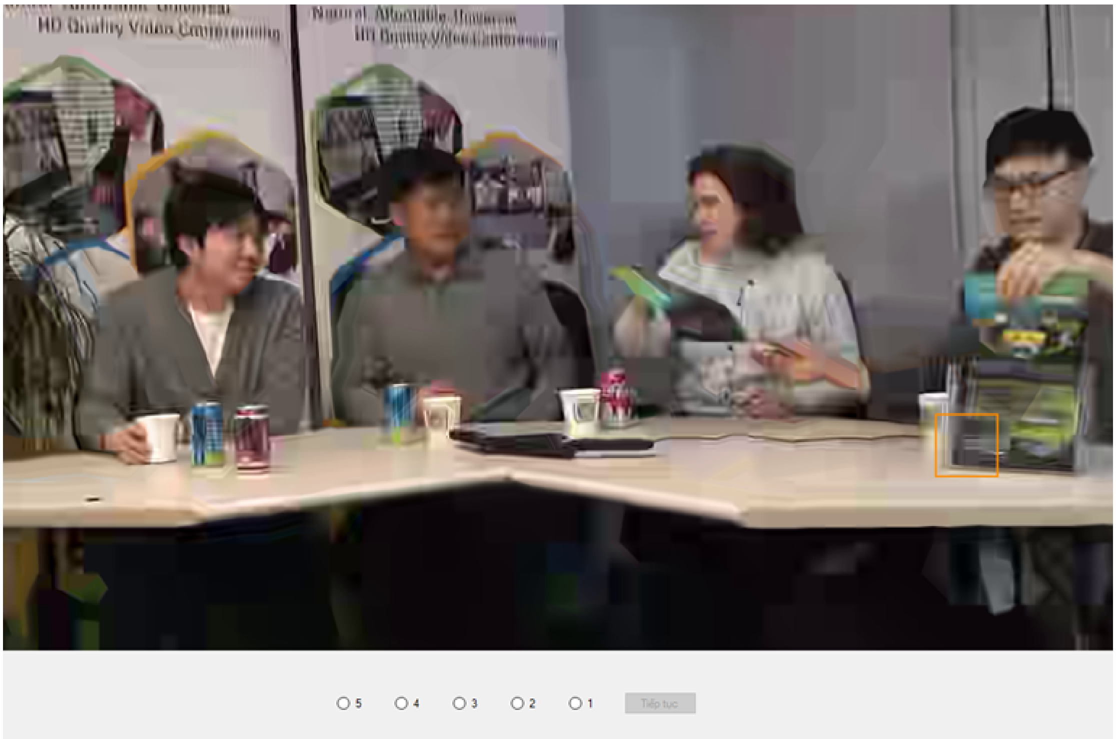


FIGURE 3.2: Testing software screenshot.

At the end of experiment, each pair is scored by the mean of the DMOS that up to 20 subjects give during the experiment.

We carefully select 1511 pairs which are scored by at least 10 people and name this sub-database HMII (Human Machine Interaction Image). HMII is used to evaluate well-known IQA algorithms and our methods in the second contribution.

3.1.4 Benchmark Analyses

To analyze the efficiency of IQA algorithms, we apply 7 state-of-the-art full reference image quality assessment methods on the proposed HMII database, to investigate their performance and demonstrate the new challenges in fine-grained image-patch quality assessment problem. The FR-IQA algorithms include PSNR, SSIM [24], RFSIM [12], FSIM [11], SRSIM [38], UQI [39], VSI [40]. The implementations of all algorithms are obtained from the public websites.

3.2 Deep Image-Patch Quality Assessment

3.2.1 Architecture

Being known as a designed architecture to learn the similarity relations between two given inputs, Siamese network has been applied for face verification [19] and signature [18] tasks. The main concept is processing two networks that share the same architecture and weights parallel. In this work, we employ Siamese network for feature extraction. Before feeding the extracted features as input to the regression layers, feature extraction is followed by a feature fusion step. The proposed architecture of DIPQA is sketched in Fig.3.3

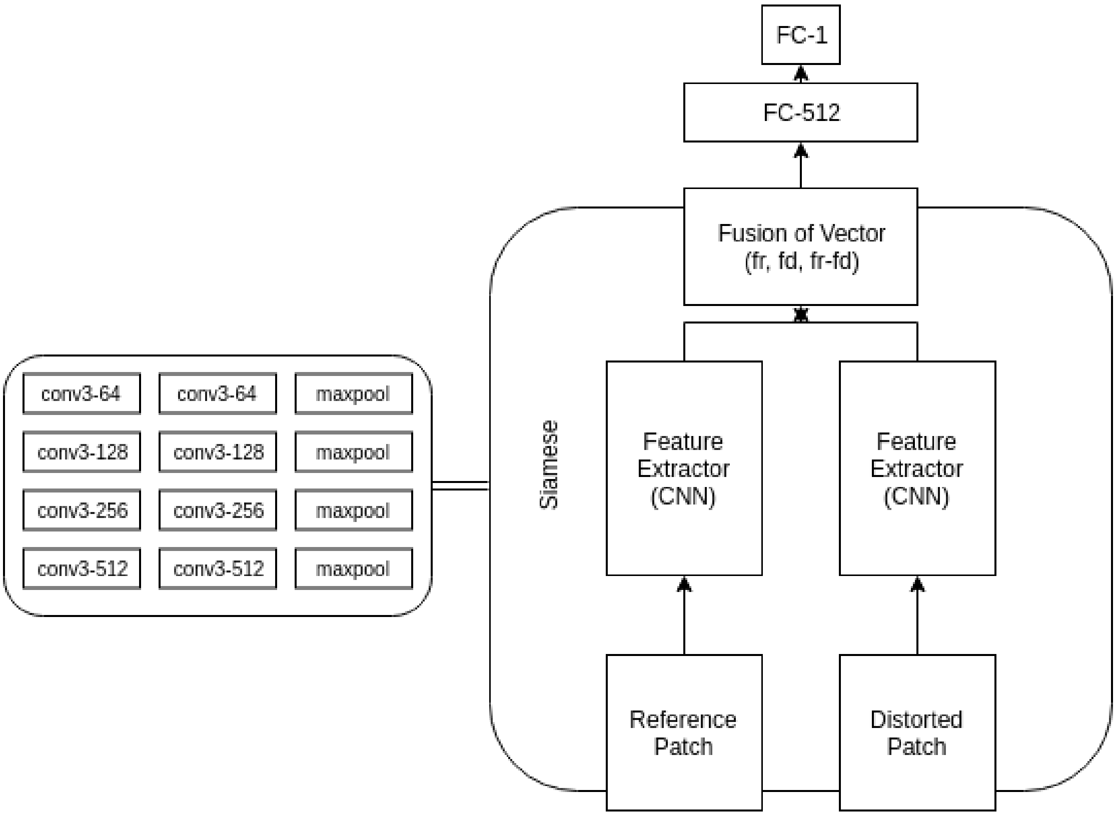


FIGURE 3.3: Deep Image-Patch Quality Assessment Network Architecture.

With the successful adaptation for various computer vision tasks [41, 42], especially in image quality assessment [13], VGGnet [43] was chosen as a base network for the feature extraction. The input of the VGG network is the size of 224 x 224 pixels. For the purpose of adjusting the network for 64 x 64 and 128 x 128 pixels, we have tried to change the architecture of VGG network such as: extend the network by 3 layers [13], cut last 3 layers, last 6 layers or even replace VGG with Resnet. Finally, we choose to cut the last 3 layers of VGGnet and achieve the best performance comparing to the other approaches. Our VGGnet-inspired DCNN comprised 12 weight layers as a feature extraction module and a regression module. The features are extracted in a series of conv3-64, conv3-64, maxpool, conv3-128, conv3-128, maxpool, conv3-256, conv3-256, maxpool, conv3-512, conv3-512, maxpool, layers. The fused features are regressed by a sequence of two fully connected layers (FC-512, FC-1). This results in about 17.3 million trainable network parameters. All convolutional layers apply 3x3 pixel-size convolution kernels and are activated through a rectified linear unit (ReLu) [44] activation function after being normalized with batch normalization. All max-pool layers have 2 x 2 pixel-sized kernels. In order to prevent overfitting, dropout regularization [45] is applied to the fully connected layers with a ratio of 0.5.

3.2.2 Feature Fusion

The feature extraction layers extract f_r and f_d which are the feature vectors of reference and distorted patch respectively. The regression layers require the network to combine these two vectors in a feature fusion step. A simplest strategy is concatenating f_r and f_d to an unique vector (f_r, f_d). Beside, $f_r - f_d$ is known as a meaningful representation for distance in feature space. Therefore, concatenating $f_r - f_d$ is expected to contribute to learning to relations between reference and distorted patch. The final output of this state is $(f_r, f_d, f_r - f_d)$

3.2.3 Training Method

For better convergence of the optimization, the feature extraction parameters are initialized with VGG13-batchnorm weights which is trained on ImageNet dataset. Our network is trained end-to-end by backpropagation, over a number of epochs. The adaptive moment estimation optimizer (ADAM) [46] is employed to alter the regular stochastic gradient descent method. Parameters of ADAM are chosen as recommended in [46] $\beta_1 = 0.9, \beta_2 = 0.999, \varepsilon = 10^{-8}$ and the learning rate α is initially set to 5×10^{-4}. The mean loss, PCC, SRCC over images during validation is computed in evaluation mode after each epoch.

Chapter 4

Evaluation

4.1 Evaluation Method

Dataset: This database comprises 1511 quality annotated images based on 1511 source reference image patches that are subject to different distortion levels of compression. Differential mean opinion score (DMOS) for this dataset were computed for each pair, which is in the range 1 to 5.

Evaluation Metrics: To evaluate the performances of the IQA algorithms, we used two standard measures, i.e., Spearman's rank order correlation coefficient (SRCC) and Pearson's linear correlation coefficient (PLCC).

Experiment Setup: Both the experiments in this thesis are performed on HMII database.

For the first experiment, the purpose is to evaluate how well an objective metric agrees with subjective preferences of subjects. We carefully select the Mathlab implementations of 7 algorithms to predict object scores for the entire database.

For the second one, different models are competed to find the best 'ground truth' predictor for patch quality. Results reported are based on the average performance of 10 folds cross-validation. Deep learning models converge after 50 epochs.

4.2 Experiment results

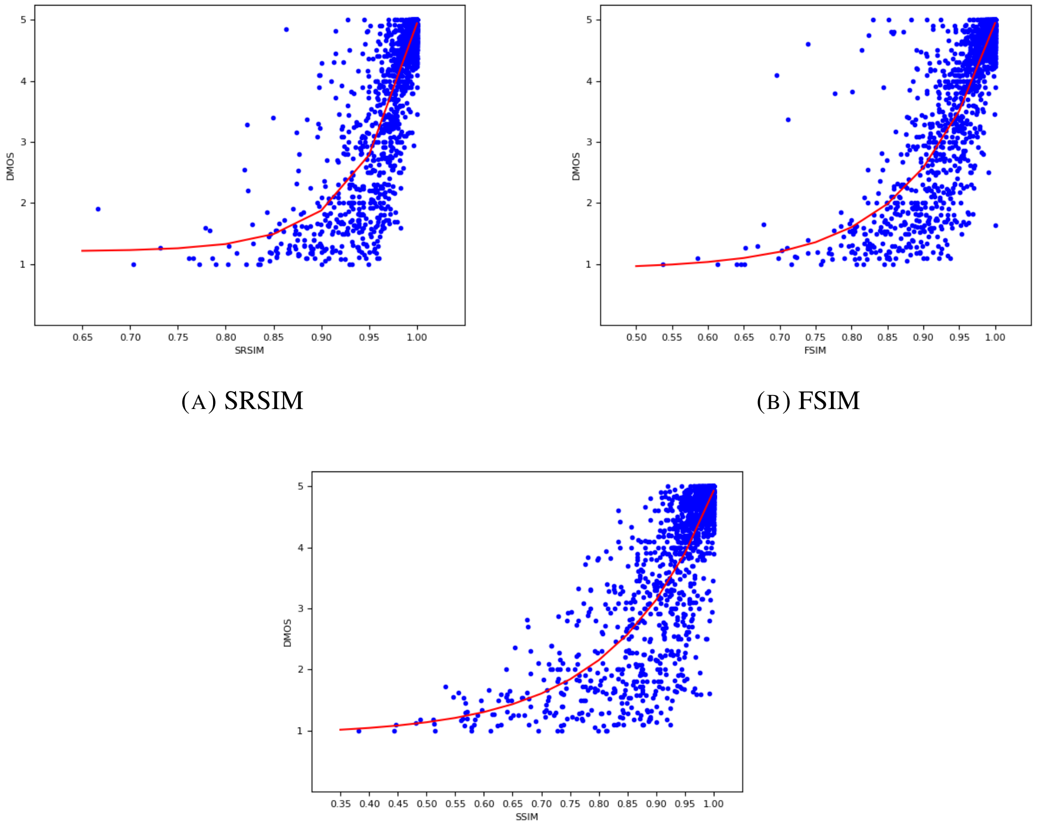
4.2.1 HMII Benchmark Analysis

First, we evaluate the pairwise preference consistency using the classic correlation coefficients SRCC and PLCC, as shown in Table 4.1. The SRCC and PLCC are the average values for the distorted images of the same reference image, and the top 2 correlation coefficient values are highlighted. We can see that the PSNR and UQI are poorly correlated with human perceptual quality, and even contrary to subjective results. This defective performance of PSNR is also mentioned in the work of Zhang *et al.*[15] about Fine-Grained Quality Assessment. Although VSI combine the HVS features and achieve more consistent results than PSNR in global image assessment, it is poorly correlated with human perceptual quality in fine-grained patch quality assessment. For the two correlation coefficients, these IQA methods shows quite similar characteristics. As a whole, FSIM achieves top 2 performance for all the cases and the SSIM achieves better performance with PLCC while SRSIM performs better with SRCC.

	HMII (64x64)		HMII (128x128)	
	PLCC	SRCC	PLCC	SRCC
SSIM[24]	**0.785**	0.787	**0.795**	0.797
RFSIM[12]	0.774	0.757	0.789	0.759
FSIM[11]	**0.794**	**0.799**	**0.824**	**0.815**
PSNR	0.200	0.737	0.194	0.752
UQI[39]	0.023	0.621	0.012	0.589
VSI[40]	0.765	0.765	0.768	0.786
SRSIM[38]	0.777	**0.803**	0.718	**0.803**

TABLE 4.1: PLCC and SRCC for different IQA algorithms

In addition to visualize the objective score by the top 3 IQA algorithms, we plot the distributions of subjective scores and objective scores on a 2-D graph and also plot the fitted curve on the same figure. The following figure shows the scatter distributions of subjective DMOS versus the predicted scores obtained by the SSIM, FSIM and SRSIM on the proposed database.



(A) SRSIM

(B) FSIM

(C) SSIM

FIGURE 4.1: Objective Score by top 3 IQA on HMII

From the plots, we can see that these IQA algorithms tend to predict higher score for patches. SRSIM and FSIM frequently predict score which is higher than $0.9_{[0-1]}$ for the image with DMOS is greater than $2_{[1-5]}$. Although FSIM achieves highest performance with the two correlation coefficients, SSIM achieves more consistent results with subjective results on the diagrams. These results prove that some existing IQA models perform poorly in distinguishing the fine-grained distortion levels, which are feasible to determine by human visual system. Therefore, these metrics may not be suitable for perceptual-based image compression because the distortion differences between various coding modes are usually marginal. Moreover, the fine-grained image-patch quality assessment is demanded and should be evaluated on the HMII databases.

4.2.2 Image-Patch Models

In this experiment, we use the following models to evaluate with our proposed DIPQA:

- *IPM*: Zhang *et al.*[15] assume the the curve model to predict image-patch quality is a cubic polynomial function:

$$f(\Phi(\mathbf{d});\theta) = a\Phi(\mathbf{d})^3 + b\Phi(\mathbf{d})^2 + c\Phi(\mathbf{d}) + d$$

  where $\theta = a, b, c, d$ are the parameters for the non-linear function of Image-Patch model and $\Phi(\mathbf{d})$ represents the feature of patch $\mathbf{d}$. MSE and SSIM are chosen for the design of features. In our work, we tried top 3 FR-IQA methods from the first experiment: SSIM, FSIM and SRSIM.

- *DIQaM*: Bosse *et al.*[13] present a Deep Neural Networks for No-Reference and Full-Reference Image Quality Assessment which obtains superior performance on different IQA benchmarks. We utilize the extractor architecture from this paper to train a Deep Neural Network on our database.

First, we use the previous works to extract the SSIM, FSIM and SRSIM feature for IPM. Then, the above curve model is fitted using the least square method to obtain the parameters that best fit the learning set. DIQaM, DIPQA (VGG extractor) and DIPQA (VGG finetuning) is built with the similar architecture which share them same regression part. With the proposed DIPQA, we approach with two different tuning strategies: one is fine-tuned with the VGGNet weight and then retrained with HMII; the other one use VGGNet as a feature extractor, this part is not trained with the entire network.

	HMII (64x64)		HMII (128x128)	
	PLCC	SRCC	PLCC	SRCC
IPM (SSIM)	0.836	0.784	0.843	0.794
IPM (FSIM)	0.848	0.795	0.871	0.810
IPM (SRSIM)	0.854	0.802	0.857	0.798
DIQaM	0.916	0.824	0.905	0.819
DIPQA (VGG extractor)	0.802	0.754	0.830	0.760
DIPQA (VGG finetuning)	**0.921**	**0.848**	**0.955**	**0.871**

TABLE 4.2: Comparing different Full-Reference Image-Patch approaches

The Table 4.2 summarizes the performance of the proposed models in comparison to other methods on HMII database in terms of PLCC and SRCC. With any of the two correlation coefficients, DIPQA (VGG finetuning) achieve superior performance to the others. From the results of this project, we can also see that the larger size of the patch seem to be more accurate when assessing image-patch quality by Objective models.

Chapter 5

Conclusions

5.1 Conclusions

This project presents a new subject quality rating database considering local image quality assessment. Due to the lack of ‘ground truth’ quality of patches, we expect HMII to be a useful database for patch-based approaches. We also introduce a simple effective patch-based deep neural network that allows for feature learning and regression in an end-to-end framework. We believe that this proposed approach could achieve better result if we enlarge HMII database.

5.2 Future work

HMII Database: There are still some limitations on the proposed database to improve.

- Enlarge database to increase the number of image and them number of subject per image
- Generate more images to cover more type of distortions
- Filter with different outlier detection methods

Image-Patch model: In the future, we are planning to design more Image-Patch models and do experiments to evaluate on different databases. With DIPQA, there are some applications that we also consider:

- Applied in Image and Video Compression

- Associated a pooling state to compete with other Image and Video Quality Assessment (VQA) algorithms
- Improve current IQA/VQA algorithms

References

[1] W. Lin and C. C. Jay Kuo, "Perceptual visual quality metrics: A survey," *Journal of Visual Communication and Image Representation*, 2011.

[2] H. Wang, X. Zhang, C. Yang, and C. C. Kuo, "Analysis and Prediction of JND-Based Video Quality Model," in *2018 Picture Coding Symposium, PCS 2018 - Proceedings*, 2018.

[3] S. Wang, K. Gu, X. Zhang, W. Lin, S. Ma, and W. Gao, "Reduced-Reference Quality Assessment of Screen Content Images," *IEEE Transactions on Circuits and Systems for Video Technology*, 2018.

[4] Y. Zhang, W. Lin, X. Zhang, Y. Fang, and L. Li, "Aspect Ratio Similarity (ARS) for image retargeting quality assessment," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 2016.

[5] S. S. Channappayya, A. C. Bovik, and R. W. Heath, "Rate bounds on SSIM index of quantized images," *IEEE Transactions on Image Processing*, 2008.

[6] Z. Chen, W. Lin, and K. Ngi Ngan, "Perceptual video coding: Challenges and approaches," in *2010 IEEE International Conference on Multimedia and Expo, ICME 2010*, 2010.

[7] S. Wang, A. Rehman, Z. Wang, S. Ma, and W. Gao, "SSIM-motivated rate-distortion optimization for video coding," *IEEE Transactions on Circuits and Systems for Video Technology*, 2012.

[8] X. Zhang, S. Wang, K. Gu, W. Lin, S. Ma, and W. Gao, "Just-Noticeable Difference-Based Perceptual Optimization for JPEG Compression," *IEEE Signal Processing Letters*, 2017.

[9] X. Zhang, R. Xiong, W. Lin, J. Zhang, S. Wang, S. Ma, and W. Gao, "Low-Rank-Based Nonlocal Adaptive Loop Filter for High-Efficiency Video Compression," *IEEE Transactions on Circuits and Systems for Video Technology*, 2017.

[10] S. Ma, X. Zhang, J. Zhang, C. Jia, S. Wang, and W. Gao, “Nonlocal in-loop filter: The way toward next-generation video coding?” in *IEEE Multimedia*, 2016.

[11] L. Zhang, L. Zhang, X. Mou, and D. Zhang, “FSIM: A feature similarity index for image quality assessment,” *IEEE Transactions on Image Processing*, 2011.

[12] L. Zhang, L. Zhang, and X. Mou, “RFSIM: A feature based image quality assessment metric using Riesz transforms,” in *Proceedings - International Conference on Image Processing, ICIP*, 2010.

[13] S. Bosse, D. Maniry, K. R. Müller, T. Wiegand, and W. Samek, “Deep Neural Networks for No-Reference and Full-Reference Image Quality Assessment,” *IEEE Transactions on Image Processing*, vol. 27, no. 1, pp. 206–219, 2018.

[14] L. Kang, P. Ye, Y. Li, and D. Doermann, “Simultaneous estimation of image quality and distortion via multi-task convolutional neural networks,” in *Proceedings - International Conference on Image Processing, ICIP*, 2015.

[15] X. Zhang, W. Lin, S. Wang, J. Liu, S. Ma, and W. Gao, “Fine-Grained Quality Assessment for Compressed Images,” *IEEE Transactions on Image Processing*, vol. 28, no. 3, pp. 1163–1175, 2019.

[16] L. Kang, P. Ye, Y. Li, and D. Doermann, “Convolutional neural networks for no-reference image quality assessment,” in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2014.

[17] P. Ye, J. Kumar, L. Kang, and D. Doermann, “Unsupervised feature learning framework for no-reference image quality assessment,” in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2012.

[18] J. BROMLEY, J. W. BENTZ, L. BOTTOU, I. GUYON, Y. LECUN, C. MOORE, E. SÄCKINGER, and R. SHAH, “SIGNATURE VERIFICATION USING A “SIAMESE” TIME DELAY NEURAL NETWORK,” *International Journal of Pattern Recognition and Artificial Intelligence*, 2004.

[19] S. Chopra, R. Hadsell, and Y. LeCun, “Learning a similarity metric discriminatively, with application to face verification,” in *Proceedings - 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2005*, 2005.

[20] ITU, "Methods for the subjective assessment of video quality, audio quality and audiovisual quality of Internet video and distribution quality television in any environment," *Recommendation ITU-T P.913*, 2014.

[21] ITU-T Recommendation P.910, "Subjective video quality assessment methods for multimedia applications," *International Telecommunication Union, Geneva*, 2008.

[22] I.-r. Bt, Q. Itu-r, T. Itu, and R. Assembly, "RECOMMENDATION ITU-R BT . 500-11 Methodology for the subjective assessment of the quality of television pictures ANNEX 1 Description of assessment methods Common features," *Methodology*, 2002.

[23] Z. Wang and X. Shang, "Spatial pooling strategies for perceptual image quality assessment," in *Proceedings - International Conference on Image Processing, ICIP*, 2006.

[24] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Transactions on Image Processing*, 2004.

[25] S. Atsawaraungsuk and P. Horata, "Evolutionary circular-ELM for the reduced-reference assessment of perceived image quality," *Lecture Notes in Electrical Engineering*, 2015.

[26] J. A. Redi, P. Gastaldo, I. Heynderickx, and R. Zunino, "Color distribution information for the reduced-reference assessment of perceived image quality," *IEEE Transactions on Circuits and Systems for Video Technology*, 2010.

[27] D. M. Chandler, "Seven Challenges in Image Quality Assessment: Past, Present, and Future Research," *ISRN Signal Processing*, 2013.

[28] G. K. Wallace, "The JPEG still picture compression standard," *IEEE Transactions on Consumer Electronics*, 1992.

[29] V. Ratnakar and M. Livny, "An efficient algorithm for optimizing DCT quantization," *IEEE Transactions on Image Processing*, 2000.

[30] T. S. Ou, Y. H. Huang, and H. H. Chen, "SSIM-based perceptual rate control for video coding," *IEEE Transactions on Circuits and Systems for Video Technology*, 2011.

[31] W. S. McCulloch and W. Pitts, "A logical calculus of the ideas immanent in nervous activity," *The Bulletin of Mathematical Biophysics*, 1943.

[32] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, 1986.

[33] H. A. Rowley, S. Baluja, and T. Kanade, “Neural network-based face detection,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1998.

[34] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based learning applied to document recognition,” *Proceedings of the IEEE*, 1998.

[35] A. Krizhevsky, “Learning Multiple Layers of Features from Tiny Images,” *Technical Report, Department of Computer Science University of Toronto*, 2009.

[36] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet Classification with Deep Convolutional Neural Networks,” in *ImageNet Classification with Deep Convolutional Neural Networks*, 2012.

[37] D. Scherer, A. Müller, and S. Behnke, “Evaluation of pooling operations in convolutional architectures for object recognition,” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2010.

[38] L. Zhang and H. Li, “SR-SIM: A fast and high performance IQA index based on spectral residual,” in *Proceedings - International Conference on Image Processing, ICIP*, 2012.

[39] Z. Wang and A. C. Bovik, “A universal image quality index,” *IEEE Signal Processing Letters*, 2002.

[40] L. Zhang, Y. Shen, and H. Li, “VSI: A visual saliency-induced index for perceptual image quality assessment,” *IEEE Transactions on Image Processing*, 2014.

[41] R. Girshick, “Fast R-CNN,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2015.

[42] E. Shelhamer, J. Long, and T. Darrell, “Fully Convolutional Networks for Semantic Segmentation,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

[43] K. Simonyan and A. Zisserman, “Very Deep Convolutional Networks for Large-Scale Image Recognition,” pp. 1–14, 2014. [Online]. Available: http://arxiv.org/abs/1409.1556

[44] V. Nair and G. Hinton, “Rectified Linear Units Improve Restricted Boltzmann Machines,” in *Proceedings of the 27th International Conference on Machine Learning*, 2010.

[45] I. Sutskever, G. Hinton, A. Krizhevsky, and R. R. Salakhutdinov, “Dropout : A Simple Way to Prevent Neural Networks from Overfitting,” *Journal of Machine Learning Research*, 2014.

[46] D. P. Kingma and J. L. Ba, “Adam: A method for stochastic gradient descent,” *ICLR: International Conference on Learning Representations*, 2015.