No-reference objective quality assessment and enhancement for compressed images and videos

Yang, Gao

2011

Yang, G. (2011). No-reference objective quality assessment and enhancement for compressed images and videos. Doctoral thesis, Nanyang Technological University, Singapore.

<https://hdl.handle.net/10356/44824>

<https://doi.org/10.32657/10356/44824>

No-Reference Objective Quality
Assessment and Enhancement for
Compressed Images and Videos

Yang Gao

School of Electrical and Electronic Engineering

A thesis submitted to the Nanyang Technological University in
partial fulfilment of the requirement for the degree of
Doctor of Philosophy

2011

Declaration of Authorship

I, Yang Gao, declare that this thesis titled, *No-Reference Objective Quality Assessment and Enhancement for Compressed Images and Videos* and the work presented in it are my own, and have not been submitted for a higher degree to any other university or institute.

Date

Yang Gao

To my parents.

Acknowledgements

I would like to express my deepest appreciation and sincere gratitude to my supervisor, Associate Professor Tan Yap Peng, for his invaluable inspiration, guidance and support. This work would not have been possible without him. My thanks also go to Dr. Wang Ci and Dr. Chen Zhenzhong, for their help and support.

I would like to acknowledge School of EEE, NTU, for awarding me the research scholarship. Its excellent facilities and resources have also been essential to this work.

Finally, I would like to thank my parents for their love, understanding, and support through my years of study away from home.

Contents

Acknowledgements	i
Summary	vii
List of Figures	ix
List of Tables	xiii
Abbreviations	xv
1 Introduction	1
1.1 Motivation and Objective	1
1.2 Scope and Main Contributions of the Thesis	5
1.3 Thesis Organization	7
2 Background and Literature Review	9
2.1 Image and Video Compression	10
2.1.1 Digital Image and Video	10
2.1.2 Spectral and Temporal Redundancy	12
2.1.3 Discrete Cosine Transformation	17
2.1.4 Quantization	20
2.1.5 Image and Video Quality	23
2.2 Related Work	28

2.2.1	Quality Assessment	28
2.2.1.1	Reduced-reference Techniques	28
2.2.1.2	No-reference HVS-based Techniques	31
2.2.1.3	No-reference Statistical Techniques	32
2.2.2	Quality Enhancement	34
2.2.2.1	Single-input Enhancement and Restoration	34
2.2.2.2	Multiple-description Coding and Other Related Work	36
2.2.3	Projection onto Convex Sets	38
2.3	Conclusion	39
3	No-reference PSNR Estimation for Compressed Pictures	41
3.1	Existing Methods for NR PSNR Estimation	43
3.1.1	Distribution Models for DCT Coefficients	43
3.1.2	Laplace-distribution-based Methods	45
3.2	Shifted Block Boundaries for Parameter Estimation	46
3.2.1	Distribution of Coefficients	47
3.2.2	The Effect of Quantization	49
3.2.3	The Case of Residue Coefficients (Video)	51
3.2.4	The Number of Shifted Locations	53
3.2.5	Neural Network Regression	53
3.2.6	Alternative: Enlarged Blocks	56
3.3	Experimental Results	57
3.3.1	JPEG	59
3.3.1.1	Shifted Blocks	59
3.3.1.2	Enlarged Blocks	62
3.3.2	H.264	63
3.4	Conclusion	65
4	Relative Compression Quality Assessment	67

4.1	Trivial and Nontrivial Cases	69
4.2	Conservative Estimate of the Difference of MSE	71
4.3	Experimental Results	77
4.3.1	Standard vs. Alternative JPEG Quantization Matrices . . .	78
4.3.2	JPEG vs. JPEG2000	80
4.3.3	Region-based Image Quality Comparison	80
4.3.4	The Effect of Compression Ratio	83
4.3.5	Other Possible Measures	85
4.4	Conclusion	86
5	Reconstruction from Multiple Compressed Copies	87
5.1	Reconstruction Using Constraints	89
5.1.1	Quantization Constraints	89
5.1.2	Smoothness Constraint	91
5.1.3	Projection onto Convex Sets	93
5.1.3.1	Initial Estimation	93
5.1.3.2	Projection Proper	98
5.2	Experimental Results	101
5.2.1	Still Images	101
5.2.1.1	Images of the Same Size and Format	101
5.2.1.2	Images of Different Sizes	105
5.2.2	Video Sequences	107
5.2.2.1	Effect of GOP Structures	108
5.2.2.2	Effect of Input Video Quality	112
5.2.2.3	Videos in Different Formats	115
5.2.2.4	Effect of the Number of Input Copies	116
5.3	Conclusion	118
6	Conclusion and Future Research	121

6.1	Review of Objectives	121
6.1.1	No-reference PSNR Estimation for Compressed Pictures . .	122
6.1.2	Relative MSE Estimation for Compressed Images	123
6.1.3	Quality Enhancement Based on Multiple Compressed Copies	124
6.1.4	Summary of Contributions	125
6.2	Future Research	125
Publications		127
Bibliography		131

Summary

This thesis proposes new and effective methods to assess the objective quality of compressed digital images and videos, and to enhance the reconstruction quality of them using multiple compressed copies.

Digital images and videos are shared over the Internet by millions of users every day. Such contents are usually highly compressed for efficient transmission at the expense of fidelity. The end users may want to know the quality of these images and videos after lossy compression, should she wish to archive or further process them. There may also be several differently compressed copies of the same content residing at different locations on the web. The end user may like to compare the quality between them so as to pick the best one, or even generating a “super copy” with image quality superior to any of the copies available.

The objective quality of a compressed picture can be assessed by estimating the coefficient distributions of the corresponding unknown original picture. In the thesis, we identify the primary source of inaccuracy in the existing methods, and present our observations on the behavior of block-transform-domain statistics with modified block boundaries and quantization. Based on these, we propose a new no-reference quality assessment method using alternative block boundaries to generate unique observations from the compressed picture. By combining this technique with neural network regression, we are able to significantly reduce the

quality estimation errors compared with the state-of-the-art methods. The proposed technique allows the end user to more accurately assess the objective quality of individual compressed pictures.

When several image or video copies with the same content are available, the comparison of their relative quality becomes more meaningful than assessing their quality one by one. We show that the out-of-bound portion of a reconstructed coefficient relative to the quantization range of another copy can be used as a conservative indicator for relative quality. Based on this mechanism, we propose a method to make qualitative comparison between compressed copies at picture area level. The method is based on the knowledge that the original picture value lies within the quantization range of every copy, and we experimentally show it to be more robust to different compression formats and settings than assessing the quality of each copy individually. The method is useful for the end user to pick the best copy when multiple compressed copies are available online.

Furthermore, each of these copies contains unique information about the same original that other copies do not have. We explain the scenarios giving rise to the varieties in information, and that such varieties make it possible to reconstruct the picture more faithfully by using all compressed copies than any individual copy. We propose a method based on projection onto convex sets to combine multiple nonlinear constraints to generate pictures superior in quality to any of the input copies. Through simulations, we show that our projection-based method is able to make use of copies of different formats and frame sizes as well as image priors such as smoothness condition to reduce visual artifacts. Higher objective and subjective quality are achieved at the same time.

List of Figures

1.1	Illustration of the problems considered	4
2.1	A digital image, an enlarged 8-by-8 pixel area, and the values of a pixel	11
2.2	A digital image, expressed in RGB intensities (upper row) and Y- CbCr component values (lower row). Brighter means larger in value.	13
2.3	Illustration of motion estimation and compensation	13
2.4	Examples of common GOP structures. Arrows show directions of reference.	16
2.5	Basis functions of 8-by-8 DCT	18
2.6	Effect of 2D DCT: top - image block, bottom - coefficients	19
2.7	Effect of quantization on DCT coefficients: top - inverse-transformed image block, bottom - representative coefficient values	21
2.8	Quantization with step-size q and half-deadzone $\Delta q > 0.5q$	22
3.1	The default block divisions and a shifted one	42
3.2	Probability distributions of 8-by-8 block-DCT coefficients at each frequency, Lena image. Normalized for presentation.	44
3.3	Log-probability density distribution of the default and shifted block- DCT coefficients at frequency (6,7) of the Lena image, 8×8 blocks	48
3.4	Log-probability density distribution of the block-DCT coefficients at frequency (6,7) of the Lena image, JPEG Quality = 75	50

3.5	A neural network with a single hidden layer (video parameters in parentheses)	55
3.6	The default block divisions and the enlarged ones	57
3.7	Test images from LIVE database	58
3.8	Scatter plots of PSNR estimation results, JPEG	61
3.9	Scatter plots of PSNR estimation results, H.264	63
4.1	Quantization Matrices: IJG Quality Level 44 (left), Photoshop Quality Level 3 (right)	71
4.2	Coefficient X and the effect of quantization	72
4.3	Conservative error of B with respect to A (left) and that of A with respect to B (right). Unknown or unused values are shown as dotted lines.	74
4.4	Conservative errors between A and B, when the pictures are not compressed from the same original	74
4.5	Test images used, from left: Lena, Barbara, Mandrill and Tiffany .	77
4.6	Difference in MSE estimated with four different methods	79
4.7	Difference in MSE estimated by the newly proposed method, JPEG	79
4.8	Difference in MSE estimated using pairs of JPEG and JPEG 2000 images	81
4.9	Regional quality maps by 8-by-8 blocks. Left: estimated, right: ground truth	82
4.10	Visual examples. Left to right: thresholded quality maps, Copy A, Copy B.	83
4.11	Estimated zero-crossings at various compression levels	84
4.12	Estimated difference in MSE: using near bound and far bound . . .	85

5.1	Illustration of possible relations among the true value, QCS, and NQCS (intersection) for video copies of the same format and frame dimensions	91
5.2	Effect of initial estimation in POCS.	95
5.3	Quantization noise in the spatial domain	95
5.4	Example block diagram of the proposed algorithm with two copies of different sizes and formats.	100
5.5	Comparison between proposed method and simple averaging	104
5.6	Two quantization step sizes, one is twice the other	104
5.7	Barbara image, input copies and reconstruction using the proposed method. left: full-size coarsely quantized; right: 3/4-size finely quantized (enlarged for presentation); bottom: reconstruction . . .	107
5.8	Effect of different GOP structures.	110
5.9	Subjective comparison; top: cropped area; left to right: input, ML initial estimate, POCS output	114
5.10	PSNR results obtained with different number of input copies. . . .	117
5.11	Subjective comparison, from left: input video, 3-copy POCS, 12-copy POCS. Same cropped area used as Fig. 5.9	118

List of Tables

2.1	Performance comparison between PSNR and mean SSIM	26
3.1	λ estimated with the default and shifted boundary locations, original image without quantization	49
3.2	λ estimated with the default and shifted boundary locations, JPEG-quantized image	51
3.3	λ estimated from quantized coefficients based on different boundary shifting, JPEG	54
3.4	Average relative estimation error of $\lambda(\%)$	60
3.5	PSNR estimation results, JPEG	62
3.6	PSNR estimation results, JPEG, using enlarged block sizes	62
3.7	PSNR estimation results, H.264	65
5.1	PSNR comparison, JPEG image in the same size (dB)	103
5.2	PSNR comparison, JPEG image in different sizes (dB)	105
5.3	PSNR (dB) results obtained from video copies with different GOP combinations. (Note: The number of B-pictures between adjacent I-/P-pictures is equal to M-1 in each copy.)	111
5.4	PSNR (dB) results obtained at different quality settings	113
5.5	PSNR (dB) results obtained with mixed input formats	116

Abbreviations

ABBREVIATIONS	FULL EXPRESSIONS
AC	alternating current
DCT	discrete cosine transform
DC	direct current
FR	full-reference
GOP	group of pictures
HVS	human visual system
IJG	Independent JPEG Group
ML	maximum likelihood
MSE	mean squared error
NQCS	narrowed quantization constraint set
NR	no-reference
PSNR	peak signal-to-noise ratio
QCS	quantization constraint set
RR	reduced-reference

Chapter 1

Introduction

In this chapter, we discuss the motivation to our research and what we aim to achieve through our investigations on the topic of objective quality assessment and enhancement for compressed images and videos. We also provide the scope of our study and summarize the main contributions we have made.

1.1 Motivation and Objective

We have seen the rise of social media and various image and video sharing websites during the past few years, such as Flickr, Youtube and Vimeo. With the rapidly growing bandwidth of household Internet connections and constantly dropping prices of digital cameras and camcorders, more people have started sharing their photos and videos over World Wide Web. Central to all these activities is the technology of compression, which allows much more efficient transmission and storage of images and videos.

To maximize this efficiency, most image and video compression methods utilize, among other techniques, quantization of information and trades fidelity for

reduction in data size. However, these digital images and videos are served over the Internet without clear indications of their compression quality. The person at the receiving end usually has no definite idea about how well these compression-distorted pictures resemble their uncompressed originals. This poses a problem when a content or service provider wants to monitor the quality of streaming video, when a user is browsing the Internet for an image of sufficiently high quality for her project, or when a video hosting site is trying to determine how to best transcode a user-uploaded video recording for web presentation.

Besides the quality of a single compressed image or video, given the web's open-access nature, it is not uncommon to find differently compressed versions of the same image or video hosted at different web addresses. Many of these variations may be compressed from the same source image or video using different formats, encoders or compression settings, resulting in similar-looking versions of different qualities. A user may be interested in finding one copy that has the best compression quality so that she could work on the image further without having to worry about amplifying the compression artifacts to a disturbing level. In fact, having multiple compressed copies may prompt the user's desire to combine them all to get a "super copy" that possesses all the merits of the copies plainly available and is of better quality than any of them.

Of course, the visual quality can be assessed by comparing the compressed pictures against the uncompressed originals. The technique sounds fool-proof except for the fact that in the context of Internet applications, the uncompressed original is nowhere to be found by the receiver. While the above-mentioned problems can also be solved by getting a human observer to stare at the computer screen for hours and holding a mouse in her hand to mark different areas of images or video frames and take notes of their quality; an average user cannot afford the time or effort to go through dozens of images one by one, or thousands of video frames as videos usually runs at 25 or 30 frames per second.

As a result, there has been a shared interest among researchers in developing techniques to automatically assess and enhance the quality of compressed images and videos. Methods have been proposed to estimate the perceived quality of compressed images by identifying and measuring the compression artifacts [1–6]. There are also research projects done on estimating the objective quality, in terms of mean squared error (MSE) or peak signal-to-noise ratio (PSNR), without referring to the original pictures [7–11]. Postprocessing techniques have also been explored, ranging from various forms of adaptive filtering of image content [12–19] to optimized decoding of compressed images [20–24]. These research works have provided important knowledge and inspiration for our own investigation.

However, there are still problems unsolved and improvements to be made based on existing techniques in this research area and especially in the context of Internet applications. One such problem is no-reference (NR) quality assessment, which keeps the end user informed about the objective compression quality of a particular image or video sequence, based on the received pictures alone. The existing methods all base their estimation directly on the statistics of the quantized coefficients, and their performance is therefore limited by the distortion introduced by quantization. While these methods do provide feasible means to estimate the objective quality without referring to the originals, their accuracy can be further improved by the use of previously undiscovered information in the compressed pictures. Another problem is when multiple pictures having the same content are available on the Internet, the user may want to find one with the best compressed quality. The aforementioned NR quality estimation methods may very well be used, but they may not be accurate enough to correctly decide between a pair of images with very similar quality measures and may require more computation than what is really necessary considering the similarity between these images. A simpler estimator for their relative compression quality is preferred. Such a new technique could reduce estimation errors and reduce complexity at the same time. In addition, we also realize that the best image or video among a few may not

carry all the useful information in all the available copies, and the end user should be able to generate a image or video sequence with quality even better than that supposedly best copy. An illustrative summary of these problems is shown in Fig. 1.1.

FIGURE 1.1: Illustration of the problems considered

Therefore, the aim of our research is to better understand the effect of lossy compression on objective image quality and to use this understanding to develop new automated methods for the estimation of the visual quality of compressed images and videos, without having to refer to the uncompressed source pictures, and for enhancing the visual quality of compressed images and videos when multiple copies of the same content are available. We hope to further improve the usefulness of the Internet by providing these tools to the users. To meet this objective, the following are to be addressed:

- To investigate the techniques for no-reference objective image quality assessment, to identify and overcome their limitations.
- To show that statistical modeling is not necessarily the best way for relative quality assessment and propose a more robust alternative.
- To illustrate that multiple compressed copies can be used to jointly reconstruct pictures of superior quality.
- To experimentally demonstrate the effectiveness of the methods proposed.

1.2 Scope and Main Contributions of the Thesis

In this thesis, we focus our attention on assessing and enhancing the quality of images and videos that are compressed using block-transform techniques, assuming the unavailability of the original pictures.

The quality of interest is objectively measurable in terms of the numerical differences between the compressed and the original pictures, although the latter is only referred to for verification purposes. In addition, it does not concern the quality of the original image prior to compression, such as noise, distortion, aliasing, and other artifacts introduced during image acquisition or generation. The assessment of quality therefore refers to the estimation of the loss in signal fidelity solely due to compression, and the enhancement refers to the partial recovery of such a loss and the improvement in objective quality.

There are also many ways in which the pictures may be processed before compression. The techniques for the detection and quantification of re-compression, histogram manipulation, and resampling have been well documented in the literature [25–29] and this thesis shall not dig deep into such cases as their combinations are impossible to enumerate and overly complex to summarize. Therefore

the pictures are assumed to be directly compressed from the original with no other processing unless otherwise noted.

The thesis consists of three main contributions within the scope just described, and they can be summarized as follows.

No-reference PSNR estimation - We identify the common technique used to blindly estimate the PSNR of compressed images and videos and its limitation in fully exploiting the information carried in the compressed pictures. We present our observations on the statistics of block-DCT coefficients, their quantized counterparts, and those based on different block boundary configurations. These observations lead to the proposal of an improved method with two variants for estimating the probability density functions of the original coefficients based on the quantized ones, using either shifted or enlarged block boundaries instead of those defined during compression.

Relative quality estimation - We point out that while no-reference PSNR estimation performance can be improved, it may not be the best approach when multiple compressed copies of the same image or video are available and their qualities are compared. We propose to compare in the transform domains the reconstruction levels of one copy against the quantization bounds of another copy, forming a conservative estimate of the difference in MSE. This method focuses on the minute differences in compression quality at coefficient level. The result is a more robust estimation of relative quality in term of which copy is better, while the degree to which it is better is also positively related to the ground truth.

Reconstruction from multiple compressed copies - We realize the variety of information contained in differently compressed pictures of the same content, and propose to reconstruct pictures using the intersection of the quantization constraints from all input copies, so that the reconstructed picture

is superior in quality to all inputs. We base our method on the projection onto convex sets technique, in order to allow the use of inputs in different compression formats and different picture dimensions. We also demonstrate that it is possible to include image priors in the process to further improve the quality of the reconstructed images.

1.3 Thesis Organization

In Chapter 2, the fundamentals of digital image and video compression techniques and the research on related topics are reviewed. Specifically, research on quality assessment for image and video is discussed. Studies related to the reconstruction of images based on multiple inputs, such as multi-descriptive coding and multi-channel processing, are also reviewed.

Realizing the common difficulty faced by researchers in no-reference PSNR estimation, we propose a new method in Chapter 3, which significantly improves the estimation accuracy based on observations generated by shifting or enlarging the block boundaries. This is followed by another application scenario where multiple compressed copies of the same image may be available over the Internet, for which we propose a relative quality assessment technique to reliably pick the best picture out of the bunch in Chapter 4. In Chapter 5 we further discuss that having multiple compressed copies of the same image or video not only allows the user to pick the best of them, but also to use our proposed method to reconstruct a version superior to all of them.

The thesis is concluded in Chapter 6 with remarks on future research.

Chapter 2

Background and Literature Review

Digital images usually consist of large amounts of data, often too large to be conveniently stored or transmitted. Videos are essentially sequences of still images, with dozens of them in each second of runtime, and are even larger in volume. Lossless compression does reduce the amount of data, but that alone is still insufficient for the majority of the applications. By exploiting the properties of natural images and the human visual system, it is possible to drastically reduce the amount of data by introducing distortions that do not significantly affect the perceived quality. This is known as lossy compression, since the signals have been irreversibly modified during compression and some fidelity has been lost. These lossy compression technologies have permitted much more efficient use of disk quota and bandwidth limits, allowing the access of images and videos through relatively slow connections, with shorter delays, and have given rise to a vast variety of Internet multimedia applications.

Besides lossy compression, there are many factors in the acquisition and transmission of digital pictures that lead to loss of fidelity. These factors have inspired

research on the quality of digital images and videos. Image quality has been studied as a special form of signal quality, and numerical measures have been proposed to quantify such a quality. It has also been studied as a visual stimulus by engaging human subjects, because the images and videos are mostly meant to be looked at after all. The quality enhancement techniques often take advantage of knowledge specific to the application scenario and therefore vary greatly from one to another.

In this chapter, we will review the basics of image and video compression techniques, upon which our discussion in the latter chapters shall base. With that covered, we will also give a roundup of research on topics related to image quality assessment and enhancement.

2.1 Image and Video Compression

2.1.1 Digital Image and Video

Digital pictures are numbers. A common full-color picture consists of N rows by M columns of elements, known as pixels or pels (short for picture elements). Each pixel consists of 3 numbers, representing the intensities of red, green, and blue primary colors because most humans see color through lights of these three different ranges of wavelengths. Fig. 2.1 shows a digital image and a breakup of its pixels. Video is essentially a sequence of still images captured at consecutive time instances. When shown in quick succession, these images give the impression of motion to the viewer. A pixel in a picture can be identified as $x(m, n)$, where m , n indicate the coordinate of that pixel, and the (scalar or vector) value of $x(m, n)$ indicates the look of it.

Such images and videos are in a form ready for display on screen, but are usually too large in data volume to be stored or transmitted. For example, the

FIGURE 2.1: A digital image, an enlarged 8-by-8 pixel area, and the values of a pixel

image in Fig. 2.1 has $M = N = 512$, and each pixel is represented by 3 bytes (each byte for the intensity of one primary color, sufficient for a smooth gradient). This works out to $512 \times 512 \times 3 = 786,432$ bytes and such a resolution only fills up a quarter of a computer screen by today's standards, and is insufficient for anything printed larger than a passport-sized photo. A more practical example is full high-definition television (HDTV) videos. With each picture having a size of 1920 wide by 1080 tall, 3 bytes per pixel, and the sequence running 30 pictures per second, each second a total of $1920 \times 1080 \times 3 \times 30 = 186,624,000$ bytes run through whatever channel that carries this data flow. A full-length movie of 90 minutes will amount to just over 1 tera (10^{12}) bytes.

Such humongous volumes of data are clumsy to deal with at the best, and a lot of applications such as digital publishing, digital cameras, and multimedia over the Internet would have been prohibitively expensive simply for their data storage and transmission requirements. Fortunately, research has led to effective compression

techniques to represent the likeliness of the original pictures with much fewer bytes and have made many applications possible.

2.1.2 Spectral and Temporal Redundancy

Compression is about reducing redundancy. It expresses the information in a compact way and does away with whatever unnecessary or insignificant.

It is necessary to display an image in red, green and blue primaries in order for us to see it in full color. However, there is a large correlation between these three color components and this is inefficient for storage. In order to reduce this mutual redundancy, each pixel undergoes a color transformation, from RGB to a luma-chroma color space (YCbCr) using:

$$\begin{aligned} Y &= 0.299 \cdot R + 0.587 \cdot G + 0.114 \cdot B \\ Cb &= 128 - 0.169 \cdot R - 0.331 \cdot G + 0.500 \cdot B \\ Cr &= 128 + 0.500 \cdot R - 0.419 \cdot G - 0.081 \cdot B, \end{aligned} \tag{2.1}$$

where R , G and B refer to the red-, green- and blue-intensities respectively; Y is the luminance; Cb is the blue-difference; Cr is the red-difference. Each component is assumed to be recorded in 1 byte (8-bit unsigned integer, range $[0, 255]$)¹. An example is shown in Fig. 2.2. Given the same color image, YCbCr has less variation in two out of three components. This color transformation not only makes data more compact, but also allows different treatment to Y and Cb/Cr channels. Because Y is perceptually more important, it can be better preserved while Cb and Cr can be less precisely represented without greatly impacting the visual quality. This operation is pixel-based and is invertible (save for the effect

¹Formulae are for JPEG [30] image compression. Video follows the same principle but backward compatibility calls for footroom and headroom of 16 levels each and the formulae are slightly different. See [31] for details.

FIGURE 2.2: A digital image, expressed in RGB intensities (upper row) and YCbCr component values (lower row). Brighter means larger in value.

FIGURE 2.3: Illustration of motion estimation and compensation

of limited precision) using:

$$\begin{aligned} R &= Y + 1.402 \cdot (Cr - 128) \\ G &= Y - 0.344 \cdot (Cb - 128) - 0.714 \cdot (Cr - 128) \\ B &= Y + 1.772 \cdot (Cb - 128). \end{aligned} \tag{2.2}$$

A video sequence consists of hundreds of still images known as video frames, taken at consecutive time instances. A few dozens of such images are taken within a second of time. It is unsurprising to find neighboring frames fairly similar and the objects in the images are only slightly displaced. Motion estimation and compensation exploit the similarity between neighboring frames to reduce redundancy. Fig. 2.3 illustrates the concept of motion-based coding of a video frame.

Some video frames are encoded in the same way as still images. These frames are known as the intra-frames (I-frames), meaning they are treated independently of other frames and can be decoded on their own. They serve as the anchors for the beginning of playback, quick seek positions, and references for other frames during encoding and decoding. The other type of frames are known as the inter-frames, which are based on the information already presented in other frames. These frames exploit the temporal redundancy between adjacent frames to improve the compression ratio.

To encode an inter-frame, a set of reference frames are usually used, denoted as Θ . The video frame being encoded is partitioned into disjoint blocks. The encoder searches each reference frame j for an image area that best resembles the current block x_i . The similarity may be measured by sum of squared differences, sum of absolute differences, or some other criteria. These best matching image areas $\hat{x}_{i,j}$ form the set of reference blocks, and the displacements between the current block and the reference blocks are recorded as motion vectors. By taking some weighted average of these blocks, an image block that very closely resembles the current block is formed, and is named the prediction. However, because all the changes

in the scene are modeled as planar translations, a lot of other possible types of motion, and differences such as rotation, translation in the third dimension, deformation, occlusion, and change in lighting, are only approximated. This means the prediction is in general different from the current block. It is therefore necessary to record the difference between the current block and the prediction, resulting in a residue. Formally, the residue can be expressed as:

$$r_i(m, n) = x_i(m, n) - \sum_{j \in \Theta} w_{ij} \cdot \hat{x}_{ij}(m, n), \quad (2.3)$$

where $x_i(m, n)$ refers to a pixel at coordinates m, n in the i th frame and w_{ij} 's are the weights. x_{ij} are the picture area in frame j that best resembles x_i . The weighted sum of all references form the prediction. Based on the reference frames and using both motion vectors and residues, the current frame can be reconstructed. The residues can be further compressed in a manner similar to still images, as will be shown later. The reconstruction of the current block is:

$$\hat{x}_i(m, n) = \hat{r}_i(m, n) + \sum_{j \in \Theta} w_{ij} \cdot \hat{x}_{ij}(m, n), \quad (2.4)$$

where $\hat{r}_i$ indicates that some distortion has been introduced to r_i . The use of notation $\hat{x}_{ij}$ is because the reference frames themselves are also compressed and are not the original². If the unaltered residue r_i were used, the expression above would equate to the original picture x_i .

There are two further sub-types of inter-frames, known as the forward predicted (P-frames) and bidirectionally predicted (B-frames) frames. P-frames are predicted from a single reference frame of either type I or P preceding itself in time. B-frames are predicted from other frames before and after itself in time, and hence offer further redundancy reduction compared with P-frames. Each video sequence

²This is in order to prevent error propagation. Because only the compressed frames are transmitted, it would introduce errors if the original pictures were referred to during encoding.

may have its own unique way of assigning frame types, and such a structure is known as the structure of group of pictures (GOP), and two examples are given in Fig. 2.4. Different GOP structures offer different trade-offs among compression ratio, delay, and complexity. For example, a sequence having a GOP structure of IPPP..., incur less encoding and decoding delays, because only one reference frame needs to be buffered at any given time; on the other hand, a GOP structure of IBBPBBP... is likely to offer better compression ratio because B-frames are less redundant, while computation needed for motion estimation is increased considerably because for each B-frame, at least two references are searched for similar image areas.

FIGURE 2.4: Examples of common GOP structures. Arrows show directions of reference.

It is important to note that motion estimation varies greatly between encoders and parameters used during encoding. A different set of reference frames, weighting scheme, block size for motion estimation, and search method used can all affect the predictions produced and hence the residues calculated by subtracting the predictions from the originals. For example, H.264 [32] uses blocks as small as 4-by-4 for motion estimation, with the precision of a quarter of a pixel, and supports multiple references blended together as a reference. In contrast, the older MPEG-2 [33] format uses larger blocks of 16-by-16 and only one reference frame. As a result of this and many other differences, H.264-compressed videos usually have less signal in the residues than those compressed with MPEG-2 and have greater compression ratios for the same quality.

2.1.3 Discrete Cosine Transformation

After color transformation (and motion compensation for video), each component in the YCbCr color space is then compressed separately with almost identical procedures. We shall discuss the Y (luminance) picture component from this point onwards as the conclusions generally extend to Cb and Cr components [34].

Images in the spatial domain, that consist of pixel values at coordinate locations in the $m - n$ plane, tend to have similar values in a small neighborhood. Such correlation among pixels means there must be quite a bit of redundancy in recording them individually. Two-dimensional discrete cosine transform (2D DCT) is thus employed to reduce the correlation. Formally, the normalized 2D DCT is defined as:

$$X(u, v) = \alpha(u)\alpha(v) \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} x(m, n) \cos \left[\frac{\pi u}{M} \left(m + \frac{1}{2} \right) \right] \cos \left[\frac{\pi v}{N} \left(n + \frac{1}{2} \right) \right],$$

$$\alpha(u) = \begin{cases} \sqrt{1/M}, & \text{if } u = 0 \\ \sqrt{2/M}, & \text{otherwise} \end{cases} \quad \alpha(v) = \begin{cases} \sqrt{1/N}, & \text{if } v = 0 \\ \sqrt{2/N}, & \text{otherwise} \end{cases}$$
(2.5)

where x refers to the pixel value in the spatial domain, or the residue pixel for video (i.e., r); X refers to the coefficient in the frequency domain; m and n are pixel coordinates, u and v are frequency indices; M and N are the dimensions of the image or image area. This transformation is unitary: the inner product of two vectors before the transformation is equal to their inner product after the transformation.

Because the transformation has a complexity on the order of $MN \log MN$, most of the time the transformation is performed on sub-blocks of images instead of their entirety to reduce the computation cost. For example, in JPEG [30] and MPEG-2 [33] standards, pictures are partitioned into 8-by-8 disjoint blocks, while

FIGURE 2.5: Basis functions of 8-by-8 DCT

in H.264 [32], the block size of 4-by-4 is also used. This transformation computes the correlation between the image block and each of a set of basis functions (the cosine terms in (2.5)), and generates a coefficient. The 8-by-8 DCT bases are depicted in Fig. 2.5. An image block becomes a block of coefficients of the same dimensions. For the Y component of the cropped block in Fig. 2.1, a comparison between the representations before and after transformation is given in Fig. 2.6.

The top-left coefficient is called the DC coefficient, named in analogy to the direct-current component of a varying electric current. The rest are known as the AC coefficients, recording the change in intensity in various manners. It can be seen from Fig. 2.6 that there are only a few large coefficients occupying the upper-left corner, while the others are comparatively small in magnitude. This achieves compactness in signal representation. In JPEG, the DC coefficient is subtracted by 1024 to reduce the dynamic range; in videos, because the residues are differences between very similar image blocks, their DC coefficients have a mean value of 0 in general hence no further manipulation is necessary.

In H.264 compression, efforts have been devoted to simplifying the computation

FIGURE 2.6: Effect of 2D DCT: top - image block, bottom - coefficients

of 2D DCT. The resulting transform approximates DCT but uses only integers throughout the process and the actual number of multiplications is greatly reduced compared with DCT, by replacing them with bit shifting and inversion.

Besides DCT, there are also other representations available. A class of transforms known as the discrete wavelet transforms (DWT) is also used for image compression, such as in JPEG2000 [35]. Wavelet transforms use spatially narrowly-confined waveforms at different scales as basis functions to represent signals. Since the bulk of the current compression standards use DCT or its variants, we shall not discuss these other transforms in detail.

A very important property of unitary transforms is that the sum of the square of a function is equal to the sum of square of its transform. This allows quantities to be measured in either domain and is known as Parseval's theorem. The wavelet transform used in JPEG2000 and the integer transform in H.264 loosely observe this property too.

2.1.4 Quantization

All the procedures described earlier reduce redundancy by representing the same information in a different manner without losing the ability to get back to its original form. If the compression process stops here, it would have been almost lossless, subject to the errors due to limited precision.

To achieve higher compression ratios, however, quantization is commonly performed in the transform domain to further reduce the data volume. It usually consists of a division followed by rounding, resulting in much fewer output levels than input levels. This allows a coarser but more compact representation of the original signal. In JPEG compression, a transform coefficient is quantized as

$$Y(u, v) = \text{round} \{X(u, v)/Q(u, v)\} \quad (2.6)$$

where Q is the quantization matrix containing step-sizes, and X and Y are the original image and the quantized image respectively, both in the DCT domain. The representative value (reconstruction level) of the coefficient can be obtained as

$$\hat{X}(u, v) = Y(u, v) \times Q(u, v) \quad (2.7)$$

This process maps all possible values between $\hat{X}(u, v) - Q(u, v)/2$ and $\hat{X}(u, v) + Q(u, v)/2$ onto a single level $\hat{X}(u, v)$. Using the same image block in Fig. 2.6 as an example and applying uniform quantization with step-size of 10, we show an example of representative values and reconstructed spatial image block in Fig. 2.7. Compared with Fig. 2.6, there are a lot of zeros in representative values, which are easily compressed; at the same time, the reconstructed spatial image is very similar to the original.

FIGURE 2.7: Effect of quantization on DCT coefficients: top - inverse-transformed image block, bottom - representative coefficient values

In some compression standards, the quantization has a dead-zone around 0, and can be expressed with generality as:

$$Y(u, v) = \begin{cases} 0, & \text{if } |X(u, v)| \leq \Delta Q(u, v) \\ \text{sign}[X(u, v)] \cdot \left\lceil \frac{|X(u, v)| - \Delta Q(u, v)}{Q(u, v)} \right\rceil, & \text{otherwise} \end{cases} \quad (2.8)$$

where $\Delta Q(u, v)$ is the half-width of the deadzone, and $\lceil \cdot \rceil$ rounds up the number enclosed. Replacing X with R , and $\hat{X}$ with $\hat{R}$ in Equations (2.6), (2.7), and (2.8) gives us the expressions for quantization of video frame residues. Different video compression formats use different deadzones: for MPEG-2, $\Delta Q(u, v) = Q(u, v)/2$, and for H.264, the deadzone can be customized in the encoder. A half-deadzone greater than $Q(u, v)/2$ effectively shifts a reconstruction value closer to zero relative to the range of values it represents. This approximates the optimal reconstruction based on the Laplace distribution model, which will be discussed later in Section 2.2. It also quantizes more small-valued coefficients to zero, gaining extra compression ratio at the expense of a small drop in quality.

A quantizer with a deadzone is illustrated in Fig. 2.8. As shown in the figure, quantization maps a range of values onto a single value. It is not possible to pinpoint the original coefficient values given the quantized ones. This many-to-one mapping discards information and creates distortion. All we can know from a quantized coefficient is that it must have come from a certain range of original values. The upper and lower bounds of such a range corresponding to a positive dequantized coefficient of a still image can be expressed as:

$$\begin{aligned} X^{\text{upper}}(u, v) &= \hat{X}(u, v) + \Delta Q(u, v) \\ X^{\text{lower}}(u, v) &= \hat{X}(u, v) + \Delta Q(u, v) - Q(u, v). \end{aligned} \quad (2.9)$$

For a negative coefficient the expression is a mirror image of (2.9), and a zero-valued coefficient must lie within the range $[-\Delta Q(u, v), \Delta Q(u, v)]$. The collection of the ranges for all coefficients in a picture is known as the quantization constraint set (QCS), for the unknown original coefficient values are constrained within it.

In the case of video frames, the QCS of the residues are not directly comparable because they are the results of different motion estimation processes. A meaningful

FIGURE 2.8: Quantization with step-size q and half-deadzone $\Delta q > 0.5q$

definition of QCS for video frames is obtained by adding the transform of the prediction to the range of the residue coefficients. For a positive residue coefficient $R(u, v)$:

$$\begin{aligned} X_i^{\text{upper}}(u, v) &= \hat{X}_i^{\text{pred}}(u, v) + \hat{R}_i(u, v) + \Delta Q_i(u, v) \\ X_i^{\text{lower}}(u, v) &= \hat{X}_i^{\text{pred}}(u, v) + \hat{R}_i(u, v) + \Delta Q_i(u, v) - Q_i(u, v), \end{aligned} \quad (2.10)$$

where $\hat{R}_i(u, v)$ is the dequantized residue coefficient; $\hat{X}_i^{\text{pred}}$ is the 2D DCT representation of the prediction $\hat{x}_i^{\text{pred}} = \sum_{j \in \Theta} w_{ij} \cdot \hat{x}_{ij}$.

In JPEG2000, an optional quantization procedure can be carried out similarly to that of JPEG and MPEG-2 but with an enlarged dead-zone around zero. Alternatively, JPEG2000 could skip the explicit quantization and use embedded block coding with optimal truncation (EBCOT) [36] to control the output bitrate. EBCOT truncates signal by individual bit and is equivalent to quantization with a step-size of $2^{q(u,v)}$, where $q(u, v) + 1$ is the last preserved bit for coefficient u, v . The QCS can therefore be determined accordingly.

Following quantization, entropy coding is usually performed to make sure the final representation is as compact as practically possible without introducing further distortion. Knowing that quantization is the only step in the chain of compression procedures that trades fidelity for compactness, we will set our focus on quantization in investigating the compression quality.

2.1.5 Image and Video Quality

There are two types of quality for digital image or video. One is the objective quality, which is computed from the pixel values. The other is the subjective quality, obtained by engaging human observers to look at the pictures and give marks according to their opinion as to how good the qualities of the pictures are, and the

result is quantified as mean opinion score (MOS). Arguably, the latter of the two is more important because the images and videos are generally meant to be consumed visually in the end. However, opinions can vary greatly from one individual to another and having a bunch of people staring at monitors to continuously assess the quality of some pictures is too time consuming to be practical for everyday applications. Therefore, a useful solution would be to formulate objective measures that take into consideration the characteristics of human visual system, so that the quality of an image can be assessed using only the digital numbers, while the resulting score always corresponds to the average opinion of viewers.

Because digital images and videos are both digital signals, it is natural to assess the quality of them in signal processing terms. Conventionally, the quality of a distorted signal is measured by its difference from the original: the larger the error, the lower the quality. A common measure used is the mean squared error (MSE):

$$\text{MSE} = \frac{1}{MN} \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} (x(m, n) - \hat{x}(m, n))^2, \quad (2.11)$$

where M and N are the total number of columns and rows of the picture; x and $\hat{x}$ are the original and distorted images. Since DCT is a unitary transform, by Parseval's theorem, MSE can be expressed equivalently in the frequency domain as:

$$\text{MSE} = \frac{1}{UV} \sum_{u=0}^{U-1} \sum_{v=0}^{V-1} \left(X(u, v) - \hat{X}(u, v) \right)^2 \quad (2.12)$$

where U and V are the total number of coefficients in the two respective directions. With transforms that are approximately unitary, the equivalent MSE can also be approximated in the frequency domain. Besides MSE, there are other measures, such as correlation coefficient, mean absolute error, etc.. However, they are rarely used in the context of image quality assessment and will not be discussed.

MSE can measure the distortion of a processed image compared to the original but it is a negative indicator for quality. Better quality is associated with smaller

MSE. Furthermore, the numerical values of MSE are inconvenient as indications of the quality of images, because of two reasons. First, the range of pixel values may be different from one picture to another, and the MSE measured in a picture with maximum intensity of 1 is not comparable with another with 255. Second, human perception works in a logarithmic fashion. Therefore, a derivative of MSE called peak signal-to-noise-ratio (PSNR) is formulated to better reflect the perceived quality of an image, setting the original as infinitely good:

$$\text{PSNR} = 10 \cdot \log_{10} \left(\frac{\text{MAX}^2}{\text{MSE}} \right), \quad (2.13)$$

where MAX is the maximum allowed pixel value in a picture and the unit of PSNR is decibel (dB). PSNR has been widely used as a objective quality measure because it is simple to compute, and gives a reasonably good indication of the visual quality of an image. However, it has not been without criticism, and research has been conducted to develop new measures that fit better with MOS.

Emphasizing the significance of the threshold below which minute errors become unperceived, Jayant et al. [37] proposed to compress images at a bit rate offering just-noticeable distortion (JND). Lubin described in his 1997 paper a metric called Sarnoff JND Model [38], in which a image or video sequence is decomposed into a pyramid of scales, normalized, and the contrasts in orientation, flickering (video), and color are measured. These contrasts are masked with the JND thresholds so that the unnoticeable distortions do not go into the final calculation of quality.

In their 2004 paper [39], Wang et al. proposed a new objective image quality measure, and named it structural similarity index (SSIM). Like PSNR, it also measures the distorted image against the original, and does not go through the decomposition in Sarnoff JND. Unlike PSNR, it takes into account not only the

pixel-by-pixel difference, but also the neighborhood similarity. It is expressed as

$$\text{SSIM} = \frac{(2\mu_x\mu_{\hat{x}} + c_1)(2\sigma_{x\hat{x}} + c_2)}{(\mu_x^2 + \mu_{\hat{x}}^2 + c_1)(\sigma_x^2 + \sigma_{\hat{x}}^2 + c_2)}, \quad (2.14)$$

with μ_x and $\mu_{\hat{x}}$ the mean of x and $\hat{x}$; σ_x and $\sigma_{\hat{x}}$ the variance of x and $\hat{x}$; $\sigma_{x\hat{x}}$ the covariance between x and $\hat{x}$; c_1 and c_2 for stabilizing the division should the denominator be small, and determined by the dynamic range of the image. This measure is simpler compared with Sarnoff JND and offers similar performance within small image areas.

Although PSNR is not designed upon sophisticated visual modeling, it does correlate with MOS, especially when there is only one specific type of distortion. In Table 2.1 we quote the performance comparisons published in [39], based on JPEG and JPEG2000 compression, and fitted to MOS with non-linear regression. It can be seen that PSNR does offer a good indication of the perceived quality, in the context of compression distortions, despite being outperformed by SSIM.

TABLE 2.1: Performance comparison between PSNR and mean SSIM

Performance relative to MOS	PSNR	SSIM
Correlation Coefficient	0.905	0.967
Mean Absolute Error	6.53	3.95
Root Mean Squared Error	8.45	5.06
Outlier Ratio	0.157	0.041
Spearman Rank-Order Correlation Coefficient	0.901	0.963

It is great that the more advanced objective measures can reflect human opinions in image quality better than PSNR for a wider range of distortions. Unfortunately, all of these measures mentioned above depend on the knowledge of both the distorted and the original pictures.

In our scenario of interest where quality assessment is performed at the receiving end, the original pictures are not available. This makes pixel-wise comparison

impossible. However, given enough number of pixels or coefficients, their statistics can be established, and an area-based assessment is still possible. From the discussion earlier on quantization we know that the error due to quantization is constrained within the QCS. By identifying the probability distribution $p_{u,v}(X)$ of each frequency for the block-based 2D DCT coefficients, it is possible to compute MSE in a probabilistic manner. For a certain coefficient belonging to frequency u, v , its expected squared error is:

$$e = \int_{X^{\text{lower}}}^{X^{\text{upper}}} (X - \hat{X})^2 p_{u,v}(X) dX \quad (2.15)$$

where X^{lower} and X^{upper} are the lower and upper bounds of the QCS and $p_{u,v}(X)$ is the probability density function of the coefficient X for frequency u, v . The interpretation of the formula is as follows: from QCS we know that between X^{lower} and X^{upper} must the true value X lie. With the probability distribution $p_{u,v}(X)$, we can integrate all the possible squared errors for this particular coefficient, weighted by probability density, to get the expected squared error for this coefficient. Taking the average of e over all the coefficients gives the mean squared error of the image and is equivalent to (2.12). The last obstacle towards calculating MSE by this method is to obtain the probability distribution $p_{u,v}$ for every frequency. Besides an alternative means to measure MSE, equation (2.15) also shows that by narrowing QCS and finding the optimal value for $\hat{X}$, MSE can be reduced.

We will review the models concerning $p_{u,v}$, as well as techniques to reduce MSE in Section 2.2 below, among other related work. We will use MSE (and PSNR derived from it) as the quality metric for the remaining chapters of the thesis, because it correlates with perceived image quality and there is an established framework for no-reference quality assessment based on it.

2.2 Related Work

In this section, we review the image quality assessment techniques in the literature, including those based on reduced references, those inspired by the properties of HVS, and models related to the statistical perspective of MSE (2.15) mentioned in the previous section. We also review the techniques developed to enhance the quality of compressed images and videos.

2.2.1 Quality Assessment

We have already mentioned a few image quality metrics, including PSNR, SSIM, and Sarnoff JND in the previous section. Their definitions are classified under the full-reference (FR) methods for quality assessment, because their application requires the full knowledge of the original pictures. Liu et al. was among the earliest to propose an HVS-based FR quality assessment scheme specifically for JPEG-compressed images [4]. In their work, the strengths of blocking and blurring caused by JPEG compression are quantitatively analyzed. These two strengths are then added to form a single quality measure. Lee et al. [3] took into consideration the foveation effect of HVS, and weigh different parts of video frames accordingly in distortion computation.

All the techniques discussed up till now depend on the availability of the original image or video as reference. This requirement restricts the usefulness of these methods to the content provider, who has access to the original images, while the receiver is left with no means to assess the quality in a similar fashion.

2.2.1.1 Reduced-reference Techniques

To address the intrinsic practicality problem with the FR methods, where the original image is required, there have been efforts in developing reduced-reference

(RR) methods [40–43]. These schemes require sending an additional piece of message specially prepared for quality assessment together with the compressed image content, and use the message to support the quality assessment process. The assessment makes no direct use of the original image data and have demonstrated good consistency.

Wang et al. proposed an RR image quality assessment method based on natural image statistics in the wavelet-domain [40]. They observed that the distribution of wavelet coefficients change when the image is distorted, and proposed to record the marginal distributions of the oriented bandpass filter responses (i.e., wavelet coefficients). The directional wavelets are known to have a close connection to the behavior of HVS and are expected to give a good indication of its response to changes in image. By quantizing 18 features into merely 162 bits for each image, a compact reference is generated. The marginal distributions of wavelet coefficients between the distorted image and the original can then be compared using Kullback-Leibler distance. Experiments have been performed on images distorted by additive noise, blurring, simulated transmission errors, and JPEG and JPEG2000 compressions, with consistent results comparable to FR methods including PSNR, Sarnoff JND and SSIM. In a later development in 2006 [41], the term *quality-aware image* has been coined, and the aforementioned wavelet-based RR side-information is embedded in the image by means of wavelet-domain watermarking. The advantage is that normal decoding is not affected and no separate effort is required in transmitting the RR information.

Another RR technique, called hybrid image quality metric[42], was proposed by Kusuma et al. in 2005. Like the oriented-wavelet statistics, Kusuma’s metric also takes into account the characteristics of HSV, but in an even more deliberate manner, targeting specifically blurring, ringing, blocking, and loss of image blocks. These artifacts are common in image compression and transmission, and are the main culprits for visual quality degradation. The metric is a weighted sum (and

hence the word hybrid in its name) of individual metrics each addressing one of the above artifacts. The weights are derived from subjective tests. The RR information sent to the user is the metric computed on the original image, and can be compared with the same metric computed on the received image to evaluate the loss in visual quality. The limitation of this metric is that only a few specific types of distortions are taken into consideration. However, the metric also has its additional merit as it can be applied in a no-reference context as well, although the resulting score reflect the visual quality in an absolute sense instead of relative to the original.

More recently, a more elaborate and generic HSV-based RR method called C4 has been proposed by Carnec et al.[43]. In contrast to Kusuma's method though, C4 is not limited to a few types of artifacts, and attempts to fully exploit the known HSV features. RGB colors are transformed into Krauskopf's color space via an intermediate step of simulated cone response. In this perceptual color space, a contrast sensitivity function is applied to identify the visually significant contents and subband decomposition is performed to provide orientational division, after which the differential visibility threshold elevation is computed locally. Subsequently, characteristic points are selected and the perceptual features are extracted. The features from the original image and the received image are compared and the C4 metric is computed. It has been shown experimentally that C4 fits MOS better than PSNR and SSIM. One possible drawback of C4 is that the RR description could be longer than those of the other two methods, but because of the feature selection stage, it is scalable with fine granularity.

Despite the promising performance of RR methods, the generation, transmission and consumption of the supporting messages require changes in all parts of the application pipeline, and their application to the existing compression formats and content services is not trivial as these facilities currently do not enforce or even support such side information.

2.2.1.2 No-reference HVS-based Techniques

No-reference (NR), or blind, methods for image quality assessment is applicable to digital images where no extra information about the original is available other than the compressed pictures themselves. They generally make use of prior knowledge on specific artifacts associated with compression formats, and their impact on image quality through HVS.

JPEG treats an image as 8-by-8 disjoint blocks, often resulting in discernable boundaries between blocks. Wang et al. modeled such artifacts as a set of 2D step functions superimposed on a smoothly varying image [6]. The strength of the blocky functions are then measured and this gives an indication of the compression quality of the image. By applying a visual threshold on these blocky functions, the measured strength corresponds better to the perceivable distortion in the compressed image. In another related work, Bovik and Liu [1] estimated the blocky functions in the DCT domain instead of the spatial domain. The 2D step functions are clearly identifiable by taking an image block spanning across an existing block boundary and transforming it into the DCT domain. HVS characteristics have also been exploited in the measurement of directional activities.

Besides JPEG, the quality of JPEG2000-compressed images has also been investigated in the context of NR assessment. Because of the difference in choice of block size, transform and quantization techniques, JPEG2000 images exhibit a different nature in visible distortion. Sheikh et al.[5] proposed to use natural scene statistics to measure the quality of such compressed images. The method looks at the coefficient magnitudes and their linear predictions, and compares the significant-insignificant relationship among adjacent subbands. Based on prior knowledge on how such relationships are normally observed in natural images, an estimated MOS is computed. After calibration, the results show a good conformance to the the real MOS values, but image content does appear to be an

important factor in some instances and the method therefore lacks robustness.

As of late, we have also seen the application of neural networks in NR problems, which is natural because the perception of visual quality is a neural process and artificial neural networks were developed to mimic those of human. Babu et al. [44] proposed a method based on a sequential learning network specifically for JPEG images. HVS-based features are extracted, including edge amplitude, edge length, background activity, and background luminance. The features are masked and processed to obtain the horizontal and vertical difference profiles, and consequently the final difference profile. The histogram of this final profile is used to train the neural network targeted at the MOS values. Gastaldo et al. proposed a more generic method using a circular back-propagation network [2]. However, it is aimed at assessing the quality after enhancement procedures instead of deteriorations.

2.2.1.3 No-reference Statistical Techniques

As mentioned earlier in Section 2.1.5, it is possible to calculate the MSE value if the probability density function (pdf) of the block-DCT coefficients are known. Practically, it means either sending these functions (or some approximated form of them) as RR side information to the user, or trying to estimate these functions based on the quantized coefficients upon receiving the image or video.

Many models have been proposed to describe the distribution of block-DCT coefficients. In their 1983 paper [45], Reininger et al. showed with block sizes of 8, 16 and 32 that among various proposed models for DCT coefficients at that time, it is best that the AC coefficients are modeled as Laplacian random variables, and the DC coefficients as Gaussian, even though the latter part is a poorer fit and not widely used in practice. Later in 1993, Mueller [46] proposed that generalized Gaussian, of which Gaussian and Laplacian are special cases, describes DCT coefficients better than either of them. This finding was echoed by Joshi et

al. in 1995 [47] that generalized Gaussian indeed is the better model. Adopting the less accurate but simpler Laplacian model, Smoot and Rowe [34] investigated the relationship among distribution parameters. In it the authors remarked the surprising stability of such parameters throughout a video sequence, which will be shown later in Chapter 3 to be corroborated by our findings. Lam and Goodman [48] gave a mathematical analysis of the DCT coefficient, deriving using a doubly stochastic model and arriving at the conclusion of a Laplacian model. More recently, Bhuiyan et al. [49] proposed to use the symmetric normal inverse Gaussian model to describe the DCT coefficients and demonstrated its superior ability to approximate the actual distributions compared with the generalized Gaussian model.

When the original image is available, the parameters for these distributions can be easily and accurately estimated. However, in the context of NR quality assessment, where only the quantized DCT coefficients are available, the primary difficulty becomes how to estimate these parameters. Among these distribution models, the Laplace model is the simplest and its parameter estimation only makes use of first order statistics. Even with this relative simplicity, the majority of the NR methods [7–11] for PSNR estimation make great efforts in order to obtain a more accurate estimation of the Laplace distribution parameters, yet the error in estimation still significantly outweighs the inaccuracy of the Laplace model. The more sophisticated models, although fit DCT coefficients better than the Laplace model, involve the use of second- and fourth-order statistics, which are highly sensitive to quantization and are of no particular value to our problem before the first order statistics of the original picture can be accurately estimated from the compressed picture. The prior work and the justification for the choice of Laplace model will be detailed in Chapter 3 for better cohesion in discussion.

2.2.2 Quality Enhancement

Lossy compression of images and videos lead to degradation in image quality. Many researchers have devoted their effort to improving the quality of the decoded images. Based on a single compressed image or video sequence, it is possible to improve the visual quality by filtering out the obvious artifacts caused by compression. It is also possible to recover part of the loss in quality when prior knowledge is applied to optimize the decoding process. We also review the proposals to compress the source content into self-complete and complementary versions, so that the quality of the decoded pictures are enhanced when multiple versions are received, as well as other related work.

2.2.2.1 Single-input Enhancement and Restoration

Given a single compressed image or video, one could simply decode it as described in the compression standard. However, realizing that there are common artifacts that lead to the impression of quality degradation, and that there is more information implicitly contained in the compressed data, researchers have proposed various methods to further enhance the image quality. The work in this area can be put into two categories, those aiming at reducing the visual artifacts and those trying to restore the signal.

Arguably the most simple way to enhance visual quality is to filter out the artifacts in the decoded images. As reviewed earlier, the majority of image and video compression standards encode pictures in disjoint square blocks. As a result, the block boundaries often become distractingly abrupt after compression. Applying space-invariant low-pass filtering removes or at least softens these unnatural edges, but inevitably smudges the image details at the same time. To reduce artifacts while protecting details, adaptive filtering methods have been proposed to manipulate image segments differently. In [12, 18], the authors tune the filter

depending on the activity of the image area, allowing image details to be better preserved while block boundaries smoothed. Chen's method [12] also takes into account the masking effect of HVS, by which blockiness is more visible in flat areas than in those with lots of detail. In more recent video compression standards, the blocking effect has been explicitly addressed using an in-loop adaptive filter [16]. Besides spatial-domain filtering, Liew and Yan [14] have proposed to represent the image in the wavelet domain and remove the artifacts by spotting their unusual behaviors.

These artifact-reduction methods certainly improve the visual quality of the pictures. However, they generally do not explicitly attempt to improve the measurable signal fidelity and may in some cases further deteriorate it without being aware. Knowing that the loss of quality happens during quantization, it is natural to look at the pictures in the transform domain and consider the implications of quantization. Since the true value of the original coefficients must lie within QCS, as reviewed in Section 2.1.4, it has been suggested that the reconstructed image should not only be smooth, but also conform to the QCS. The technique of projection onto convex sets (POCS) has been used in still image and video enhancement [15, 23, 50]. The resulting pictures generally have measured signal distortion slightly better than that of the directly decoded ones, but with improved visual quality.

The original coefficient is bounded by QCS but it does not lie within different parts of QCS with equal probability. Price and Rabbani [22] proposed to dequantized the coefficients so that they are closer to zero than the center value of QCS. Because the probability density function for the DCT coefficients have larger values towards zero, the biased dequantization means lower MSE. In fact, center of QCS corresponds to an oversimplified model of uniform distribution of coefficients. Storkey [24] took the model further to include not only the probability density for coefficients, but also the probability of boundary pixel values. The

combined model not only promises improved PSNR, but also visual quality.

With a single compressed image or video, the information for its reconstruction is limited to whatever comes with this picture. In the following, we will review research works related to the reconstruction of pictures based on multiple versions of the same original as well those making use of highly similar picture contents.

2.2.2.2 Multiple-description Coding and Other Related Work

In multiple-description coding (MDC)[51], a video sequence is encoded into several individually decodable versions. The decoder of an MDC scheme performs reconstruction based on these versions. The user needs only one version to see the video at its lowest quality. With the reception of more versions or descriptions, the quality of the decoded video is improved.

There are primarily three classes of MDC schemes, subsampling-based, secondary-transform-based, and complimentary-quantization-based. The first class sub-samples the video spatially or temporally, and recombines these samples when multiple versions are received to recreate the original resolution. In the second class, the DCT coefficients of the readily compressed video undergo a secondary transform to form a pair of complimentary coefficients. Upon reception of multiple versions, the decoder is able to obtain DCT coefficients of improved accuracy through the inverse secondary transform. These two classes, however, are based on deliberate designs and are not applicable to a more generic context. The third class is complimentary-quantization-based MDC. Tillo and Olmo [52] proposed to use different quantization step-sizes for different versions of the descriptor. By intersecting the QCS of multiple versions, a higher quality picture can be decoded. Others have made use of staggered quantization ranges [53] (different deadzone size for each version). The principles for quality recovery are very similar between

these two techniques. The latter scheme provides more uniform image quality among different versions.

In general, though, each MDC scheme is a system for video encoding and decoding by itself. The original video is compressed into multiple versions following the specifications of the scheme, distributed, and are received and jointly decoded by the decoder, again following the same specifications. Each video sequence has to be encoded specifically for this system and the decoder can only deal with videos encoded for it. With a maximum number of version of N , the design target of MDC is thus to maximize rate-distortion performance for any $1 \leq n \leq N$ copies received. The problem we are trying to address in this thesis, however, is to get readily available pictures encoded without joint reconstruction in mind, and make something better out of them. The differences between these pictures arise naturally through heterogenous user preferences instead of by design. This brings our attention to some other related works on reconstruction based on available picture data.

In multi-channel reconstruction [54, 55], the similarity between color channels are exploited. Because image noise usually has low correlation between these channels but image content has a high correlation, it is possible to identify and remove noise when the channels are compared. Utilizing the autocorrelation of the original and that of the noise, Galatsanos et al. [54] approximate the Wiener solution to the multi-channel system to obtain the enhanced image. The performance of the method relies on the precise modeling of both signal and noise.

Super-resolution aims to enhance the resolution of pictures after they have been acquired, usually based on multiple similar inputs. Gunturk et al.[56] proposed a method for compressed video super-resolution. They constructed a DCT-domain Bayesian estimator to enhance resolution in the presence of both quantization and additive noise, making use of the difference in image details carried by adjacent frames.

2.2.3 Projection onto Convex Sets

Projection onto Convex Sets (POCS) is a well-established technique for solving constrained optimization problems. We make use of this technique in Chapter 5 for the reconstruction of pictures from multiple compressed copies, and give a brief introduction here to it so that our discussion is more self-contained. More details of POCS can be found in Youla and Webb's paper [57].

Suppose we need to estimate some signal f , which is known a priori to belong to the intersection of I generally non-linear constraints.

$$f \in S_0 = \bigcap_{i=1}^I S_i \quad (2.16)$$

Suppose, in addition, that each constraint S_i is closed and convex. Being closed means S_i should lie on and to one side of a hyperplane, and being convex requires any line segment connecting a pair of points inside S_i to be entirely inside S_i too. Because of the nonlinear nature of the constraints, S_0 cannot be solved analytically. Therefore, the estimation of f is obtained through iterative projection of some initial estimate x onto these convex sets.

The nonlinear projection operator P_S moves x to the nearest point in S if x is not already in S , and does not alter x otherwise.

$$x' = P_S x = \arg \min_g \|x - g\|, g \in S \quad (2.17)$$

If the constraint intersection is nonempty: $S_0 \neq \emptyset$, then sequentially applying $P_{S_i}, 1 \leq i \leq I$ iteratively will eventually bring the process to convergence, where $x \in S_0$ and further iterations do not alter x any more.

The solution x approximates f in that they both satisfy S_0 . However, the final estimation is not unique and is dependent on the initial x value and the order of

projection.

2.3 Conclusion

In this chapter, we have reviewed in brief the basics of digital image and video representation and compression, followed by an introduction to the work related to compression quality assessment and enhancement. These will serve as the foundation of our further discussion on this topic and proposal of new and improved methods in the chapters to follow.

Chapter 3

No-reference PSNR Estimation for Compressed Pictures

Images and videos are available over the Internet, compressed and almost never with their originals alongside. No-reference (NR) PSNR estimation allows the content receiver to assess the objective quality of the compressed image or video and inspires confidence in using these pictures.

Existing methods for NR PSNR estimation based on the Laplace distribution model have achieved various degrees of success. The model describes the distribution of DCT coefficients sufficiently well for this specific purpose. As reviewed in the previous chapter, a single Laplace parameter needs to be estimated for each frequency band, and the PSNR can be calculated from the set of parameters. A common problem they all tried to address is the estimation of the original Laplace parameters. Compression distorts the distribution of DCT coefficients and how to get good estimates of the original parameters is still a challenge, and by improving the estimates, more accurate PSNR values can be obtained.

In this chapter, we put forth a new method for estimating PSNR of compressed still and motion pictures based on the Laplace distribution model of block-DCT

FIGURE 3.1: The default block divisions and a shifted one

coefficients. Unlike previous work in the area, we do not estimate the Laplace parameters directly from the dequantized block-DCT coefficients. Instead, we observe that block-DCT coefficient distributions are insensitive to the way in which the block boundaries are defined relative to the image, and divide the fully decompressed image into square blocks with altered boundaries, as illustrated in Fig. 3.1. These newly divided blocks are then transformed into the DCT domain and their coefficients are collected to obtain the initial estimates for Laplace parameters of the original image. These initial values are further processed using neural network regression to achieve closer resemblance to the true values. By shifting block boundaries, we are able to extract information about the original coefficients previously hidden by the default block boundary locations, and reduce PSNR estimation error.

Section 1 summarizes the existing methods for NR PSNR estimation. In Section 2 we describe our observations on the block-DCT coefficients, which lead to the proposal of our new method. Experimental results are shown in Section 3 and the chapter is concluded in Section 4.

3.1 Existing Methods for NR PSNR Estimation

3.1.1 Distribution Models for DCT Coefficients

As reviewed in Section 2.1.5, no-reference estimation of PSNR requires the knowledge of the probability distribution of the DCT coefficients. Several models have been proposed [46–49] to describe the block-DCT coefficient distribution of natural images. The zero-mean Laplace distribution has the simplest form of them all, which is

$$p_{u,v}(X) = \frac{1}{2\lambda_{u,v}} \exp\left(-\frac{|X|}{\lambda_{u,v}}\right), \quad (3.1)$$

where u, v are the frequency indices for the coefficients in the block-DCT domain. The maximum-likelihood (ML) estimator for λ is

$$\lambda_{u,v}^{ML} = \mathbf{E}[|X_{u,v}|], \quad (3.2)$$

where $\mathbf{E}$ is the expectation operator. Although generalized Gaussian [46–48] and symmetric normal inverse Gaussian [49] can likely describe the distribution of original block-DCT coefficients better, they have more parameters that require the use of higher order statistics. Hence, they are more sensitive to quantization errors than the simple Laplace model. The authors of [8] have shown in their paper that kurtosis (fourth-order statistics) is very sensitive to quantization and cannot be reliably estimated with known methods. Our preliminary experiments employing Mallat’s estimators [58] have shown that the generalized Gaussian distributions estimated based on quantized coefficients have poorer resemblance of those based on the original coefficients than using the Laplace model. Most of the existing no-reference PSNR estimation methods [7–10] use the Laplace model. We shall also see later through experiments, that even with the improved method proposed, the estimation accuracy is still limited by the process rather than the model.

FIGURE 3.2: Probability distributions of 8-by-8 block-DCT coefficients at each frequency, Lena image. Normalized for presentation.

3.1.2 Laplace-distribution-based Methods

One major and common problem all the existing works try to address is the estimation of the Laplace parameters, because quantization distorts the distributions. The effect can be seen in Fig. 3.2, where the shapes of the curves are distorted and many of them become all-zero. Turaga et al.[7] derive a second-moment expression in terms of the quantization step-size and Laplace parameter λ , and then equate it to the definition of second-moment. They noticed the tendency to over-estimate the parameter, and stated that empirically the first 24 and last 40 DCT frequencies in an 8-by-8 DCT block are to be treated differently in order to achieve a good final estimated PSNR. However, this scheme is specific to compression format and content. The magic number of 24 is neither generally applicable nor intrinsic to the model. In addition to the parameter estimation method, [7] also provided techniques to determine video picture types from a fully decoded sequence.

Ichigaya et al. [8] took note of the effect of zero-valued coefficients on the estimation of the Laplace parameters, and proposed to use two estimators, whose results are added with weights to form the final estimate. Both estimators measure the standard deviation of the DCT coefficients at each frequency, but one of them excludes all zero-valued coefficients. This two-part estimator improves the estimation accuracy for frequencies where zeros dominate, but nonzero values are still occasionally observed. For frequencies where all the coefficients have been quantized to zero, this scheme does not help with the estimation. The paper also investigated the feasibility of using higher-order statistics by looking at the estimated kurtosis, and came up with the conclusion that the distortions due to quantization is so great that the inclusion of higher-order statistics is counterproductive in PSNR estimation.

Eden proposed [10] to use the maximum-likelihood estimator based on quantized coefficients, making use of the quantization step-size and the proportion of

zero-valued coefficients, to better estimate λ for H.264 [32] video residues. Brandao and Queluz [9] also proposed to use a similar maximum-likelihood estimator for JPEG [30] coefficients. In addition, they applied a linear extrapolator to predict the λ values of higher frequencies based on those of the lower frequencies. This method generates nonzero λ values even when all coefficients in a particular frequency have been quantized to zero. However, the progressive extrapolation could lead to propagation of error, and the relationship of λ values at neighboring frequencies do not always follow a particular linear pattern. Nevertheless, Brandao's method stands as the state of the art in NR PSNR estimation for JPEG images.

Like in the existing work, we will use the Laplace model for AC coefficients for its simplicity and robustness to compression. According to Wu [59], difference between DC coefficients of adjacent blocks can be modeled using Laplace distribution. We therefore process the differences between adjacent DC coefficients in JPEG images as if they were AC coefficients. For video coding, the recorded DC coefficients are in fact residue values after prediction, which can be directly fitted using the Laplace model.

Unlike the others, we do not estimate λ values from the dequantized coefficients directly, following the observations we have made and described in the following.

3.2 Shifted Block Boundaries for Parameter Estimation

In this section, we present three key observations that lead to our new Laplace parameter estimation method based on quantized coefficients. The first is the relation between the location of the block boundaries and the resulting DCT coefficient statistics; the second is the effect of quantization on these statistics; the

last is the relationship among the distributions of the coefficients of the full image, the predicted image and the residue image for video frames.

3.2.1 Distribution of Coefficients

Pictures are usually compressed in disjoint blocks, but we realize that the first block does not have to coincide with the first pixel, although it is the default behavior in all compression formats.

Given the block size of N -by- N , there are N^2 unique ways in which the block boundaries can be disjointly placed in an image, based on a square grid. We take as an example the Lena image and decompose it with block boundaries aligned with the image boundaries as per normal for JPEG encoding ($N = 8$), and transform the blocks into the DCT domain to form the default-location dataset. Then, the block boundaries are shifted by one pixel distance to produce a new set of blocks—the global image content remains the same, but each block now is different from any of those obtained earlier. This shift-and-divide is repeated for the $N^2 - 1$ unique locations (i.e., minus the default). All $N^2 - 1$ sets of such image blocks are then transformed into the DCT domain to form the shifted-location dataset. A small number of pixels near the image borders are discarded since they do not form complete image blocks.

We show in Fig. 3.3 the log-probability density distributions of the coefficients at frequency (6,7) for the Lena image, of the default and shifted-location sets of DCT coefficients respectively. It can be seen that the shifted set carries a remarkable resemblance of the default set. The estimated Laplace parameters are shown in Table 3.1, with the top figure of each cell showing $\lambda_{u,v}$ estimated based on the default locations and the bottom based on the shifted ones. The former are considered the true values in the context of this chapter. A similar observation has been made by Smoot and Rowe during their investigation into the

FIGURE 3.3: Log-probability density distribution of the default and shifted block-DCT coefficients at frequency (6,7) of the Lena image, 8×8 blocks

distribution of block-DCT coefficients. They found that [34] “... surprising result was the stability of distribution values over images in a video sequence.” Indeed, the planar motion of a scene with respect to the fixed block boundaries is similar to shifting the block locations against a stationary image.

The statistics of block-DCT coefficients are determined by the image content, not by the division of the blocks. This is because natural images are random signals without fixed patterns of every N pixels. An artificially generated image with N -by- N patterns will certainly violate this observation.

With this, we can estimate the block-DCT coefficient statistics based on different block locations. The shifting of block location provides an easy means to obtain abundant observations of block-DCT coefficients for parameter estimation. It also alters the effect of quantization on the statistics, as will be discussed in the following subsection.

TABLE 3.1: λ estimated with the default and shifted boundary locations, original image without quantization

$u \downarrow, v \rightarrow$ each cell:							
				default		shifted	
82.07	48.86	20.83	11.29	7.55	5.31	3.94	3.27
81.06	51.08	21.55	12.16	7.84	5.69	4.39	3.57
27.75	19.19	12.47	8.77	5.84	4.42	3.52	2.95
27.66	19.53	13.18	8.60	6.03	4.45	3.52	2.92
10.72	10.68	9.21	6.85	4.94	3.94	3.16	2.70
10.79	10.75	8.95	6.79	5.05	3.89	3.15	2.69
6.14	6.13	5.93	5.07	4.11	3.31	2.80	2.43
6.15	6.20	5.84	5.09	4.12	3.31	2.75	2.43
4.11	4.10	4.10	3.90	3.28	2.78	2.42	2.22
4.11	4.15	4.08	3.86	3.35	2.79	2.42	2.20
3.10	3.01	3.01	2.93	2.67	2.36	2.21	2.08
3.07	3.00	3.01	2.94	2.70	2.39	2.16	2.02
2.45	2.46	2.33	2.33	2.27	2.14	2.03	1.93
2.47	2.43	2.35	2.31	2.25	2.12	2.00	1.90
2.25	2.18	2.08	2.04	2.03	1.98	1.88	1.82
2.26	2.14	2.06	2.02	2.00	1.95	1.87	1.80

3.2.2 The Effect of Quantization

Quantization in the block-DCT domain reduces the precision of data. Coefficients of small magnitudes tend to be quantized to 0, and these 0 values result in a smaller value of parameter λ estimated from the quantized coefficients. The extreme case happens where the quantization step-sizes are large while the signals are weak; all these coefficients may become 0 after quantization, and little information about their statistics in the original unquantized image can be obtained.

However, when the fully decompressed images are segmented into blocks with shifted boundaries and transformed into the DCT domain, these coefficients are no longer 0. An example is shown in Fig. 3.4 based on frequency (6,7) of the Lena image compressed with JPEG quality = 75 (quantization matrix recommended in the JPEG standard [30] scaled by 0.5). The dequantized coefficients are all

FIGURE 3.4: Log-probability density distribution of the block-DCT coefficients at frequency (6,7) of the Lena image, JPEG Quality = 75

0 at this frequency, giving very little information about its original distribution, while the coefficients with the shifted block locations provide a distribution that better resembles the original. Table 3.2 shows the λ values estimated from the quantized Lena image at quality level 75 using (3.2), where the upper number in each cell indicates the estimate based on the default block locations and the lower one indicate that based on the shifted blocks. Even at this relatively moderate compression setting, the estimates based on the default block locations already have a lot of 0 values. Compared with the figures shown in Table 3.1, it can be seen that based on a compressed image, the estimated λ values using the shifted block locations are closer to the true values than those using the default block locations, although the differences are still large.

In general, any shift in block boundaries results in completely different DCT coefficients for individual blocks, yet the global image content remains the same. Based on the observation above, the signals at zero-valued frequencies are not completely wiped out by quantization, as the zeros might suggest. Instead, they

TABLE 3.2: λ estimated with the default and shifted boundary locations, JPEG-quantized image

$u \downarrow, v \rightarrow$ each cell:					default shifted		
82.04	48.86	20.74	10.91	6.40	3.08	1.38	0.55
82.02	50.42	21.00	11.66	7.31	4.76	3.24	2.35
27.66	19.01	12.10	7.97	4.50	1.79	0.97	0.47
28.01	19.74	13.30	8.56	5.70	3.78	2.62	1.94
10.33	10.26	8.61	5.60	2.87	1.44	0.38	0.25
10.99	10.89	8.98	6.60	4.52	3.05	2.15	1.63
5.65	5.38	4.87	3.44	1.78	0.42	0.22	0.10
6.25	6.19	5.66	4.65	3.36	2.32	1.66	1.33
3.20	2.90	2.19	1.51	0.66	0.11	0.01	0.06
4.01	3.88	3.58	3.08	2.31	1.64	1.23	1.03
1.64	0.98	0.57	0.44	0.11	0.01	0.00	0.02
2.69	2.47	2.26	1.99	1.56	1.16	0.90	0.76
0.21	0.10	0.01	0.03	0.00	0.00	0.00	0.00
1.73	1.60	1.44	1.29	1.06	0.82	0.63	0.53
0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1.25	1.18	1.10	1.03	0.87	0.68	0.53	0.43

are attenuated across the whole image. This observation provides us with extra information about the original coefficients and a good initial estimate for the Laplace parameters to work from.

3.2.3 The Case of Residue Coefficients (Video)

Compared with still images compressed in JPEG, which are described using one set of coefficients, video frames usually comprise two parts of information, namely the prediction and the residue. A motion-predicted block is in general some weighted sum of similar but translated image areas in other video frames, which closely resembles the current image block. The prediction plus the residue gives the full image. For lossy coding, the residue image blocks are approximated by quantization after transformation.

Because the errors are in the residue coefficients, their distribution parameters λ_{res} need to be estimated in order to calculate the PSNR. However, there are two complications. Firstly, the residue blocks do not form a contiguous image. Since motion estimation is performed independently for each block (or small groups of blocks), the predictions and residues may not line up at the boundaries and shifting the new block boundary across the default boundaries create errors.

Secondly, the amount of information in the residue depends on how well the prediction resembles the original image. This could vary greatly depending on the nature of the video scene and the capability of the encoder. For example, a hand-held device usually produces videos with poorer predicted pictures due to its limited computational power compared to an studio encoder; a video sequence with zooming effects will also result in poorer predictions compared with panning because the motion fits poorly with the translation assumption for motion estimation.

The first complication means that $\hat{\lambda}_{res}$ estimated based on quantized residue coefficients alone is unreliable. The second one suggests that the solution should involve the predicted image. Knowing the relationship that the prediction adding the residue yields the full image, the information from the these three types of images can all contribute to the estimation of λ_{res} . Because the DCT coefficients of the prediction and the full image are not independently distributed, taking the difference of their λ values does not yield λ_{res} . We treat the fully reconstructed video frames in the same way as JPEG images and use the proposed shift-and-transform method to obtain its Laplace parameters $\hat{\lambda}_{full}$. Because each macroblock may have its own quantization step-size, the shifting is restricted within the macroblock. The initial estimates for the prediction and residue, $\hat{\lambda}_{pred}$ and $\hat{\lambda}_{res}$, are estimated without shifting because the images are not contiguous. All three $\hat{\lambda}$ values are to be used later to generate a refined estimate of λ_{res} .

3.2.4 The Number of Shifted Locations

In the above discussion, all N^2 alternative block boundary locations have been used for estimating λ . This can be computationally demanding. When computational time is of concern, fewer locations may be favorable. Preliminary experiments show that λ estimated from the original coefficients are almost identical regardless of the number or choice of locations. However, the estimates can vary greatly when quantized coefficients are used, as shown in Table 3.3, where results based on four different boundary locations are tabulated. It can be seen that the estimated values behaves differently at different frequencies, and there is no single location that produces universally optimal estimates. In contrast, using all $N^2 - 1$ locations not only removes the need to find the best compromise, but also provides us with as many observations as possible. To reduce computation, N -by- N blocks can be picked from random locations, so that the resulting image blocks enumerate all shifted locations and cover the whole image with some overlaps. This could bring the computational cost down to the same level of decoding the picture.

3.2.5 Neural Network Regression

In this subsection, we take the initial estimates of the Laplace parameter $\hat{\lambda}$ obtained earlier and further process them for the estimation of PSNR.

Because the aforementioned attenuation effect cannot be completely modeled as independent of the original image content and that the relation between the initial estimates and true λ values can be quite complex, we use a learning-based technique to correct the estimation of λ . A simple neural network as illustrated in Fig. 3.5 with a single hidden layer is used. Each input vector consists of the initial estimates $\hat{\lambda}$, their corresponding quantization step-size q , and the frequency indices u and v . The output is a single value of the corrected $\tilde{\lambda}$. For JPEG images,

TABLE 3.3: λ estimated from quantized coefficients based on different boundary shifting, JPEG

$u \downarrow, v \rightarrow$ each cell, shifted by:															
<div><div>$(4,4)$ $(0,4)$</div><div>$(4,0)$ $(2,2)$</div></div>															
82.3	82.0	50.8	50.8	20.3	20.9	12.0	11.9	7.39	7.47	5.14	5.17	2.72	2.71	3.07	3.04
81.9	81.9	50.2	49.7	20.8	20.8	11.3	11.6	7.28	6.93	4.86	3.40	3.68	1.75	2.45	0.88
28.3	27.6	19.2	20.3	13.8	12.6	8.53	8.42	5.84	5.49	4.18	4.34	2.15	1.97	2.60	2.70
27.8	28.2	19.4	19.7	13.1	13.7	8.53	8.51	5.72	5.51	3.90	2.75	3.16	1.44	2.05	0.73
10.7	10.9	11.1	10.6	8.89	9.22	6.62	6.61	4.28	4.14	3.40	3.31	1.49	1.31	2.19	2.22
11.0	10.9	10.8	10.8	8.80	8.98	6.63	6.42	4.52	3.79	3.23	1.80	2.66	0.81	1.81	0.47
6.24	6.23	6.16	6.07	5.60	5.35	4.66	4.41	3.33	2.82	2.83	2.62	1.37	0.74	2.02	1.77
6.23	6.37	6.20	6.15	5.73	5.57	4.56	4.58	3.14	3.17	2.38	1.67	2.23	0.88	1.43	0.48
4.01	3.71	3.99	3.64	3.33	2.88	3.05	2.77	1.99	1.40	1.97	1.80	0.51	0.26	1.33	1.23
3.90	4.12	3.80	3.83	3.49	3.23	2.85	2.82	1.79	1.53	1.54	0.43	1.53	0.19	1.00	0.15
2.95	1.90	2.81	1.82	2.63	1.07	2.43	1.25	1.89	0.40	1.80	0.82	0.79	0.09	1.40	0.59
2.66	2.95	2.50	2.84	2.43	2.71	2.05	2.39	1.51	1.75	1.27	0.92	1.29	0.52	0.87	0.29
1.23	0.29	1.11	0.25	0.78	0.08	0.76	0.14	0.38	0.02	0.49	0.09	0.11	0.01	0.34	0.07
2.26	1.21	2.15	0.96	2.05	0.71	1.88	0.57	1.50	0.27	1.30	0.08	1.28	0.04	0.89	0.03
1.95	0.02	1.89	0.02	1.72	0.00	1.59	0.01	1.31	0.00	1.28	0.00	0.59	0.00	0.96	0.00
1.29	1.96	1.27	1.82	1.23	1.72	1.13	1.67	0.94	1.25	0.79	0.70	0.76	0.40	0.53	0.22

FIGURE 3.5: A neural network with a single hidden layer (video parameters in parentheses)

$\hat{\lambda}$ are extracted from the shifted decoded image. For video frames, $\hat{\lambda}$ represents three parameters – $\hat{\lambda}_{full}$, $\hat{\lambda}_{pred}$ and $\hat{\lambda}_{res}$, and the output is $\tilde{\lambda}_{res}$.

After obtaining the $\tilde{\lambda}$ values, Equations (3.3)(3.4) are used to estimate the MSE, and PSNR can be computed from MSE according to definition (2.13). Equation (3.3) is identical to that defined in [9], where K is the total number of coefficients in the image and ε_k is the quantization error at coefficient k . The estimation of ε is based on the representative value X_k (i.e. the dequantized coefficient value), quantization step-size q_k , and the Laplace distribution density function $p_k(x)$. Equation (3.3) integrates the squared error weighted by the density function over the possible range of the original coefficient value between X^{lower} and X^{upper} , and normalizes the integral with the total probability within the range to obtain the expected squared error for the coefficient k . For video inter-frames, X , X_k , X^{upper} , and X^{lower} in (3.3) need to be replaced with R , R_k , R^{upper} , and R^{lower} respectively, to reflect the fact that the computations are based on residue signals.

$$\hat{\varepsilon}_k^2 = \frac{1}{\int_{X^{\text{lower}}}^{X^{\text{upper}}} p_k(X) dX} \int_{X^{\text{lower}}}^{X^{\text{upper}}} p_k(X) (X_k - X)^2 dX \quad (3.3)$$

$$\mathbf{MSE} = \frac{1}{K} \sum_{k=1}^K \hat{\varepsilon}_k^2 \quad (3.4)$$

Note that JPEG [30] and MPEG-2 [33] use normalized 2D-DCT and the noise power is equal in the spatial and transform domains. With compression standards such as H.264 [32] where an approximation to the normalized 2D-DCT is used, the equality is also approximated.

To summarize, the proposed PSNR estimation method includes the following steps:

1. Fully decode the compressed picture.
2. Shift the block boundary to all $N^2 - 1$ unique locations and perform DCT.
3. Estimate $\hat{\lambda}$ ($\hat{\lambda}_{full}$ for video) using (3.2) and the DCT coefficients obtained in Step 2. For video, also estimate $\hat{\lambda}_{pred}$ and $\hat{\lambda}_{res}$ based on the default block boundaries.
4. Correct $\hat{\lambda}$ using a pre-trained neural network to obtain $\tilde{\lambda}$ ($\tilde{\lambda}_{res}$ for video).
5. Apply (3.3), (3.4) and (2.13) to obtain PSNR.

3.2.6 Alternative: Enlarged Blocks

Before going into the experiments, we would like to mention an alternative scheme. Instead of using shifted blocks with the same size as used in compression, we may also divide the image into non-overlapping square blocks of size $(N + m)^2$, where $m = 1, 2, 3, \dots$. Fig 3.6 illustrates such a block arrangement.

The purpose of this operation is similar to that of shifting, namely breaking the grids upon which quantization is performed. After transforming these $(N + m)^2$ -sized blocks, the upper-left N^2 coefficients can be treated as if they were obtained

FIGURE 3.6: The default block divisions and the enlarged ones

by transforming N^2 -sized image blocks. Similarly, neural network regression is applied to correct the errors. By using a larger block-size, we are actually moving up in the scale space and exploiting the similarity between scales. Naturally, as we move further from the original scale (larger m), the resemblance of the enlarged blocks to the original image blocks gets weaker. Therefore, $m = 1$ or a block-size of $(N + 1)^2$ should best serve our purpose. The advantage of this scheme over the shifting scheme is that there are fewer blocks to compute, and the disadvantage is that if N is small (e.g. $N = 4$ for H.264), $(N + 1)^2$ blocks may have a poor resemblance to the originals.

3.3 Experimental Results

The experiments consist of two parts. The results from JPEG images are presented first, followed by the results from H.264 videos. According to the study by Smoot and Rowe [34], the chrominance channels exhibit the same distribution as the luminance channel, but with smaller λ values due to higher spatial correlation. Only luminance components are used in our experiments.

For the simulations with JPEG images, we used all 29 images from the LIVE database [60]. They are compressed using JPEG quality settings from 5 to 90 with an interval of 5. Note that the proposed method does not depend on the knowledge of this quality factor. In fact, the quantization matrices can be arbitrarily defined. Both shifted- and enlarged-block-based methods are examined.

For H.264 videos, 9 QCIF (176-by-144 pixels) sequences are used, including: bus, city, crew, football, foreman, harbor, ice, mobile and soccer. Each sequence is compressed using X264 [61] and JM 16.2 [62] encoders with main profile and adaptive quantization (per macroblock) at 100 kbps and 200 kbps, resulting in 36 encoded sequences. Group-of-pictures (GOP) structure is determined by the encoder for X264 and IBPBP for JM 16.2. Training uses X264 samples only, and the resulting neural networks are directly applied to JM 16.2 samples. For H.264 videos only shifted-block method is used.

FIGURE 3.7: Test images from LIVE database

The neural network we used has 5 neurons in its hidden layer, and is trained using 75% of the available samples and 15% each for testing and validation. Video training samples are further reduced by uniform random sub-sampling to a manageable size. The implementation was done using Matlab Neural Network Toolbox. The $\hat{\lambda}$ values are normalized to the range of $[-1 \ 1]$ for fitting and scaled back at the output. Experiments show that the number of neurons has minimal impact

on the regression results. Using 5 neurons and a starting learning rate of 0.001, it normally takes 50 iterations before the gain in performance becomes negligible, and about 150 iterations before over-fitting begins.

We use the leave-one-out cross-validation scheme to verify the performance of the proposed algorithm and compare it with that of the state-of-the-art NR PSNR estimation methods. For each test image or sequence, all other samples in the database are used for learning. Leave-one-out is a suitable evaluation scheme given the relatively small dataset of 29 images and 9 video sequences. The results on JPEG images are compared with those obtained using λ -prediction by Brandao [9] and those on H.264 videos are compared with Eden’s method [10]. True PSNR values are calculated using definition (2.13), since this is a simulated environment and the original pictures are available for verification purposes. This is not to be confused with PSNR values estimated using “true λ ”. The latter refers to λ values estimated based on the original pictures. The PSNR values estimated based on “true λ ” have limited resemblance to true PSNR because of the limitations of the Laplace model.

3.3.1 JPEG

3.3.1.1 Shifted Blocks

Table 3.4 lists the average percentage error in the estimated $\hat{\lambda}$ and $\tilde{\lambda}$ with respect to the true values using the proposed shifted-block-based method. The upper figure of each cell refers to the initial estimates based on shifted block boundaries, and the lower one is the corrected estimate output from the neural network. As can be seen, the initial estimates can be as different as 40% from the true values on average, while the corrected estimates are within 8% for individual frequencies. From earlier discussion we know that although shifting block boundaries recovers some extra information, it does not completely remove the effect of quantization

TABLE 3.4: Average relative estimation error of λ (%)

$u \downarrow, v \rightarrow$ each cell:				<div style="border: 1px solid black; padding: 2px; display: inline-block;"> initial estimate $\hat{\lambda}$ corrected estimate $\tilde{\lambda}$ </div>			
0.85	-2.38	-3.11	-5.82	-9.67	-12.99	-13.23	-5.52
-0.34	0.11	-1.55	0.48	-1.45	-2.48	-2.69	-4.50
-3.59	0.75	-0.18	-3.72	-10.23	-14.86	-16.65	-8.41
2.60	-2.60	-2.43	-1.01	1.68	-4.73	-1.80	-0.82
-3.40	-0.62	-3.11	-7.40	-14.90	-20.66	-21.81	-13.68
0.02	-2.88	-0.92	-0.24	-0.22	-0.86	-2.33	-0.12
-5.03	-2.90	-6.52	-12.62	-20.58	-26.45	-27.78	-18.69
-0.12	-3.08	-0.90	2.09	0.59	-4.35	-1.50	-0.61
-7.39	-7.49	-13.22	-19.83	-29.32	-33.73	-34.73	-25.08
1.88	-1.71	-0.08	0.38	3.20	-4.20	-2.97	-1.27
-9.13	-11.67	-18.28	-25.14	-33.81	-37.42	-37.69	-28.58
0.57	-2.61	-1.22	1.45	2.89	-2.19	-4.90	-4.45
-12.48	-17.06	-23.07	-29.77	-36.81	-40.06	-40.37	-31.24
-1.31	-4.17	-1.84	-0.10	-0.42	-5.91	-5.95	-4.09
-8.28	-14.07	-20.15	-25.97	-32.83	-35.75	-36.12	-29.06
-5.00	-7.23	-4.42	-3.81	-5.55	-3.75	-4.40	-0.56

on λ . Through neural network regression, this extra information is used to better estimate the original λ .

Fig. 3.8 and Table 3.5 compares the PSNR estimation results using our proposed method and that in [9]. It can be seen from the plots in Fig. 3.8 that our proposed method generates results closer to the true values than [9]; the slope is closer to 1 and the points are more closely packed. A slope of 1 indicate that the estimated values are accurate at all quality levels, while closely packed points indicate high precision. The better PSNR estimates are a direct result of better λ values estimated.

The results from [9] shows a tendency to underestimate PSNR especially at higher quality levels. We have communicated with the authors and verified this observation. This behavior is probably due to the propagation of λ estimation error from lower frequencies to the higher ones. At higher quality levels, the error

FIGURE 3.8: Scatter plots of PSNR estimation results, JPEG

due to direct estimation is expected to be small, while the extrapolated estimation from lower frequencies may cause greater error. Our method does not exploit the correlation between different frequency bands thus avoids this potential problem.

Table 3.5 numerically illustrates the advantage of the proposed method, which not only halves the errors of [9] but also approaches the performance of estimating using the “true” Laplace parameters. These “true” parameters are estimated from the uncompressed original images and can be considered ground truth in the Laplace model context. There is some error in the estimated PSNR using these true parameters, due to the limitation of the Laplace model. While our proposed method is close to reaching this limit of the model of choice, we still lack the necessary knowledge and tools to successfully estimate higher order statistics in the absence of the original and are unable to go for more sophisticated models such as generalized Gaussian. Note that the results reported in [9] (shown in small font in the table) are slightly better than we could reproduce, possibly due to the difference in testing strategy since they used half of the samples for learning.

TABLE 3.5: PSNR estimation results, JPEG

	Brandao's[9]*	Proposed	True λ
Mean Absolute Error (dB)	0.750 <small>0.660</small>	0.360	0.280
Root Mean Square Error (dB)	0.900 <small>0.789</small>	0.452	0.367
Correlation Coefficient	0.992 <small>0.992</small>	0.996	0.998

* Numbers in small font size are as reported in [9].

3.3.1.2 Enlarged Blocks

Having seen the comparison between the proposed shifted-block method and the state-of-the-art method, we shall take a brief look at our alternative method using enlarged blocks. In Table 3.6, we list the test results obtained using the same experimental strategy but with various block sizes. It can be seen that with block size 9-by-9, the results are comparable to those obtained using the shifting method. As m increases, the estimation errors grow. This is because the enlarged-block-based method relies on the texture similarity between different scales of the same image. As the difference in block size grow larger, this similarity is expected to reduce. If the default block size is $N = 4$, as is the case with H.264, and $m = 1$, the results based on the enlarged-block method should be analogous to $N + m = 10$ in the table¹, which is clearly not as good as the shifted-block method.

TABLE 3.6: PSNR estimation results, JPEG, using enlarged block sizes

$N + m$	MAE (dB)	RMSE (dB)	Corr.
9	0.357	0.460	0.996
10	0.404	0.531	0.995
11	0.451	0.560	0.994
12	0.456	0.567	0.994
13	0.483	0.606	0.993
14	0.523	0.655	0.983
15	0.526	0.647	0.991
16	0.576	0.711	0.990

¹They are of the same distance in the scale space: $2/8 = 1/4 = 0.25$

FIGURE 3.9: Scatter plots of PSNR estimation results, H.264

3.3.2 H.264

The results based on H.264 video sequences are shown in Table 3.7 and Fig. 3.9. The results from Eden's method were obtained using the same "calibration factors" mentioned in his paper [10].

The estimated PSNR values are closer to the true values (the diagonal lines in Fig. 3.9) with the proposed method than Eden's. The points are also more closely packed, indicating a higher precision of estimation. Our leave-one-out training scheme guarantees that no parameter has been tuned to suit the test samples. Eden's method may have become "miscalibrated" with our testing samples. Also, Eden's method show a distinct tendency to underestimate PSNR values, where the points lie mostly to the left of the diagonal line. The direct cause of this phenomenon is overestimated λ values when all coefficients of certain frequencies are quantized to zero. In such situations, the upper limit of λ based on the quantization step size is used in [10]. This also explains why Eden's method has smaller errors for higher quality pictures, where the upper-right portions of the scatter plots are tighter than the lower-left. In lower quality pictures, the quantization is coarser, and it is more likely to encounter the all-zero situation. Our proposed method, on the other hand, takes on the all-zero problem differently and is less affected by it. There is less variation in estimation accuracy between high and low quality pictures with our method. With JM 16.2 samples, the proposed method does overestimate PSNR for a few pictures of very low quality. This is possibly due to the lack of information in some of the B-pictures where most blocks have been skipped.

From Table 3.7 it can be seen that our method has lower error figures and higher correlation coefficients than Eden's. Between the two sets of samples, JM 16.2 samples show better results. This is because JM only switches quantization parameters once or twice within each frame even though the rate control basic unit is set to 1 macroblock. Under the same conditions, X264 chooses to switch between several parameters every few macroblocks. Because λ needs to be estimated for each quantization parameter, there are far fewer blocks to use for X264 samples. The smaller number of samples leads to larger estimation variance and thus poorer results.

TABLE 3.7: PSNR estimation results, H.264

	X264			JM 16.2		
	Eden's [10]	Proposed	True λ	Eden's	Proposed	True λ
MAE (dB)	2.219	1.057	0.450	1.538	1.056	0.473
RMSE (dB)	2.851	1.291	0.558	2.076	1.407	0.682
Corr.	0.890	0.950	0.994	0.944	0.972	0.991

The results using “true λ ” are not perfect, showing the limitation of the Laplace model. The test video sequences arguably represent the most challenging scenario for NR PSNR estimation because they combine very small frame size with adaptive quantization, giving very few samples per frequency-quantization parameter combination. With larger frame sizes or without adaptive quantization, the estimation results could be even better.

3.4 Conclusion

In this chapter, we have proposed an improved method for estimating the PSNR of block-DCT-based compressed pictures without referring to their originals. Realizing that the artificially assigned block boundaries are hiding information about block-DCT domain statistics, we estimate the Laplace parameters from the fully decompressed image with shifted or enlarged block boundaries. This method, used in conjunction with a neural-network-based regression procedure, greatly reduces the impact of the zero-valued quantized coefficients and produces better estimation results. The mean absolute error of the estimated PSNR is only 0.36 dB for JPEG images and 1 dB for H.264 videos of very small frame size, notably better than the state-of-the art algorithms [9, 10] and approaches the limit of the Laplace distribution model.

Chapter 4

Relative Compression Quality Assessment

In the context of the Internet, it is common for one image or video sequence to exist in several different locations and in differently compressed forms. An end user may have access to a few of such compressed pictures with essentially the same content, not knowing which of them most closely resembles the uncompressed original.

We could directly apply the method proposed in Chapter 3, blindly estimate the objective quality for each individual copy, and compare the numbers. As the results from the previous chapter suggest, this scheme should work out well under most circumstances. However, there are still a few limitations to the technique. Firstly, some errors inevitably remain in the estimation, and when the pictures have similar PSNR measures, it may become difficult to tell which is better. Secondly, it is limited to DCT-based compression formats. Lastly, the computation is complex as it involves many transform operations, and one may not want to know how well each picture is compressed but simply which copy is the best among a few. By eliminating the need for statistical parameter estimations and

instead focusing on the well-defined quantization constraints, more robustness can be expected in relative quality assessment.

In this chapter, we propose to estimate the relative objective quality in terms of the difference in mean squared error (ΔMSE) among differently compressed copies of the exact same original picture by cross-referencing them in their respective transform domains. With the aim of making accurate qualitative assessment, we circumvent the estimation of MSE or PSNR of individual pictures and try to reduce the error of our assessment by estimating ΔMSE instead. The implication of this approach, though, is that ΔPSNR values do not entail, as their computation requires the MSE values of individual pictures (or at least the ratio between them because the peak values are the same for the same original image):

$$\begin{aligned}\Delta\text{PSNR}_{A,B} &= 10 \cdot \log_{10} \left(\frac{\text{MAX}^2}{\text{MSE}_A} \right) - 10 \cdot \log_{10} \left(\frac{\text{MAX}^2}{\text{MSE}_B} \right) \\ &= 10 \cdot \log_{10} \left(\frac{\text{MAX}^2 \cdot \text{MSE}_B}{\text{MAX}^2 \cdot \text{MSE}_A} \right)\end{aligned}\tag{4.1}$$

Therefore, the difference in PSNR mentioned in this chapter are not from the estimation, but rather are computed based on ground truths and serve as a context for discussion.

This new method does not rely on the accurate estimation of statistical properties of the transform coefficients but rather makes use of the knowledge on the quantization of the coefficients. The method also works when the pictures are compressed using different transforms. The estimated ΔMSE values highly correlate with the ground truth and the method indicates the relative quality between visually similar compressed pictures. We also demonstrate the performance advantage compared with the NR assessment methods discussed earlier in Chapter 3 in the context of relative quality assessment.

The remaining of the chapter is organized as follows. Section 1 gives an introduction to the trivial and nontrivial cases of relative quality assessment. Section 2 details the proposed method. Section 3 demonstrates the capability of the proposed method with experimental results, and Section 4 gives our concluding remarks.

4.1 Trivial and Nontrivial Cases

Before going further into the discussion of assessing the relative compression quality among image copies, it is necessary to point out that there are some situations where such an assessment can be trivially performed. We would like to discuss the conditions for these trivial cases and also explain the nontrivial and more general cases.

Knowing that quantization is the only step in compression that introduces irreversible loss of fidelity, the comparison of compression quality becomes trivial if the data before quantization are equal and the quantization schemes used are similar. In such cases, the comparison of quality becomes the comparison of quantization step-sizes. Let us discuss the equality of data first, followed by an explanation of what are considered similar quantization schemes.

Having assumed that the pictures are directly compressed from the exact same original, the equality in data requires the copies of a still image to use the same block size and transformation. For video, GOP structure, motion vectors, prediction weights, etc. have to be identical as well. As soon as any one of them is violated, quantization step-sizes cannot be directly compared because the foundation of comparison is lost. For example, an image may be compressed in the JPEG [30] format while the other in JPEG2000 [35]. The different transformation applied (DCT vs. DWT) makes direct comparison of quantization step-sizes meaningless. Another example where the conditions are violated would be two

H.264-encoded [32] video sequences, one with half-pixel motion estimation precision and the other quarter-pixel precision. In the latter example, we do not actually know the encoding settings but from the motion vectors we know that the data before quantization are different. There are too many example scenarios to enumerate.

Even if the data before quantization are the same, the comparison of quantization step-sizes may not give any indication which picture copy is better. A quantization matrix may introduce different amount of distortion to each frequency band, and the energy concentration in each band varies with the image content. An example is the quantization matrix for JPEG images. The JPEG standard [30] recommended a quantization matrix and the Independent JPEG Group (IJG) [63] proposed a scaling scheme for this base matrix using a user-specified quality level between 1 and 100 for more flexible compression. This scheme is widely used because of the free distribution of the IJG's JPEG library, and is often mistaken as "the standard JPEG tables"¹. If all copies of an image are compressed using the IJG quantization matrices, their relative quality assessment could be as trivial as extracting the matrices and see which has the smallest scaling factor, and hence the highest quality. However, since the JPEG standard leaves the choice of quantization matrices to the encoder, there are situations where the comparison is more difficult. For example, Adobe Photoshop uses a set of different quantization matrices for a range of quality settings between 0 and 12. Fig. 4.1 shows two possible quantization matrices, one from the IJG and the other from Adobe Photoshop. It can be seen that even though both matrices have similar DC and low-frequency AC quantization step-sizes, the Photoshop version uses very fine quantization for high-frequency coefficients. How each of these quantization matrices fair with a particular picture depends on the frequency characteristics of its content.

¹In fact, we also used this term in Chapter 3 for simplicity's sake.

18	13	11	18	27	45	58	69	18	14	14	21	30	35	34	17
14	14	16	22	30	66	68	63	14	16	16	19	26	23	12	12
16	15	18	27	45	65	78	64	14	16	17	21	23	12	12	12
16	19	25	33	58	99	91	70	21	19	21	23	12	12	12	12
20	25	42	64	77	124	117	88	30	26	23	12	12	12	12	12
27	40	63	73	92	118	128	105	35	23	12	12	12	12	12	12
56	73	89	99	117	138	136	115	34	12	12	12	12	12	12	12
82	105	108	111	127	114	117	113	17	12	12	12	12	12	12	12

FIGURE 4.1: Quantization Matrices: IJG Quality Level 44 (left), Photoshop Quality Level 3 (right)

In the following, we deal with the non-trivial, or rather, the general case of relative compression quality assessment, posing no condition other than our originally stated assumption that the pictures are compressed from the same original. The method is therefore in principle applicable to all block-transform-based image and video compression formats, and is computationally less intensive and more accurate than directly applying the method proposed in Chapter 3.

4.2 Conservative Estimate of the Difference of MSE

Suppose we have two corresponding DCT coefficients, each from a differently compressed picture of the same content. The relative squared error $E_{A,B}(u, v)$ is

$$E_{A,B}(u, v) = \left[\hat{X}_A(u, v) - X(u, v) \right]^2 - \left[\hat{X}_B(u, v) - X(u, v) \right]^2 \quad (4.2)$$

and the difference in MSE is

$$\Delta \text{MSE}_{A,B} = \frac{1}{UV} \sum_{u=0}^{U-1} \sum_{v=0}^{V-1} E_{A,B}(u, v). \quad (4.3)$$

Unfortunately, we do not have access to the original and thus do not know $X(u, v)$. As reviewed in Chapter 2, each quantized coefficient of a block-transform-compressed picture corresponds to a known range of values, the quantization constraint set (QCS), among which the original value must lie.

FIGURE 4.2: Coefficient X and the effect of quantization

Fig. 4.2 illustrates the the original coefficient value X and its quantization bounds X^{upper} and X^{lower} , and the reconstruction level $\hat{X}$. Given multiple compressed copies from the same original image, the reconstruction levels are generally different, and the reconstruction level of one copy may fall outside the QCS of another copy. Assuming there is no other source of error but quantization, the true value of each coefficient cannot reside outside any QCS. Therefore, the distances $[\hat{X}_A(u, v) - X(u, v)]$ and $[\hat{X}_B(u, v) - X(u, v)]$ can be looked at in two parts

$$\begin{aligned} \hat{X}_A(u, v) - X(u, v) &= [\hat{X}_A(u, v) - X_B^{\text{near}}(u, v)] \\ &\quad + [X_B^{\text{near}}(u, v) - X(u, v)], \end{aligned} \quad (4.4)$$

where

$$X_B^{\text{near}}(u, v) = \begin{cases} X_B^{\text{lower}}(u, v) & \text{if } \hat{X}_A(u, v) < X_B^{\text{lower}}(u, v) \\ X_B^{\text{upper}}(u, v) & \text{if } \hat{X}_A(u, v) > X_B^{\text{upper}}(u, v) \\ \hat{X}_A(u, v) & \text{otherwise} \end{cases} \quad (4.5)$$

is nearest point of Copy B's QCS to $\hat{X}_A$. X_B^{upper} and X_B^{lower} are the upper and lower bounds of the QCS of Copy B, respectively, and $\hat{X}_A$ is the reconstruction level of Copy A. If $\hat{X}_A < X_B^{\text{lower}}$, X_B^{lower} is the near bound and X_B^{upper} is the far bound; if $\hat{X}_A > X_B^{\text{upper}}$, X_B^{upper} becomes the near bound instead. We denote the distance

between the near bound and $\hat{X}_A$ as

$$\Delta X_{A,B}(u, v) = \hat{X}_A(u, v) - X_B^{\text{near}}(u, v). \quad (4.6)$$

Similarly, the distance $\Delta X_{B,A}$ of Copy B can be computed with respect to Copy A. Again, we do not know X , but now we know that the two summation terms in (4.4) take the same sign. By dropping the term concerning the unknown X from (4.4) and plugging in (4.6), we get

$$E_{A,B}(u, v) \approx \Delta X_{A,B}^2(u, v) - \Delta X_{B,A}^2(u, v). \quad (4.7)$$

Putting it into (4.3) gives us the estimated ΔMSE between a pair of pictures. When more than two copies are present, the comparison can be performed in a pair-wise manner to rank the pictures according to their objective quality. This measure is found to be highly correlated with the true ΔMSE when verified against the original source image or video.

As illustrated in Fig. 4.3, the squared distance between $\hat{X}_A$ and X , the original, is at least $\Delta X_{A,B}^2$, and this measure can be viewed as a conservative estimate of the squared error of the reconstructed value with respect to the original value X . The original X is usually further away from the reconstructed signal than the near bound defined by the other compressed pictures but the distance is unknown. Also, when the reconstruction level falls within the bounds defined by another copy, the error is assumed to be zero. Therefore, the procedure underestimates the magnitude of the errors at coefficient level and because of its symmetry, underestimates the magnitude of ΔMSE .

If the pair of pictures are indeed compressed from the same original picture, at least one of $\Delta X_{A,B}$ and $\Delta X_{B,A}$ will be small, showing one copy superior to another at this particular coefficient. In case the coefficients have similar distortions, $\Delta X_{A,B}$ and $\Delta X_{B,A}$ will be similarly small. However, if the paired pictures

FIGURE 4.3: Conservative error of B with respect to A (left) and that of A with respect to B (right). Unknown or unused values are shown as dotted lines.

FIGURE 4.4: Conservative errors between A and B, when the pictures are not compressed from the same original

are not from the same original picture, both $\Delta X_{A,B}$ and $\Delta X_{B,A}$ will be large, as the coefficients generally have different values in different pictures and the QCSs do not overlap, as shown in Fig.4.4. This property serves as an intrinsic check against incorrectly identified picture copies and prevents the method from giving invalid results when invalid inputs are presented.

Conceivably, there are a few more options for the reference point other than the near bound. These include the distance to the far bound, the reconstruction level or somewhere inside the intersection of the two QCSs. The problem with using the far bound is that it is often outside the QCS of the other Copy and hence grossly over-estimates the error. Also, since we only accumulate such errors when the reconstruction levels are out of the bounds, using the far bound would result in a mixture of over-estimation and under-estimation, for which the behavior is not easy to predict. The reconstruction level is not found to be a useful estimate because it measures the difference between the two reconstructed pictures but offers no information about their relative quality (i.e., $|\hat{X}_A - \hat{X}_B| = |\hat{X}_B - \hat{X}_A|$). Lastly, knowing that the true value lies within the intersection of the two QCSs,

it is possible to derive a maximum-likelihood (ML) reconstructed value from all available image copies and use this value as a common reference in place of the original value to estimate the relative errors. This takes the ML reconstruction as if it were the original picture. However, the computational cost is expected to be high and there is no guarantee that such a common reference is indeed close to the original and allows a more accurate judgment. The use of a conservative estimation is expected to result in smaller estimated ΔMSE values than the true values, but offers a more robust estimate in determining which picture is better as well as a more meaningful numerical indication of the difference in quality, as verified in our experimental results later.

The above discussions assume the input pictures are compressed using the same format. Note that QCS is defined in the specific transform domain used in the compression format and cannot be readily represented in any other domain. However, the reconstructed signals can be transformed from one domain to another. When the pictures are in different formats, the $\hat{X}$ s are to be obtained by transforming the decoded spatial image $\hat{x}$ into the target domain (e.g., transforming $\hat{x}_A$ into the domain of X_B). While JPEG [30] and MPEG-2 [33] use DCT, which is energy-conservative, JPEG2000 [35] uses biorthogonal discrete wavelet transform (DWT), which is approximately energy-conservative [64, 65]. The integer transform used in H.264 [32] also approximates this property.

Because the frequency correspondence is lost between different transform domains, $E_{A,B}(u, v)$ is undefined when inputs are of different formats. The approximated ΔMSE could be alternatively calculated in the spatial domain as

$$\begin{aligned}
 E_{A,B}(m, n) &\approx \Delta x_{A,B}^2(m, n) - \Delta x_{B,A}^2(m, n) \\
 \Delta\text{MSE}_{A,B} &= \frac{1}{MN} \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} E_{A,B}(m, n),
 \end{aligned} \tag{4.8}$$

where M and N are the number of pixels in the row and column direction, respectively, and $\Delta x_{A,B}$ and $\Delta x_{B,A}$ are the inverse transforms of $\Delta X_{A,B}$ and $\Delta X_{B,A}$, respectively. Of course, ΔMSE could also be calculated by summing ΔX^2 in each domain first thus skipping the $E(u, v)$ step. However, doing it in the spatial domain has an added advantage, as discussed below.

Different compression formats and different quantization schemes lead to different artifacts and the output quality is dependent on the image content. Therefore, a global quality measures may not reflect the quality of local image regions. For example, some compression schemes may preserve details better in textured regions, while introducing artifacts in smooth regions. Some other schemes may produce compressed pictures that appear smooth but lack of details. By calculating the quality differences in the spatial domain, we are able to estimate the relative quality among spatial regions in addition to that of the whole pictures. Instead of averaging over the entire picture as in (4.8), we can average over different regions of interest and they can be in any shape. This capability gives us the added flexibility in judging the quality among compressed pictures.

The proposed method is summarized as follows for multiple compressed still images:

1. Calculate the QCS and the decoded image $\hat{x}_i$ for each copy in its own domain;
2. For each pair: Copy i and Copy j , $i \neq j$:
 - (a) Transform $\hat{x}_i$ into the domain of X_j .
 - (b) Calculate the out-of-bound error $\Delta X_{i,j}$ using (4.5).
 - (c) Transform $\hat{x}_j$ into the domain of X_i .
 - (d) Calculate the out-of-bound error $\Delta X_{j,i}$ using (4.5).
 - (e) Inverse transform $\Delta X_{i,j}$ and $\Delta X_{j,i}$ into the spatial domain.
 - (f) Calculate the approximated ΔMSE using (4.8).

3. Rank all images according to their pairwise relative quality.

For video sequences, the same procedure applies to each set of corresponding frames.

4.3 Experimental Results

In this section, we apply the proposed method to pairs of compressed images in both JPEG and JPEG2000 formats to judge their relative quality and compare the experimental results with those obtained by directly applying the NR PSNR estimation methods in [7–9] as well as our method proposed in Chapter 3. The paired images are compressed such that they have similar PSNR values to reflect the most demanding scenario. Our test images include Lena, Barbara, Mandrill, and Tiffany. Only luminance images are used in the experiments and the results can be extended to color images as the chrominance channels are compressed independent of the luminance channel using the same techniques. Visual examples are also included for region-based quality assessment.

FIGURE 4.5: Test images used, from left: Lena, Barbara, Mandrill and Tiffany

4.3.1 Standard vs. Alternative JPEG Quantization Matrices

Since the IJG quantization matrices offer finer steps of quality setting, we use one image compressed using Photoshop at quality level 3 (Copy A) and another image compressed using the IJG matrix (Copy B). The compression quality setting is varied for Copy B. In the figures, the estimated and true $\Delta\text{MSE}_{A,B}$ are plotted against the MSE of Copy B. $\Delta\text{MSE}_{A,B} = 0$ means both copies are the same in terms of MSE, while a positive difference indicates that Copy B is better than Copy A, and vice versa. Naturally, the closer the estimated $\Delta\text{MSE}_{A,B}$ curve resembles the true one, the better the estimation is. More importantly, since we are trying to judge which copy is better, the estimated $\Delta\text{MSE}_{A,B}$ should cross zero close to where the true $\Delta\text{MSE}_{A,B}$ does.

We first present in Fig. 4.6 the results obtained by applying the methods described in [7–9], referred to as Method 1, 2, and 3, respectively and in order, to individual test images. Ideally, the estimated curve should coincide with the true curve. For binary relative quality assessment (which is better), the estimated curve should have a positive slope and cross $\Delta\text{MSE}_{A,B}$ close to 0. As can be seen from the figure, Methods 1 and 3 consistently judge Copy B having better quality than Copy A, except for the Mandrill image where Method 3 performs well, while Method 2 does the opposite. The problem with these methods is with their inability to correctly estimate the MSE of the Photoshop-compressed image. As a result, none of these methods is able to judge correctly which image has a better quality when their qualities are close. Both Methods 1 and 2 implicitly assume certain DCT quantization matrix design, upon which their estimation for the probability density function is based. Because quantization distorts the estimation of the Laplace parameters, these methods attempt to compensate for such effects. Consequently, the images quantized using IJG matrix usually have a smaller estimation error, while those compressed using Photoshop tend to have a

FIGURE 4.6: Difference in MSE estimated with four different methods

FIGURE 4.7: Difference in MSE estimated by the newly proposed method, JPEG

larger error. Method 4 is the shifted-block-based method we propose in Chapter 3, whose neural networks are trained using the LIVE database images [60] and do not involve the use of the test images. This method shows less error than the reference methods but still may fail to judge the relative quality reliably when the PSNR values are very close.

The results from our method proposed in this chapter are shown in Fig. 4.7. It is evident that the estimation error, ranging between -0.7 and 0.7, is much smaller compared with those obtained by directly applying the existing methods [7–9] and the one we propose in Chapter 3. Also, the points where the estimated curves cross 0 are very close to those of the true ΔMSE , i.e., within 2.2 in MSE (about

0.5 dB in PSNR in the current context). The method is able to correctly judge the relative quality among JPEG images using different quantization matrix designs even when their MSE values are close.

The slopes of the estimated curve are less steep than those of the true values. This is due to the conservative nature of the proposed method. However, since the estimated curves have the same upward trend as the true value curves, images with even larger quality difference can also be correctly judged by this method.

4.3.2 JPEG vs. JPEG2000

This part of the experiment demonstrates the capability of the proposed method when the images are compressed in mixed formats. Again, two copies of compressed images are used. Copy A is a JPEG2000 image with 5 decomposition levels and a uniform quantization factor of 24, while Copy B is a JPEG image compressed with varying quality settings. The estimated and true Δ MSE values are shown in Fig. 4.8, where the x-axis shows the true MSE values of Copy B. The distances between the estimated and the true zero-crossing are less than 5.0 in MSE. As with the previous experiment, the conservatively estimated Δ MSE has a gentler slope than the true value, but it nevertheless follows a consistent upward trend. Our method is capable of assessing the relative quality between pictures compressed in different formats. No direct comparison can be made with existing NR methods, because they only work in bDCT domain.

4.3.3 Region-based Image Quality Comparison

In this part of the experiment, we demonstrate the region-based quality assessment by calculating the average of the estimated squared difference in non-overlapping

FIGURE 4.8: Difference in MSE estimated using pairs of JPEG and JPEG 2000 images

8-by-8 blocks. We compare the estimated quality difference against the ground truth and present visual examples.

In Fig. 4.9, we show the visual maps of the 8-by-8 block-based relative quality measures. The input image copies are compressed from the Barbara image. Copy A in both Fig. 4.9(a) and (b) uses the standard JPEG quantization matrix at quality level 78 and has a PSNR of 36.44 dB; Copy B in Fig. 4.9(a) uses the aforementioned Photoshop quantization matrix and has a PSNR of 36.46 dB; Copy B in Fig. 4.9(b) uses JPEG2000 with a uniform quantization step-size of 17 and has a PSNR of 36.56 dB. Middle gray means the two copies are of the same quality. A brighter tone means Copy A is better than Copy B in that area, and a darker tone means Copy A is worse than Copy B. As can be seen from the figure, the bright and dark areas of the estimated map generally agree with those in the ground truth map, indicating that the proposed method is able to decide which image area is of better quality. The contrast is lower in the estimated map due to the conservative nature of our proposed measure.

Fig. 4.10 shows two 64-by-64-pixel areas cropped from the estimated quality map in Fig. 4.9(a) and the reconstructed images. The quality map has been thresholded at 128 (middle gray) for easier visual inspection: white means Copy A is better and black means Copy B is better. Areas with perceivable visual

(a) JPEG IJG vs. JPEG Photoshop

(b) JPEG vs. JPEG2000

FIGURE 4.9: Regional quality maps by 8-by-8 blocks. Left: estimated, right: ground truth

differences have been highlighted with circles. It can be seen that Copy A has less artifacts in smooth areas, while Copy B retains more subtle texture details. This can be explained by the larger quantization step-sizes at low frequencies and smaller step-sizes at high frequencies used in Copy B compared with Copy A. Similar observations have also been made on image pairs consisting of one JPEG image and one JPEG2000 image. Because the overall quality are close for each pair of images, such quality difference may only be obvious in certain image areas.

FIGURE 4.10: Visual examples. Left to right: thresholded quality maps, Copy A, Copy B.

4.3.4 The Effect of Compression Ratio

In the experiments presented in Section 4.3.1, the Copy B images are compressed using Adobe Photoshop with JPEG quality set to Level 3. Here in Fig. 4.11 we present the results based on various other quality settings using more image samples. Our test image set consists of 24 768-by-512 images from the Kodak Photo CD [66], and only the luminance components of the images are compared. The Copy B images are compressed at Photoshop quality levels 0, 1, 2, 3, and 6, respectively. The zero-crossing points are linearly interpolated using two estimated ΔMSE points, one above and the other below zero, obtained by using different IJG quality levels for Copy A. Ideally, the zero-crossings should coincide with the MSE (or PSNR) of Copy B. It can be seen from Fig. 4.11 that the zero-crossings are very close to the true PSNR of Copy B. The overall correlation coefficient is $r = 0.9932$. With moderate to good quality images, the error is less than 0.7 dB while with low quality images the error may approach 1.4 dB. The estimation error

FIGURE 4.11: Estimated zero-crossings at various compression levels

is larger with highly compressed images because their QCSs are larger. Since our method calculates only the error outside such constraint sets, a larger set implies that a larger error within the set is likely unaccounted for. From the experimental results based on pairs of compressed images we can see that the proposed method still has a small error when judging images that are very close in quality. When more than two compressed images are present, there may be situations where some of the pairs are judged incorrectly, resulting in pairwise: $A > B$, $B > C$, and $C > A$ hence no winner can be picked out among these images. Under the proposed scheme, the solution to this problem requires more accurate assessment of individual image pairs.

4.3.5 Other Possible Measures

FIGURE 4.12: Estimated difference in MSE: using near bound and far bound

In the Section 4.2, we mentioned a few possible ways to estimate the relative error other than using the proposed near bound of the QCS. Unfortunately, the ML-based method is not readily applicable. Also, the reconstruction levels themselves alone do not give any information about relative quality among compressed images. Here we show the experimental results based on the far-bound measure relative to the proposed near-bound measure.

In Fig. 4.12, the estimation results based on the distance from the near bound and far bound are shown. The compressed images used are the same as those in Fig. 4.8. When using the far bound, the estimated curves of ΔMSE generally have a positive slope and are much steeper than those of the true values. Moreover,

the curves do not cross zero where the true values do. Instead, it consistently underestimates ΔMSE near the actual zero-crossing point. Therefore, it cannot be used to estimate the relative quality of compressed images.

4.4 Conclusion

We have presented in this chapter a method to judge the relative quality among pictures compressed from the same original picture. The method is automatic and the measure is objective. Based on the QCSs defined by these compressed copies, we proposed a conservative estimate for the difference in mean squared error of the quantized coefficients as an indication of the relative qualities. The proposed measure correlates well with the difference in MSE and the method is able to identify the better image out of a pair less than 0.7 dB apart in PSNR under nontrivial scenarios. The estimated difference at zero-crossing is typically within $\pm 1.5 \Delta\text{MSE}$ and compares favorably against directly applying methods that estimate PSNR image by image, including the one we proposed earlier in Chapter 3.

Chapter 5

Reconstruction from Multiple Compressed Copies

In the previous two chapters, we have proposed methods that allow the end user to estimate the absolute and relative compression quality of images and videos without referring to the original pictures. By applying the methods, the best compressed picture copy may be identified.

Since each compressed copy contain unique pieces of information, namely the quantization constraint sets (QCS), there is more to be done than simply picking the best and forgetting the rest. Knowing that the true value must lie within the QCS's, the intersection of all the available sets will provide a narrowed constraint set that has a lower quantization noise variance, leading to a picture more closely resembling the original.

Practically, not all compression formats utilize the same transform of the same block size, and the narrowed constraint set may not take an explicit form. Therefore, we propose to project the reconstructed pictures onto individual QCS and impose an additional smoothness constraint set, so that the final output has reduced quantization artifacts and approaches the quality of the original. We also

take into consideration of resized pictures, because it is common for users to resize the image to suit the layout of a particular web page and for a video sharing website to provide a video sequence at reduced resolutions. The use of constrained optimization guarantees the results do not deviates further away from the available information and the output quality is at least as good as the best input. The method of POCS is particularly suitable for this task because the constraints are naturally nonlinear but closed and convex. The algorithm makes no other specific assumption about the nature of the original sequence, and consistently produces output with improved quality over all existing copies.

In Chapter 2, we reviewed multiple-description coding (MDC). The key difference between MDC and the reconstruction using multiple Internet videos is that in MDC, compressed versions are deliberately encoded to have complimentary information, while the various copies obtained from the Internet were compressed by different individuals with no consideration on their joint reconstruction. In addition, the joint decoding of multiple MDC versions is well-defined in the codec, while the joint reconstruction of Internet videos has many practical issues to address and lacks off-the-shelf solutions. This is simply because these similar copies were not created with the intention to be jointly used at all.

The rest of the chapter is organized as follows. Section 1 details the picture reconstruction constraints and our solution. Section 2 follows up with experimental results to demonstrate the effectiveness of the proposed method. The chapter concludes in Section 3.

5.1 Reconstruction Using Constraints

5.1.1 Quantization Constraints

Following earlier discussion, we know that for a certain compressed copy (denoted as i in the subscript), the true coefficient value $X_i(u, v)$ lies within a well-defined range, namely the quantization constraint set (QCS)

$$\begin{aligned} X_i(u, v) &\in S_{X_i}(u, v) \\ S_{X_i}(u, v) &= \{X_i(u, v) : X_i^{\text{lower}}(u, v) \leq X_i(u, v) \leq X_i^{\text{upper}}(u, v)\}. \end{aligned} \quad (5.1)$$

For JPEG [30] and MPEG-2 [33] compressed pictures, the above can be simply written as

$$S_{X_i}(u, v) = \{X_i(u, v) : |X_i(u, v) - \hat{X}_i(u, v)| \leq \epsilon_i(u, v)\} \quad (5.2)$$

where $\epsilon_i(u, v) = \frac{Q_i(u, v)}{2}$ is half the quantization step-size, and Q_i and $\hat{X}_i$ are the quantization matrix and representative value for the corresponding block in the i th copy. Suppose that there are I copies available, a series of similar constraint sets $S_{X_1}(u, v) \dots S_{X_I}(u, v)$ can be obtained in the same manner, and the true value $X(u, v)$ must lie within the intersection of all such sets, which is generally smaller than the individual constraint sets owing to the differences in quantization step-sizes as well as motion estimation and compensation. Formally,

$$X(u, v) \in S_X(u, v) = \bigcap_{i=1}^I S_{X_i}(u, v) \quad (5.3)$$

We shall refer to set S_X as the narrowed quantization constraint set (NQCS) in the following text. It should be noted that without any other specific prior knowledge, the reduced size of S_X guarantees smaller quantization noise variance compared

with that from a single copy. Such a reduction in noise variance depends on the number of copies available, as well as the way they are encoded.

It is trivial if two copies are compressed with the same encoder and the same parameter set. The output will be identical and no gain of quality should be expected.

For image and video copies in the same format, however, if the quantization parameters are different, due to different quality settings, the QCSs will be different among the copies. Furthermore, for video sequences, different group-of-picture (GOP) structures lead to different predicted pictures (predictions) for the same original frame, and in turn define different shifted QCSs even if the quantization step-sizes are identical. Hence, a single picture from the original sequence may be coded in different modes in different copies. In addition, encoder implementation varies from one to another, and this also adds to the variation in the QCSs. A graphical illustration is shown in Fig. 5.1, where in case (a) the two video copies have different quantization step-sizes; (b) the predictions are different; and (c) both prediction levels and quantization step-sizes are different. All three scenarios offer some possible reduction in quantization noise variance. The same principle in case (a) is also applicable to intra-frames and still images by setting the predictions to zero.

When multiple copies are compressed in different formats, the information carried by each copy is even more different from the others, and offers further possibility of enhancement.

When the inputs are of different frame dimensions, regardless of format, it is usually the case that the larger pictures are also of high quality compared with the smaller ones enlarged to the same size. However, it is also possible that the larger ones are more heavily compressed and the smaller ones may contain more valid information especially at lower frequencies.

FIGURE 5.1: Illustration of possible relations among the true value, QCS, and NQCS (intersection) for video copies of the same format and frame dimensions

Note that the NQCS can be obtained explicitly by (5.7) only when all input copies share the same picture dimensions, block size, and the same transform function, e.g., 8-by-8 and DCT. Otherwise, there is no explicit expression for the NQCS, since the QCSs are defined only in their respective domains, e.g., 4-by-4 integer transform for H.264 [32], which is different from 8-by-8 DCT for MPEG-2.

5.1.2 Smoothness Constraint

In addition to the quantization constraints, we also impose an ellipsoidal smoothness constraint S_C on the reconstructed picture, knowing that natural images are generally continuous and smooth, while compression could introduce discontinuities. This is a rather simplistic constraint for natural images and the reason we chose it was mainly to demonstrate that prior knowledge can be used in our reconstruction process with little difficulty and their incorporation can help to enhance the visual quality.

We shall refer to this constraint as the smoothness constraint set (SCS) hereafter. Let x be the lexicographically ordered representation of a picture in the

spatial domain, the SCS can be expressed as

$$x \in S_C = \{x : \|Dx\|^2 \leq \xi\} \quad (5.4)$$

where D is the matrix representation of a high-pass filter measuring the lack of smoothness in the picture, and ξ is a positive scalar depending on the original picture. However, in practical application, the original picture is not available and the value of ξ cannot be obtained reliably. An alternative way to incorporate the smoothness assumption is to solve for x which

$$\begin{aligned} &\text{minimizes } \|Dx\|^2 \\ &\text{subject to } \|X_i - \hat{X}_i\|^2 \leq \epsilon_i^2 \end{aligned} \quad (5.5)$$

where $\|\bullet\|$ measures the Euclidean distance. Using Lagrange multipliers, the solution to (5.5) together with the strict quantization constraint can be obtained by minimizing

$$\begin{aligned} \Phi(x, \lambda) &= \sum_{i=1}^I \left(\|X - \hat{X}_i\|^2 - \epsilon_i^2 \right) + \lambda \|Dx\|^2 \\ &\text{subject to } X \in S_X \end{aligned} \quad (5.6)$$

The Lagrange multiplier λ controls the tradeoff between smoothness and the conformity to the NQCS. The selection of λ has been discussed in detail in [55] for multichannel image restoration and is not reiterated here.

The solution to (5.5) can be obtained using maximum a posterior (MAP) estimation in the spatial domain, and thus applicable to picture copies coded with different standards. However, the method is not able to incorporate the strict condition in (5.6) and may produce results outside the NQCS. In fact, the spatial-domain condition does not carry all the information in the transform domain. In the transform domain, every coefficient has its upper and lower bounds clearly

defined. In contrast, the spatial domain condition only specify that the total difference should not exceed a certain limit. Obviously the latter is much less restrictive compared with the former. In our proposed method, we adopt the method of projection onto convex sets (POCS) to obtain an enhanced picture, as detailed below.

5.1.3 Projection onto Convex Sets

Given several picture copies encoded from the same original source, for each transform coefficient in the i th copy, the upper and lower bounds of its QCS can be obtained as discussed in Section 2.1.4. If all copies share the same dimensions, block size and transform function, the bounds of the NQCS can be explicitly obtained as

$$\begin{aligned} X^{\text{upper}}(u, v) &= \min_i X_i^{\text{upper}}(u, v) \\ X^{\text{lower}}(u, v) &= \max_i X_i^{\text{lower}}(u, v) \end{aligned} \tag{5.7}$$

As mentioned earlier, images and videos coded using different block sizes and transform functions have their QCSs defined in different domains and cannot be directly combined to form the NQCS. Using POCS, the constrained optimization is implemented as an iterative process. In each cycle of the iteration, the reconstructed picture is projected onto all QCSs successively, as well as onto the SCS. In this way, pictures of different sizes or compressed using different formats can be incorporated in the reconstruction.

5.1.3.1 Initial Estimation

Our empirical results suggest that the initial picture estimation has substantial impact on the final output quality. This effect cannot be modeled as a linear process

since POCS itself is nonlinear and depends on the picture content. Graphically, the effect is illustrated in Fig. 5.2. It can be seen that initial estimation “2” results in a final reconstruction closer to the true picture value, even though both final outputs satisfy both the QCS and SCS.

When the NQCS can be directly computed, the initial picture estimation can be obtained by taking the point that has the highest probability inside the NQCS, so that the initial estimate has the smallest noise variance

$$\hat{X}^{init} = \arg \min_{\tilde{X}} \int \left(X - \tilde{X} \right)^2 p_{u,v}(X) dX \quad (5.8)$$

The probability density function $p_{u,v}$ employed is either Laplace distribution, which better describes the distribution; or uniform, which greatly simplifies computation at the expense of accuracy. However, using the former would require the knowledge of the Laplace parameters. It is possible to employ the technique proposed in Chapter 3 to estimate them from the compressed copies, but it is unnecessarily complicated for initial estimates. Using the simplified assumption that the transform coefficients follow uniform distribution within the QCS, the solution to (5.8) lies at the center of the NQCS. For each DCT coefficient,

$$\hat{X}^{init}(u, v) = \frac{X_i^{\text{upper}}(u, v) + X_i^{\text{lower}}(u, v)}{2} \quad (5.9)$$

In more general cases where the NQCS cannot be directly computed due to differences in picture size or format, the initial picture can be constructed in the spatial domain using decoded pictures from all copies. As noted earlier, the MAP estimation does not necessarily satisfy the QCSs but it can nevertheless serve as a good initial estimation. To simplify the computation, we remove the smoothness prior from MAP and leave it to POCS, effectively reducing it to a maximum likelihood (ML) estimator. The quantization errors have zero mean and finite variances in the DCT domain for each coefficient, and inverse DCT sums this

FIGURE 5.2: Effect of initial estimation in POCS.

FIGURE 5.3: Quantization noise in the spatial domain

error for every pixel within each block, scaled by the cosine functions. Therefore, the error of a pixel is the sum of many independent random variables of uniform distributions with zero mean and scaled variances. If the variances were the same, these random variables would be identically distributed and their sum, by central limit theorem, would be Gaussian distributed. With different variances, the sum approximates Gaussian distribution. Fig. 5.3 shows such an approximation, using an image with uniform noise as content, compressed at JPEG quality 50, decoded, and its error measured against the original in the spatial domain. It can be seen that although not an exact fit, the Gaussian model describes the quantization noise in the spatial domain very well.

Assuming the quantization noise is independent among all compressed copies in the spatial domain¹, the likelihood function can be expressed as the product those of individual pictures:

$$p(\hat{x}_1, \hat{x}_2, \dots, \hat{x}_I | x) \propto \exp \left[\sum_{i=1}^I (\hat{x}_i - x)^T \mathbf{K}_i^{-1} (\hat{x}_i - x) \right] \quad (5.10)$$

where x and $\hat{x}$ are vector representations of the true and decoded pictures in the spatial domain; $\mathbf{K}_i$ is the covariance matrix between x and $\hat{x}$, obtained by performing inverse transform to the variance matrix, $\mathbf{K}_{i,T}$, in the frequency domain, i.e.,

$$\mathbf{K}_{i,T} = \begin{bmatrix} \sigma_1^2 & & & \\ & \sigma_2^2 & & \\ & & \ddots & \\ & & & \sigma_K^2 \end{bmatrix} \quad (5.11)$$

$$\mathbf{K}_i = \mathbf{C}^T \mathbf{K}_{i,T} \mathbf{C} \quad (5.12)$$

where σ_m , $m \in [1, K]$ is the noise variance of the m th coefficient in lexicographical order, and $\mathbf{C}$ is the 2D-DCT transform matrix corresponding to the lexicographical representation. Using the uniform distribution assumption, the variance is calculated with quantization step-size $Q(m)$ as

$$\sigma_k^2 = \frac{Q^2(k)}{12} \quad (5.13)$$

Maximizing (5.10) yields the maximum likelihood estimation for the initial picture

$$\hat{x}_{ML} = \arg \max_x \exp \left[\sum_{i=1}^I (\hat{x}_i - x)^T \mathbf{K}_i^{-1} (\hat{x}_i - x) \right] \quad (5.14)$$

¹This assumption is valid if the copies use different transforms, and is a close approximation if they use the same transform but different predictions among video sequences.

Since the quantization matrices are usually scaled versions of a predefined base matrix in JPEG and uniform for video residues, a computationally simpler approximation can be obtained as

$$\hat{x}_{ML} \approx \frac{\sum_{i=1}^I \frac{1}{q_i^2} \cdot \hat{x}_i}{\sum_{i=1}^I \frac{1}{q_i^2}} \quad (5.15)$$

for pictures compressed in the same format, where q_i is the scaling factor applied to the base quantization matrix or the quantization step-size for the residue signals. In either case, the initial estimate is computed using all available copies and makes use of the quantization information, and the output is improved over the input copies. However, the ML estimation does not enforce the QCSs rigorously and the initial picture may not satisfy the NQCS conditions.

When the input pictures have different dimensions, it is conceivable that the reconstructed picture should be the same size as the biggest of them all, as smaller pictures can carry less details. When it comes to the initial estimate, it should be noted that a picture does not carry valid information beyond its native resolution. Therefore, when merging the image content of all inputs, it is necessary to use high-frequency contents from the larger pictures, while taking a weighted sum of the low-frequency contents as described earlier. For example, when two image copies in the same compression format are available, and Copy A is larger in dimensions than B, the ML-like estimator is

$$\hat{x}_{ML} \approx \frac{\frac{1}{q_A^2} \cdot LPF[\hat{x}_A] + \frac{1}{q_B^2} \cdot LPF[\hat{x}'_B]}{\frac{1}{q_A^2} + \frac{1}{q_B^2}} + HPF[\hat{x}_A], \quad (5.16)$$

where the LPF and HPF are a pair of complimentary low-pass and high-pass filters, and $\hat{x}'_B$ is the enlarged version of $\hat{x}_B$. The normalized cut-off frequency of the low-pass filter is M_B/M_A , which is the ratio of the dimensions between Copies B and A. The transition of such filters need not be very sharp because this is after all an initial estimate, and computation should be kept simple.

When the input pictures are compressed in different formats, the quantization step-sizes are not directly comparable. In such cases, simply averaging the pixel values often give satisfactory results when the objective quality are similar, because the compression noise caused by different compression techniques are usually independent. The NR PSNR estimation method proposed in Chapter 3 and the relative quality assessment method in Chapter 4 can also be used if applicable.

5.1.3.2 Projection Proper

We use the picture estimated above as the starting point of the POCS process. In each cycle, the SCS is first applied to reduce compression artifacts. The smoothing is performed based on (5.17) across the whole picture so that all the artifacts can be addressed, regardless of how close to or far from the block boundaries they are. The Laplacian kernel is used for the high-pass filter D , and the iteration index is shown as superscript. For the a th iteration, the smoothing process can be expressed as

$$\hat{x}_1^a = \hat{x}^a - \lambda^a D^T D \hat{x}^a, \quad (5.17)$$

where the subscript 1 indicates it is to be projected onto the 1st QCS subsequently. The positive parameter λ^a is gradually reduced as the iteration goes on, so that the legitimate high-frequency components emerged through POCS are not removed.

After applying the SCS, the projection onto the NQCS is performed. When the input copies have different dimensions, block sizes or transform functions, the NQCS is not directly available. The reconstructed picture is thus resized, partitioned and transformed according to each individual input copy, and compared against its QCS. The out-of-bound coefficient values are forced to the nearest bound in the spatial domain. The process is repeated for all input copies. The error in the corresponding block with respect to the QCS in the i th copy is obtained

as

$$\Delta X_i^a(u, v) = \begin{cases} \hat{X}_i^a(u, v) - X_i^{\text{upper}}(u, v) & \text{if } \hat{X}_i^a(u, v) > X_i^{\text{upper}}(u, v) \\ \hat{X}_i^a(u, v) - X_i^{\text{lower}}(u, v) & \text{if } \hat{X}_i^a(u, v) < X_i^{\text{lower}}(u, v) \\ 0 & \text{otherwise} \end{cases} \quad (5.18)$$

and the correction to the estimated picture is performed as

$$\hat{x}_{i+1}^a = \hat{x}_i^a - \Delta x_i'^a(m, n), \quad (5.19)$$

where $\Delta x_i'^a$ is the spatial domain representation of ΔX_i^a , after resizing if necessary. The aforementioned reduction of λ^a guarantees that the iteration comes to a stop within finite cycles.

To summarize, our proposed algorithm performs the following steps:

Step 1 Compute the QCSs.

Step 2 Obtain an initial reconstruction using ML estimation in the spatial domain with (5.15), (5.16), or (5.7) if all input copies are of the same format and size.

Step 3 Project the reconstructed picture onto SCS and correct it using (5.17).

Step 4 For each input copy:

- (a) Resize the reconstructed picture to the same size as the current copy.
- (b) Partition the reconstructed picture into blocks and transform them into the domain according to the current copy.
- (c) Project the transform coefficients onto the QCS of the current copy and calculate the errors as in (5.18).

FIGURE 5.4: Example block diagram of the proposed algorithm with two copies of different sizes and formats.

- (d) Transform the errors into spatial domain and resize to the same as the reconstructed picture.
- (e) Correct the error in the spatial domain using (5.19).

Step 5 Repeat Step 3 and 4 until the result converges.

The process normally converges in 2 or 3 iterations and the overall computational cost is similar to other common POCS-based algorithms. Fig. 5.4 gives an example block diagram of the proposed algorithm when using two copies in different formats as inputs.

5.2 Experimental Results

In this section, we demonstrate the effectiveness of the proposed method on both still images and videos. Only the luminance channel is used and presented for easier visual evaluation, and the method is directly applicable to the chrominance channels. Besides two input copies of identical picture dimensions and compression format, we also take JPEG images of different sizes, and videos in mixed formats as inputs. The effects of GOP structure as well as the number of copies on output quality are also investigated using video sequences.

5.2.1 Still Images

In the following, simulation results are shown with four natural images: Lena, Barbara, Mandrill and Tiffany. Pairs of compressed images are used in the experiment. All images are compressed in the JPEG [30] format. The first part of the experiments deals with images of the same dimensions, and the second deals with those of different dimensions.

5.2.1.1 Images of the Same Size and Format

We use images of the same pixel dimensions and format to demonstrate the basic case of reconstruction with multiple compressed copies. Three pairs of test samples are used for each image: The first pair consists of two images compressed in very different quality settings – specifically, quality settings 20 and 70 using the IJG quantization matrices; the second pair has two images at quality settings 45 and 50 respectively; the last pair has one image compressed using an alternative table (Adobe Photoshop quality=3) and the other using the IJG quantization matrix, which is set so that both have similar PSNR measures. These quantization matrices are the same set we have shown earlier in Section 4.1. The final outputs

are compared with the initial estimates and the pixel-averaging method for noise reduction, where the corresponding pixels in the two images are averaged to produce the output. Table 5.1 shows the PSNR figures from the simulation. Column “average” lists the results of denoising by averaging; “initial” is the initial estimation and “final” is the final output. ΔPSNR is the difference in PSNR between “final” and the better of the two inputs calculated against the original.

From Table 5.1, it can be observed that simple averaging is able to achieve some gain in PSNR when the two copies are very close in quality, as shown in the 2nd and 3rd row for each image. However, when the two copies have relatively large difference in compression quality (1st row for each image), the averaged image has a lower PSNR value than the better of the two inputs. When the two images are of similar quality, because the quantization noise can be either positive or negative for any pixel, they tend to partially cancel out and reduce the noise. When they are of very different quality, the poorer one has such large errors compared to the better one, and the errors cannot cancel out. The resulting picture therefore has quality in between of the two inputs. In contrast, the proposed method consistently delivers higher PSNR measures than any of the input images, even when their qualities differ greatly. It does not count on the noise to cancel out, but use both of them to narrow down the range in which the true value must exist.

With the proposed method, the highest gain in PSNR over the better copy is obtained when the two input images are of similar PSNR but use different quantization matrices (the third row for each image in Table 5.1), where the improvement can be as high as 5 dB. This is due to the difference in quantization matrix used - the alternative table has all the mid- and high-frequency quantization step-sizes identical, while the IJG table uses larger step-sizes towards higher frequency. The smallest gain is seen where the two input images are very different in quality. When one copy has much smaller quantization step-sizes than the other, the intersection of the sets is often simply the smaller set. Hence the reduction of quantization

TABLE 5.1: PSNR comparison, JPEG image in the same size (dB)

image	input copies		average	proposed		Δ PSNR
	A	B		initial	final	
Lena	33.0	37.3	36.4	37.4	37.8	0.5
	35.5	35.8	36.1	36.2	36.3	0.5
	35.9	35.8	37.9	38.4	38.4	2.5
Barbara	28.3	35.0	32.9	35.1	35.4	0.4
	32.0	32.5	32.9	33.1	33.1	0.6
	36.5	36.4	38.9	40.3	40.3	3.8
Mandrill	26.2	31.1	30.1	31.1	31.7	0.6
	28.6	29.0	29.2	29.6	29.6	0.6
	34.5	34.3	37.3	39.7	39.7	5.2
Tiffany	33.5	37.5	36.8	37.6	37.9	0.4
	35.7	36.0	36.3	36.5	36.5	0.5
	37.9	38.0	38.0	40.9	40.9	2.9

noise variance is insignificant. The PSNR difference between the initial estimate and the final output is very small. This is because the initial estimate is already accurate with respect to the quantization constraints, and the marginal gain comes only from the smoothness constraint. As we shall see in the next subsection, when the images are of different sizes, the initial estimate is no longer guaranteed to fall inside the intersected quantization constraint set.

Figure 5.5 shows the PSNR from both the proposed method and simple averaging across a large range of quality differences. Two copies of the Lena images were used: Copy A fixed at quality setting of 50 and Copy B with varying quality setting from 5 to 95 with an increment of 5. This further demonstrates the limitation of the averaging method when the copies differ in quality, contrasting the robustness of the proposed method. Judging by the distance between the curve of the output and the better input, it appears that the highest gain is not obtained when the two copies are very close in PSNR but at quality settings 25 and 75 for Copy B. This does not contradict the results shown earlier in the third row for each picture in Table 5.1. In this experiment, the two quantization matrices are scaled versions of each other, and there the high and low frequency contents do

FIGURE 5.5: Comparison between proposed method and simple averaging

not complement each other between copies. Copy A was fixed at quality setting of 50. According to the IJG scaling method [63], quality of 25 leads to double the quantization step sizes of 50, and 50 is double of 75. Since JPEG quantization has no deadzone, every other quantization range of the better copy will have its center aligned with a range boundary of the worse copy. This is illustrated in Fig. 5.6. Such intersections (shaded areas) are half the size of even the better copy and lead to superior reconstruction quality.

FIGURE 5.6: Two quantization step sizes, one is twice the other

TABLE 5.2: PSNR comparison, JPEG image in different sizes (dB)

image	size of B (pixels)	input copies		proposed		Δ PSNR
		A	B	initial	final	
Lena	384	37.8	37.8	38.5	39.1	1.3
		36.3	37.8	38.0	38.6	0.8
		39.3	37.8	39.6	40.1	0.8
	256	33.8	33.9	34.3	35.3	1.4
		35.3	33.9	35.8	36.0	0.7
		32.4	33.9	34.2	34.8	0.9
Barbara	384	29.1	29.1	30.7	33.2	4.1
		27.5	29.1	29.9	32.2	3.1
		30.6	29.1	31.8	33.6	3.0
	256	25.4	25.3	25.7	27.4	2.0
		23.8	25.3	25.3	26.3	1.0
		26.7	25.3	27.5	28.2	1.5
Mandrill	384	28.3	28.2	29.2	30.3	2.0
		26.7	28.2	28.3	29.6	1.4
		29.7	28.2	30.5	31.4	1.7
	256	24.8	24.9	25.1	26.4	1.5
		23.4	24.9	24.9	25.8	0.9
		26.4	24.9	26.7	27.2	0.8
Tiffany	384	36.5	36.5	37.3	38.4	1.9
		34.9	36.5	36.7	37.8	1.2
		37.9	36.5	38.5	39.2	1.3
	256	32.5	32.5	33.4	34.8	2.3
		31.1	32.5	33.0	34.2	1.7
		34.0	32.5	34.5	35.4	1.4

5.2.1.2 Images of Different Sizes

In this part of the simulation, we use pairs of images of different dimensions. Every pair consists of one image in its original dimensions – 512-by-512, while the other is resized to 384-by-384 or 256-by-256 before compression. The smaller images are compressed with JPEG quality set to 95, and the larger image compressed so that the two have similar PSNR measures when compared with the 512-by-512 original. All resizing are performed using the bicubic sampling algorithm.

The results are shown in Table 5.2. For each image, the first three rows show

Copy B resized to 384-by-384, and the next three rows show 256-by-256. The finely quantized smaller copies provide more accurate information than the larger copies in the mid- and low- frequency band, while the larger copies contribute to most of the high-frequency information in the final image. The first row in each set has a pair of images of very close PSNR. For images with less high-frequency content, such as Lena, the gain in PSNR is relatively small at around 1 dB; while images with moderate high-frequency content get about 2 dB improvement. The Barbara image is unique in that it has very strong and regular line patterns, which become significantly weakened during downsizing. In this case, the gain in PSNR can reach 3 to 4 dB, and the visual improvement is quite noticeable.

The second and third row of every three rows are pairs of samples with PSNR difference of about ± 1.5 dB. The gain in PSNR over the better copy is clearly reduced compared with cases where both copies are of nearly identical PSNR. As the difference in PSNR increases, there is less information to contribute by the low quality copy.

Fig. 5.7 visually demonstrates that the smoothness in the finely quantized low-resolution image and the details in the coarsely quantized high-resolution image can be combined into the final output. The larger copy suffers from severe compression artifacts near the boundaries between patterned and smooth regions and the pattern in some parts look unnaturally rough. Some of the line patterns become disconnected across block boundaries. On the other hand, the smaller copy look blurry and aliasing is visible in the original finely patterned area. The reconstructed image has sharp but natural textures and is mostly free of visible artifacts.

FIGURE 5.7: Barbara image, input copies and reconstruction using the proposed method. left: full-size coarsely quantized; right: $3/4$ -size finely quantized (enlarged for presentation); bottom: reconstruction

5.2.2 Video Sequences

Further experiments were conducted using CIF-sized (352×288 pixels) video sequences. Unless otherwise noted, three test sequences foreman, bus and crew were used. The crew sequence has relatively low contrast in the Y-component and the presence of camera flashes in the scene makes motion estimation difficult; the bus sequence contains lots of fine details as well as fast moving objects; the

part of the foreman sequence being used has large stationary background areas and changing foreground. Compression was done using the reference MPEG-2 [33] encoder except for the third experiment where H.264 [32] was also used.

The H.264 video sequences were encoded with the deblocking filter disabled. According to List *et al.* [16], the filter improves the visual quality of the compressed video at low bit-rates, while offering a 9% saving in bit rate. This filtering procedure does not affect the extraction of QCSs from the sequences and the gain in PSNR through the proposed reconstruction procedures is almost the same with the filter switched on or off. The size of GOP was chosen to be 12, which is commonly used for MPEG-2 compression. The number of input copies was two except for the last experiment, and the number of frames used was 24. The initial pictures were generated using the simplified ML estimator. The sequences were encoded in noninterlaced mode except in the third experiment. The Δ PSNR shown in the tables are measured with respect to the best input copy. Subjective quality assessment is also included in the form of cropped regions from pictures.

This subsection is further divided into four parts. Firstly, the effect of various group-of-picture (GOP) structure combinations is looked into; secondly, videos of different quality settings are tested; then, the algorithm is applied to videos of mixed picture modes and mixed coding standards; lastly, the effect of the number of input copies is examined.

5.2.2.1 Effect of GOP Structures

In this experiment, the input video sequences were compressed using different GOP structures. The effect of different combinations was compared based on the output video quality. In practice, the choice of GOP structure depends on the intended application and is usually determined by allowable coding delay, random access requirement, and bit budget. However, such variations result in a large

number of possible combinations and it is impossible to enumerate all of them. The settings were therefore reduced to three, including video pairs with identical and aligned GOP structures, identical GOP length but different number of B-frames, and identical GOP structures but different starting frames. The PSNR values from the “foreman” sequence are plotted in detail in Fig. 5.8.

Fig. 5.8(a) shows the result from two sequences with perfectly aligned I-, P- and B-pictures. The number of B-frames between adjacent I-/P-pictures is 2 (i.e., IBBPBBP...). The bit-rates of the two copies were set to be slightly different at 0.9 Mbps and 1.1 Mbps, respectively, so that the QCSs are different. Otherwise, the two copies would be identical in all aspects and no gain is achievable through the construction of the NQCS. It can be observed that the two input copies follow the same trend in the PSNR plots throughout the entire sequence, with the 1.1-Mbps copy having superior quality of about 1 dB over the 0.9-Mbps copy. The proposed algorithm is able to achieve about 0.5 dB gain in PSNR over the better copy. This is consistent with the results shown earlier on JPEG images of different PSNR values.

Fig. 5.8(b) shows the results from two sequences with different numbers of B-pictures between adjacent I-/P-pictures. Both sequences were encoded at 1 Mbps. The two sequences have similar PSNR figures, and the proposed algorithm achieves about 1.4 dB gain in PSNR for most pictures. However, a closer look reveals that pictures No. 1 and 13 have very little gain of around 0.2 dB only. The reason is that both of them are intra-coded frames and do not make use of inter-frame prediction. In this case, the NQCS only benefits from the difference in quantization step-sizes. Since both sequences were coded at 1 Mbps, such a difference comes from the rate control mechanism of the encoder and is unlikely to be significant considering the fact that the video content is identical.

Fig. 5.8(c) shows the results from two sequences with identical GOP structures but shifted starting frames. In the experiment, we have assumed that the frame

(a) Identical and aligned GOP

(b) Different number of B-pictures

(c) Shifted GOP

FIGURE 5.8: Effect of different GOP structures.

alignment is known. In practical scenarios, the frames can be aligned using video matching techniques available from content-based video analysis literature [67, 68]. The sequence with 0 starting shift has I-pictures at indices 1 and 13, and the one with 1 starting shift has I-pictures at indices 2 and 14. Three GOPs (36 pictures) were coded for each sequence and the corresponding 24 pictures were re-aligned and used for reconstruction. Again, both sequences were coded at 1 Mbps and their PSNR plots are very similar. Unlike in Fig. 5.8(b), the gain in PSNR from the proposed algorithm in Fig. 5.8(c) is always around 1.4 dB. In this setup, each pair of input pictures are of different types. They generally have different predictions and sometimes quantization parameters. Such differences likely produce narrower NQCS, and are the main reason why a relatively high gain is sustained through the sequence.

TABLE 5.3: PSNR (dB) results obtained from video copies with different GOP combinations. (Note: The number of B-pictures between adjacent I-/P-pictures is equal to M-1 in each copy.)

Video	Copy A		Copy B		Output PSNR	Δ PSNR
	Coding	PSNR	Coding	PSNR		
Foreman	1.1 Mbps, 0 Shift, M=3	36.96	0.9 Mbps, 0 Shift, M=3	35.95	37.48	0.52
	1.0 Mbps, 0 Shift, M=3	36.50	1.0 Mbps, 0 Shift, M=2	36.39	37.53	1.03
	1.0 Mbps, 0 Shift, M=3	36.50	1.0 Mbps, 1 Shift, M=3	36.57	37.90	1.33
Bus	1.1 Mbps, 0 Shift, M=3	30.07	0.9 Mbps, 0 Shift, M=3	28.92	30.57	0.50
	1.0 Mbps, 0 Shift, M=3	29.54	1.0 Mbps, 0 Shift, M=2	29.31	30.65	1.11
	1.0 Mbps, 0 Shift, M=3	29.54	1.0 Mbps, 1 Shift, M=3	29.55	31.06	1.51
Crew	1.1 Mbps, 0 Shift, M=3	37.64	0.9 Mbps, 0 Shift, M=3	36.53	38.01	0.37
	1.0 Mbps, 0 Shift, M=3	37.12	1.0 Mbps, 0 Shift, M=2	37.15	38.26	1.11
	1.0 Mbps, 0 Shift, M=3	37.12	1.0 Mbps, 1 Shift, M=3	36.93	38.49	1.37

Table 5.3 lists the average PSNR values obtained from the above experiments. At around 1 Mbps, the two-copy scenario suggests that shifting the GOP structure gives the highest output PSNR as well as the highest gain in PSNR over the better input copy. This combination guarantees that the pictures of different modes are not aligned and resembles real-world cases where different copies are coded using different parameters and may have different starting points. This configuration is used through the following experiments for its consistency. It can also be observed that when two input sequences have different quality settings (i.e. 1.1 Mbps and 0.9 Mbps in our case), the output from the proposed algorithm is only slightly better than the high-quality copy. This is further investigated in the next experiment.

5.2.2.2 Effect of Input Video Quality

In this part of the experiment, compressed video pairs of various bit-rates are tested, as well as pairs with mixed bit-rates. The results are shown in Table 5.4.

In Table 5.4(a), each pair has two copies coded at the same bit-rate, and the rate is varied across pairs. Bit-rates from 0.6 Mbps to 1.4 Mbps were used with a 0.2-Mbps increment. Copies A and B were coded with their starting points shifted by one picture apart. The proposed method is able to provide between 1 to 1.5 dB gain in PSNR, depending on the input PSNR. It should be noted that the gain values at different bit-rates are not to be directly compared. This is because when the base quality rises, the same difference in PSNR actually corresponds to a smaller visual quality improvement. Since it is beyond the scope of our discussion, we shall not delve into the quantitative analysis and comparison on the subjective quality.

In Table 5.4(b), the video copies in each pair have different bit-rate settings and thus quality measures. The Pel Avg column provides PSNR measures of

TABLE 5.4: PSNR (dB) results obtained at different quality settings

(a) Same bit-rate for the two input copies

Video	bit-rate (Mbps)	Copy A PSNR	Copy B PSNR	Output PSNR	Δ PSNR
Foreman	0.6	33.52	33.59	34.55	0.96
	0.8	35.31	35.39	36.53	1.14
	1.0	36.50	36.57	37.90	1.33
	1.2	37.40	37.45	38.85	1.40
	1.4	38.16	38.26	39.72	1.46
Bus	0.6	26.19	26.19	27.25	1.06
	0.8	28.18	28.21	29.53	1.32
	1.0	29.54	29.55	31.06	1.51
	1.2	30.60	30.58	32.22	1.62
	1.4	31.44	31.46	33.19	1.73
Crew	0.6	33.80	33.62	34.71	0.89
	0.8	35.85	35.77	37.11	1.26
	1.0	37.12	36.93	38.49	1.37
	1.2	38.06	37.83	39.51	1.45
	1.4	38.84	38.61	40.38	1.54

(b) Different bit-rates for the two input copies

Video	Copy A		Copy B		Pel Avg	ML Init	POCS	Δ PSNR
	Mbps	PSNR	Mbps	PSNR	PSNR	PSNR	PSNR	
Foreman	1.0	36.50	1.0	36.57	37.80	37.85	37.90	1.33
	0.8	35.31	1.2	37.45	37.47	38.00	38.02	0.57
	0.6	33.52	1.4	38.26	36.89	38.40	38.41	0.15
Bus	1.0	29.54	1.0	29.55	30.90	31.04	31.06	1.51
	0.8	28.18	1.2	30.58	30.84	31.17	31.20	0.62
	0.6	26.19	1.4	31.46	29.91	31.57	31.60	0.14
Crew	1.0	37.12	1.0	36.93	38.45	38.46	38.49	1.37
	0.8	35.85	1.2	37.83	38.23	38.51	38.54	0.71
	0.6	33.80	1.4	38.61	37.34	38.79	38.84	0.23

the output by pixel-averaging the two inputs. With the proposed method, the 0.6-Mbps and 1.4-Mbps pair offers the highest output PSNR, but the gain over the 1.4-Mbps copy is only about 0.2 dB. The highest gain is observed when two input copies have roughly the same qualities, with both coded at 1 Mbps. This observation corroborates with that in the still image experiment presented earlier.

FIGURE 5.9: Subjective comparison; top: cropped area; left to right: input, ML initial estimate, POCS output

When two input copies are of very different quality – in this case, over 5 dB difference in PSNR, the low quality copy simply does not carry any information that the high quality copy does not already have. The 0.2 dB gain in PSNR is due to the smoothing of blocking artifacts, which also offers better visual quality than the initial reconstruction even with the same PSNR measure, as can be seen in Fig. 5.9, where the final output retains all the structural details while containing less artifacts.

In the following experiments, video copies were encoded at similar quality settings, so that the gain in PSNR may be more consistently observed.

5.2.2.3 Videos in Different Formats

As discussed earlier, it is not trivial to obtain an explicit form of the NQCS from videos coded with different block sizes and transform functions. To demonstrate the capability of the proposed algorithm in dealing with such videos, four test pairs were prepared for each video sequence. The first pair has two MPEG-2 encoded videos. Copy A was coded in noninterlaced mode (frame-picture mode, 25 frames per second) and Copy B was coded in interlaced mode (field-picture mode, 50 fields per second). In the remaining pairs, Copy A was coded using H.264 with forced 4-by-4 integer transform, and Copy B using MPEG-2 with 8-by-8 DCT. The bit-rates were adjusted in order for the compressed copies in each pair to have similar average PSNR values, within the range of ± 1.5 dB. The results are shown in Table 5.5.

For the interlaced-noninterlaced pairs and the H.264-MPEG-2 pairs with nearly identical PSNR, the proposed algorithm is capable of delivering notable gain in PSNR, ranging from 1.5 to 2 dB depending on the contents and formats. This shows that the proposed algorithm is capable of utilizing picture information in different representations. Note that the results are better compared with the 1.0 Mbps pairs in Table 5.4(a) even though the input sequences all have nearly identical PSNR. This is not unexpected because distortions take up more distinctly different forms between video formats than within a single format. Each input copy thus has more information unknown to the other copy, and contributes more to the reconstruction process. For H.264-MPEG-2 pairs with ± 1.5 dB difference, the gain in PSNR over the better copy is reduced, and this is inline with the trend seen in Table 5.4(b). There is an apparent asymmetry in ΔPSNR between the third and fourth pairs. This is partly because we have held the H.264 copy constant and varied the quality of the MPEG-2 copy. This results in a situation similar to that shown earlier in Table 5.4(a) where the base PSNR values are different and the gain values may not be directly comparable.

TABLE 5.5: PSNR (dB) results obtained with mixed input formats

Video	Copy A		Copy B		Output PSNR	Δ PSNR
	Coding (Mbps)	PSNR	Coding (Mbps)	PSNR		
Foreman	nonint. 1.0	36.50	int. 1.3	36.45	38.49	1.99
	H.264 0.30	36.51	MPEG-2 1.0	36.57	38.20	1.63
	H.264 0.30	36.51	MPEG-2 0.8	35.31	37.36	0.85
	H.264 0.30	36.51	MPEG-2 1.2	37.45	38.59	1.14
Bus	nonint. 1.0	29.54	int. 1.3	28.91	31.16	1.61
	H.264 0.45	29.83	MPEG-2 1.0	29.55	31.57	1.74
	H.264 0.45	29.83	MPEG-2 0.8	28.18	30.59	0.76
	H.264 0.45	29.83	MPEG-2 1.2	30.58	32.08	1.50
Crew	nonint. 1.0	37.12	int. 1.3	37.31	38.88	1.57
	H.264 0.40	37.08	MPEG-2 1.0	36.93	38.76	1.68
	H.264 0.40	37.08	MPEG-2 0.8	35.85	37.99	0.91
	H.264 0.40	37.08	MPEG-2 1.2	37.83	39.31	1.48

5.2.2.4 Effect of the Number of Input Copies

The results of all the experiments presented so far are mainly based on two copies of video inputs. With more copies of compressed videos, it is likely that the size of the NQCS can be further reduced, resulting in even higher output quality. In the last part of the experiment, we prepared a total of 12 copies of compressed videos using the “foreman” sequence, and every next copy was coded with its starting point delayed by one more frame. The chosen GOP length of 12 enabled us to produce 12 copies without coinciding I-pictures, thus allowing maximum coding variations among the copies. The average PSNR values are plotted for different number of copies in Fig. 5.10. The value at 1-copy is the average PSNR of the input video copies.

When 3 copies are available, the gain in PSNR is about 2 dB, while the extra 9 copies offers about 1.5 dB extra gain. Each compressed copy does have some distinct information to contribute. However, there is much in common between them. As the number of copies increase, every next copy may share a bit more in common with the existing copies. Judging from the slope of the curve, the gain

FIGURE 5.10: PSNR results obtained with different number of input copies.

will further increase with even more copies for videos with longer GOPs, but that number would be far from practical and the marginal gain for each extra copy will be negligible.

To show the subjective quality improvement produced by the proposed algorithm, three 64-by-64-pixel crops are shown in Fig. 5.11. They are crops from the sequences of the last experiment. The crop from the input video shows visible compression artifacts near the collar and blocking artifacts near chin area, and the edge of the collar is somewhat distorted. The 3-copy output has a much smoother appearance and a better defined edge of the collar, although some mild blocking is still noticeable. The 12-copy output has very little artifact and better low-contrast details, but the difference in visual quality improvement is not as visually significant as between an untreated copy and the 3-copy version.

FIGURE 5.11: Subjective comparison, from left: input video, 3-copy POCS, 12-copy POCS. Same cropped area used as Fig. 5.9

5.3 Conclusion

We have addressed in this chapter the problem of reconstructing image and video of better quality based on multiple compressed copies from the same source content. Specifically, we reason that a compressed pictures provide quantization constraint sets instead of simply decoded values, and that differently coded copies contain different such sets. These sets provide extra information for image quality enhancement when multiple compressed copies of the same original are available. We then proposed an effective algorithm based on POCS to exploit the intersection of these quantization constraint sets to reconstruct pictures from the compressed copies. By doing so, we managed to reduce the quantization noise variance in the reconstructed pictures, and gracefully handled pictures of mixed formats, dimensions and mixed scanning modes including progressive and interlaced videos. Knowing that natural images are smooth and compression introduces discontinuity, we also imposed a smoothness constraint to further improve the quality of the reconstruction. We have shown empirically that the proposed algorithm can consistently enhance the quality of the reconstructed pictures over all available copies. For JPEG images, when similar quantization matrices are used, the gain in PSNR is about 0.5 dB over the better input copy, while up to 5 dB of gain is observed when drastically different quantization matrices are used. Under the video scenario where the GOP structures are misaligned among the input copies,

the proposed algorithm can obtain an average gain in PSNR up to 1.5 dB with just 2 copies and up to 3.5 dB with 12 copies, even though the gain diminishes as the number of copy increases.

Chapter 6

Conclusion and Future Research

The recent rise of video content sharing over the Internet and the sustained popularity of compressed images as a part of the essential web contents have given the end users plenty of options when it comes to which content to consume. However, given the choice among several similar versions of the same image or video content, the users need to be empowered with the proper tools so that the selection process is easier and the end results are better.

We have investigated the problem of quality assessment and enhancement of compressed images and videos. We summarize our research work and sketch future research directions in the following.

6.1 Review of Objectives

In this thesis, we have focused on assessing the compression quality of images and videos and enhancing it with the presence of multiple differently compressed copies. The objectives include:

- To investigate the techniques for no-reference objective image quality assessment, to identify and overcome their limitations.
- To show that statistical modeling is not necessarily the best way for relative quality assessment and propose a more robust alternative.
- To illustrate that multiple compressed copies can be used to jointly reconstruct pictures of superior quality.
- To experimentally demonstrate the effectiveness of the methods proposed.

In each of the following subsections, we summarize our contributions to one of the first three objectives listed above. Experiments have been carried out for each individual proposed method to show that it indeed serves the purpose.

6.1.1 No-reference PSNR Estimation for Compressed Pictures

Most images and videos available through the Internet and digital cable are compressed, and the end user generally does not have access to the original content to assess their compression quality. In order to allow objective quality assessment at the user end, a few methods have been proposed in the literature to blindly estimate the PSNR of compressed images and videos. All of these methods rely on estimating the distributions of the original DCT coefficients based on the quantized ones and the Laplace distribution assumption. Under these premises, the problem becomes how to minimize the estimation error for the Laplace parameters. Existing methods try to reduce this quantization-induced error by deriving maximum-likelihood estimators and linear prediction between frequencies, to various degrees of success.

However, we have realized in our investigation that quantization is closely bonded to the way the image blocks are defined, and that the block partitioning

is independent of the image contents. Since block-based DCT gives completely different frequency-domain representations upon even the slightest change in window location, we have proposed in Chapter 3 to estimate the Laplace parameters based on shifted block boundaries. Using the shifted or enlarged block boundaries, the DCT coefficients gathered from the decompressed images reveals more information about the original image than what the directly dequantized coefficients might suggest. We have also proposed to use a simple neural network to establish a numerical connection between the estimated and the original Laplace parameters, with the help of the quantization step-sizes. We have shown using experimental results that the proposed methods are able to achieve accurate and precise estimations of PSNR of compressed images and videos, and outperform the state-of-the-art methods.

6.1.2 Relative MSE Estimation for Compressed Images

In Chapter 3 we have proposed a blind PSNR estimation method to assess the compression quality of an individual image or a video frame on its own. However, there are situations where multiple copies of the same image or video may be available to an end user, especially when we consider those shared over the Internet by various users. In such cases, the original question on the absolute compression quality of a single image or video may become a relative one - which among these few copies offers the best compression quality?

Although we have shown that the block-shifting and the block-enlargement methods from Chapter 3 is capable of accurately estimating the PSNR value without reference, it has limitations in this new context of relative quality assessment. Knowing that the loss in quality is primarily due to quantization, and assuming that the images are different only in their compression, we proposed in Chapter 4 to compare compression quality pairwise by projecting one reconstructed image

onto the quantization constraints of another image and vice versa. The method take into account only the out-of-bound distances when calculating the difference in MSE. It is therefore a conservative measure that systematically underestimates this difference. However, it is able to pick the better copy with much higher accuracy and is simple in computation. As an added advantage, this projection-based method can produce localized relative quality estimates. In other words, it not only tells which image is better compressed overall, but also indicate areas where one image has higher quality than another.

6.1.3 Quality Enhancement Based on Multiple Compressed Copies

Using the method proposed in Chapter 4, an end user can determine which copy of the image or video has the best compression quality among a few, then we asked ourselves whether is this really the best the user can get.

As it turned out, we could indeed get better quality than the best copy. Each compressed copy contains slightly different information, and the copy with the best quality may not contain all the information in those inferior copies. We therefore proposed in Chapter 5 to make use of all the information available from these compressed copies to reconstruct an image or video sequence with higher quality than any of the inputs. Knowing that the true value of each quantized coefficient must lie within a computable range, and that the range defined by different copies may partially intersect to form a narrower range, we proposed to take values within this intersected range for reconstruction, so as to minimize the mean squared error in the final output. The projection-based method we used make it possible to combine quantization information from different compression formats. It also guarantees that the output is, in the worst case, not worse than any of the inputs. We demonstrated with image and video examples and in different

compression formats that the proposed method is able to produce outputs that are superior to any single input.

Combining the ability to enhance the quality with that to assess the quality as discussed earlier, we have proposed techniques that allow the end users to evaluate and improve the quality of compressed images and videos. Hopefully this and other research in this field will lead to better multimedia experience for everyone accessing the Internet.

6.1.4 Summary of Contributions

This thesis identified and addressed problems in the area of no-reference objective quality assessment and enhancement for compressed pictures. It has shown that it is possible to improve the statistical parameter estimation for PSNR estimation by breaking the artificial block boundaries. It has demonstrated the robustness of a new conservative relative quality assessment method, without relying on the statistical properties. It has also shown that multiple compressed copies of the same picture can be jointly used to reconstruct a picture of improved quality.

While there are still practical limitations to overcome, the thesis has shown the potential of further improving the usability of the Internet through automated image quality assessment and enhancement.

6.2 Future Research

In the buildup to this thesis, our research has been focused on the objective compression quality of still images and videos, including its no-reference assessment and enhancement. We looked for novel and feasible solutions for these problems,

and inevitably have left something to be desired and more related topics to be explored.

For the quality assessment of compressed images and videos in general, we may move our focus from PSNR to other objective quality measures that better correlate with MOS. After all, the pictures are consumed by being looked at, and the PSNR values of the pictures may not correspond very well with the perceived quality across different compression formats. Future studies may also make use of reduced reference for more informed estimation, as blind methods has little information besides the already compressed pictures and more recent RR methods have shown promising performance. With the rise of high dynamic range (HDR) imaging, we may also look into the perceived quality of compressed HDR images and videos.

From quality enhancement based on multiple compressed copies, we may also consider forms of distortion other than compression. For example, two images may have the same content but different tonal characteristics and color balances. And from them we can trace back to a common image which has the finest gradations and the largest dynamic range, and is more suitable than either of the inputs for further processing. Ultimately, this could also lead to the notion of “virtual master copy”, which contains all the information of all the copies of the same image or video online. This “virtual master copy” can then be used to either for processing, or as a form of archive since all those copies can be generated with some cues from the master.

With the arrival of a new generation of 3D televisions, an interesting problem is how we can quantitatively analyze the perceived image quality in a stereoscopic context. Certainly this concerns not only the compression process but also the generation of content as well as the presentation techniques. The new data and their corresponding compression techniques used in 3D TV also create new possibilities for video quality enhancement. No-reference and reduced-reference

techniques may be specifically tailored around 3D TV based on quality measures such as SSIM, VQM [69], or even more advanced metrics.

Publications

1. G. Yang, Y. P. Tan *Blind PSNR Estimation Using Shifted Blocks for JPEG Images*, accepted to 2011 IEEE International Symposium on Circuits and Systems
2. G. Yang, C. Wang, Y. P. Tan *Estimating Relative Objective Quality Among Images Compressed from the Same Original*, accepted to 2011 IEEE International Symposium on Circuits and Systems
3. G. Yang, Y. P. Tan *No-Reference PSNR Estimation Using DCT Coefficients of Shifted Blocks*, revised and re-submitted to IEEE Transactions on Image Processing
4. C. Wang, G. Yang, Y. P. Tan *Reconstructing Videos From Multiple Compressed Copies*, p1342-1351 IEEE Trans. Circuits and Systems for Video Technology, Sept 2009
5. G. Yang, C. Wang, Y. P. Tan *Regularized Dequantization for Image Copies Compressed with Different Quantization Parameters*, p.3172-3175 IEEE International Conference on Image Processing, Oct 2008

Bibliography

- [1] A. C. Bovik and S. Liu, “DCT-domain blind measurement of blocking artifacts in DCT-coded images,” in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, vol. 3, 2001, pp. 1725–1728.
- [2] P. Gastaldo, R. Zunino, I. Heynderickx, and E. Vicario, “Objective quality assessment of displayed images by using neural networks,” *Signal Processing: Image Communication*, vol. 20, no. 7, pp. 643–661, 2005.
- [3] S. Lee, M. S. Pattichis, and A. C. Bovik, “Foveated video quality assessment,” *IEEE Transactions on Multimedia*, vol. 4, no. 1, pp. 129–132, 2002.
- [4] C.-M. Liu and J.-Y. Lin, “Objective image quality measure for block-based DCT coding,” in *Digest of Technical Papers - IEEE International Conference on Consumer Electronics*, 1997, pp. 172–173.
- [5] H. R. Sheikh, A. C. Bovik, and L. Cormack, “No-reference quality assessment using natural scene statistics: JPEG2000,” *IEEE Transactions on Image Processing*, vol. 14, no. 11, pp. 1918–1927, 2005.
- [6] Z. Wang, A. C. Bovik, and B. L. Evans, “Blind measurement of blocking artifacts in images,” in *IEEE International Conference on Image Processing*, vol. 3, 2000.

-
- [7] D. S. Turaga, Y. Chen, and J. Caviedes, “No reference PSNR estimation for compressed pictures,” *Signal Processing: Image Communication*, vol. 19, no. 2, pp. 173–184, 2004.
 - [8] A. Ichigaya, M. Kurozumi, N. Hara, Y. Nishida, and E. Nakasu, “A method of estimating coding PSNR using quantized DCT coefficients,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 16, no. 2, pp. 251–258, 2006.
 - [9] T. Brandao and M. P. Queluz, “No-reference image quality assessment based on DCT domain statistics,” *Signal Processing*, vol. 88, no. 4, pp. 822–833, 2008.
 - [10] A. Eden, “No-reference estimation of the coding PSNR for H.264-coded sequences,” *IEEE Transactions on Consumer Electronics*, vol. 53, no. 2, pp. 667–674, 2007.
 - [11] S. Y. Shim, J. H. Moon, and J. K. Han, “PSNR estimation scheme using coefficient distribution of frequency domain in H.264 decoder,” *Electronics Letters*, vol. 44, no. 2, pp. 108–110, 2008.
 - [12] T. Chen, H. R. Wu, and B. Qiu, “Adaptive postfiltering of transform coefficients for the reduction of blocking artifacts,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 11, no. 5, pp. 594–602, 2001.
 - [13] Z. Li and E. J. Delp, “Block artifact reduction using a transform-domain Markov random field model,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 15, no. 12, pp. 1583–1593, 2005.
 - [14] A. W. C. Liew and H. Yan, “Blocking artifacts suppression in block-coded images using overcomplete wavelet representation,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 14, no. 4, pp. 450–461, 2004.

- [15] A. W. C. Liew, H. Yan, and N. F. Law, "POCS-based blocking artifacts suppression using a smoothness constraint set with explicit region modeling," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 15, no. 6, pp. 795–800, 2005.
- [16] P. List, A. Joch, J. Lainema, G. Bjontegaard, and M. Karczewicz, "Adaptive deblocking filter," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 13, no. 7, pp. 614–619, 2003.
- [17] J. Luo, C. W. Chen, K. J. Parker, and T. S. Huang, "Artifact reduction in low bit rate DCT-based image compression," *IEEE Transactions on Image Processing*, vol. 5, no. 9, pp. 1363–1368, 1996.
- [18] Y. Luo and R. K. Ward, "Removing the blocking artifacts of block-based DCT compressed images," *IEEE Transactions on Image Processing*, vol. 12, no. 7, pp. 838–842, 2003.
- [19] M. Y. Shen and C. C. Kuo, "Review of postprocessing techniques for compression artifact removal," *Journal of Visual Communication and Image Representation*, vol. 9, no. 1, pp. 2–14, 1998.
- [20] R. Prost, Y. Ding, and A. Baskurt, "JPEG dequantization array for regularized decompression," *IEEE Transactions on Image Processing*, vol. 6, no. 6, pp. 883–888, 1997.
- [21] W. Philips, "Corrections to 'JPEG dequantization array for regularized decompression'," *IEEE Transactions on Image Processing*, vol. 7, no. 12, p. 1725, 1998.
- [22] J. R. Price and M. Rabbani, "Biased reconstruction for JPEG decoding," *IEEE Signal Processing Letters*, vol. 6, no. 12, pp. 297–299, 1999.

-
- [23] J. J. Zou and H. Yan, "A deblocking method for BDCT compressed images based on adaptive projections," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 15, no. 3, pp. 430–435, 2005.
 - [24] A. Storkey and M. Allan, "Cosine transform priors for enhanced decoding of compressed images," *Intelligent Data Engineering and Automated Learning—IDEAL 2004*, pp. 533–539.
 - [25] T. Pevny and J. Fridrich, "Detection of double-compression in JPEG images for applications in steganography," *IEEE Transactions on Information Forensics and Security*, vol. 3, no. 2, pp. 247–258, 2008.
 - [26] W. Wang and H. Farid, "Exposing digital forgeries in video by detecting double MPEG compression," *Proceedings of the Multimedia and Security Workshop 2006, MM and Sec'06*, vol. 2006, pp. 37–47, 2006.
 - [27] A. Popescu and H. Farid, "Exposing digital forgeries by detecting traces of resampling," *IEEE Transactions on Signal Processing*, vol. 53, no. 2 II, pp. 758–767, 2005.
 - [28] M. Stamm and K. Ray, "Blind forensics of contrast enhancement in digital images," *Proceedings - International Conference on Image Processing, ICIP*, pp. 3112–3115, 2008.
 - [29] D.-Y. Hsiao and S.-C. Pei, "Detecting digital tampering by blur estimation," *Proceedings - First International Workshop on Systematic Approaches to Digital Forensic Engineering*, vol. 2005, pp. 264–278, 2005.
 - [30] ITU-T Recommendation T.81, *Digital compression and coding of continuous-tone still images - Requirements and guidelines*, 1992.
 - [31] ITU-R BT.601-6, *Studio encoding parameters of digital television for standard 4:3 and wide screen 16:9 aspect ratios*, 2007.

-
- [32] ITU-T Recommendation H.264, *Advanced video coding for generic audiovisual services*, 2005.
 - [33] ITU-T Recommendation H.222.0, *Information technology - Generic coding of moving pictures and associated audio information: Systems*, 1996.
 - [34] S. R. Smoot and L. A. Rowe, "Study of DCT coefficient distributions," in *Proceedings of the SPIE Symposium on Electronic Imaging*, vol. 2657, 1996, pp. 403–411.
 - [35] ISO/IEC 15444-1:2004, *Information technology – JPEG 2000 image coding system*, 2004.
 - [36] D. Taubman, "High performance scalable image compression with EBCOT," *IEEE Transactions on Image Processing*, vol. 9, no. 7, pp. 1158–1170, 2000.
 - [37] N. Jayant, J. Johnston, and R. Safranek, "Signal compression based on models of human perception," *Proceedings of the IEEE*, vol. 81, no. 10, pp. 1385–1422, 1993.
 - [38] J. Lubin, "Human vision system model for objective picture quality measurements," *IEE Conference Publication*, no. 447, pp. 498–503, 1997.
 - [39] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.
 - [40] Z. Wang and E. P. Simoncelli, "Reduced-reference image quality assessment using a wavelet-domain natural image statistic model," in *Proceedings of SPIE - The International Society for Optical Engineering*, vol. 5666, 2005, pp. 149–159.
 - [41] Z. Wang, G. Wu, H. Sheikh, E. Simoncelli, E.-H. Yang, and A. Bovik, "Quality-aware images," *IEEE Transactions on Image Processing*, vol. 15, no. 6, pp. 1680–1689, 2006.

- [42] T. Kusuma, Z. H.-J., and M. Caldera, "On the development of a reduced-reference perceptual image quality metric," *ICW '05 Proceedings of the 2005 Systems Communications*, vol. 2005, pp. 178–184, 2005.
- [43] M. Carnec, P. Le Callet, and D. Barba, "Objective quality assessment of color images based on a generic perceptual reduced reference," *Signal Processing: Image Communication*, vol. 23, no. 4, pp. 239–256, 2008.
- [44] R. V. Babu, S. Suresh, and A. Perkis, "No-reference JPEG-image quality assessment using GAP-RBF," *Signal Processing*, vol. 87, no. 6, pp. 1493–1503, 2007.
- [45] R. C. Reininger and J. D. Gibson, "Distribution of the two-dimensional DCT coefficients for images," *IEEE Transactions on Communications*, vol. COM-31, no. 6, pp. 835–839, 1983.
- [46] F. Mueller, "Distribution shape of two-dimensional DCT coefficients of natural images," *Electronics Letters*, vol. 29, no. 22, pp. 1935–1936, 1993.
- [47] R. L. Joshi and T. R. Fischer, "Comparison of generalized Gaussian and Laplacian modeling in DCT image coding," *IEEE Signal Processing Letters*, vol. 2, no. 5, pp. 81–82, 1995.
- [48] E. Y. Lam and J. W. Goodman, "A mathematical analysis of the DCT coefficient distributions for images," *IEEE Transactions on Image Processing*, vol. 9, no. 10, pp. 1661–1666, 2000.
- [49] M. I. H. Bhuiyan, M. Omair Ahmad, and M. N. S. Swamy, "Modeling of the DCT coefficients of images," in *Proceedings - IEEE International Symposium on Circuits and Systems*, 2008, pp. 272–275.
- [50] A. Zakhor, "Iterative procedures for reduction of blocking effects in transform image coding," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 2, no. 1, pp. 91–95, 1992.

-
- [51] V. K. Goyal, "Multiple description coding: Compression meets the network," *IEEE Signal Processing Magazine*, vol. 18, no. 5, pp. 74–93, 2001.
 - [52] O. G. Tillo, T., "Improving the performance of multiple description coding based on scalar quantization," *IEEE Signal Processing Letters*, vol. 15, pp. 329–332, 2008.
 - [53] H. S. Tian, C., "A new class of multiple description scalar quantizer and its application to image coding," *IEEE Signal Processing Letters*, vol. 12, no. 4, pp. 329–332, 2005.
 - [54] N. P. Galatsanos and R. T. Chin, "Digital restoration of multichannel images," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 37, no. 3, pp. 415–421, 1989.
 - [55] N. P. Galatsanos, A. K. Katsaggelos, R. T. Chin, and A. D. Hillery, "Least squares restoration of multichannel images," *IEEE Transactions on Signal Processing*, vol. 39, no. 10, pp. 2222–2236, 1991.
 - [56] B. K. Gunturk, Y. Altunbasak, and R. M. Mersereau, "Multiframe resolution-enhancement methods for compressed video," *IEEE Signal Processing Letters*, vol. 9, no. 6, pp. 170–174, 2002.
 - [57] D. Youla and H. Webb, "Image restoration by the method of convex projections: Part 1 Theory," *Medical Imaging, IEEE Transactions on*, vol. 1, no. 2, pp. 81–94, 1982.
 - [58] S. Mallat, "A theory for multiresolution signal decomposition: The wavelet representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 11, no. 7, pp. 674–693, 1989.
 - [59] X. Wu, "Lossless compression of continuous-tone images via context selection, quantization, and modeling," *IEEE Transactions on Image Processing*, vol. 6, no. 5, pp. 656–664, 1997.

-
- [60] H. R. Sheikh, Z. Wang, L. Cormack, and A. Bovik, "LIVE image quality assessment database release 2," 6 April, 2010.
- [61] x264 - a free H.264/AVC encoder. [Online]. Available: <http://www.videolan.org/developers/x264.html>
- [62] JM software download, H.264/AVC software coordination. [Online]. Available: <http://iphone.hhi.de/suehring/tml>
- [63] Independent JPEG group website. [Online]. Available: <http://www.ijg.org/>
- [64] F. M. De Saint-Martin, P. Siohan, and A. Cohen, "Biorthogonal filterbanks and energy preservation property in image compression," *IEEE Transactions on Image Processing*, vol. 8, no. 2, pp. 168–178, 1999.
- [65] B. Usevitch, "Optimal bit allocation for biorthogonal wavelet coding," in *Data Compression Conference Proceedings*, 1996, pp. 387–395.
- [66] Kodak lossless true color image suite. [Online]. Available: <http://r0k.us/graphics/kodak/>
- [67] C. Kim and B. Vasudev, "Spatiotemporal sequence matching for efficient video copy detection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 15, no. 1, pp. 127–132, 2005.
- [68] Y.-P. Tan, S. R. Kulkarni, and P. J. Ramadge, "Framework for measuring video similarity and its application to video query by example," in *IEEE International Conference on Image Processing*, vol. 2, 1999, pp. 106–110.
- [69] ITU-T J.144, *Objective perceptual video quality measurement techniques for digital cable television in the presence of a full reference*, 2004.