STRUCTURAL SIMILARITY BASED IMAGE QUALITY ASSESSMENT: POOLING STRATEGIES AND APPLICATIONS TO IMAGE COMPRESSION AND DIGIT RECOGNITION

by

XINLI SHANG

Presented to the Faculty of the Graduate School of

The University of Texas at Arlington in Partial Fulfillment

of the Requirements

for the Degree of

MASTER OF SCIENCE IN ELECTRICAL ENGINEERING

THE UNIVERSITY OF TEXAS AT ARLINGTON

August 2006

ACKNOWLEDGEMENTS

The author is deeply indebted to Dr. Zhou Wang, who has been his resource of inspiration and a guiding force during his study at the University of Texas at Arlington, both as an instructor and thesis supervisor. The author would like to thank Dr. K. R. Rao and Dr. Mingyu Lu for their interest in his research and for taking time to serve in his thesis committee.

The author is grateful to all the teachers who taught him during the years he spent in school, both in China and in the Unites States.

Finally, the author would like to express his immense gratitude to his family members for being his motivation with their continuous love and support.

April 14, 2006

ABSTRACT

STRUCTURAL SIMILARITY BASED IMAGE QUALITY ASSESSMENT: POOLING STRATEGIES AND APPLICATIONS TO IMAGE COMPRESSION AND DIGIT RECOGNITION

Publication No. _____

Xinli Shang, M.S.

The University of Texas at Arlington, 2006

Supervising Professor: Zhou Wang

This thesis studies a recently proposed perceptual image similarity measure − the structural similarity (SSIM) index. Although still in its early stage, the SSIM index has demonstrated superior performance in a large number of tests as compared to the currently most widely used image distortion/quality measures, the mean squared error (MSE) and the peak signal-to-noise-ratio (PSNR). This motivates us to further investigate the SSIM method and extend it to other image processing and pattern recognition applications. Specifically, three topics have been studied in this thesis: spatial pooling strategies for perceptual image quality assessment, structural similarity-guided perceptual image compression, and handwritten digit recognition using complex wavelet structural similarity index.

Recently, a number of objective image quality assessment algorithms have been proposed to predict human perception of image quality. Many of these algorithms are

implemented in two stages. In the first stage, image quality is evaluated within local regions. This results in a quality/distortion map over the image space. In the second stage, a spatial pooling algorithm is employed that combines the quality/distortion map into a single quality score. While great effort has been devoted to developing algorithms for the first stage, little has been done to find the best strategies for the second stage (and simple spatial average is often used). We have investigated three spatial pooling methods for the second stage: Minkowski pooling, local quality/distortion-weighted pooling, and information content-weighted pooling. Extensive experiments with the LIVE database (developed at the Laboratory of Image and Video Engineering at The University of Texas at Austin) show that all three methods may improve the prediction performance of perceptual image quality measures, but the third method demonstrates the best potential to be a general and robust method that leads to consistent improvement over a wide range of image distortion types.

State-of-the-art image compression techniques, such as the set partitioning in hierarchical trees (SPIHT) algorithm and JPEG2000, apply a wavelet decomposition to the image followed by a bitplane coding scheme. The bitplane coding technique allows for continuously rate scalable coding and can help to order the encoded bitstreams according to their importance, where a bit's importance is directly related to its contribution to the MSE between the original and the decoded image. Since MSE is not an adequate measure of perceptual image quality, here we attempt to replace the role of MSE with the SSIM index. By applying such a new optimization goal into the bitplane coding scheme, we obtain visually improved images using the same bit rate. In comparison with SPIHT and JPEG2000 compressed images, the perceptual quality of the decoded images is more evenly distributed over the image space.

The structural similarity method has also been used for handwritten digit recognition. In particular, an extended version of the SSIM index, namely the complex wavelet

structural similarity (CW-SSIM), has been employed. One of the major advantages of the CW-SSIM index is its robustness to small geometric distortions such as translation, scaling and rotation of the images. This is an extremely important feature for many pattern recognition problems. As an initial attempt for the application of the CW-SSIM method for pattern recognition, we use it as the similarity measure for handwritten digit recognition. Experiments with the MNIST database [1] show that our algorithm results in an error rate of 5.06% with 200 templates for each digit. This is achieved without any sophisticated preprocessing stages that normalize and register the test images before comparison.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

Digital imagery is a pervasive way to describe the world. In recent years, there is a dramatic growth in the use of digital images as a means for representing and communicating information. In the literature, a large amount of effort has been devoted to developing methods for improving images appearance, or for maintaining the appearance of images that are processed. To maintain, control, and enhance the quality of images, it is important for images acquisition, management, communication, and processing systems to be able to identify and quantify image quality degradations [2]. For the purpose of visual perception, the human eyes are the ultimate receivers, hence the most reliable way of assessing the quality of an image is subjective evaluation [3]. In a standard subjective test, a number of human observers are asked to give scores on a test image and the mean value of all the subjective scores (the mean opinion score, or MOS) can then be used as an indicator of image quality. However, MOS method has two drawbacks: it is very expensive, and it is usually too slow to handle a large number of images in real time. Therefore, it is imperative to develop effective automatic image quality assessment systems.

Objective image quality assessment techniques can be classified into full-reference (FR), reduced-reference (RR), and no-reference (NR) method according to the availability of the "original image", which is considered to be distortion-free or perfect quality [2]. Most of the proposed objective quality measures in the literature adopt the FR method, i.e., assuming that the undistorted reference image exists and is fully available. But in many practical applications, an image quality assessment system does not have access

to the reference images. Therefore, it is desirable to develop measurement approaches that can evaluate image quality blindly (NR). NR image quality assessment turns out to be a very difficult task, although human observers usually can effectively and reliably assess the quality of distorted images without using any reference at all. In the third type of image quality assessment method (RR), the reference image is not fully available. Instead, certain features are extracted from the reference image and employed by the quality assessment system as side information to help evaluate the quality of the distorted image. This thesis mainly focuses on FR image quality assessment.

Mean squared error (MSE) is widely used in objective image quality assessment as well as many other fields. MSE is simple and straightforward. However, it has long been pointed out that MSE is a poor model for visual perception of image distortion. Since the human eye is the ultimate receiver in most real world applications, the characteristics of the human visual system (HVS) would need to be taken into consideration in the design of image quality evaluation systems. Pioneering work in the field of objective image quality assessment was done by Mannos and Sakrison, who proposed an image fidelity criteria that takes into account the human visual sensitivity as a function of spatial frequency [4]. Other important early work was documented in an edited book by Watson [5]. Reviews of the most recent development of objective image quality assessment can be found at [2, 3]. In [6] and [7], the structural similarity (SSIM) measure was proposed as a tool for image quality assessment. Experiment results showed that this approach achieves better correlations with subject evaluations. The complex wavelet version of the SSIM method (CW-SSIM) is provided in [8], which gives better performance than SSIM in the existence of small geometric distortions.

Many recently proposed perceptual image quality assessment algorithms are implemented in two stages. In the first stage, image quality is evaluated within local regions. This results in a quality/distortion map over the image space. In the second stage, a

spatial pooling algorithm is employed that combines the quality distortion map into a single quality score. While great effort has been devoted to developing algorithms for the first stage, little has been done to find the best strategies for the second stage. Typically, simple spatial average is used.

Image compression has been an active topic in the past two to three decades. However, most of the current image codecs, such as the set partitioning in hierarchical trees (SPIHT) algorithm and the JPEG2000 codecs (VM [9], JASPER[10] and JJ2000 [11]), adopt a rate-based MSE minimization encoding approach [12]. The involvement of perceptual optimization has been minimal and its effectiveness is very limited. This leaves space for further perceptual improvement by incorporating new image quality assessment algorithm such as the SSIM index.

Handwritten digit recognition is widely desirable in many real applications such as automatic bank check reading and automatic postal code recognition. Since FR image quality assessment algorithms can be directly used to evaluate the similarity between two images, it would be interesting to see how they perform in digit recognition problems. In particular, the CW-SSIM index would be useful because it can provide adequate image comparison without a precise registration stage at the front end.

The goal of this thesis is threefold: 1) Study the pooling strategies of objective image quality assessment; 2) Design an structural similarity - guided perceptual image compression algorithm; 3) Design an algorithm for handwritten digit recognition using CW-SSIM.

1.1 Objective Image Quality Assessment

MSE is a simple and straightforward objective image quality measure. Let X and Y be two matrices representing two images being compared. Let M and N be the numbers

of rows and columns in the images (we have assumed that the two images have the same size). MSE is defined as Eq. 1.1 [6]:

$$MSE = \frac{1}{M \times N} \sum_{i=1,j=1}^{M,N} (x_{ij} - y_{ij})^2 \tag{1.1}$$

Assuming that X is the original image, then the MSE value can be used as a measure of the distortion in the distorted image Y. Another related and often-used measure for image quality is the peak signal-to-noise Ratio (PSNR), which is defined as follows.

$$PSNR = 10 \log_{10} \frac{L^2}{MSE} \tag{1.2}$$

Here L is the dynamic range of allowable image pixel intensities. For gray-scale images of 8 bits/pixel, L = 2^8-1 = 255. Note that for a given L, an MSE value can be uniquely mapped to a PSNR value, and vice versa.

MSE is often the most convenient for the purpose of algorithm optimization, since it is differentiable and when combined with linear algebra tools, closed-form solutions can often be found for real problems. In addition, MSE often has a clear physical meaning − the energy of the error signal (an error signal is defined as the difference signal between the two images being compared). These are the major reasons why MSE (and PSNR) is extensively used throughout the literature of image processing, communication, and many other signal processing fields.

Nevertheless, MSE has long been criticized for its poor correlation with perceived image quality. An instructive example is shown in Figure 1.1, where the original "Einstein image" (a) is altered by several different types of distortions: (b) contrast stretch; (c) mean luminance shift; (d) Gaussian noise contamination; (e) impulsive noise contamination; (f) JPEG compression; (g) blurring. It is important to note that several of these

images have nearly identical MSE values relative to the "original", yet these same images present dramatically different visual quality.

A common explanation is that MSE does not reflect the way that the human visual systems perceive images. A number of important psychophysical and physiological features of Human Visual System (HVS) are not accounted for by the MSE. Notice that MSE is defined using one dimensional formulation, thus it does not make use of any spatial structural information in the image at all, which may be essential for the evaluation of perceptual image quality.

The first notable work in the field of objective *perceptual* image quality assessment is the pioneering work of Mannos and Sakrison [4], who proposed an image fidelity criteria that takes into account human visual sensitivity as a function of spatial frequency.

The Daly model [13], or visible differences predictor (VDP), is intended to be used for high-quality imaging systems, in which the probability of whether the difference between two images can be discerned is evaluated. The output of this model is a probability-of-detection map between the reference and the distorted images. This model includes a number of processing stages, including a point-wise nonlinearlity, filtering based on the contrast sensitivity function (CSF), space-frequency channel decomposition, contrast calculation, masker calculation, and a probability-of-detection calculation [2]. The model uses a modified version of Watson's cortex transform for the channel decomposition, which separates the image signal into five scale levels followed by six orientations. For each channel, a threshold elevation map is computed from the contrast in that channel. There are two distinct features of Daly model. One is that it allows for mutual masking, which includes not only the reference image, but also the distorted image in the calculation of the masking factor. The second is that a psychometric function is used to convert the strengths of the normalized error into a probability-of-detection map before the pooling stage.

Another model that attempts to estimate the probability-of-detection of the differences between the reference and distorted images is by Lubin [14]. Lubin model starts by filtering the images using a low-pass PSF that simulates eye optics. The filtered images are then re-sampled according to the retinal photoreceptor sampling. Next, the images are decomposed using a Laplacian pyramid into seven resolutions, followed by band-limited contrast calculations. To reflect orientation selectivity, the signal is further decomposed into four orientations using a bank of steerable filters. The decomposed signal is normalized using the subband base-sensitivity determined by the CSF. A point-nonlinearity of sigmoid shape is implemented to account for intra-channel masking. The normalized error signal is convolved with disk-shaped kernels before a Minkowski pooling stage across scale. The pooled error at each spatial location is then converted into a probability-of-detection map. An additional pooling stage may be finally applied to obtain a single number for the entire image.

The Safranek-Johnston model [15] is designed for perceptual image coding. It decomposes the image signal using the generalized quadrature mirror filter(GQMF)transform, a separable decomposition that equally divides the frequency space into 16 subbands. At each subband, a base sensitivity factor is determined by the noise sensitivity on a mid-gray image and was obtained by subjective experiment.

In the Teo-Heeger model [16], the channel decomposition is applied after a front-end linear filtering stage. In an earlier version of the model, a hex-QMF transform, which is quadrature mirror filter implemented on a hexagonally-sampled image, is used to accomplish the channel decomposition. Later, the authors adopted a steerable pyramid decomposition with six orientations, which is a polar separable wavelet design that avoids aliasing in the subbands.

Some researchers consider spatially varying distortion metrics which attempt to exploit the masking phenomenon of the visual systems. Watson's work [17] [18] and others work such as [19] on visual optimization are noteworthy in this regard.

Watson's DCT model first divides the image into distinct blocks and a visibility threshold is calculated for each coefficient in each block. Three factors determine the visibility threshold. The first is the baseline contrast sensitivity associated with the DCT component, which is determined empirically. The second factor is luminance masking, which only affects the DC coefficient in the DCT. The third factor is contrast/texture masking, in which the masking adjustment is determined by all the coefficients within the same block.

Watson's wavelet model is based on direct measurement of the human visual sensitivity threshold for individual wavelet coefficients. It constructs a mathematical model for DWT noise detection thresholds that is a function of level, orientation, and display visual resolution. This allows for calculation of a "perceptually lossless" quantization matrix, for which all errors are in theory below the visual threshold. The model can be used as the basis for adaptive quantization schemes or the bit plane coding rate distortion scheme.

[19] presented an algorithm that locally adapts the quantizer step size at each pixel according to an estimate of the masking measure. This estimate is based on the pixels already coded for the prediction of the pixels not yet coded. This algorithm exploits the spatiotemporal masking properties of the human visual system, based on psychophysical masking phenomena, to establish thresholds of just-noticeable distortion (JND) or minimally noticeable distortion (MND). The central ideas are: 1) to "hide" coding distortion beneath spatial and temporal JND thresholds, and 2) to augment the classical coding paradigm of redundancy removal with elimination of irrelevant signal

information, i.e., discarding those signal components which are imperceptible to the human receiver.

The structural similarity (SSIM) method [6] [7] is a recently proposed approach for image quality assessment. An important observation is that natural image signals are highly structured: their pixels exhibit strong dependencies, especially when they are spatially proximate, and these dependencies carry important information about the structure of the objects in the visual scene. The motivation of SSIM is to find a more direct way to compare the structures of the reference and the distorted signals.

The system diagram of the SSIM method is shown in Figure 1.2. Suppose x and y and are two nonnegative image signals, which have been aligned with each other (e.g., image patches extracted from the same location in two images). Suppose that one of the signals has perfect quality, then the similarity measure can serve as a quantitative measurement of the quality of the second signal. The system separates the task of similarity measurement into three comparisons: luminance, contrast and structure. First, the luminance of each signal is compared. Assuming discrete signals, this is estimated as the mean intensity

$$\mu_x = \frac{1}{N}\sum_{i=1}^{N} x_i \tag{1.3}$$

and

$$\mu_y = \frac{1}{N}\sum_{i=1}^{N} y_i \tag{1.4}$$

Secondly, signal contrast is estimated as the standard deviation (the square root of variance). The contrast comparison is then the comparison of σ_x and σ_y.

$$\sigma_x = (\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \mu_x)^2)^{\frac{1}{2}} \tag{1.5}$$

$$\sigma_y = (\frac{1}{N-1}\sum_{i=1}^{N}(y_i - \mu_y)^2)^{\frac{1}{2}} \tag{1.6}$$

Thirdly, the signal is normalized (divided) by its own standard deviation, so that the two signals being compared have unit standard deviation. The structure comparison s(x,y) is conducted on these normalized signals $\frac{(x-\mu_x)}{\sigma_x}$ and $\frac{(x-\mu_y)}{\sigma_y}$.

$$\sigma_{xy} = (\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \mu_x)(y_i - \mu_y))^{\frac{1}{2}} \tag{1.7}$$

Finally, the comparisons are combined and the SSIM index is defined as:

$$SSIM(x,y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \tag{1.8}$$

where C_1, C_2 are small positive constants.

Although the spatial domain SSIM index gives superior performance in a wide range of image distortion types and levels, it is highly sensitive to translation, scaling and rotation of images, as demonstrated in images (h)-(l) of Figure 1.3. [8] extended the SSIM method into the complex wavelet transform domain, so that it is insensitive to these "non-structural" image distortions that are typically caused by the movement of the

image acquisition devices, rather than the changes of the structures of the objects in the visual scene. This new image similarity measure does not require a precise registration process in the front, and naturally combines a number of invariants into one simple measurement.

In the complex wavelet transform domain, suppose $c_x = \{c_{x,i} | i = 1, ..., N\}$ and $c_y = \{c_{y,i} | i = 1, ..., N\}$ are two sets of coefficients extracted at the same spatial location in the same wavelet subbands of the two images being compared, respectively. The complex wavelet SSIM (CW-SSIM) index is defined as:

$$S\ (c_x, c_y) = \frac{2|\sum_{i=1}^{N} c_{x,i} c_{y,i}^*| + K}{\sum_{i=1}^{N} |c_{x,i}|^2 + \sum_{i=1}^{N} |c_{y,i}|^2 + K} \tag{1.9}$$

Here c^* denotes the complex conjugate of c and K is a small positive constant. CW-SSIM is based on the following two assumptions [8]:

1)The structural information of local image features is mainly contained in the relative phase patterns of the wavelet coefficients.

2)Consistent phase shift of all coefficients does not change the structure of the local image feature.

Figure 1.3 demonstrates the CW-SSIM measure for image quality assessment. A 2-scale, 16-orientation steerable pyramid decomposition is constructed and the 16 subbands at the second scale are used by the CW-SSIM measure. It can be seen that images with almost the same MSE values but different distortion types (Images (b)-(g)) have drastically different visual quality, which is better predicted by SSIM and CW-SSIM. However, the SSIM method fails to provide useful quality prediction when the images are slightly shifted, scaled or rotated (Images (h)-(l)). These are effectively accounted for by CW-SSIM, which gives significantly higher scores to Images (b), (c) and (h)-(l) than to Images (d)-(g).

1.2 Spatial Pooling for Image Quality Assessment

Many of the image quality assessment algorithms (especially full-reference algorithms) adopted a two-stage implementation: In the first stage, image quality/distortion is evaluated locally within small regions, resulting in a quality/distortion map. In the second stage, a spatial pooling algorithm is employed to combine the quality/distortion map into a single quality score. Such a two-stage approach may be applied directly in image pixel domain or after channel decompositions (e.g., applied to a wavelet subband).

A pixel-domain full-reference example is shown in figure 1.4, where the goal is to evaluate the quality of image (b) with a given perfect-quality reference image (a). Two methods are used to compute local quality/distortions - absolute difference and the structural similarity (SSIM) index. The resulting quality/distortion maps are shown in figure 1.4(c) and (d), respectively. For easy comparison, we have adjusted the quality/distortion map representations so that brighter indicates better quality in both maps. Careful inspection shows that the SSIM index (computed within a local window that slides across the image space) better reflects the spatial variations of perceived image quality, for example, the blockiness in the sky is clearly indicated in figure 1.4(d) but not in figure 1.4(c). However, the major concern here is not on how to create a better quality map but on how to convert a quality map into a scalar quality score.

Surprisingly, in the literature, little investigation and careful comparison have been devoted to developing and testing spatial pooling methods. In practice, spatial pooling has often been treated superficially, e.g., using a simple spatial average. Some methods incorporate human interactions or automatic object detections and segmentations to define the regions-of-interest or points-of- fixations before spatial pooling (e.g., [20] [21]), but these methods may not be easily applied to general-purpose image quality assessment because for many images, it may not always be easy to find obviously outstanding objects that attract visual attention. On the other hand, problems arise with the direct spatial

average approach when the distortion is highly non-uniform over the image space. For example, when only a small region in an image is corrupted with extremely annoying artifacts, but all other regions have high quality, human subjects tend to pay more attention to the low quality region and give an overall quality score lower than the average of the quality/distortion map.

This thesis studies three strategies for spatial pooling - Minkowski pooling, local quality/distortion-weighted pooling, and information content-weighted pooling.

1.3 Image Coding and Perceptual Optimization

Uncompressed image data requires considerable storage capacity and transmission bandwidth. Despite the rapid progress in mass-storage density, processor speed, and digital communication system performance, the demand for data storage capacity and data-transmission bandwidth continue to outstrip the capabilities of available technologies. The recent growth of data intensive multimedia-based web applications have not only sustained the need for more efficient ways to encode signals and images but have made compression of such signals central to storage and communication technology [22].

JPEG (Joint Photographic Experts Group) standard has been established by ISO (International Standards Organization) and IEC (International Electro-Technical Commission) for still image compression. JPEG is based the block discrete cosine transform (DCT) and its performance significantly degrades at low bit-rates. In recent years, the wavelet transform has been an attractive technology in image compression field. Wavelet-based coding provides considerable improvements in image quality at high compression ratios (low bit rates). In the last few years, several wavelet-based schemes for image compression, such as embedded zerotree wavelet (EZW), SPIHT and JPEG2000, have been developed.

Shapiro [23] introduced the EZW image coding that uses a tree-like data structure to encode the coefficients of wavelet decomposition followed by an continuously scalable bitplane coding algorithm. The zerotree idea is based on the hypothesis that if a wavelet coefficient at a coarse scale is insignificant with respect to a given threshold T, then all wavelet coefficients of the same orientation in the same spatial location at a finer scales are also likely to be insignificant with respect to threshold T. Said and Pearlman [24] provided a new and more effective implementation of a modified EZW algorithm based on a set partitioning in hierarchical trees (SPIHT) algorithm. They also presented a scheme for progressive transmission of the coefficient values that incorporates the concepts of ordering the coefficients by magnitude and transmitting the most significant bits first. They used a uniform scalar quantizer and claimed that the ordering information made this simple quantization method more efficient. An efficient way to code the ordering information is also proposed.

In recent years, JPEG2000 (i.e.,ISO/IEC15444) has been approved as a new international standard. It supports both lossy and lossless compression. In addition to improved compression performance,a number of other attractive features are provided, including: 1) progressive recovery of an image by fidelity or resolution; 2) region of interest coding, whereby different parts of an image can be coded with differing fidelity; 3) random access to particular regions of an image without decoding the entire code stream; 4) a flexible file format with provisions for specifying opacity information and image sequences; and 5) good error resilience [25].

JPEG2000 codec is based on wavelet/subband coding techniques [26] [27]. In the encoder, after the image data has been transformed, the resulting coefficients may be quantized. Quantization is the first primary source of information loss in the coding path. There are two coding stages in JPEG2000 - tier-1 coding and tier-2 coding. In tier-1 coding, The quantizer indices for each subband are partitioned into code blocks.

Code blocks are rectangular in shape, and their nominal size is a free parameter of the coding process. After a subband has been partitioned into code blocks, each of the code blocks is independently coded. For entropy coding, a context-based adaptive binary arithmetic coder is used [28]. For each code block, an embedded code is produced, comprised of a number of coding passes. The output of the tier-1 encoding process is, therefore, a collection of coding passes (significance pass, refinement pass, and cleanup pass) for the various code blocks. In tier-2 encoding, the coding pass information is partitioned into data units called packets, a process typically referred to as packetization. Each coding pass is either assigned to one of the L layers or discarded. The coding passes containing the most important data are included in the lower layers, while the coding passes associated with finer details are included in higher layers. The rate control algorithm must decide in which layer each coding pass is to be included. Since some coding passes may be discarded, tier-2 coding is the second primary source of information loss in the coding path. To decide which coding pass should be discarded before tier-2 coding in order to compress image data, many distortion metrics/criteria has been put forward in the literature.

Since MSE is known to be a poor model for visual perception of image distortion, some authors [29] have considered a relatively straightforward extension of MSE to a weighted MSE in the frequency domain, where the weights were derived from studies of the contrast sensitivity function (CSF) of the human visual system. A number of spatially varying distortion metrics have been proposed in the literature to exploit the visual masking effect. Watson's work [17] on visual optimization of JPEG compressed images is noteworthy in this regard, as is the work of [19]. The EBCOT algorithm is provided by [30], which provides improved image quality when compared with SPIHT. Specifically, EBCOT coded images exhibit substantially less ringing artifacts around

edges and superior rendition of textures. This thesis tries to use SSIM as an objective criteria in the context of visual optimization of wavelet-based image compression.

1.4 Handwritten Digit Recognition

Pattern recognition is one of the main tasks of biological perception and information processing systems, and it is also a major challenge in computer science and engineering. The problem of pattern recognition is to classify objects into categories, given that objects in a particular category may vary widely, while objects in different categories may be very similar. A typical example is handwritten digit recognition [31]. Automatic handwritten digit recognition is desirable in many real world applications, including automatic bank check reading and automatic postal code reading. Digit characters, typically represented as binary images must be classified into one of 10 (0 to 9) categories using a classification function. Building such a classification function is a major technological challenge.

Handwritten digit recognition algorithms can be roughly divided into two camps: image-based matching and feature-based matching. In image-based matching, prototype image patterns (or templates) for each category are stored. Each incoming pattern can then be compared to all the stored prototypes, and the label associated with the prototype that best matches the input will be output. Rather than trying to keep an image representation of the training set, feature-based matching algorithms to learn a set of feature parameters from the training set and stored these feature parameters for each category. When recognizing a new pattern, its feature parameters are calculated and compared with the features for each category, and the final recognition result is determined by the closeness of patterns in the feature space.

Many methods are proposed for handwritten digits recognition [32] [33], and several representative methods are discussed as follows.

Linear classifier: It is probably the simplest classifier, where each input pixel value contributes to a weighted sum for each output unit. The output unit with the highest sum is then recognized as the class of the input digit. For the MNIST database (http://yann.lecun.com/exdb/mnist/), an error rate of 8.4% is obtained with a deskewing preprocessing stage in the front.

Baseline Nearest Neighbor Classifier: K-nearest neighbor classifier is proposed in [32]. It is carried out by calculating an Euclidean distance measure between input images. The error rate for the MNIST database is 1.22%, which is obtained with a set of front end preprocessing steps, including deskewing, noise removal and blurring.

LeNet 4 [32]: It is an improved version of LeNet 1 [34] that has a 32x32 input layer. It includes more feature maps and an additional layer of hidden units that is fully connected to both the last layer of features maps and to the output units. LeNet 4 contains about 260,000 connections and has about 17,000 free parameters. The error rate for the MNIST database is 1.1 %.

Convolutional Net with cross-entropy [35]: This classifier attempts to expand the training set by adding a new form of distorted data. Unlike many other advanced approaches, the convolutional neural network method does not require sophisticated computations such as momentum, weight decay, structure dependent learning rates, averaging layers, tangent propagation, or even finely-tuning the architecture. It achieved an error rate of 0.4% for the MNIST, which is the best performance reported in the literature.

Most of the existing handwritten digit recognition methods are complex in structure and are difficult to implement. This motivated us to apply CW-SSIM for handwritten digit recognition, which may provide a completely different and simplified approach with an acceptable error rate.

1.5 Thesis Outline

The rest of the thesis is organized as follows: Detailed information about spatial pooling for image quality assessment are described in chapter 2, including Minkowski pooling, local quality/distortion-weighted pooling, and information content-weighted pooling. All methods are tested with the LIVE database [36]. Chapter 3 presents our work on structural similarity-guided perceptual image compression which is implemented by incorporating the SPIHT and JPEG2000 compression algorithms. In chapter 4, we describe handwritten digit recognition using complex wavelet structural similarity index. The method is tested with the MNIST database [1]. Chapter 5 concludes this thesis.

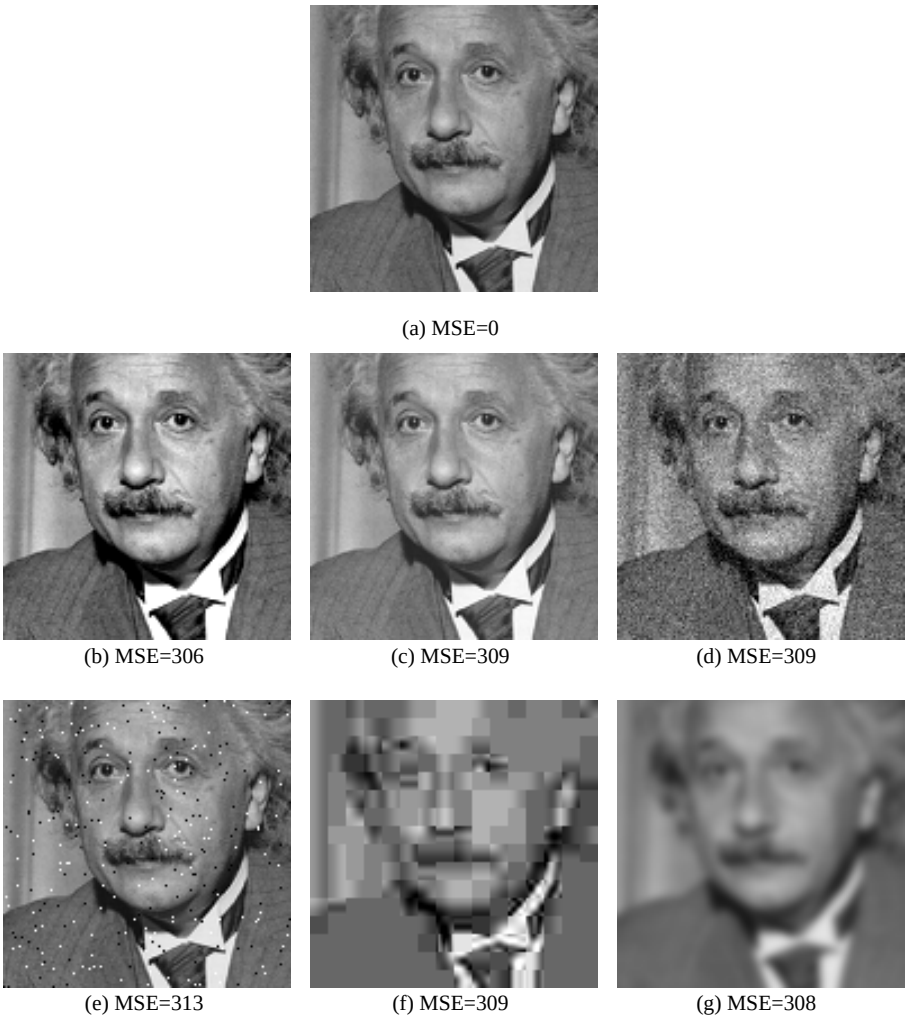


Figure 1.1. Einstein Image Altered with Different Types of Distortions. Images are ordered in raster scan. (A) Reference Image (8 Bits/Pixel, Assumed to Have Perfect Quality); (B) Contrast Stretch; (C) Mean Luminance Shift; (D) Gaussian Noise Contamination; (E) Impulsive Noise Contamination; (F) JPEG Compression; (G) Blurring.

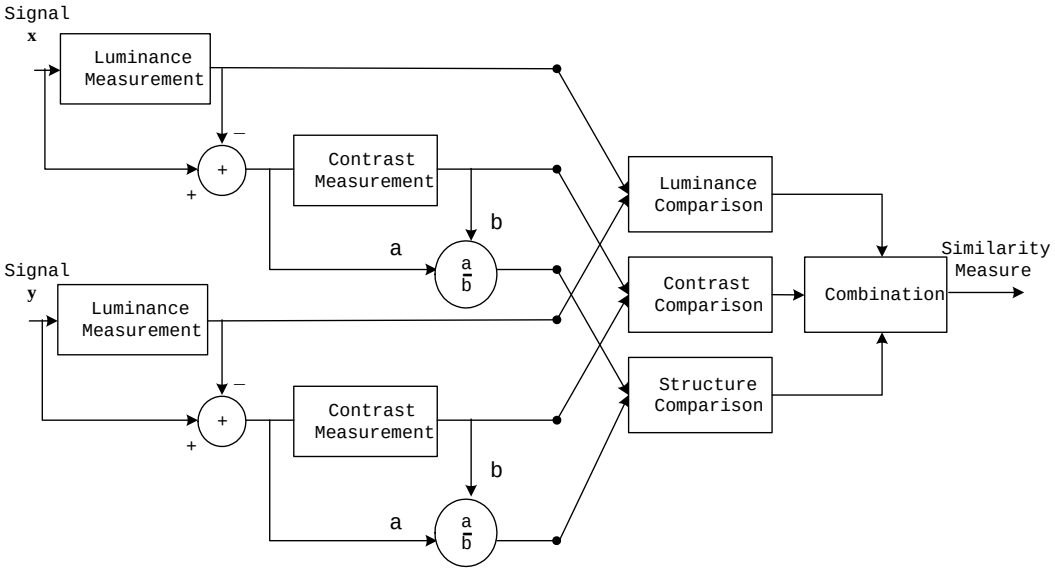


Figure 1.2. Diagram of the Structural Similarity (SSIM) Measurement System.

Figure 1.3. Comparison of Image Similarity Measures for Images with Different Types of Distortions. Images are ordered in raster scan. (A) Reference Image (8Bits/Pixel, Assumed To Have Perfect Quality); (B) Contrast Stretch; (C) Mean Luminance Shift; (D) Gaussian Noise Contamination; (E) Impulsive Noise Contamination; (F) JPEG Compression; (G) Blurring; (H) Spatial Scaling (Zooming Out); (I) Spatial Translation (To The Right); (J) Spatial Translation (To The Left); (K) Rotation (Counterclockwise); (L) Rotation (Clockwise).

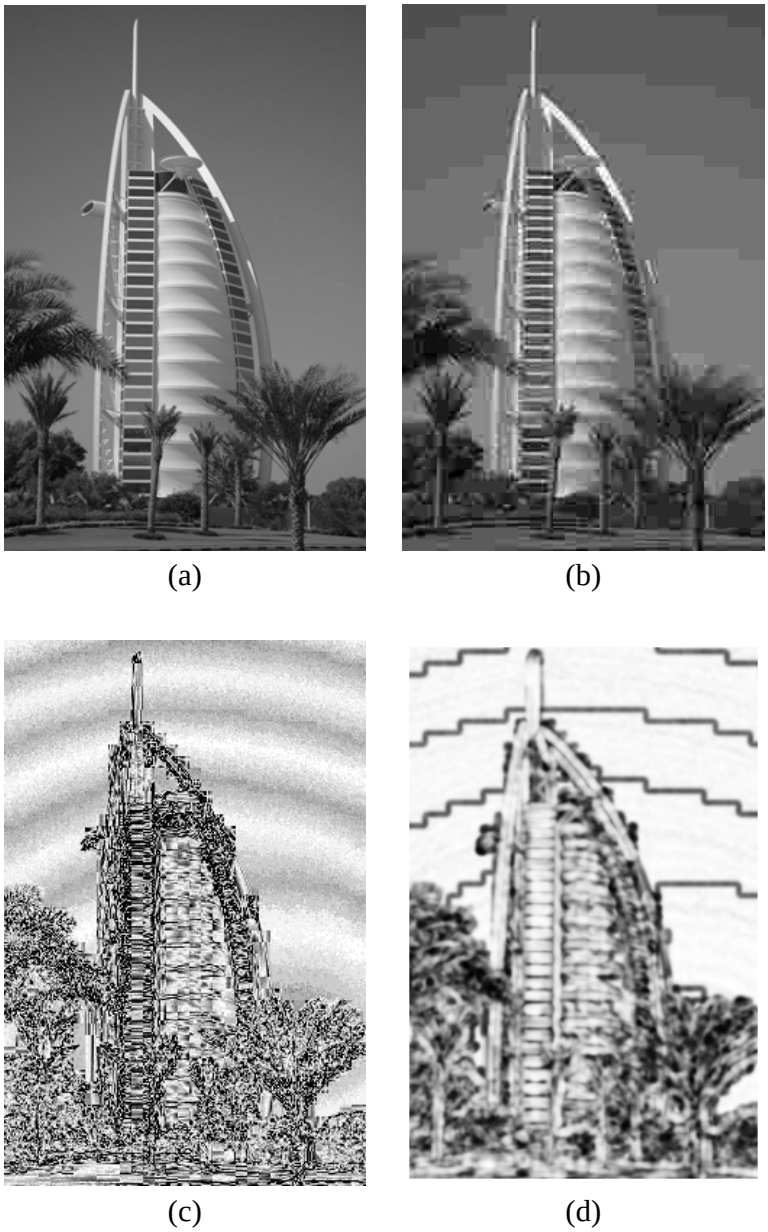


Figure 1.4. Quality/Distortion Map Example (a) Original image; (b) Distorted Image (by JPEG compression); (c) Absolute Difference Map: Brighter Indicates Better Quality (Smaller Absolute Difference between the Original and the Distorted Images); (d) SSIM Index Map: Brighter Indicates Better Quality (Larger SSIM Value).

CHAPTER 2

IMAGE QUALITY ASSESSMENT POOLING STRATEGIES

In this chapter, we discuss and test a variety of spatial pooling algorithms for perceptual image quality assessment. We first state our motivation for the design and development of these algorithms, and then present detailed algorithms and the testing procedures. Finally, we discuss the experimental results of our test.

2.1 Motivation

Many recently proposed perceptual image quality assessment algorithms are implemented in two stages. In the first stage, image quality is evaluated within local regions. This results in a quality/distortion map over the image space. In the second stage, a spatial pooling algorithm is employed that combines the quality/distortion map into a single quality score. While great effort has been devoted to developing algorithms for the first stage, little investigation and careful comparison have been devoted to developing and testing spatial pooling methods. In practice, spatial pooling has often been treated superficially, e.g., using a simple spatial average. Some methods incorporate human interactions or automatic object detections and segmentations to define the regions-of-interest or points-of-fixations before spatial pooling, but these methods may not be easily applied to general-purpose image quality assessment because for many images, it may not always be easy to find obviously outstanding objects that attract visual attention. On the other hand, problems arise with the direct spatial average approach when the distortion is highly non-uniform over the image space. For example, when only a small region in an image is corrupted with extremely annoying artifacts, but all other regions

have high quality, human subjects tend to pay more attention to the low quality region and give an overall quality score lower than the average of the quality/distortion map.

Perhaps the simplest way to do pooling is to average all the samples in the quality/distortion map. The first question here is that whether each sample in the quality/distortion map has equal contribution to perceptual image quality. The answer is most likely no, given the example. The second question that follows is how to determine the weights assigned to each sample in the quality/distortion map. This is not a trivial problem and motivates us to study different strategies for spatial pooling.

2.2 Algorithm Description

2.2.1 Minkowski Pooling

Let m_i be the quality/distortion value at the i-th spatial location in the quality/distortion map. Minkowski pooling is defined as follows:

$$M = (\sum_{i=1}^{N} m_i^{\beta})^{\frac{1}{\beta}} \tag{2.1}$$

where N is the number of samples in the quality/distortion map and β is a constant exponent (typically chosen to lie between 1 and 4 in the literature of image quality assessment).

Without losing the generality and to make the expression easy to work with, we adopt the following expression:

$$M = \frac{1}{N} \sum_{i=1}^{N} m_i^{p} \tag{2.2}$$

where N is the number of samples in the quality/distortion map, and p is the Minkowski power. As a special case, when m_i represents the absolute difference as in Figure 1.4(c), then Eq.2.2 is directly related to the l_p norm (subject to a normalization constant). In

particular, when p = 1, it reduces to the mean absolute error (MAE). When p = 2, it becomes the widely used MSE. As p increases, more and more emphasis will be put at the image regions that have high distortions. It is often conjectured that an appropriate value of p should provide a reasonable estimation of how humans rate image quality.

2.2.2 Local Quality/Distortion-Weighted Pooling

The non-uniform quality distribution problem may also be solved more directly by assigning spatially varying importance (weights) over the image space. A general form of such a spatial weighting approach is given by

$$M = \frac{\sum_{i=1}^{N} w_i m_i}{\sum_{i=1}^{N} w_i} \tag{2.3}$$

where w_i is the weight assigned to the i-th spatial location. The idea of local quality/distortion-weighted pooling is to define the weight w_i by the local quality measure m_i itself, i.e.,

$$w_i = f(m_i) \tag{2.4}$$

For example, in the case that m_i represents a distortion measure (higher value indicates higher distortion) and we would like to put more emphasis on the spatial locations where the image quality is extremely bad, then we would choose $f(\cdot)$ to be a monotonically increasing function. On the other hand, if m_i is a quality measure (higher value indicates better quality), then we would prefer $f(\cdot)$ to be a monotonically decreasing function.

2.2.3 Information Content-Weighted Pooling

In information content-weighted pooling, a similar spatial weighting method as in Eq. 2.3 is employed. However, the weights are determined by the local image content (of either or both of the reference and the distorted images), rather than the measured local quality/distortion. Let x_i and y_i be the local image patches (e.g., a collection of pixels in a local window) extracted around the i-th spatial location from the reference and the distorted images, respectively. The weight w_i is computed using a function

$$w_i = g(x_i, y_i) \tag{2.5}$$

The local energy-weighted pooling method proposed in [37] may be considered as a special case of this approach, where the weighting function is given by

$$g(x, y) = \sigma_x^2 + \sigma_y^2 + C \tag{2.6}$$

Here σ_x and σ_y are the standard deviations of x and y, respectively, and C is a constant representing a baseline minimal weight. The underlying justification of using Eq. 2.6 is that the high-energy (or high-variance) image regions are likely to contain more information. If the ultimate goal of visual perception is to efficiently extract useful information from the visual scene, then the high energy regions are more likely to attract visual attention, and thus should be given more importance. While this general idea is well motivated, the specific formulation of Eq. 2.6 is not directly an information measure based on any statistical model. Here we propose a new method, in which the perceived local information content is quantified as the number of bits that can be received from a statistical image information source that passes through a noisy visual channel. To keep the algorithm tractable, we assume a local Gaussian source model and an additive Gaus-

sian channel model. Similar information communication-based models have been used previously for image quality assessment [38], though not involved in the spatial pooling stage. Assume that the source power is S and the channel noise power is C_{noise} (which is considered as an estimate of the intrinsic noise in the visual system [38]). A well-known result from information theory is that the received information can be computed as

$$I = \frac{1}{2}\log(1 + \frac{S}{C}) \tag{2.7}$$

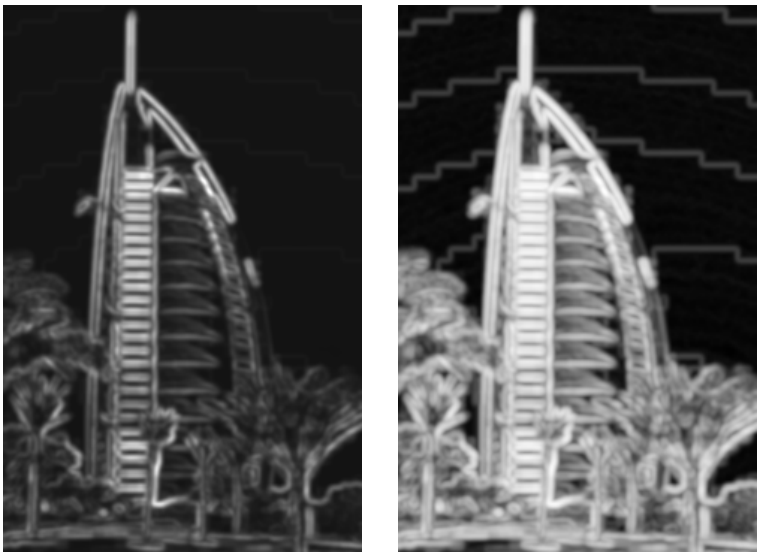
Now assume that the source power of a local image patch x can be estimated as σ_x^2, and the channel noise variance is a known parameter (as in [38]). Then the weighting function is given by

$$g(x, y) = \log[(1 + \frac{\sigma_x^2}{C})(1 + \frac{\sigma_y^2}{C_{noise}})] \tag{2.8}$$

Here we have removed the front scalar constant, which has no effect on the final pooling result because of the normalization in Eq. 2.3. We have also added the information content of both the reference and the distorted image patches, so as to make the algorithm symmetric. Figure 2.1 gives an example of an information content-based weighting function over the image space, which is computed for the images shown in 1.4. As in [7], in the computation of local σ_x^2 and σ_y^2, a sliding Gaussian window with standard deviation of 1.5 pixels is employed.

2.3 Experiment Result

We test the objective image quality measures with different spatial pooling approaches using the LIVE database (developed at the Laboratory for Image and Video Engineering at The University of Texas at Austin) [36], which contains seven subject-rated data sets, including two data sets for JPEG 2000 compression (contains 87 and



(a) (b)

Figure 2.1. Local Information Content Calculated Weighting Function of Images (a) and (b) in Figure 1.4. (a) Computed Using Eq. 2.6; (b) Computed Using Eq.2.8.

82 images, respectively), two for JPEG compression (contains 87 and 88 images, respectively), one for white Gaussian noise contamination (145 images), one for Gaussian blur (145 images), and one for transmission errors of JPEG 2000 compressed images (145 images). For each objective quality measure being evaluated, we report the Spearman rank order correlation coefficients (ROCC) between the subjective and objective scores for each data set. The ROCC is defined as

$$r = 1 - \frac{6 \sum_{i=1}^{N} d_i^2}{O(O^2 - 1)} \tag{2.9}$$

where O is the number of images in the data set, and d_i is the difference between the ith image's ranks in subjective and objective evaluations. ROCC is one of the metrics adopted by the video quality experts group (VQEG, www.vqeg.org) for the evaluation of video quality measures.

The image quality measures being evaluated are divided into two groups. The first group uses the absolute difference to create the distortion map, and the second group uses the SSIM index to generate the quality map.

2.3.1 Absolute Difference

2.3.1.1 Minkowski Pooling

Suppose that the original image is X and distorted image is Y. Let m_i be the quality/distortion value at the i-th spatial location in the quality/distortion map.

$$m_i = |x_i - y_i| \tag{2.10}$$

$$M = \frac{1}{N}\sum_{i=1}^{N} m_i^p \tag{2.11}$$

where N is the number of samples in the quality/distortion map, and p is the Minkowski power.

Figure 2.2 is the experiment result by using this method and changing p from $\frac{1}{8}$ to 8, i.e $\frac{1}{8}$, $\frac{1}{4}$, $\frac{1}{2}$, 1, 2, 4 and 8.

2.3.1.2 Local Quality/Distortion-Weighted Pooling

The result of $w_i = m_i^{\frac{1}{4}}$, $w_i = m_i^{\frac{1}{2}}$, $w_i = m_i^1$, $w_i = m_i^2$, $w_i = m_i^4$, $w_i = m_i^8$, $w_i = m_i^{16}$, $w_i = m_i^{32}$ in the Eq 2.4 is as figure 2.3 and figure 2.4:

From the figure, we can see that as p increases, the average performance (the bold blue line) keeps stable in all the Figures (a), (b), (c), (d), (e), (f), (g) and (h). And as the weight increases, the performance increases.

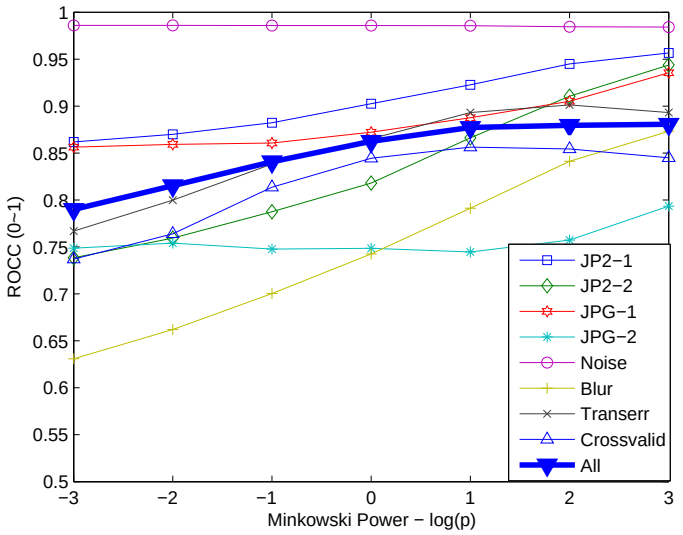


Figure 2.2. Absolute Difference Minkowski Pooling.

2.3.1.3 Information Content-Weighted Pooling

The idea of information content - weighted pooling is to define the weight w_i by the information content measure m_i itself, i.e. $w_i = g(m_i)$.

In the first approach, we set

$$g(x, y) = \sigma_x^2 + \sigma_y^2 + C \tag{2.12}$$

Here σ_x and σ_y are the standard deviations of x and y, respectively, and C is a constant representing a baseline minimal weight. The experiment result of this approach is shown in Figure 2.5.

In the second approach, we set

$$g(x, y) = (\sigma_x^2 + \sigma_y^2 + C)^{\frac{1}{2}} \tag{2.13}$$

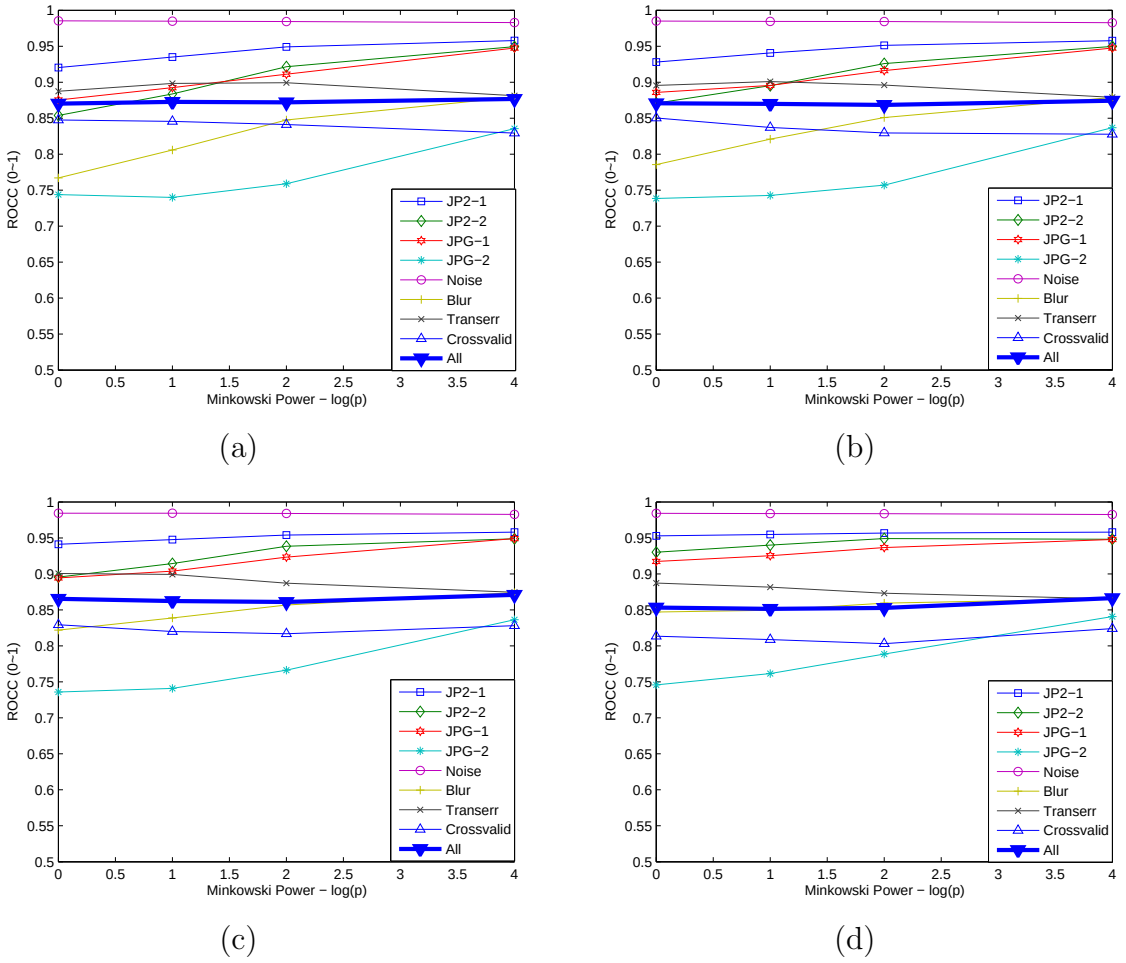


Figure 2.3. Absolute Difference Local Quality/Distortion-Weighted Pooling (a) $w_i = m_i^{\frac{1}{4}}$; (b) $w_i = m_i^{\frac{1}{2}}$; (c)$w_i = m_i^1$; (d)$w_i = m_i^2$.

Here σ_x and σ_y are the standard deviations of x and y, respectively, and C is a constant representing a baseline minimal weight (It may be the same value as in Eq. 2.5). The experiment result of this approach is shown in Figure 2.6.

In the third approach, we set

$$g(x, y) = \log[(1 + \frac{\sigma_x^2}{C}) + (1 + \frac{\sigma_y^2}{C})] \tag{2.14}$$

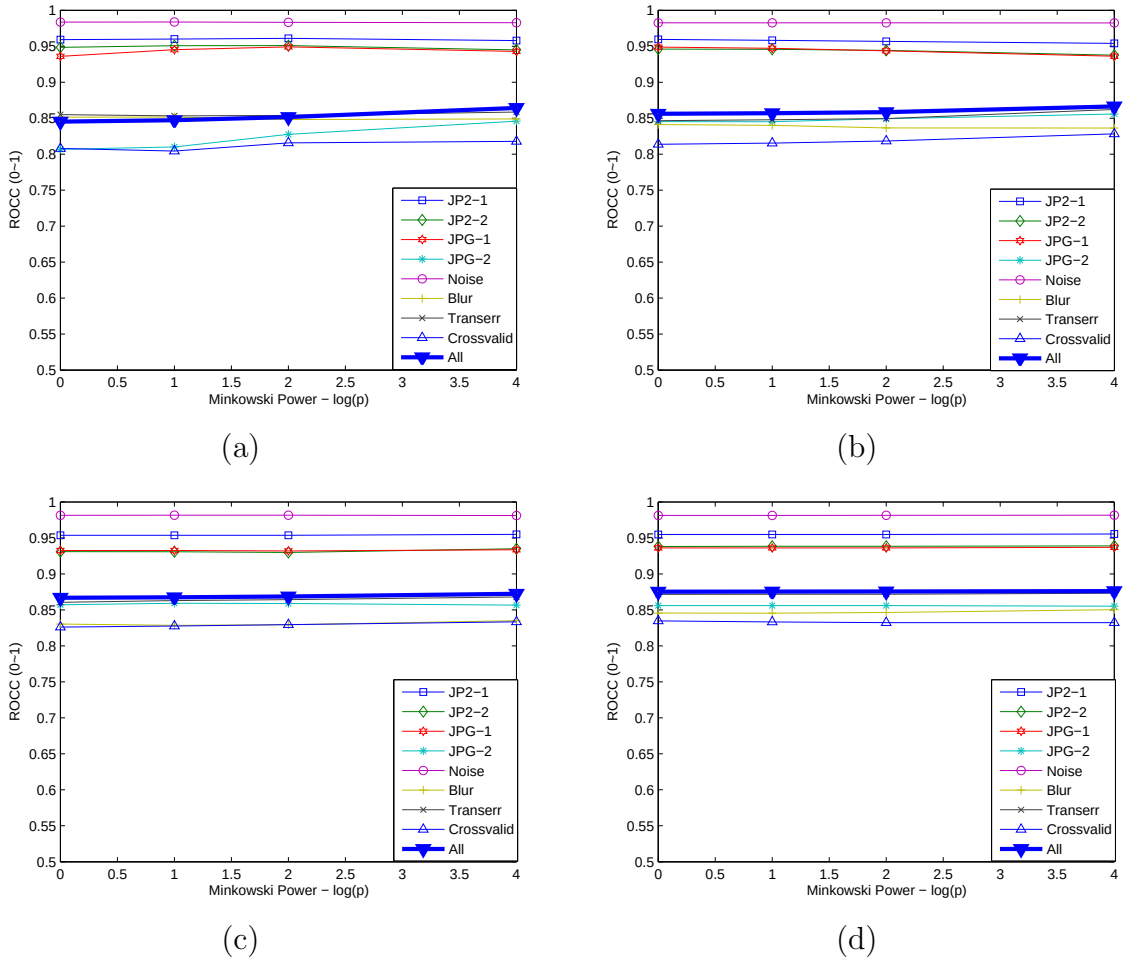


Figure 2.4. Continue with Fig. 2.3 (a) $w_i = m_i^4$; (b)$w_i = m_i^8$; (c)$w_i = m_i^{16}$; (d)$w_i = m_i^{32}$.

Here σ_x and σ_x are the standard deviations of x and y, respectively, and C is a constant representing a baseline minimal weight (It may be the same value as in Eq. 2.5). The experimental result of this approach is shown in Figure 2.7.

Figures 2.5, 2.6, 2.7 show that as p increases, the average performance may decrease a little bit or keep stable. Comparing these three figures, we find that the third approach can get better performance.

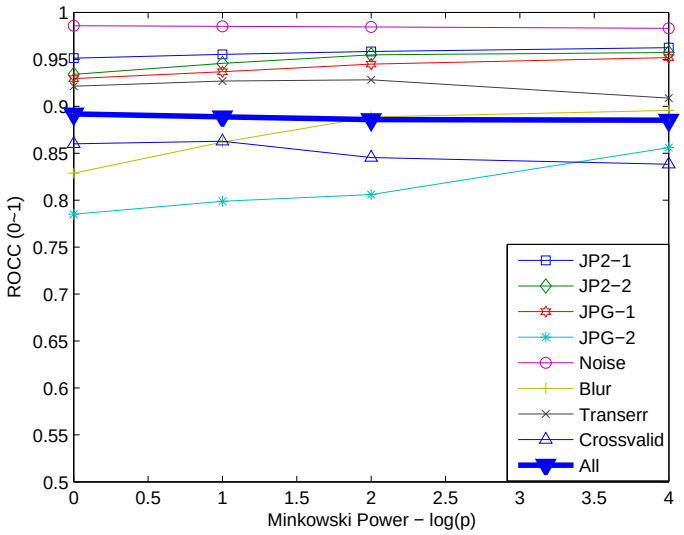


Figure 2.5. Absolute Difference Information Content - Weighted Pooling - $g(x,y) = \sigma_x^2 + \sigma_y^2 + C$.

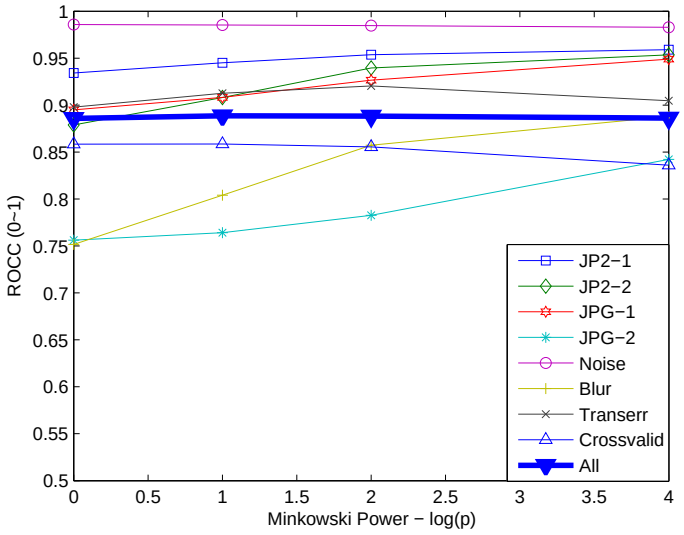


Figure 2.6. Absolute Difference Information Content - Weighted Pooling - $g(x,y) = (\sigma_x^2 + \sigma_y^2 + C)^{\frac{1}{2}}$.

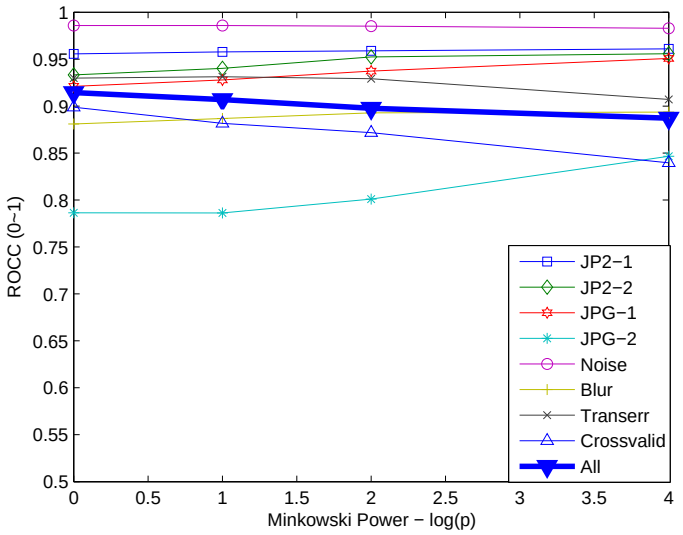


Figure 2.7. Absolute Difference Information Content - Weighted Pooling - $g(x,y) = \log[(1+\frac{\sigma_x^2}{C})+(1+\frac{\sigma_y^2}{C})]$.

2.3.2 SSIM Quality Map

In SSIM difference approach, SSIM algorithm is used to generate the quality/distortion map, instead of using absolute difference. All the other method are the same as section 2.3.1.

2.3.2.1 Minkowski Pooling

Figure 2.8 is the experimental results by using Minkowski pooling with p ranging from $\frac{1}{8}$ to 8, i.e. $\frac{1}{8}$, $\frac{1}{4}$, $\frac{1}{2}$, 1, 2, 4 and 8. Figure 2.8 shows that as p increases, the average performance decreases.

2.3.2.2 Local Quality/Distortion Weighted Pooling

Figure 2.9 is the experimental result by using local quality/distortion-weighted pooling with p ranging from 1 to 4 for $w_i = m_i^1$, $w_i = m_i^2$, $w_i = m_i^4$. From section

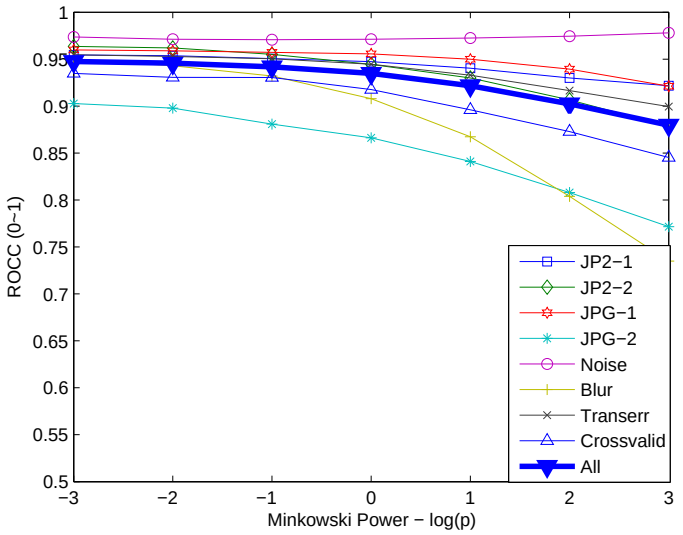


Figure 2.8. SSIM Minkowski Pooling(p is from 1 to 4).

2.3.1, we find that as p changes, the overall performance keeps almost stable. And for $w_i = f(m_i)$, as the poser of m_i changes, the average value changes insignificantly. So here we only need to test p and the power of m_i from 1 to 4. The Figure 2.9 shows that as p increases, the average performance decreases.

2.3.2.3 Information Weighted Pooling

Figure 2.10 is the experimental result by using information content-weighted pooling with p ranging from 1 to 4. It shows that as p increases, the average performance decreases.

2.3.3 Analysis of Result

The ROCC results for the two groups of objective image quality measures are shown in Tables 2.1 and 2.2, respectively. For easy visualization, we have added a $'\wedge'$, $'\vee'$ or $'-'$ behind each ROCC number to indicate an increase/decrease/no significant-change of

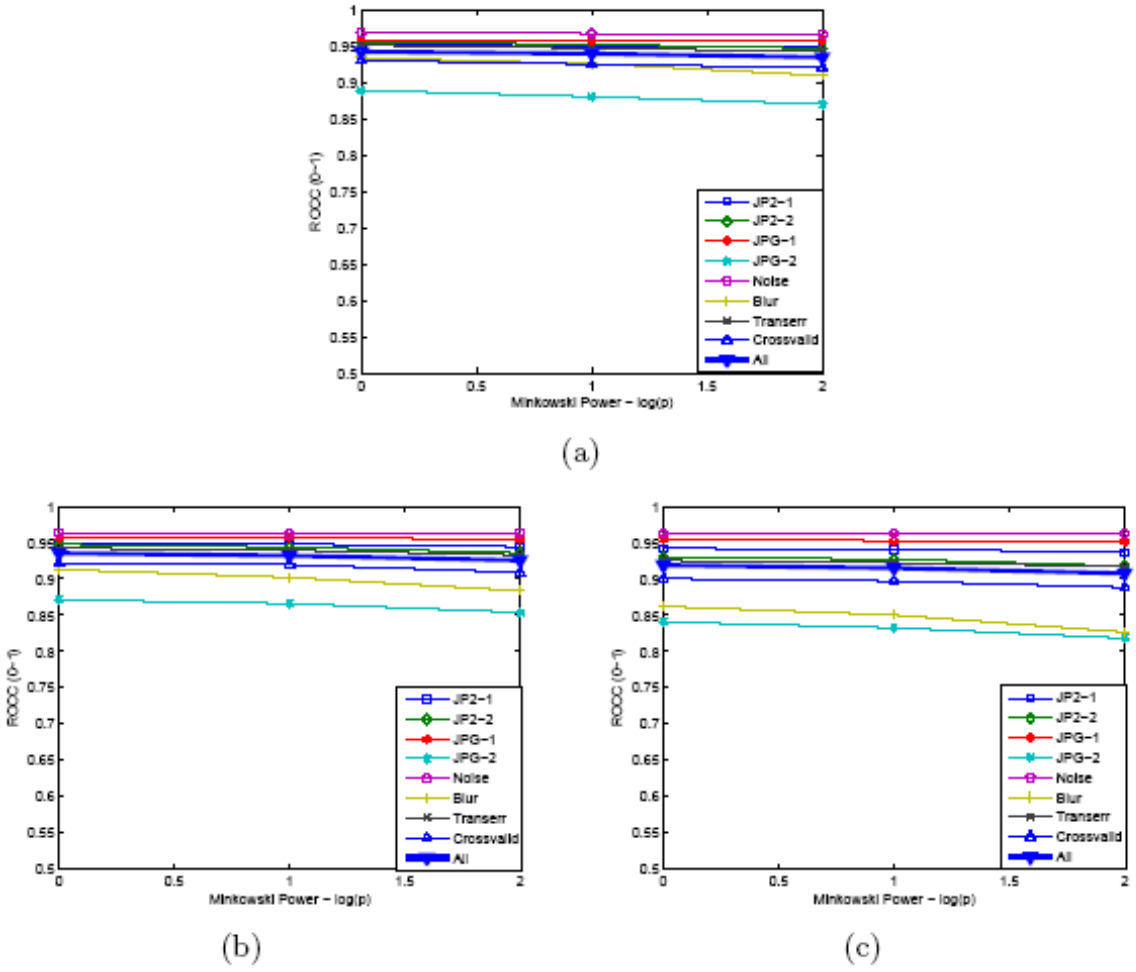


Figure 2.9. SSIM Local Weighted Pooling (a)$w_i = m_i^1$; (b)$w_i = m_i^2$; (c)$w_i = m_i^4$.

ROCC value as compared to the baseline ROCC (given by spatial average pooling). We have also added a final column that gives the average improvement of ROCC values over the baseline. It can be observed that all three pooling strategies may lead to improvement of quality prediction performance. However, the best parameter choices of the Minkowski pooling methods and the local quality/distortion-weighted pooling methods depend on the underlying specific local quality/distortion measure. For example, Minkowski pooling with $p = 4$ results in improvement when the local quality/distortion measure is the

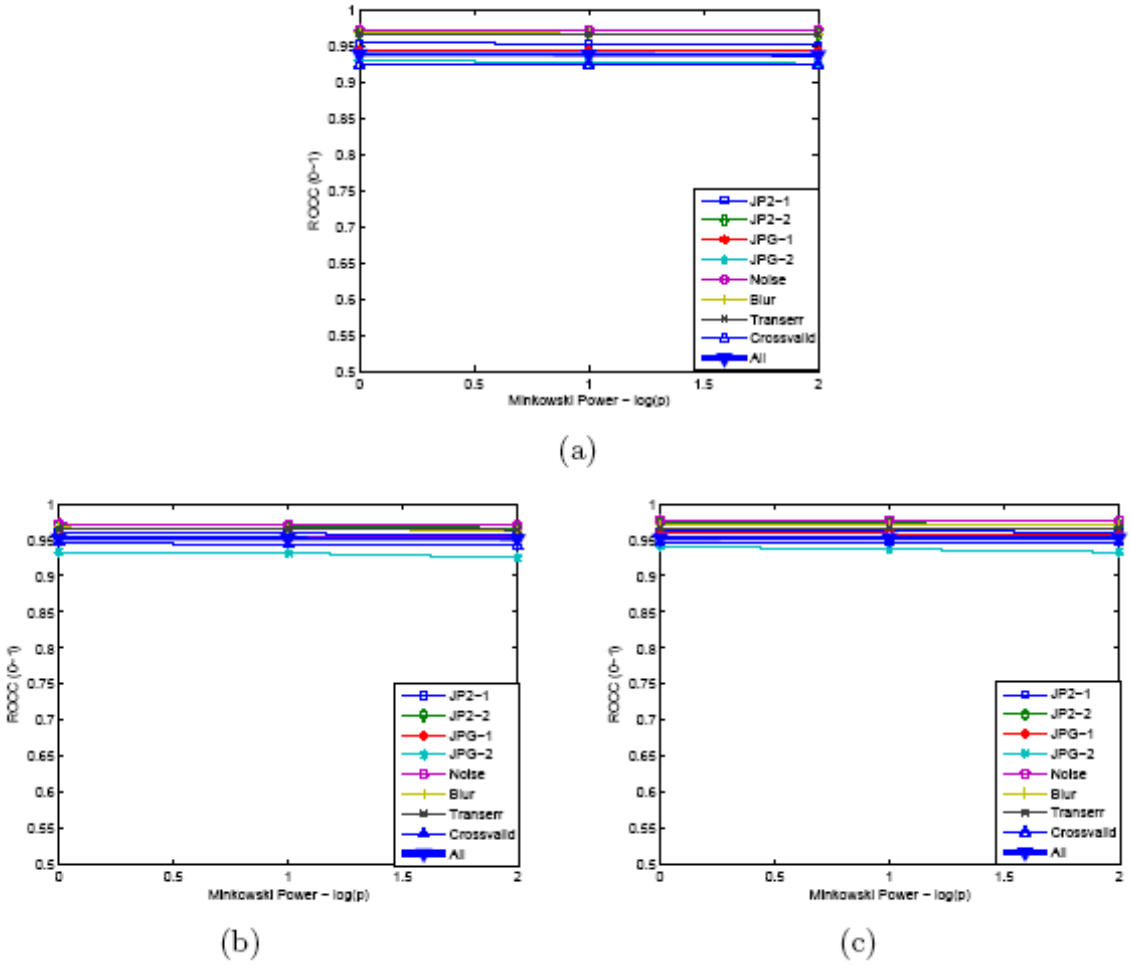


Figure 2.10. SSIM Information Weighted Pooling: (a) Eq.2.5; (b) Eq.2.6; (c) Eq.2.7.

absolute difference, but not the SSIM index. Comparatively, the information content-weighted pooling method, especially when the newly proposed Eq. 2.14 is used as the weighting function, appears to be more stable and general. It results in consistent and most of the time significant improvement over a wide range of image distortion types for both cases of local quality/distortion measures.

Table 2.1. Performance comparison of spatial pooling methods. The absolute difference is used to generate the distortion map. JP2: JPEG2000 dataset; JPG: JPEG; Noise: white Gaussian noise; Blur: Gaussian blur; Error: transmission error; AI: average improvement; PS: pooling strategy; SA: spatial average; MP: Minkowski pooling; LQDWP: local quality distortion-weighted pooling; ICWP: info. content-weighted pooling; $'\wedge'$: increase of ROCC value compared with spatial average pooling; $'\vee'$: decrease of ROCC value compared with spatial average pooling; $'-'$: no significance change of ROCC value compared with spatial average pooling

method			LIVE dataset / ROCC result							
PS	p	w_i	JP2-1	JP2-2	JPG-1	JPG-2	Noise	Blur	Error	AI
SA	1	1	0.9026	0.8180	0.8722	0.7485	0.9857	0.7425	0.8651	0
MP	1/8	1	0.8619$\vee$	0.7384$\vee$	0.8562$\vee$	0.7487$-$	0.9860$-$	0.6309$\vee$	0.7669$\vee$	-0.0494
	1/4	1	0.8700$\vee$	0.7595$\vee$	0.8593$\vee$	0.7541$\wedge$	0.9860$-$	0.6621$\vee$	0.7998$\vee$	-0.0348
	1/2	1	0.8823$\vee$	0.7875$\vee$	0.8607$\vee$	0.7478$-$	0.9858$-$	0.7003$\vee$	0.8386$\vee$	-0.0188
	2	1	0.9227$\wedge$	0.8662$\wedge$	0.8876$\wedge$	0.7446$\vee$	0.9856$-$	0.7921$\wedge$	0.8931$\wedge$	$+0.0225$
	4	1	0.9449$\wedge$	0.9105$\wedge$	0.9052$\wedge$	0.7573$\wedge$	0.9845$-$	0.8413$\wedge$	0.9012$\wedge$	$+0.0443$
	8	1	0.9566$\wedge$	0.9438$\wedge$	0.9355$\wedge$	0.7934$\wedge$	0.9843$-$	0.8731$\wedge$	0.8931$\wedge$	$+0.0636$
LQ-DWP	1	$\lvert m_i\rvert^{1/8}$	0.9152$\wedge$	0.8431$\wedge$	0.8721$-$	0.7479$-$	0.9855$-$	0.7536$\wedge$	0.8840$\wedge$	$+0.0095$
	1	$\lvert m_i\rvert^{1/4}$	0.9204$\wedge$	0.8539$\wedge$	0.8753$\wedge$	0.7438$\vee$	0.9853$-$	0.7671$\wedge$	0.8875$\wedge$	$+0.0141$
	1	$\lvert m_i\rvert^{1/2}$	0.9280$\wedge$	0.8709$\wedge$	0.8858$\wedge$	0.7385$\vee$	0.9849$-$	0.7856$\wedge$	0.8956$\wedge$	$+0.0221$
	1	$\lvert m_i\rvert^{1}$	0.9412$\wedge$	0.8956$\wedge$	0.8944$\wedge$	0.7359$\vee$	0.9844$-$	0.8218$\wedge$	0.9006$\wedge$	$+0.0342$
	1	$\lvert m_i\rvert^{2}$	0.9529$\wedge$	0.9302$\wedge$	0.9173$\wedge$	0.7457$-$	0.9841$-$	0.8470$\wedge$	0.8873$\wedge$	$+0.0471$
	1	$\lvert m_i\rvert^{4}$	0.9592$\wedge$	0.9485$\wedge$	0.9360$\wedge$	0.8068$\wedge$	0.9836$-$	0.8514$\wedge$	0.8550$\vee$	$+0.0580$
	1	$\lvert m_i\rvert^{8}$	0.9594$\wedge$	0.9461$\wedge$	0.9487$\wedge$	0.8453$\wedge$	0.9826$\vee$	0.8412$\wedge$	0.8466$\vee$	$+0.0622$
IC-WP	1	Eq. (2.13)	0.9512$\wedge$	0.9341$\wedge$	0.9294$\wedge$	0.7850$\wedge$	0.9858$-$	0.8287$\wedge$	0.9214$\wedge$	$+0.0573$
	1	Eq. (2.14)	0.9556$\wedge$	0.9332$\wedge$	0.9210$\wedge$	0.7864$\wedge$	0.9859$-$	0.8809$\wedge$	0.9299$\wedge$	$\mathbf{+0.0655}$

Table 2.2. Performance comparison of spatial pooling methods. The SSIM index is used to generate the quality map. JP2: JPEG2000 dataset; JPG: JPEG; Noise: white Gaussian noise; Blur: Gaussian blur; Error: transmission error; AI: average improvement; PS: pooling strategy; SA: spatial average; MP: Minkowski pooling; LQDWP: local quality distortion-weighted pooling; ICWP: info. content-weighted pooling; $'\wedge'$: increase of ROCC value compared with spatial average pooling; $'\vee'$: decrease of ROCC value compared with spatial average pooling; $'-'$: no significance change of ROCC value compared with spatial average pooling

method			LIVE dataset / ROCC result							
PS	p	w_i	JP2-1	JP2-2	JPG-1	JPG-2	Noise	Blur	Error	AI
SA	1	1	0.9545	0.9636	0.9598	0.9028	0.9737	0.9497	0.9546	0
MP	1/8	1	0.9549−	0.9660−	0.9609−	0.9069∧	0.9777∧	0.9559∧	0.9554−	+0.0027
	1/4	1	0.9547−	0.9652−	0.9608−	0.9063∧	0.9768∧	0.9552∧	0.9554−	+0.0022
	1/2	1	0.9542−	0.9642−	0.9605−	0.9035−	0.9755−	0.9531∧	0.9551−	+0.0011
	2	1	0.9537−	0.9620−	0.9589−	0.8978∨	0.9712−	0.9430∨	0.9529−	−0.0027
	4	1	0.9506∨	0.9551∨	0.9573−	0.8808∨	0.9707∨	0.9321∨	0.9505∨	−0.0088
	8	1	0.9473∨	0.9443∨	0.9556∨	0.8662∨	0.9712−	0.9078∨	0.9447∨	−0.0174
LQ-DWP	1	$\lvert m_i\rvert^{-1/8}$	0.9541−	0.9637−	0.9600−	0.9023−	0.9743−	0.9513−	0.9550−	+0.0003
	1	$\lvert m_i\rvert^{-1/4}$	0.9544−	0.9642−	0.9605−	0.9030−	0.9755−	0.9537∧	0.9552−	+0.0011
	1	$\lvert m_i\rvert^{-1/2}$	0.9552−	0.9661−	0.9609−	0.9083∧	0.9779∧	0.9566∧	0.9551−	+0.0031
	1	$\lvert m_i\rvert^{-1}$	0.9577∧	0.9698∧	0.9606−	0.9114∧	0.9825∧	0.9603∧	0.9492∨	+0.0047
	1	$\lvert m_i\rvert^{-2}$	0.9617∧	0.9708∧	0.9613−	0.9096∧	0.9849∧	0.9640∧	0.9381∨	+0.0045
	1	$\lvert m_i\rvert^{-4}$	0.9638∧	0.9678∧	0.9627−	0.8527∨	0.9592∨	0.9603∧	0.8797∨	−0.0161
	1	$\lvert m_i\rvert^{-8}$	0.9673∧	0.9615−	0.9629∧	0.8664∨	0.9584∨	0.9507−	0.8668∨	−0.0178
IC-WP	1	Eq. (2.13)	0.9535−	0.9671∧	0.9439∨	0.9288∧	0.9723−	0.9672∧	0.9662∧	+0.0058
	1	Eq. (2.14)	0.9612∧	0.9743∧	0.9591−	0.9401∧	0.9776∧	0.9716∧	0.9659∧	**+0.0130**

CHAPTER 3

SSIM-GUIDED PERCEPTUAL IMAGE COMPRESSION

In this chapter, we present our work on structural similarity-guided perceptual image compression. To help readers understand our algorithm more clearly, we first make a brief introduction on SPIHT and JPEG2000 encoding schemes and their rate distortion systems. We then describe our perceptual coding schemes by incorporating the SSIM index with the SPIHT and JPEG2000 coding algorithms. Experimental results demonstrates the effectiveness of our algorithms.

3.1 Motivation

Currently available JPEG2000 software (VM [9], JASPER [10] and JJ2000 [11]) all adopt the rate-based MSE minimization encoding approach [12]. As we noted above, MSE is a poor model in perceptual coding. Because the human eyes are the ultimate receiver in most real world applications, it would be preferable to take the properties of the human visual system (HVS) into considerations. Since SSIM provides a much better indication of perceptual image quality, it is desirable to use it as a new image quality criteria in the optimization of image compression algorithms such as SPIHT and JPEG2000.

Other perceptual image distortion metrics (e.g., [39]) have been used to control the bit rate in wavelet image compression. However, most of these methods are not compatible with the standard base-line decoder. It means that we have to modify each decoder in order to accommodate these algorithms. This highly limits the application scope of these methods, because in many real world environment such as digital video

broadcasting, it is impossible to modify all existing codecs. Therefore, here we attempt to develop a perceptual coding scheme that is compatible with standard baseline decoders.

3.2 SPIHT Encoding Scheme

3.2.1 Progressive Image Transmission

An original image is denoted by a set of pixel values $p_{i,j}$, where (i, j) is the pixel coordinate. The coding is actually done to the array [24]

$$c = \Omega(p) \tag{3.1}$$

where $\Omega(\cdot)$ represents a hierarchical subband transformation. The 2-D array c has the same dimensions of p, and each element $c_{i,j}$ is called transform coefficient at coordinate (i, j) . For the purpose of coding, we assume that each $c_{i,j}$ is represented with a fixed-point binary format, with a small number of bits, typically 16 or less, and be treated as an integer.

A major objective in a progressive transmission scheme is to select the most important information, which yields the largest distortion reduction, to be transmitted first. For this selection, the mean squared-error (MSE) distortion measure is used.

Information in the value of $|c_{i,j}|$ can also be ranked according to its binary representation, and the most significant bits should be transmitted first. This idea is used, for example, in the bit-plane method for progressive transmission [40].

3.2.2 Transmission of the Coefficient Values

Assume that the coefficients are ordered according to the minimum number of bits required for its magnitude binary representation, that is, ordered according to a one-to-one mapping $\eta : I \rightarrow I^2$, such that

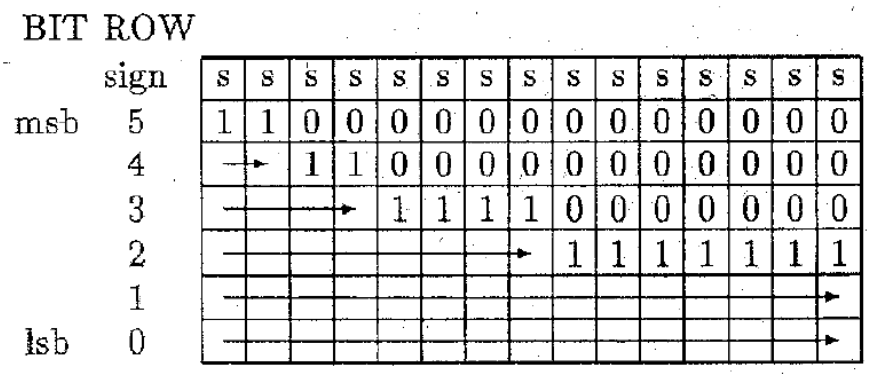


Figure 3.1. Binary Representation Of The Magnitude-Ordered Coefficient.

$$\lfloor \log_2 |c_{\eta(k)}| \rfloor \geq \lfloor \log_2 |c_{\eta(k+1)}| \rfloor \tag{3.2}$$

Figure 3.1 shows the schematic binary representation of a list of magnitude - ordered coefficients. Each column k in Figure 3.1 contains the bits of $c_{\eta(k)}$. The bits in the top row indicate the sign of the coefficient. The rows are numbered from the bottom up, and the bits in the lowest row are the least significant.

The progressive transmission method outlined above can be implemented with the following algorithm to be used by the encoder.

1) Output $n = \lfloor \log_2(max_{(i,j)}\{|c_{i,j}|\}) \rfloor$ to the decoder;

2) Output μ_n, followed by the pixel coordinates $\eta(k)$ and sign of each of the μ_n coefficients such that $2^n \leq |c_{\eta(k)}| < 2^{n+1}$;

3) Output the nth most significant bit of all the coefficients with $|c_{i,j}| \geq 2^{n+1}$ (i.e., those that had their coordinates transmitted in previous sorting passes), in the same order used to send the coordinates (refinement pass);

4) Decrement n by one, and go to Step 2).

The algorithm stops at the desired rate or distortion. Normally, good quality images can be recovered after a relatively small fraction of the pixel coordinates are transmitted.

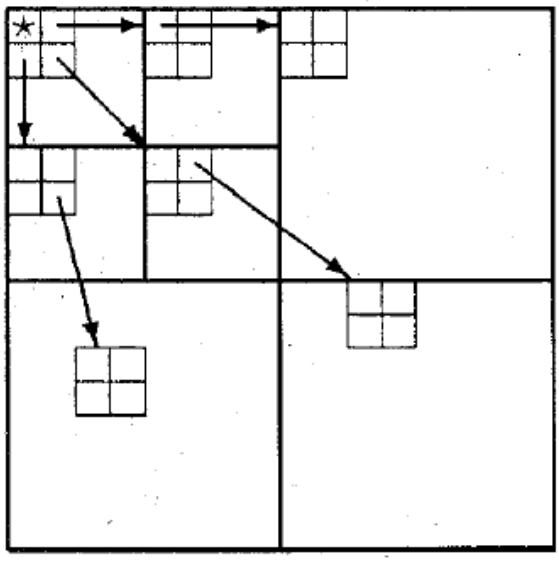


Figure 3.2. Examples Of Parent-Offspring Dependencies In The Spatial-Orientation Tree.

3.2.3 Spatial Orientation Trees

Normally, most of an image's energy is concentrated in the low frequency components. Consequently, the variance decreases as we move from the highest to the lowest levels of the subband pyramid. Furthermore, it has been observed that there is a spatial self-similarity between subbands, and the coefficients are expected to be better magnitude-ordered if we move downward in the pyramid following the same spatial orientation.

A tree structure, called spatial orientation tree, naturally defines the spatial relationship on the hierarchical pyramid. Figure 3.2 shows how our spatial orientation tree is defined in a pyramid constructed with recursive four-subband splitting. Each node of the tree corresponds to a pixel and is identified by the pixel coordinate. Its direct descendants (offspring) correspond to the pixels of the same spatial orientation in the next finer level of the pyramid.

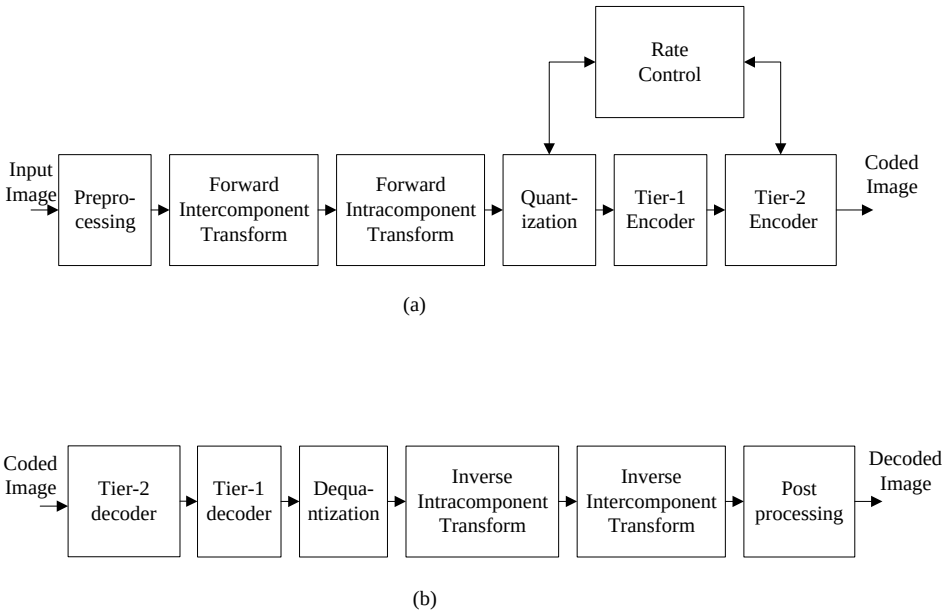


Figure 3.3. JPEG2000 Codec Structure: (a)Encoder (b)Decoder.

3.3 JPEG2000 Codec Structure

The general structure of the codec is shown in Figure 3.3 with the form of the encoder given by Figure 3.3(a) and the decoder given by Figure 3.3(b). From these diagrams, the key processes associated with the codec can be identified: 1) preprocessing/ postprocessing, 2) intercomponent transform, 3) intracomponent transform, 4) quantization/dequantization, 5) tier-1 coding, 6) tier-2 coding, and 7) rate control. The decoder structure essentially mirrors that of the encoder. That is, with the exception of rate control, there is a one-to-one correspondence between functional blocks in the encoder and the decoder [25].

3.3.1 Preprocessing

A nominal dynamic range that is approximately centered about zero of input data is expected by each codec. The preprocessing stage of the encoder is to ensure that. Assuming a particular component has P bits/pixel, each pixel leads to a nominal dynamic

range of $[-2^{P-1}, 2^{P-1}-1]$ or $[0, 2^{P-1}]$ for signed or unsigned data respectively. For unsigned sample values, the nominal dynamic range is clearly not centered about zero. The opposite case is for signed sample values. If the given sample data can not met this expectation, adjustment will be carried out. The postprocessing stage of the decoder essentially reverses the preprocessing in the encoder.

3.3.2 Intercomponent Transform

The intercomponent transform stage is performed after preprocessing stage in the forward. Such a transform operates on all of the components data, and serves to reduce the correlation between components, leading to improved coding efficiency. Baseline JPEG2000 codec defines two intercomponent transforms: Irreversible Color Transform (ICT) and Reversible Color Transform (RCT). ICT is nonreversible and real-to-real in nature, while the RCT is reversible and integer-to-integer. Both of these transforms essentially map image data from the RGB to YCrCb color space. The components on which they operate must be sampled at the same resolution (i.e., have the same size). As a result, the ICT and RCT can be employed only when the images to be coded have at least three components, and the first three components are sampled at the same resolution. The ICT can only be used in the case of lossy coding, while the RCT can be used in either lossy or lossless case. The ICT is nothing more than the classic RGB to YCrCb color space transform. The forward transform is defined as

$$\begin{pmatrix} V_0(x,y) \\ V_1(x,y) \\ V_2(x,y) \end{pmatrix} = \begin{pmatrix} 0.299 & 0.587 & 0.114 \\ -0.16875 & -0.33126 & 0.5 \\ 0.5 & -0.41869 & -0.08131 \end{pmatrix} \begin{pmatrix} U_0(x,y) \\ U_1(x,y) \\ U_2(x,y), \end{pmatrix} \tag{3.3}$$

where $U_0(x;y), U_1(x;y), and U_2(x;y)$ are the input components corresponding to the red, green, and blue color planes, respectively, and $V_0(x;y), V_1(x;y), and V_2(x;y)$ are the output components corresponding to the Y, Cr, and Cb planes, respectively. The inverse transform can be shown to be

$$\begin{pmatrix} U_0(x,y) \\ U_1(x,y) \\ U_2(x,y) \end{pmatrix} = \begin{pmatrix} 1 & 0 & 1.402 \\ 1 & -0.34413 & -0.71414 \\ 1 & -1.772 & 0 \end{pmatrix} \begin{pmatrix} V_0(x,y) \\ V_1(x,y) \\ V_2(x,y). \end{pmatrix} \tag{3.4}$$

3.3.3 Intracomponent Transform

The intercomponent transform stage is followed by the intracomponent transform stage at the encoder. In this stage, transforms that operate on individual components will be applied. The specific type of operator employed for this purpose is the wavelet transform. During the wavelet transform, a component is divided into several frequency subbands. The number of subbands is a parameter of JPEG2000. Due to the statistical properties of wavelet transform on image data, image can usually be coded more efficiently in wavelet domain than in the pixel domain.

Both reversible integer-to-integer and nonreversible real-to-real wavelet transforms are employed by the baseline codec. The basic building block for such transforms is the 1-D 2-channel perfect-reconstruction (PR) uniformly maximally - decimated (UMD) filter bank (FB) which has the general form shown in Figure 3.4. For detailed information, please refer to JPEG2000 standard.

3.3.4 Quantization/Dequantization

The resulting coefficients of intercomponent and/or intracomponent transforms on tile-component data may be quantized in the encoder. Higher compression can be

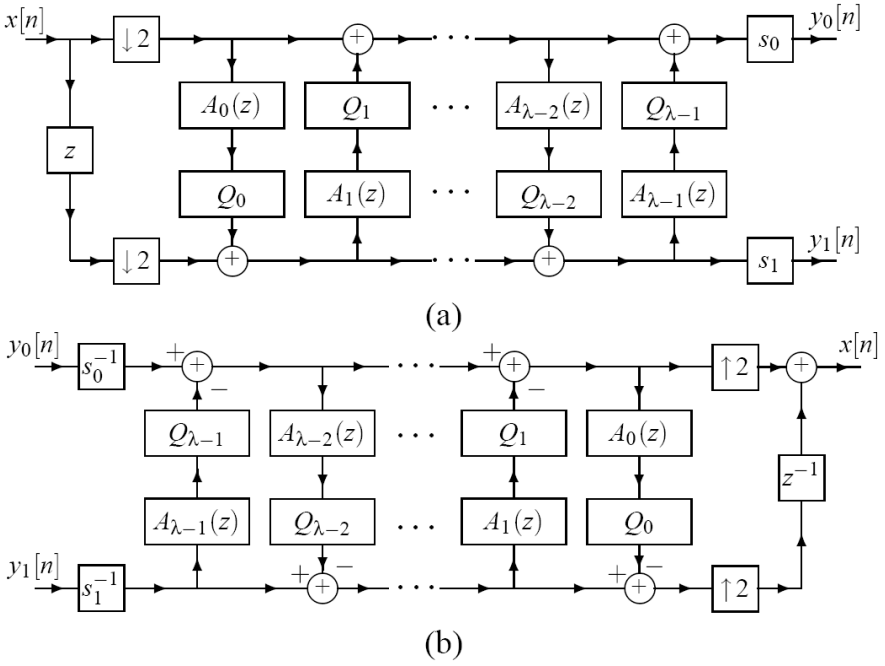


Figure 3.4. Lifting Realization of 1-D 2-Channel PR UMDFB (a) Analysis Side (b) Synthesis Side.

achieved by quantization which represents transformed coefficients with only the minimal precision required to obtain the desired level of image quality. One of the two primary sources of information loss is quantization of transform coefficients in the encoder side (The other source being the coding pass data discarded). Transformed coefficients are quantized using scalar quantization with a dead zone. Each quantizer has only one parameter, its step size. Different subband has different quantizer steps for the transformed coefficients. Mathematically, the quantization process is defined as Eq. 3.5.

$$V(x,y) = \lfloor |U(x,y)|/\Delta \rfloor sgn U(x,y) \tag{3.5}$$

where Δ is the quantizer step size, U(x, y) is the input subband signal, and V(x, y) denotes the output quantizer indices for the subband.

The baseline codec has two distinct modes of operation, referred to herein as integer mode and real mode. In integer mode, all transforms employed are integer-to-integer in nature (e.g., RCT, 5/3 WT). In real mode, real-to-real transforms are employed (e.g., ICT, 9/7 WT). In integer mode, the quantizer step sizes are always fixed at one, effectively bypassing quantization and forcing the quantizer indices and transform coefficients to be one and the same. In this case, lossy coding is still possible, but rate control is achieved by another mechanism (to be discussed later). In the case of real mode (which implies lossy coding), the quantizer step sizes are chosen in conjunction with rate control [25].

3.3.5 Tier-1 Coding

Quantization is followed by tier-1 coding. Tier-1 coding is the first of two coding stages (The other one is tier-2 coding). The quantized indices of each subband are partitioned into blocks called code blocks. Code blocks are rectangular in shape, and their nominal size is a parameter of the encoder, subject to some constraints:

1) The nominal width and height of a code block must be an integer power of two;

2) The product of the nominal width and height can not exceed 4096.

Each of the code blocks is then independently coded. The coding is performed using the bit-plane coder. For each code block, an embedded code is produced, comprised of numerous coding passes (Significance Pass, Refinement Pass and Cleanup Pass). The output of the tier-1 encoding process is, therefore, a collection of coding passes for the various code blocks [25].

3.3.6 Bit-Plane Coding

After all of the subbands have been partitioned into code blocks, a bit-plane coder is applied to each of the resulting code blocks independently. The bit-plane coding technique employed here is similar to those used in the embedded zero tree wavelet

(EZW) and set partitioning in hierarchical trees (SPIHT) codecs. However, there are still two notable differences:

1) No inter-band dependencies are exploited;

2) There are three coding passes per bit plane instead of two.

A sequence of symbols is generated by the bit-plane encoding process for each coding pass. Part or all of these symbols may be entropy coded, depending on the target bit rate. A context based adaptive binary arithmetic coder is then used for the purposes of entropy coding.

3.3.7 Tier-2 Coding

Tier-2 encoding is performed in the encoder after tier-1 encoding. The input to the tier-2 encoding process is the set of bit-plane coding passes generated during tier-1 encoding. In essence, tier-2 encoding is nothing but packetization process, in which the coding pass information is divided into data units called packets. The resulting packets are then output to the final code stream. Each packet is comprised of two parts: a header and a body. The header indicates which coding passes are included in the packet, while the body contains the actual coding pass data itself. In the code stream, the header and the body may appear together or separately, depending on the coding options in effect. Rate scalability is achieved through (quality) layers.

3.4 Rate Control and Distortion Metrics

In JPEG2000, rate control can be achieved through two distinct mechanisms: 1) the choice of quantizer step sizes, and 2) the selection of the subset of coding passes to be included in the code stream [25]. When the integer coding mode is used (i.e., when only integer-to-integer transforms are employed) only the first mechanism may be used, since the quantizer step sizes must be fixed at one. When the real coding mode is used,

then either or both of these rate control mechanisms may be employed. When the second mechanism is used, the encoder can elect to discard coding passes in order to control the rate. The encoder knows the contribution that each coding pass makes to rate, and can also calculate the distortion reduction associated with each coding pass.

The accurate rate control is achieved by the selection of the coding pass data of each code-block to be included in the code-stream [41]. In other words, the code-block bit-stream will be truncated at a particular point. JPEG2000 has no requirement on the selection of a particular rate control method. However, an optimal rate control process called PCRD optimization is recommended in the standard. This process had been described in [30]; Let $\{B_i\}_{i=1,2...}$ denote the set of code-blocks in the whole image/tile. For each code-block, an embedded bit-stream is formed by the tier-1 coding of all the bit-planes from MSB to LSB. In the bit-stream, there is a set of feasible truncation points, each of which is defined at the end of a coding pass [30]. We use n_i to identify the feasible truncation points of the ith code-block B_i, with $n_i = k$ corresponding to the kth truncation point from the MSB. For code block B_i, the bit-stream can be truncated at any feasible n_i, resulting in corresponding discrete length or bit rate $R_i^{n_i}$. The corresponding distortion incurred by reconstructing the truncated bit-stream is denoted by $D_i^{n_i}$. The rate control optimization process selects the truncation points of all code-blocks to minimize the overall reconstructed image distortion D, where

$$D = \sum_i D_i^{n_i} \tag{3.6}$$

subject to the rate constraint

$$R = \sum_i R_i^{n_i} \leq R_{budget} \tag{3.7}$$

where R_{budget} denotes the target bit rate.

Using the Lagrange multiplier technique [30], the optimization process is equivalent to minimizing the cost function

$$J = D + R\lambda = \sum_i (D_i^{n_i(\lambda)} + R_i^{n_i(\lambda)}\lambda) \tag{3.8}$$

Therefore, if we can find a value of λ such that the resulting set of truncation points $\{n_i(\lambda)\}_{i=1,2...}$ minimizes Eq. 3.6 and yields $R = R_{budget}$, both the value of λ and the set of truncation points will be optimal in the sense that we cannot reduce the distortion without increasing the bit rate beyond R_{budget}.

PCRD [30] is a simple algorithm to find the optimal truncation points. At any feasible truncation point , PCRD computes the R-D slope, which is defined as

$$S_i^{ni} = \frac{\Delta D_i^{n_i}}{\Delta R_i^{n_i}} = \frac{D_i^{n_i-1} - D_i^{n_i}}{R_i^{n_i} - R_i^{n_i-1}} \tag{3.9}$$

Assuming that the R-D slope is strictly decreasing [30] such that $S_i^{n_i+1} < S_i^{n_i}$ for any feasible truncation point n_i the optimal value of λ denoted as $\lambda_{optimal}$ is equal to the minimum value of λ which satisfies the rate constraint. Theoretically, there are infinitely many possible values. Thus, an iterative approach with fast convergence is used in PCRD to search for the $\lambda_{optimal}$. Once we know the $\lambda_{optimal}$ the optimal truncation points can be found by Eq. 3.8 with $\lambda = \lambda_{optimal}$.

However, the R-D slopes at the feasible truncation points of real images may not be strictly decreasing, especially those in the initial few bit planes. Figure 3.5 is an example of the feasible truncation point. Thus, in the PCRD implementation, the feasible truncation points at which the R-D slope are not strictly decreasing are considered "unfeasible" and PCRD would not truncate at those points.

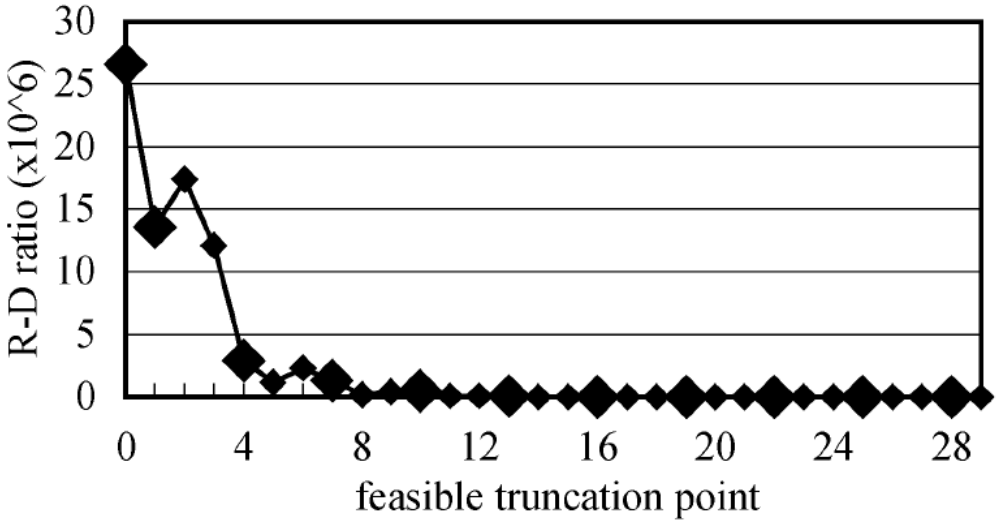


Figure 3.5. Feasible Truncation Point.

In our design, we use the SSIM algorithm as the rate distortion metric instead of the MSE. Basing on this information, the encoder can then include the coding passes in the order of decreasing distortion reduction per unit rate until the bit budget has been exhausted.

Compared with JPEG2000, the rate control in SPIHT is pretty simple. The algorithm just stops at the desired rate or distortion. In other words, the rate control in SPIHT is without any perceptual distortion control. The whole process can be described with Figure 3.6.

The encoder goes through every bit plane from the most significant bitplane to the least significant bitplane. When the target bit budget is exhausted, the encoder ceases at certain point. All of the following bit will be discarded. In Figure 3.6, assuming the target bit rate is achieved in the point where the arrow points at, all of the bits with shadow should be discarded.

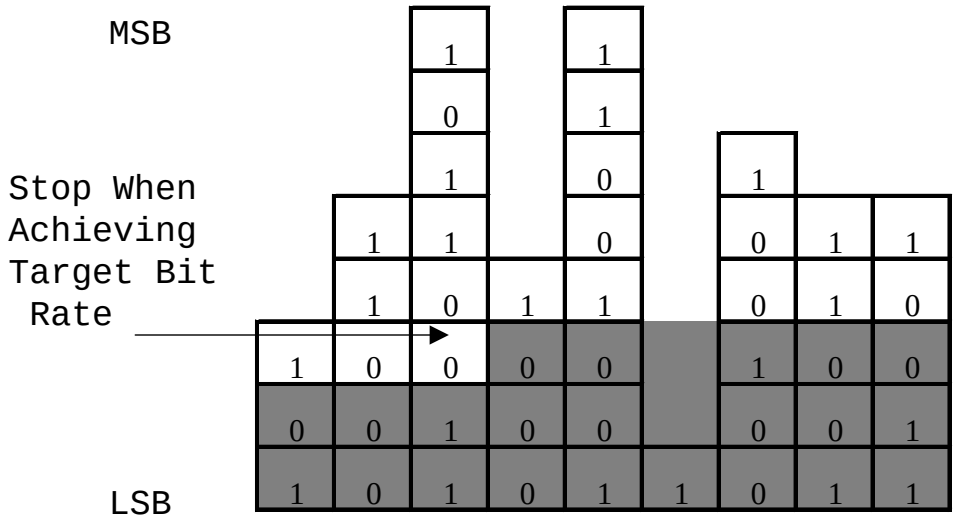


Figure 3.6. SPIHT Rate Control: Encoder Stops at Desired Bit-Rate Or Distortion.

As we know, in bit plane coding, a zero tree can cover many zeros. In other words, in each bit plane, the more number of bit '1', the bigger size of coding pass it generates. From Figure 3.6, we can see that the algorithm just stops at the desired rate or distortion. This scheme, in essence, is equivalent to the method illustrated by Figure 3.7.

Firstly, remove all the '1's in the bit planes lower than the bit plane that the coder stops at (here, 'remove' means replacing all '1's to '0's);

Secondly, Remove all the '1's behind the stoping point in the bit plane that the coder stops at (We call the line in Figure 3.7 that removes the bits '1' the 'Removing Line');

Finally, apply a bitplane coding of the 'modified bit planes' without any rate control.

Our perceptually-guided coding algorithm was motivated by the idea of change the "Removing Line" into a "Removing Curve", as demonstrated in Figure 3.8.

This is desirable because the bits in the same bit plane at different location may have different contributions to image quality. To implement the idea, however, we need

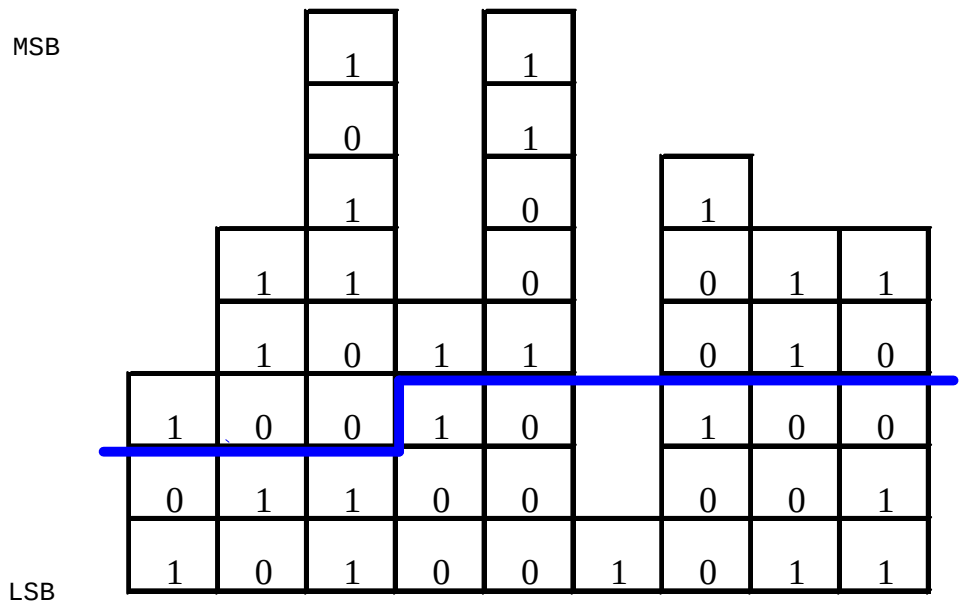


Figure 3.7. Equal Rate Control Scheme: All of the Bits Under the Line is Removed.

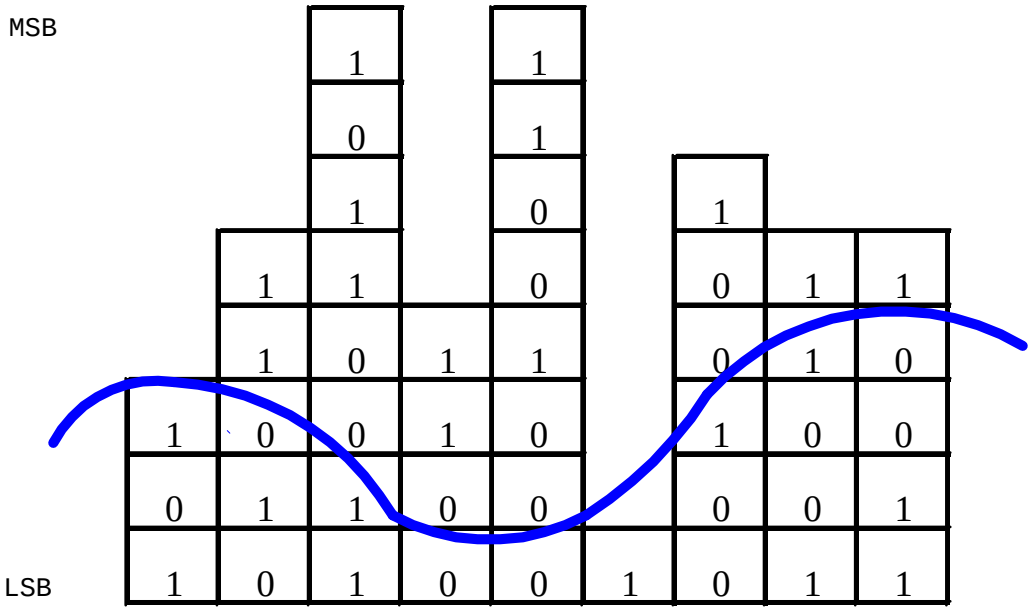


Figure 3.8. 'Removing Curve': All of the Bits under the Curve is Removed.

Original Image

Compressed Image

Figure 3.9. Example Images; Left one is Original Image; Right one is Compressed Image.

to develop an algorithm to automatically decide on which bits should be reserved and which bits need to removed.

Let that $X = x_{i=1,2...N}$ denote the original image and $Y = y_{i=1,2...N}$ denote the compressed image. Let

$$E = SSIM(X, Y) \tag{3.10}$$

Here, $E = e_{i=1,2...N}$ is the error map generated by using the SSIM algorithm on the image X and Y. An example is given in Figure 3.9 (image size is 512×512).

After calculating the SSIM index with a sliding window approach across the image, we obtain the error map between X and Y, as demonstrated in Figure 3.10.

Comparing the original image with the compressed image, we can see that the SSIM error map can reflect human's perception of error quite well. Specifically, the top and the edge of the tower is distorted severely; A piece of fake cloud is added on the top of the right house; The middle of the fence is blurred. They are all successfully indicated by the SSIM error map as the dark regions.

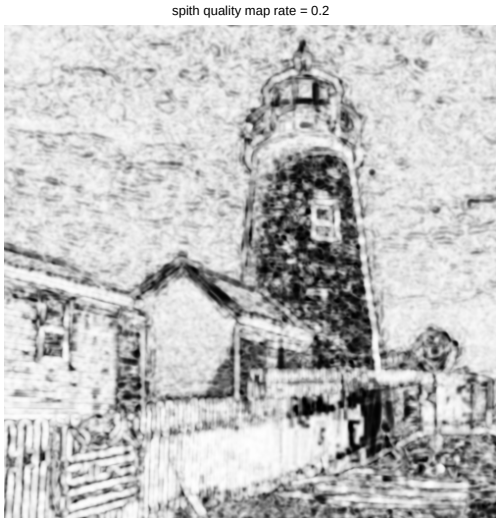


Figure 3.10. Quality/Distortion Map Example.

The contrast sensitivity function models the sensitivity of the HVS as a function of the spatial frequency content in visual stimuli [2]. A typical CSF is shown at the bottom part of Figure 3.11. In general, the CSF has a band-pass nature. It peaks at a spatial frequency around 4 cycles per degree of visual angel and drops significantly with both increasing and decreasing frequencies. This effect is demonstrated at the top part of Figure 3.11, which is widely known as the Campbell-Robson CSF chart [42]. In the chart, the pixel intensity is modulated using sinusoids along the horizontal dimension, while the modulating spatial frequency increases logarithmically. The image contrast increases logarithmically from top to bottom. Now suppose that the perception of contrast is determined solely by the image contrast. Then, the alternating bright and dark bars should appear to have equal height across any horizontal line across the image. However, the bars are observed to be significantly higher at the middle of the chart. As a matter of fact, the peak shifts with viewing distance, and it is important to note that this observed effect is a property of the HVS, but not the test image.

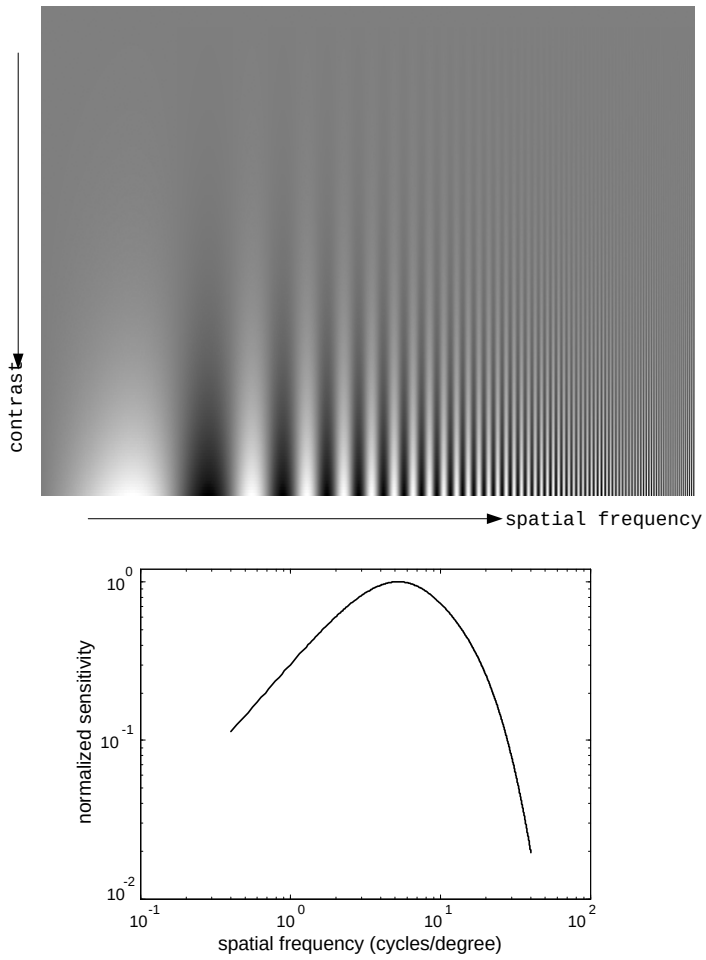


Figure 3.11. Contrast Sensitivity Function. Top: Campbell-Robson CSF chart; Bottom: Normalized Visual Sensitivity as a Function of Spatial Frequency.

Because of the non-uniform distributions of cone receptors and ganglion cells in the retina of humans, when one fixates at a point in the visual environment, the region around the fixation point is sampled with the highest spatial resolution, and the resolution decrease rapidly with distance from the fixation point. A simulation of such a "foveation process" is shown in Figure 3.12. If attention is focused at the man at the lower part of the image(where the foveal center was placed), then the foveated and the original images are almost indistinguishable. In other words, when we observe the image X and Y in Figure 3.9, we may have a higher spatial resolution of the distortion part such as the

(a) (b)

Figure 3.12. Foveated Image (a)Original Image; (b)Foveated Image, where the assumed fixation point is at the man at the lower part of the image.

top and the edge of the tower and the top of the right house. As a result, the observer may draw a conclusion that the image is very distorted, with less regard to the fact that many other areas in the image actually have pretty high quality.

If we move the higher error value to a lower error value location, in other words, if we smooth the error map, then the "most distorted" parts of the image Y may become much less annoying.

Our proposed scheme is shown in Figure 3.13. The decoder is the same as the original JPEG2000 baseline decoder. In other words, our proposed algorithm is decoder compatible. In the encoder side, a "SSIM Equalization" module is added to evenly distribution the coded errors in terms of the SSIM measure.

The purpose of "SSIM equalization" is to make the error evenly distributed if possible, or at least smooth it. To make error evenly distributed or smooth the error map, it is necessary to estimate and model the errors. Equalization is carried out after quan-

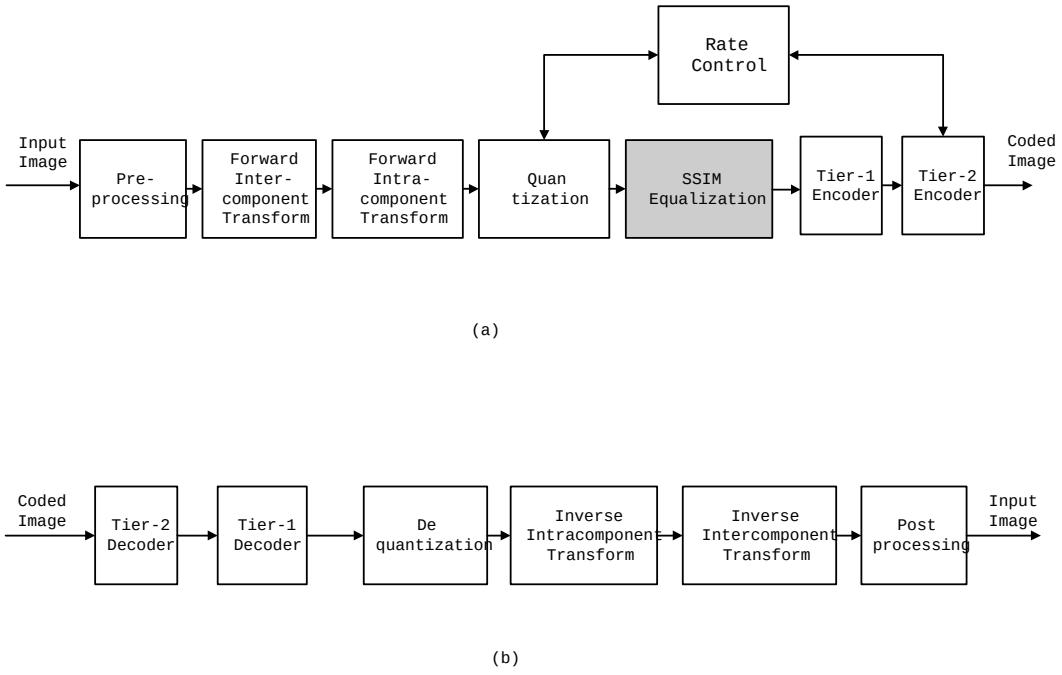


Figure 3.13. Proposed Scheme (a)Encoder Structure (b)Decoder Structure.

tization and before tier-1 coding. Our estimation and Modeling schemes are illustrated in Figure 3.14:

The task of equalization modeling is to generate a equalization map to adjust the DWT coefficients according to the estimation error map. Before we go any further, it is necessary to investigate the characteristics of the Discrete Wavelet Transform (DWT). Two specific wavelet transforms are assumed by the baseline JPEG2000 codec: the 5/3 and 9/7 wavelets. The 5/3 wavelet transform is reversible, integer-to-integer, and non-linear. This transform was proposed in [43], and is simply an approximation to a linear wavelet transform proposed in [44]. The 9/7 wavelet transform is nonreversible and real-to-real. This transform, proposed in [27], is also employed in the FBI fingerprint compression standard [45] (although the normalizations differ). The 5/3 transform and 9/7 transform are different in terms of parameters of 1-D UMDFB, while maintaining the state of art of wavelet transform.

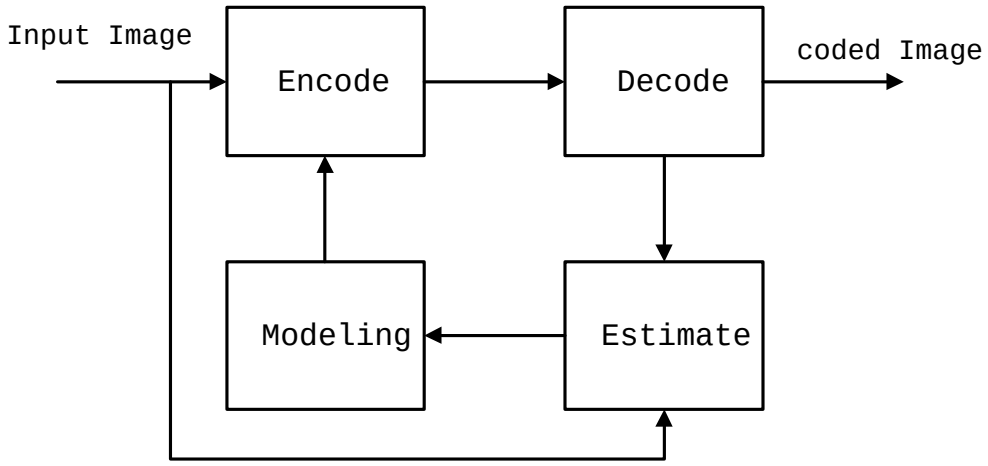


Figure 3.14. Estimation and Modelling Scheme.

The estimated error map is calculated in spatial domain, which is an IDWT translation of frequency domain. The estimated error map is blurred with a Gaussian function, which is defined as

$$G(r) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-r^2}{2\sigma^2}} \tag{3.11}$$

where σ is the standard deviation of the Gaussian function. In our scheme, we select σ = 1.5. Down sampling is carried out after the Gaussian blur to distribute the error into each resolution level. In JPEG2000, the number of resolution levels is a parameter of each transform. A typical value for this parameter is six (for a sufficiently large image). In SPIHT, the number of resolution levels depends on the image size.

The next step is to map each error in estimated error map into equalization value. In JPEG2000, the maximal number of bit planes is less than or equal to the depth of the pixel intensity. However, in the SPIHT algorithm, the number of bit planes can be extended from positive to negative. The Figure 3.15 is an example of the equalization map. In the equalization map, the bright part is the high value in the equalization map

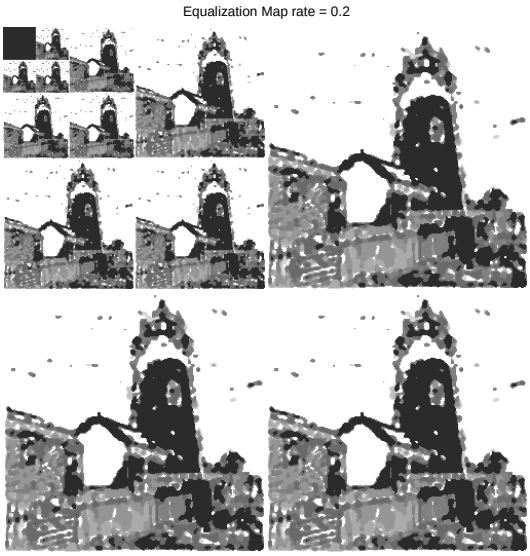


Figure 3.15. Equalization Map: The Bright Part Means the High Value in the Equalization Map; The Dark Part Means the Low Value; There Are 10 Levels in this Example.

and the dark part is the low value. More bits will be removed from bottom for higher value than lower value. There are a total of 10 levels in this example.

3.5 Experiment Result

3.5.1 Results of Proposed Scheme with SPIHT Coding

The SPIHT code is modified to incorporate the proposed encoding with perceptual SSIM distortion metric. The original image bit rate is 8 bits/pixel (bpp) and the image size is 512×512. We test the compressed image of which the bit rates are from 0.2 bpp to 0.8 bpp by using the original SPIHT and the SPIHT algorithm with our SSIM distortion metric. The quality/distortion map is calculated and presented by comparing with the original image.

Figures 3.16 and 3.17 give examples of the original image, the SPIHT compressed image and the modified SPIHT compressed image with the proposed method at a bit

rate of 0.2 bits/pixel (bpp). The images are (a)Original Image; (b) Equalization Map; (c) SPIHT image; (d) SSIM image; (e)Local Enlarged SPIHT Image; (f)Local Enlarged SSIM Image; (g)SPIHT Quality Map; and (g)SSIM Quality Map, respectively.

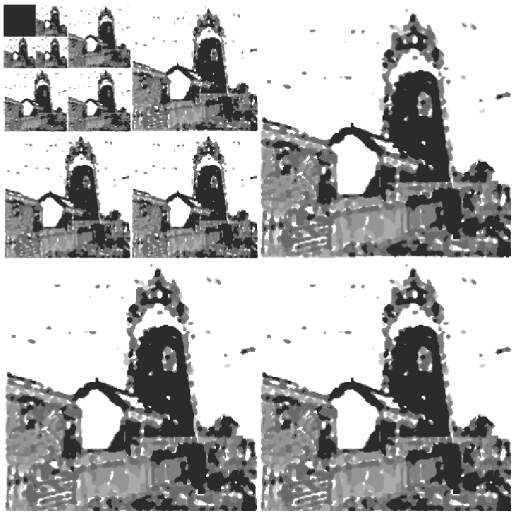
When the bit rate is 0.2bpp, the images are highly compressed (Refer to append A for images at different bit rates). It can be seen that the SPIHT image with SSIM optimization has a clearly higher quality than SPIHT image. For example, on the top of the tower, the edge is missing in SPIHT coded image, but is easily discerned in SSIM image. The edge of the top of the right house is distorted with a fabricated piece of cloud, but is still clear in the SSIM image.

Figure 3.18 shows the SSIM index and PSNR comparison between SPIHT coded images and SSIM images. We can see that our proposed scheme do have some degradation at bit rate 0.2, 0.3 and 0.4 bits/pixel in terms of PSNR. But we get much improvement in terms of SSIM index. The maximum value of the SSIM quality/distortion map is almost keep unchanged. However, the minimum value of quality/distortion map gets much improvement at bit rate 0.2, 0.3 and 0.4 bits/pixel. In other words, the quality map of our proposed scheme is smoother than that of the original SPIHT coded image. As the bit rate increases, the image quality improves accordingly, and the difference between SPIHT and SSIM images becomes imperceivable.

3.5.2 Result of Proposed Scheme with JPEG2000 Coding

The JASPER [10] implementation of JPEG2000 is modified to incorporate the proposed encoding scheme for perceptual SSIM quality control. The original image has a bit rate of 8 bits/pixel (bpp) and a size of 512×512. We test the compressed image at bit rates from 0.2 bpp to 0.8 bpp by incorporating JPEG2000 with our SSIM distortion control scheme respectively. The quality/distortion map is computed by comparing with the original image.

(a) Original



(b) Curve Map

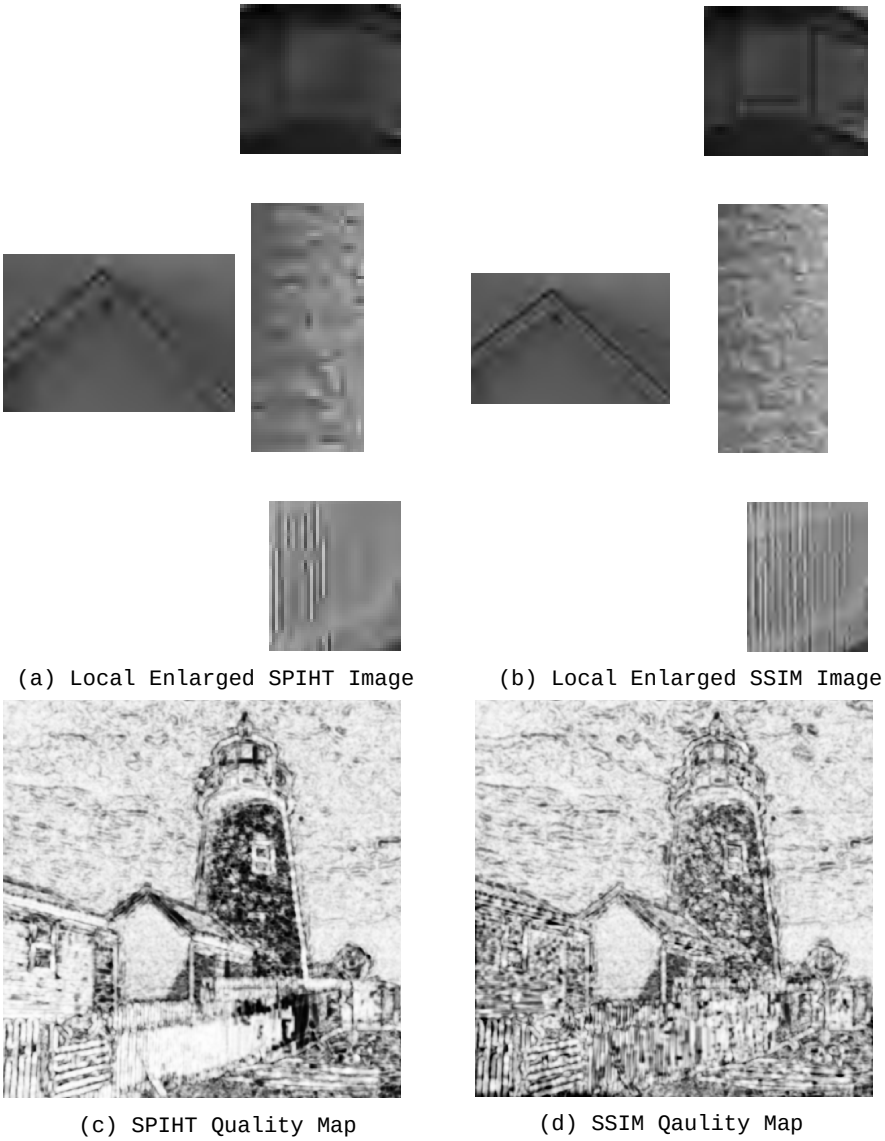
(c) SPIHT Image

(d) SSIM Image

Figure 3.16. Results of Proposed Scheme with SPIHT Coding: (a)Original Image; (b) Equalization Map: The brighter part means a higher value and more bits of coefficient will be removed; The darker part means a lower value and less bits of coefficient will be removed;; (c) SPIHT Compressed Image; (d) SPIHT Compressed Image with SSIM Optimization. Rate = 0.2 bpp.

Figures 3.19 and 3.20 are examples of the original image, JPEG2000 compressed image, and modified JPEG2000 with the proposed quality control scheme. Figure 3.21 compares the PSNR and SSIM index values of JPEG2000 compressed images and the SSIM image. It can be seen that our proposed scheme has a little degradation at 0.2, 0.3 and 0.4 bits/pixel in terms of PSNR, and little improvement in terms of the SSIM index. The maximum value of SSIM quality/distortion map is almost unchanged. The minimum value of quality/distortion map gets a little improvement at bit rates of 0.2, 0.3 and 0.4 bits/pixel. As the bit rate increases, the image quality improves accordingly, and the difference between JPEG2000 image and SSIM image becomes indistinguishable. From the result, we can see that the proposed scheme with JPEG2000 is not as effective as in SPIHT.

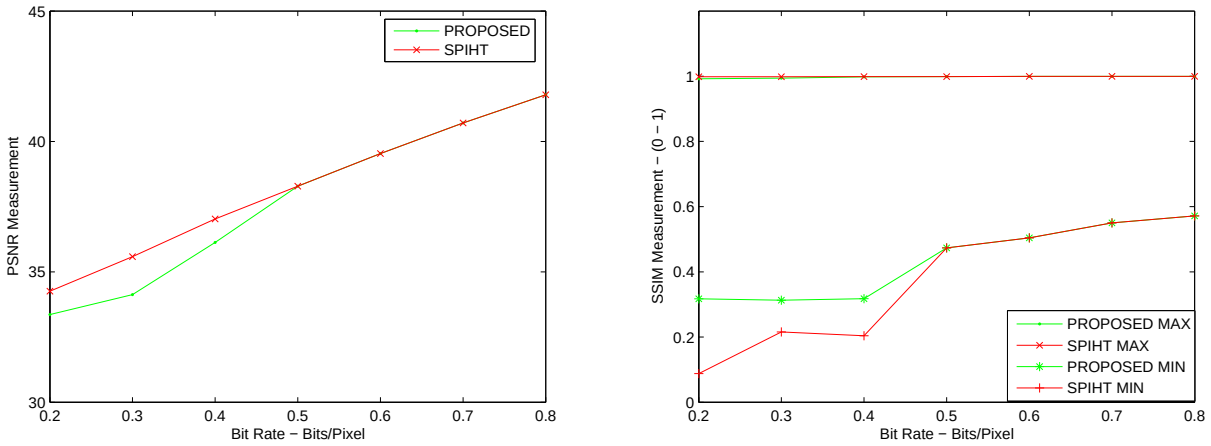
From our experiment, we conclude that our proposed algorithm can achieve better image quality in terms the SSIM index. However, the improvement is only moderate, especially in the case of JPEG2000 compression. The reason might be that the optimal spatial bit allocation may not be sufficiently reached simply by moving bits from one spatial location to another. During the bit movement, the image quality at the spatial locations where new bits are added will improve, but at the same time, the quality at the spatial locations that lose bits will degrade, and the quality improvement at one spatial location may not fully compensate the quality degradation at another spatial location. Therefore, more advanced perceptual bit allocation schemes need to be developed in the future to achieve more effective perceptual image coding result.



(a) Local Enlarged SPIHT Image (b) Local Enlarged SSIM Image

(c) SPIHT Quality Map (d) SSIM Qaulity Map

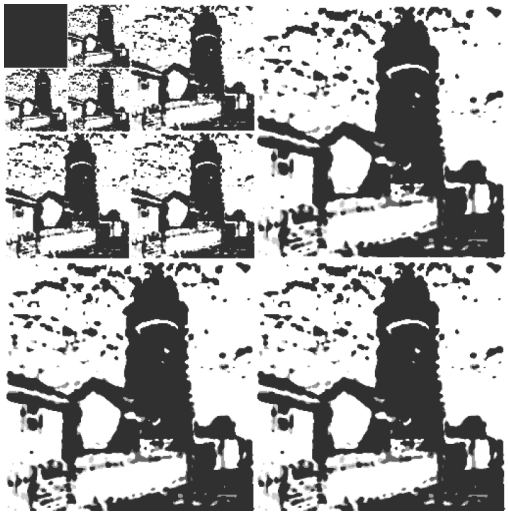
Figure 3.17. Continue with figure 3.16 (a)Local Enlarged SPIHT Image; (b)Local Enlarged SSIM Image; (c)SPIHT Quality Map; (d)SSIM Quality Map. Rate = 0.2 bpp.



(a) (b)

Figure 3.18. Numerical Comparison of Proposed Scheme with Original SPIHT (a) PSNR Comparison (b) SSIM Comparison.

(a) Original



(b) Curve Map

(c) JPEG2000 Image

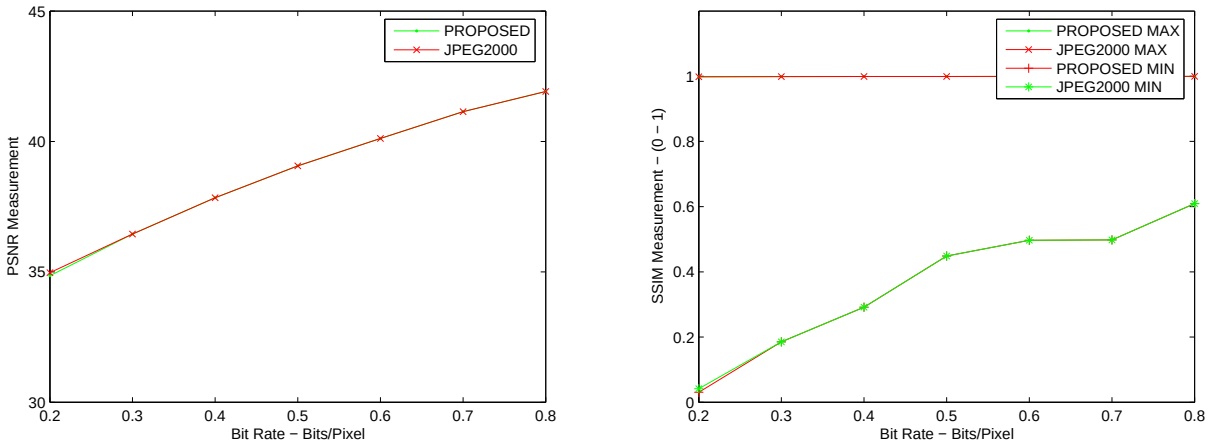
(d) SSIM Image

Figure 3.19. Results of Proposed Scheme with JPEG2000 Coding:(a)Original Image; (b) Equalization Map: The brighter part means a higher value and more bits of coefficient will be removed; The darker part means a lower value and less bits of coefficient will be removed;; (c) JPEG2000 Compressed Image; (d) JPEG2000 Compressed Image with SSIM Optimization. Rate = 0.2 bpp.

(a) Local Enlarged JPEG2000 Image (b) Local Enlarged SSIM Image

(c) JPEG2000 Quality Map (d) SSIM Qaulity Map

Figure 3.20. Continue with figure 3.19 (a)Local Enlarged JPEG2000 Image; (b)Local Enlarged SSIM Image; (c)JPEG2000 Quality Map; (d)SSIM Quality Map. Rate = 0.2 bpp.



(a) (b)

Figure 3.21. Numerical Comparison of Proposed Scheme with Original JPEG2000 (a) PSNR Comparison (b) SSIM Comparison.

CHAPTER 4

HANDWRITTEN DIGIT RECOGNITION USING CW-SSIM

In this chapter, we describe our algorithm of handwritten digit recognition using complex wavelet structural similarity (CW-SSIM) index. We first explain the motivation of using the CW-SSIM index for handwritten digit recognition. We then present in more details the CW-SSIM algorithm. Finally, we describe the procedures and test results of using CW-SSIM for recognizing handwritten digits in the MNIST database (http://yann.lecun.com/exdb/mnist/).

4.1 Motivation

A major drawback of the spatial domain SSIM algorithm is that it is highly sensitive to translation, scaling and rotation of images, as demonstrated in Images (h)-(l) of Figure 1.3. This is an undesirable feature for most image pattern recognition tasks such as handwritten digit recognition, because the images are often shifted, scaled or rotated by a small amount. A straightforward way to resolve this problem is to apply a registration process before computing the spatial domain SSIM index. This could potentially eliminate some simple parametric distortions by estimating their parameters and applying a corresponding inverse transformation to the distorted image. However, current image registration algorithms are often computationally expensive and the registration accuracy is not always satisfactory. The CW-SSIM method provides an alternative solution without a precise registration stage in the front, because it is insensitive to small translation, scaling and rotation of images by itself. Many existing digit recognition algorithms are computational complicated. The simplicity and robustness properties of the

CW-SSIM algorithm lead us to use it as a tool for handwritten digit recognition. This work is an initial attempt of using CW-SSIM for any pattern recognition applications.

4.2 Algorithm Description

The main task of image pattern recognition is to classify images into categories, so that images in a particular category are similar, while images in different categories may vary widely. In handwritten digit recognition, characters, typically represented as fixed-size images must be classified into one of 10 (0 to 9) categories using a classification function. The classification functions can be divided into two camps: image-based matching and feature-based matching, as described in Chapter 1. Our algorithm belongs to image-based matching, which uses a set of representative templates for each digit category. There are two phases in our proposed scheme: template selection and CW-SSIM based recognition.

4.2.1 Template Selection

The first step of our algorithm is to select a set of representative templates for each digit category from a large database of training digit images. The goal is to maximize the variations between the representative templates under a constraint of the total number of allowed templates.

Our template selection scheme create ten template sets, each for one digit category. The algorithm works as follows: For each digit category, we start from the first training image and include it as the first template in the template set. For newcoming images, we compute its CW-SSIM value with all existing templates in the template set. If the value is higher than a threshold T_{thesh}, we regard the new training image as redundant and simply move to the next training image; Otherwise, the image is added to the template set as a new template. This procedure continues until all the training images have been tested or

the maximum size of the template set is reached. The same procedure is applied to all ten categories sets of training images separately, resulting in ten template sets. The result of this template selection process depends on the threshold T_{thesh}, which was manually selected in our experiments.

4.2.2 CW-SSIM Based Recognition

For any given test image to be recognized, the task of the CW-SSIM based recognition process is to calculate the CW-SSIM values between the test image with all templates in all of the ten template sets. The category corresponds to the template of the highest CW-SSIM value will then be determined as the recognition result.

Suppose X and Y are two digit images to be compared. A complex version of the steerable pyramid decomposition [46] is first applied to both images. In the complex wavelet transform domain, suppose $c_x = \{c_{x,i} | i = 1, ..., N\}$ and $c_y = \{c_{y,i} | i = 1, ..., N\}$ are two sets of coefficients extracted at the same spatial location in the same wavelet subbands of the two digit images, respectively. The CW-SSIM index is defined as

$$S\ (c_x, c_y) = \frac{2|\sum_{i=1}^{N} c_{x,i} c_{y,i}^*| + K}{\sum_{i=1}^{N} |c_{x,i}|^2 + \sum_{i=1}^{N} |c_{y,i}|^2 + K} \tag{4.1}$$

Here c^* denotes the complex conjugate of c and K is a small positive constant.

To better understand the CW-SSIM index, we rewrite it as a product of two components:

$$S\ (c_x, c_y) = \frac{2\sum_{i=1}^{N} |c_{x,i}||c_{y,i}^*| + K}{\sum_{i=1}^{N} |c_{x,i}|^2 + \sum_{i=1}^{N} |c_{y,i}|^2 + K} * \frac{2|\sum_{i=1}^{N} c_{x,i} c_{y,i}^*| + K}{2\sum_{i=1}^{N} |c_{x,i} c_{y,i}^*| + K} \tag{4.2}$$

The first component is completely determined by the magnitudes of the coefficients and the maximum value one is achieved if and only $|c_{x,i}| = |c_{y,j}|$ for all i's. The second

component, on the other hand, is fully determined by the consistency of phase changes between c_x and c_y. It achieves the maximum value one when the phase difference between $c_{x,i}$ and $c_{y,j}$ is a constant for all i's. We consider this component as a useful measure of image structural similarity based on the believes that:

1) The structural information of local image features is mainly contained in the relative phase patterns of the wavelet coefficients.

2) Consistent phase shift of all coefficients does not change the structure of the local image feature.

This measure is not only insensitive to small geometrical distortions, but also to luminance and contrast changes. Note that luminance and contrast changes of images can be roughly described as a point-wise linear transform of local pixel intensities: $y_i = ax_i + b$, here a and b are constant. Due to the linear and bandpass nature of the wavelet transform, the effect in the wavelet domain is a constant scaling of all the coefficients, i.e., $c_{y,i} = ac_{x,i}$ for all i's. Substitute this into Eq. 4.1, we can see that a perfect value one is obtained for the second component and the first component gives

$$S\ (c_x, c_y) = \frac{2\alpha + K/\sum_{i=1}^{N} |c_{x,i}|^2}{1 + \alpha^2 + K/\sum_{i=1}^{N} |c_{x,i}|^2} \tag{4.3}$$

At strong image features (large coefficient magnitudes), $K/\sum_{i=1}^{N} |c_{x,i}|^2$ is small and can be ignored, leading to an insensitive measure (compared with MSE) - changing the magnitude by a factor of 10 percent (a = 1.1) only causes reduction of the SSIM value from 1 to 0.9955. The measure is even less sensitive at weaker image features (small coefficient magnitudes).

4.3 Experiment Result

We test our handwritten digit recognition algorithm with complex wavelet structural similarity index using the MNIST database [1], which contains a training set of 60,000 images, and a test set of 10,000 images. It is a subset of a larger set available from NIST. The MNIST database was constructed from NIST's Special Database 3 and Special Database 1 which contain binary images of handwritten digits. NIST originally designated SD-3 as their training set and SD-1 as their test set. However, SD-3 is much cleaner and easier to recognize than SD-1. That is because SD-3 was collected among Census Bureau employees, while SD-1 was collected among high-school students. The digits in this database have been size-normalized and centered in a fixed-size image. These images contain grey levels as a result of the anti-aliasing technique used by the normalization algorithm. The image size is 28x28 pixels.

We tested our algorithm with five different choices of the size of the template sets, ranging from 10 to 200 templates per digit category. Table 4.1 shows the results of correct recognition rate, which is defined as the percentage of the total number of correctly recognized digits divided by the total number of digits being recognized.

It can be observed that the correct recognition rate increases as the template increases. It is interesting that the digits "0" and "1" always have higher correct recognition rate than other digits and digit "5" always gets the lowest performance. The reason may be that digits "0" and "1" have less correlation with all other digits while digit "5" has strong correlation and are often confused with other similar digits such as "8", "6", and "3". To better visualize the recognition result, Append B shows all the templates (10, 30, 50, 100, 200) and the incorrectly recognized digits.

The CW-SSIM computation is the most time-consuming process in both the template selection and the CW-SSIM comparison procedures. Therefore, the overall computational complexity scales approximately linearly with the number of CW-SSIM calcula-

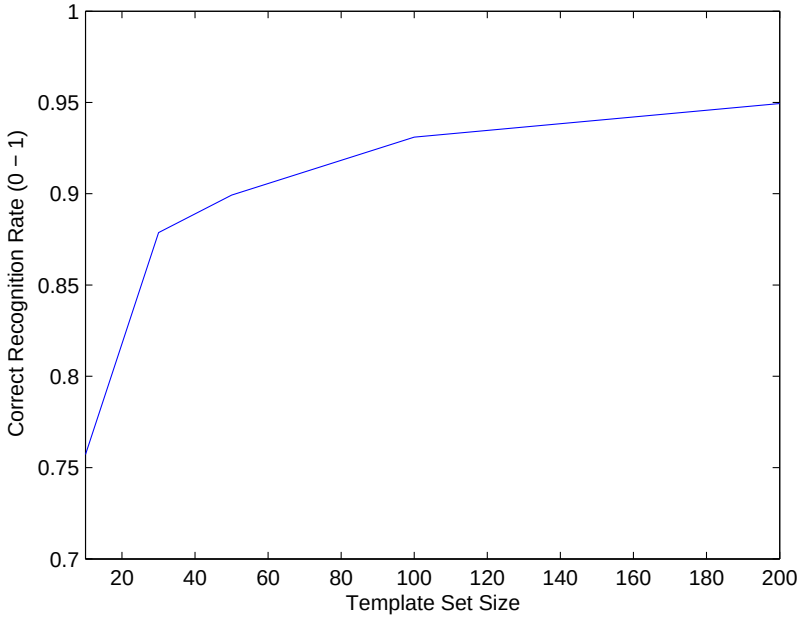


Figure 4.1. Correct Recognition Rate as a Function of the Size of the Template Sets.

Table 4.1. Correct recognition rate for ten digit categories with different sizes of the template sets.

Digits	Template Set Size				
	10	30	50	100	200
0	0.9622	0.9602	0.9633	0.9765	0.9908
1	0.8758	0.9903	0.9894	0.9877	0.9921
2	0.7209	0.8605	0.8740	0.9234	0.9477
3	0.7406	0.8851	0.8950	0.9198	0.9277
4	0.6823	0.8320	0.8737	0.9022	0.9287
5	0.5157	0.7388	0.7993	0.8901	0.9260
6	0.8925	0.9697	0.9749	0.9645	0.9791
7	0.6770	0.7996	0.8560	0.8920	0.9212
8	0.7536	0.8296	0.8398	0.8984	0.9261
9	0.7512	0.9207	0.9277	0.9554	0.9544
mean	0.7572	0.8787	0.8993	0.9310	0.9494

tions. For template selection, the number of CW-SSIM calculations depends on the total number of training images and the threshold used in the selection and thus the exact number is difficult to estimate. However, this number is bounded between $10(\frac{K}{10} - N + \frac{N^2}{2})O$ and $10(\frac{N^2}{2} + (N-1)(\frac{K}{10} - N + 1))O$, where N is the number of templates for each digit and K is the number of images in the training set. For practical applications, the speed of the CW-SSIM comparison procedure is more critical in the sense that it determines the speed of recognition. If L is the number of images in the testing set, then the total number of CW-SSIM calculations is $(L \times 10N)O$.

CHAPTER 5

CONCLUSIONS

In this thesis, we have studied several aspects of the structural similarity index, from its spatial pooling methods to its applications to image compression and pattern recognition.

5.1 Pooling Strategies for Image Quality Assessment

We have tested three spatial pooling strategies (Miknowski pooling, local quality/distortion weighted pooling, and information content-based pooling) for perceptual image quality assessment based on an extensive experiment with the LIVE database. The following conclusions can be drawn from this study:

1) SSIM index is a better indication of local image quality than the absolute difference.

2) All three pooling methods may improve the prediction performance of image quality measures, compared with simple spatial averaging.

3) The information content-weighted pooling approach demonstrates the best potential to be a general and stale approach that provides consistent improvement over a wide range of image distortion types.

Future work includes testing the pooling methods with other image quality measures (including those that involve wavelet decompositions) and developing more accurate method for the estimation of local information content by adopting improved statistical image models.

5.2 Structural Similarity-Guided Perceptual Image Compression

We have implemented a structural similarity-guided perceptual image compression algorithms by incorporating the SSIM measure with the existing SPIHT and JPEG2000 algorithms. The verification was done by testing the images at bit rate ranging from 0.2 to 0.8 bits/pixel. According to the experimental result, we can make the following conclusions:

1) The proposed algorithm achieves better compressed image quality according to the SSIM measure, which is a much better indicator of image quality than PSNR or MSE.

2) The proposed algorithm provides a better performance when incorporated with the SPIHT algorithm than with the JPEG2000 algorithm. Given the complexity of the JPEG2000 algorithm, the reason is not manifest and is still under investigation.

3) The performance of the proposed algorithm is much better at lower bit rate (0.2 to 0.4) than at higher bit rate (0.5 and above). This is not surprising because images coded with SPIHT or JPEG2000 at high bit rate have high quality and the quality is more spatially evenly distributed, leaving less space for improvement.

4) The effectiveness of the proposed algorithm is only moderate and sometimes depend on the types of the content (smooth regions, textures, sharp edges ...) in the image. During the bit redistribution process in the algorithm, the image quality at the spatial locations where new bits are added will improve, but at the same time, the quality at the spatial locations that lose bits will degrade, and the quality improvement at one spatial location may not fully compensate the quality degradation at another spatial location. Advanced perceptual bit allocation schemes are still yet to develop in the future.

5.3 Handwritten Digit Recognition

We have made an initial attempt of using the complex wavelet structural similarity index for handwritten digit recognition. With a simple implementation and experiment, the test result for the MNIST database is surprisingly good. When the size of the template set is 200, the correct recognition rate is 94.94%. Although the current result is still inferior to some existing approaches (e.g., [35]), it is still very encouraging. Be aware that handwritten digits recognition and specifically the MNIST database has been extensively studied over the years and many existing algorithms are complicated in nature and requires a sophisticated preprocessing stage and a long training process. Our method does not require any preprocessing and is applied in a straightforward way. The method can be further improved in the following perspectives:

1) Include a preprocessing stage, such as rotation of the images and thinning of the strokes.

2) Optimize the template selection process to account for more variations within each digit category.

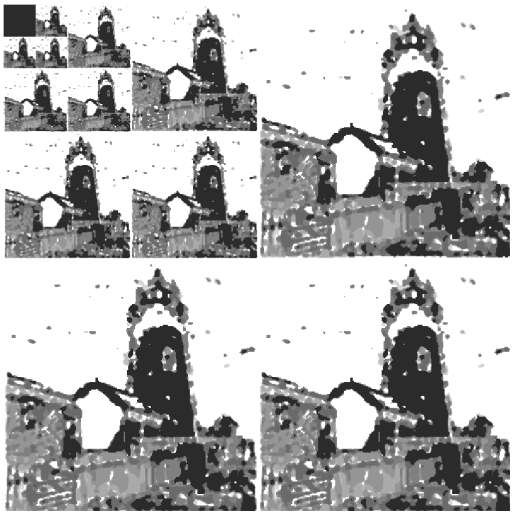
3) Improve the CW-SSIM algorithm by adopting more wavelet subbands and by comparing the CW-SSIM index at multi-scales.

4) Combine the CW-SSIM index with other pattern comparison algorithms.

APPENDIX A
EXPERIMENTAL RESULT OF PROPOSED COMPRESSION SCHEME IN TERMS OF SPIHT

In this appendix, we present the images, 'removing' curve maps, quality/distortion maps for SPIHT at bit rate 0.2, 0.3 and 0.4 bpp.

(a) Original

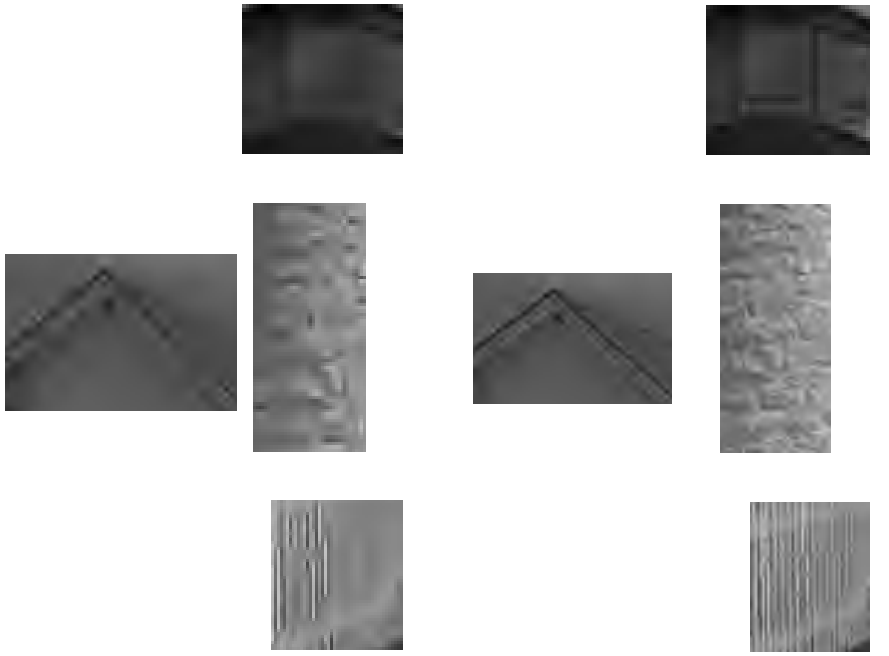


(b) Curve Map

(c) SPIHT Image

(d) SSIM Image

Figure A.1. Results of Proposed Scheme with SPIHT Coding: (a)Original Image; (b) Equalization Map: The brighter part means a higher value and more bits of coefficient will be removed; The darker part means a lower value and less bits of coefficient will be removed; (c) SPIHT Compressed Image; (d) SPIHT Compressed Image With SSIM Optimization; Rate = 0.2 bpp.



(a) Local Enlarged SPIHT Image (b) Local Enlarged SSIM Image

(c) SPIHT Quality Map (d) SSIM Qaulity Map

Figure A.2. Continue with A.1 (a)Local Enlarged SPIHT Image; (b)Local Enlarged SSIM Image; (c)SPIHT Quality Map; (d)SSIM Quality Map; Rate = 0.2 bpp.

(a) Original

(b) Curve Map

(c) SPIHT Image

(d) SSIM Image

Figure A.3. Results of Proposed Scheme with SPIHT Coding: (a)Original Image; (b) Equalization Map: The brighter part means a higher value and more bits of coefficient will be removed; The darker part means a lower value and less bits of coefficient will be removed; (c) SPIHT Compressed Image; (d) SPIHT Compressed Image With SSIM Optimization; Rate = 0.3 bpp.

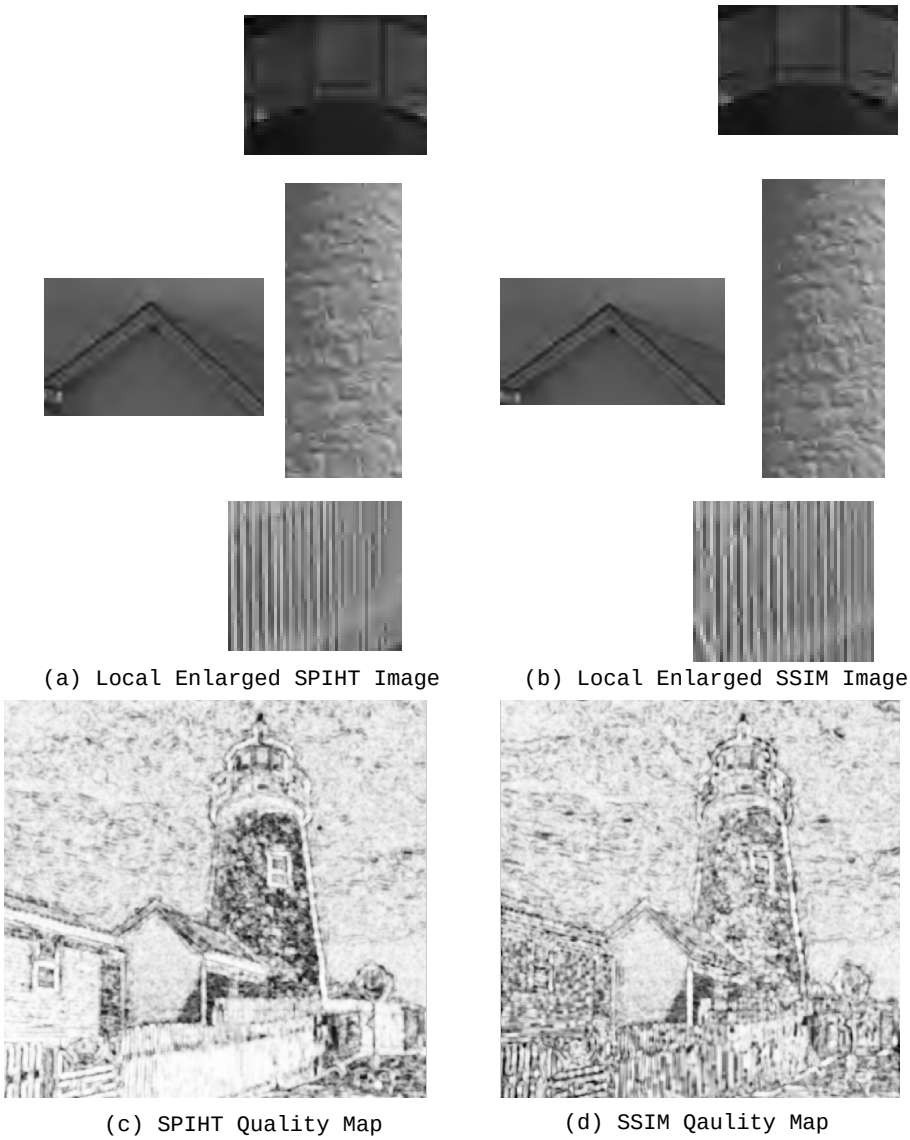
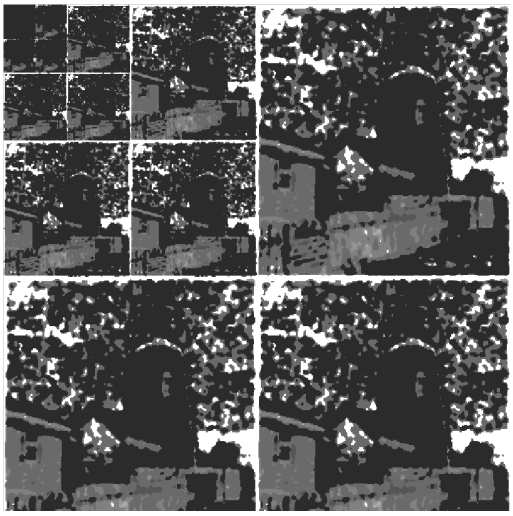


Figure A.4. Continue with A.3(a)Local Enlarged SPIHT Image; (b)Local Enlarged SSIM Image; (c)SPIHT Quality Map; (d)SSIM Quality Map; Rate = 0.3 bpp.

(a) Original

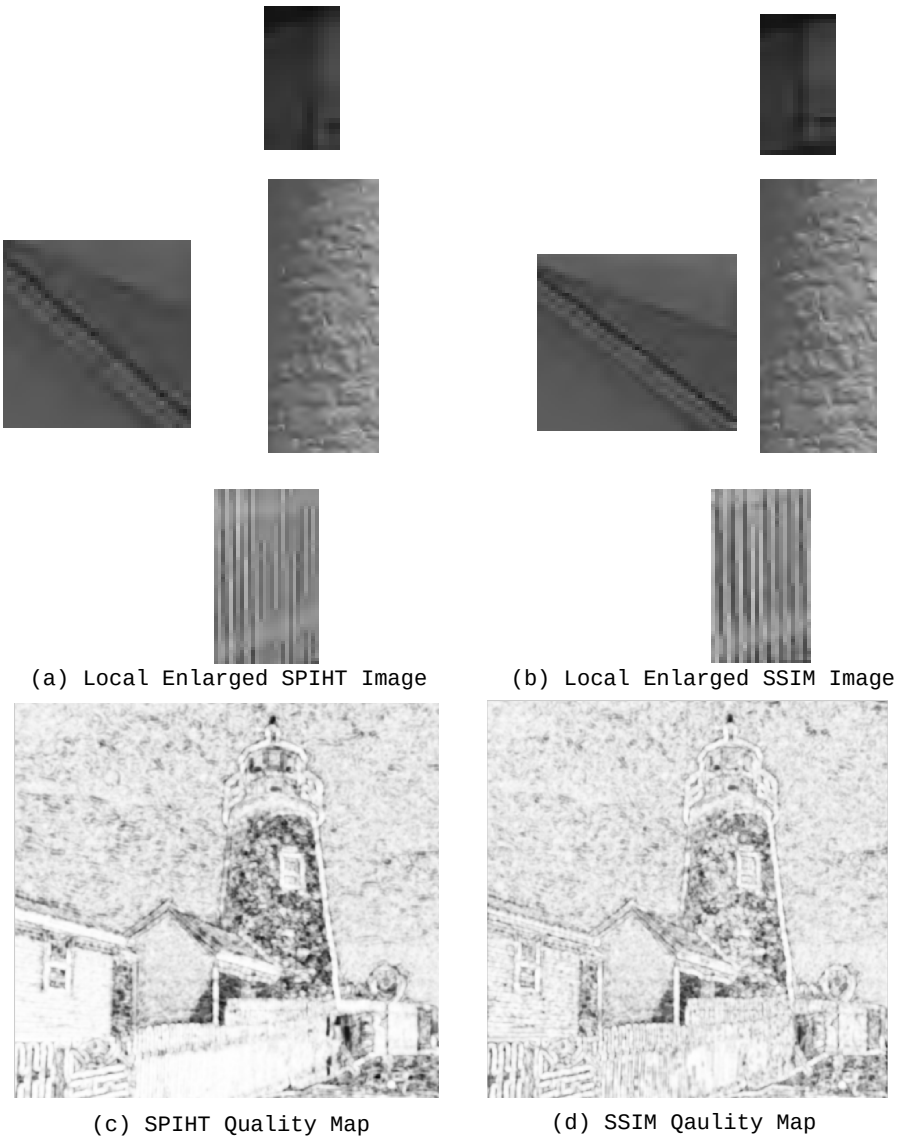


(b) Curve Map

(c) SPIHT Image

(d) SSIM Image

Figure A.5. Results of Proposed Scheme with SPIHT Coding: (a)Original Image; (b) Equalization Map: The brighter part means a higher value and more bits of coefficient will be removed; The darker part means a lower value and less bits of coefficient will be removed; (c) SPIHT Compressed Image; (d) SPIHT Compressed Image With SSIM Optimization; Rate = 0.4 bpp.



(a) Local Enlarged SPIHT Image (b) Local Enlarged SSIM Image

(c) SPIHT Quality Map (d) SSIM Qaulity Map

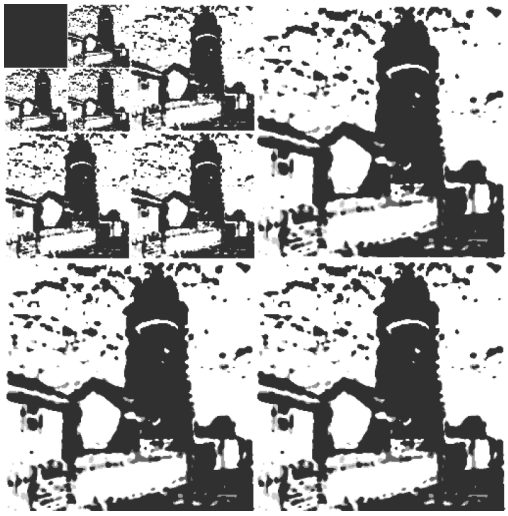
Figure A.6. Continue with A.5(a)Local Enlarged SPIHT Image; (b)Local Enlarged SSIM Image; (c)SPIHT Quality Map; (d)SSIM Quality Map; Rate = 0.4 bpp.

APPENDIX B

EXPERIMENTAL RESULT OF PROPOSED COMPRESSION SCHEME IN TERMS OF JPEG2000

In this appendix, we present the images, 'removing' curve maps, quality/distortion maps for JPEG2000 at bit rate 0.2, 0.3 and 0.4 bpp.

(a) Original

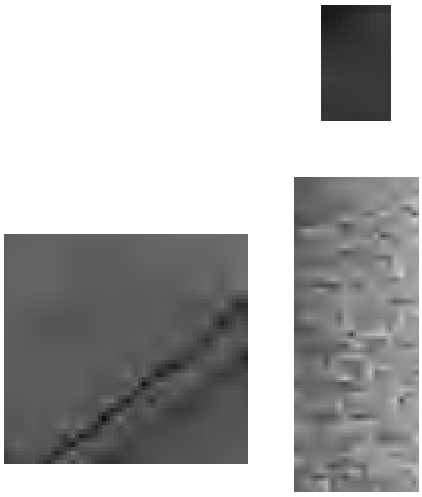


(b) Curve Map

(c) JPEG2000 Image

(d) SSIM Image

Figure B.1. Results of Proposed Scheme with JPEG2000 Coding: (a)Original Image; (b) Equalization Map: The brighter part means a higher value and more bits of coefficient will be removed; The darker part means a lower value and less bits of coefficient will be removed; (c) JPEG2000 Compressed Image; (d) JPEG2000 Compressed Image with SSIM Optimization; Rate = 0.2 bpp.



(a) Local Enlarged JPEG2000 Image

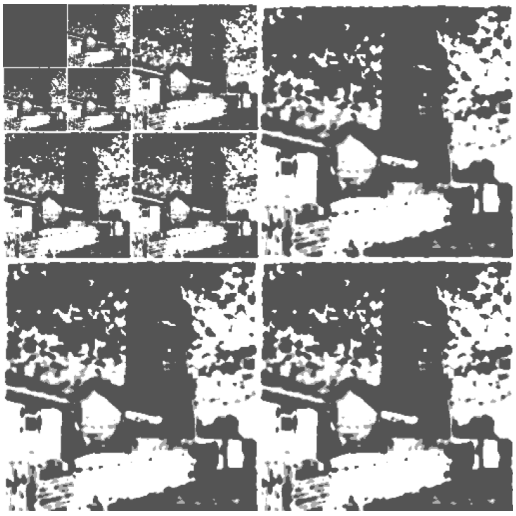
(b) Local Enlarged SSIM Image

(c) JPEG2000 Quality Map

(d) SSIM Qaulity Map

Figure B.2. Continue with B.1 (a)Local Enlarged JPEG2000 Image; (b)Local Enlarged SSIM Image; (c)JPEG2000 Quality Map; (d)SSIM Quality Map; Rate = 0.2 bpp.

(a) Original

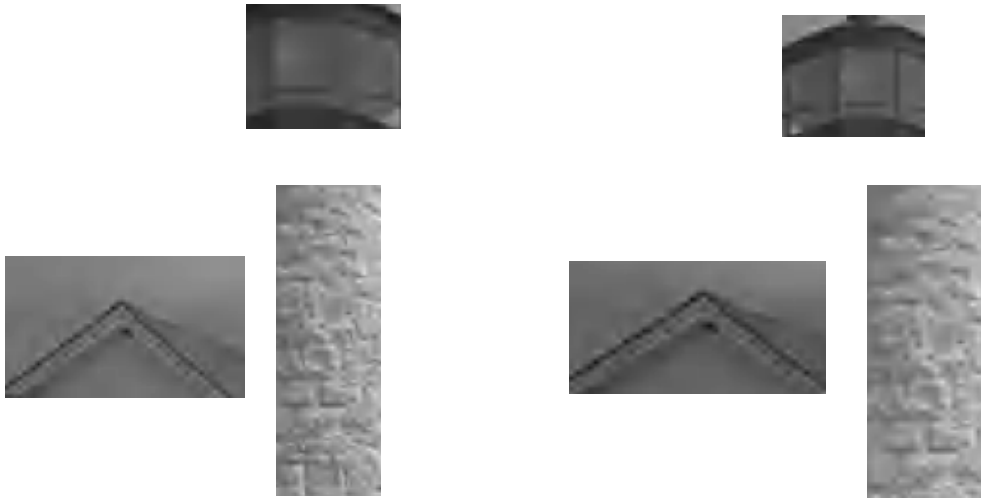


(b) Curve Map

(c) JPEG2000 Image

(d) SSIM Image

Figure B.3. Results of Proposed Scheme with JPEG2000 Coding: (a)Original Image; (b) Equalization Map: The brighter part means a higher value and more bits of coefficient will be removed; The darker part means a lower value and less bits of coefficient will be removed; (c) JPEG2000 Compressed Image; (d) JPEG2000 Compressed Image with SSIM Optimization; Rate = 0.3 bpp.



(a) Local Enlarged JPEG2000 Image

(b) Local Enlarged SSIM Image

(c) JPEG2000 Quality Map

(d) SSIM Qaulity Map

Figure B.4. Continue with B.3 (a)Local Enlarged JPEG2000 Image; (b)Local Enlarged SSIM Image; (c)JPEG2000 Quality Map; (d)SSIM Quality Map; Rate = 0.3 bpp.

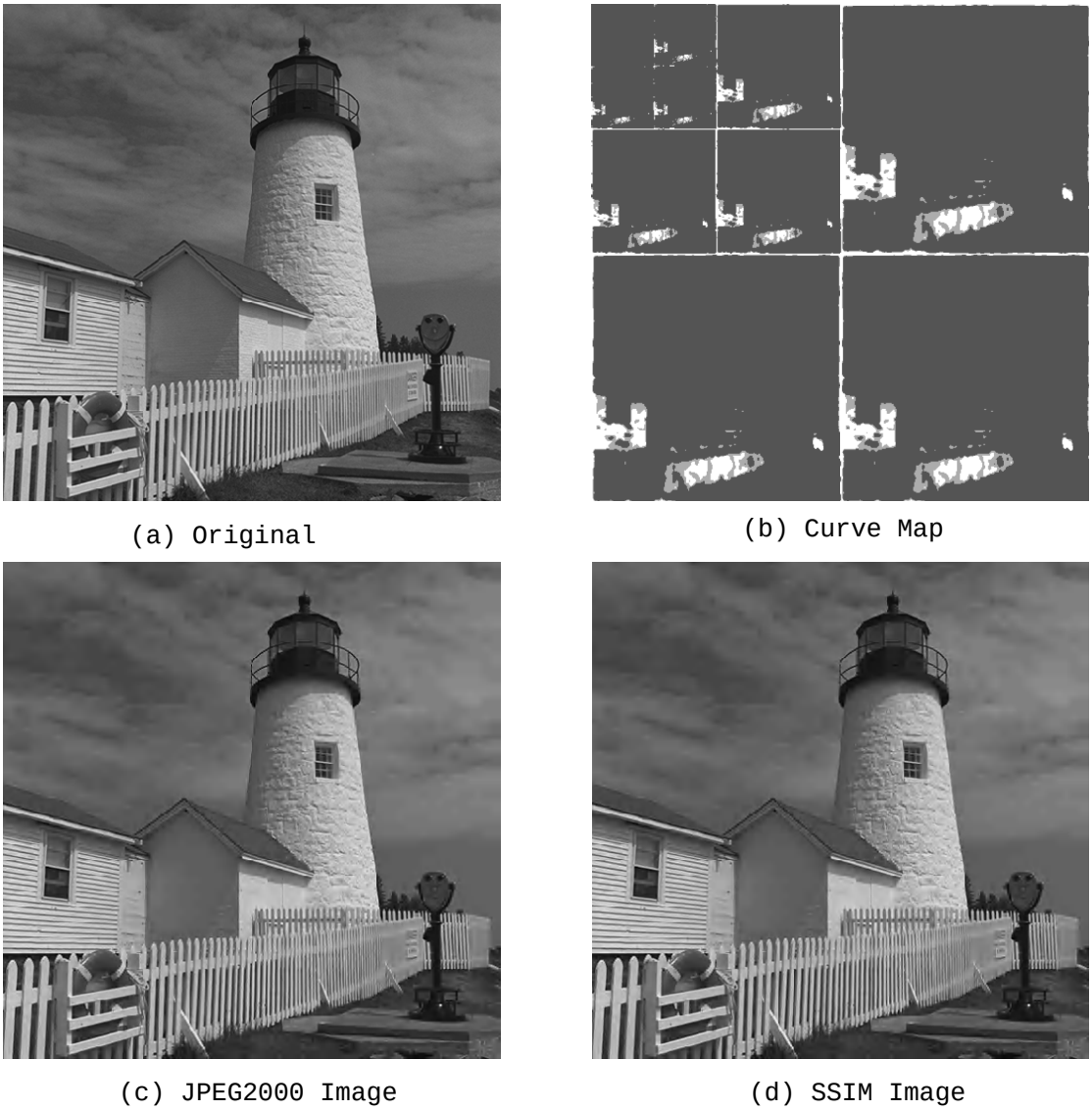
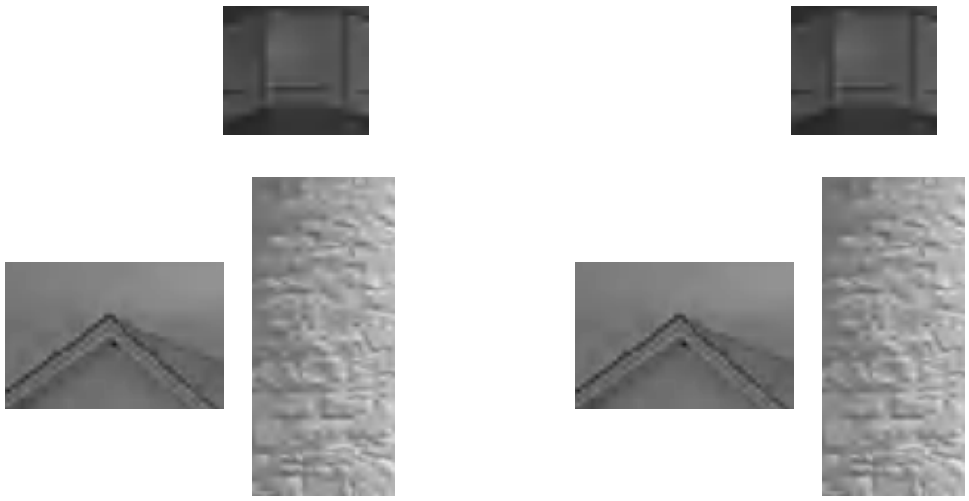


Figure B.5. Results of Proposed Scheme with JPEG2000 Coding: (a)Original Image; (b) Equalization Map: The brighter part means a higher value and more bits of coefficient will be removed; The darker part means a lower value and less bits of coefficient will be removed; (c) JPEG2000 Compressed Image; (d) JPEG2000 Compressed Image with SSIM Optimization; Rate = 0.4 bpp.

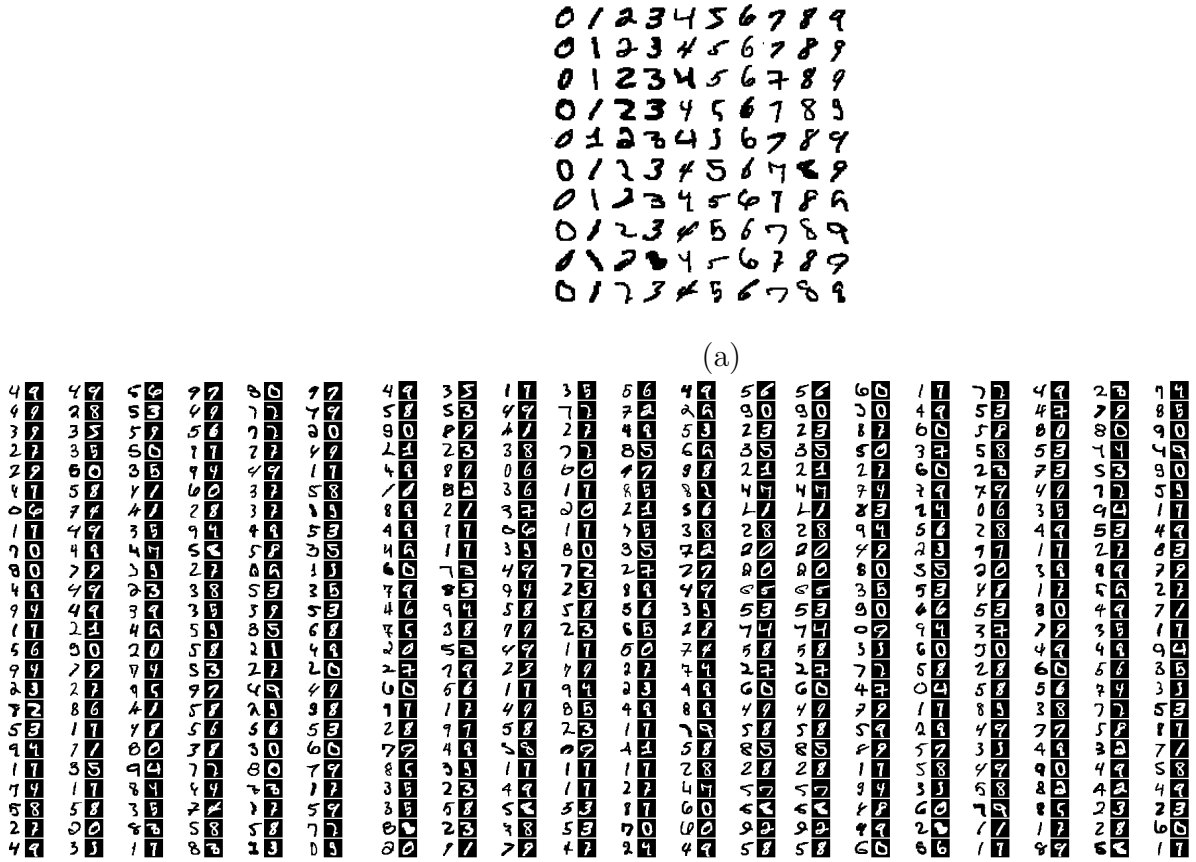


(a) Local Enlarged JPEG2000 Image (b) Local Enlarged SSIM Image

(c) JPEG2000 Quality Map (d) SSIM Qaulity Map

Figure B.6. Continue with B.5 (a)Local Enlarged JPEG2000 Image; (b)Local Enlarged SSIM Image; (c)JPEG2000 Quality Map; (d)SSIM Quality Map; Rate = 0.4 bpp.

APPENDIX C
EXPERIMENTAL RESULT OF CW-SSIM DIGIT RECOGNITION



(a)

(b)

Figure C.1. Results of CW-SSIM Digits Recognition (a) Template; (b) Incorrectly Recognized Digits; Template set size is 10.

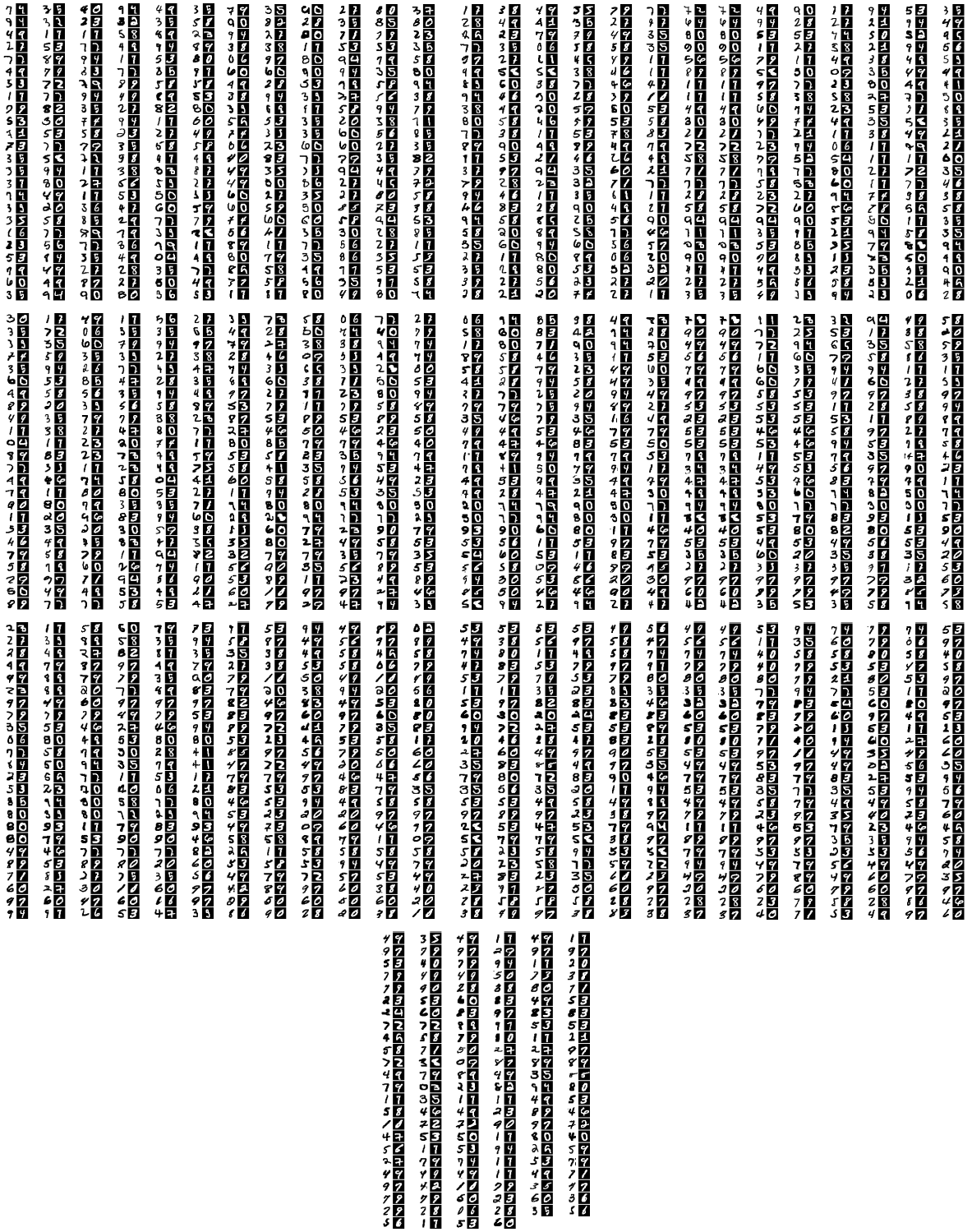
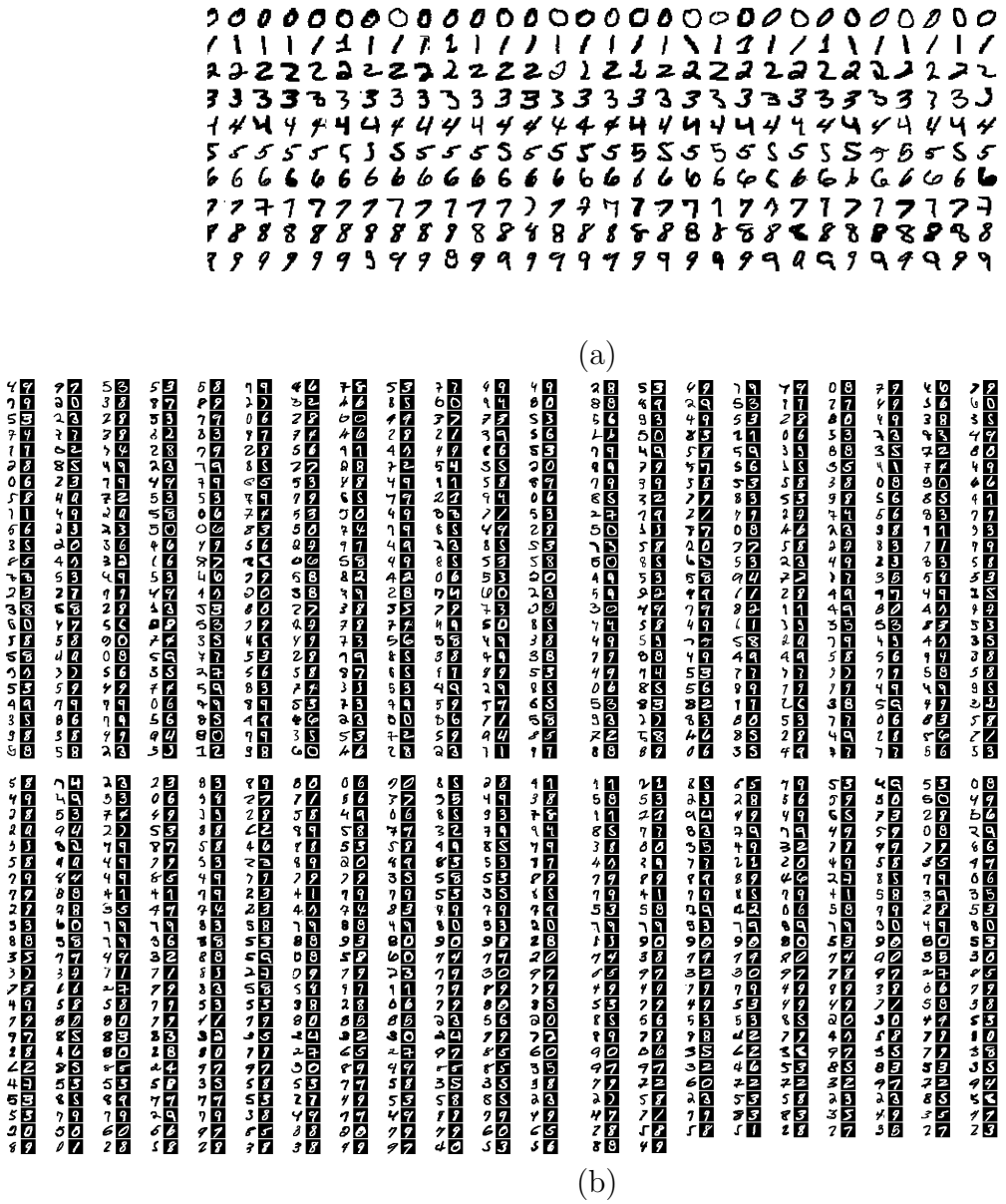


Figure C.2. Continues with Figure C.1; Incorrectly Recognized Digits; Template set size is 10.

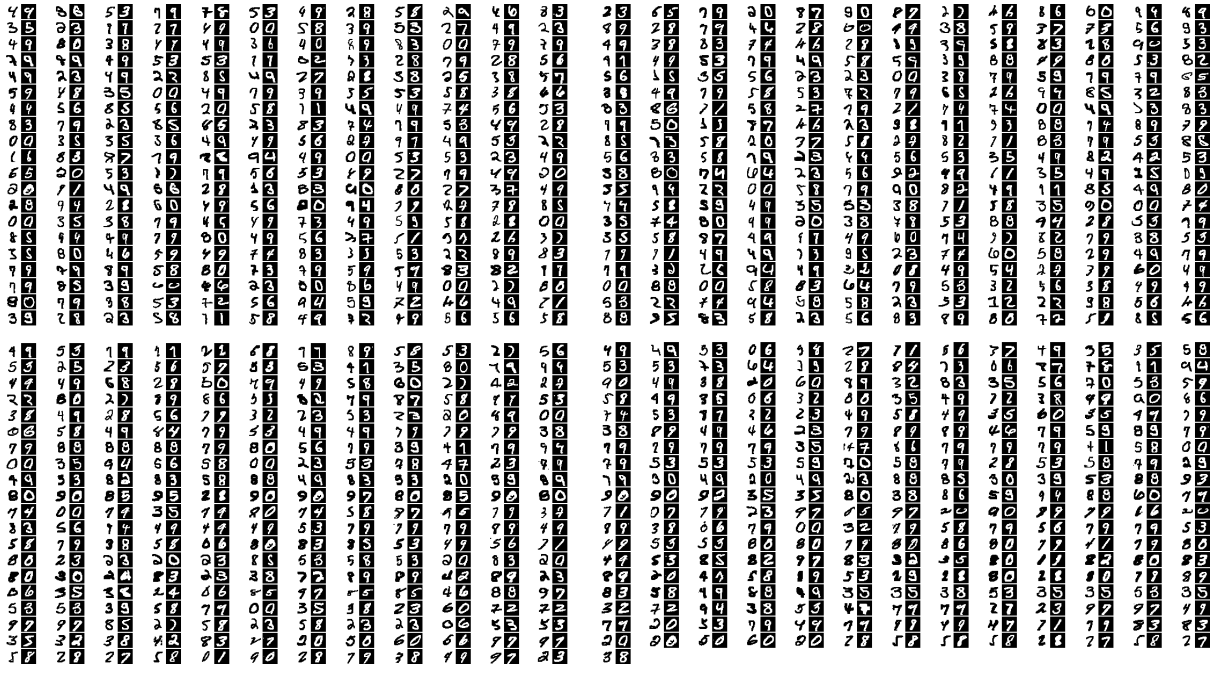


(a)

(b)

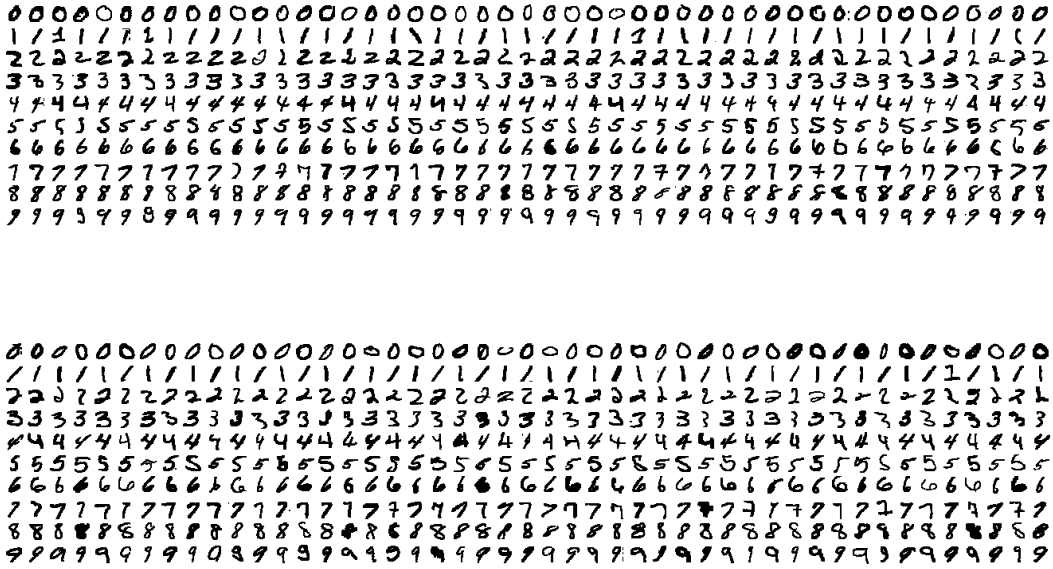
Figure C.3. Results of CW-SSIM Digits Recognition (a) Template; (b) Incorrectly Recognized Digits; Template set size is 30.

(a)

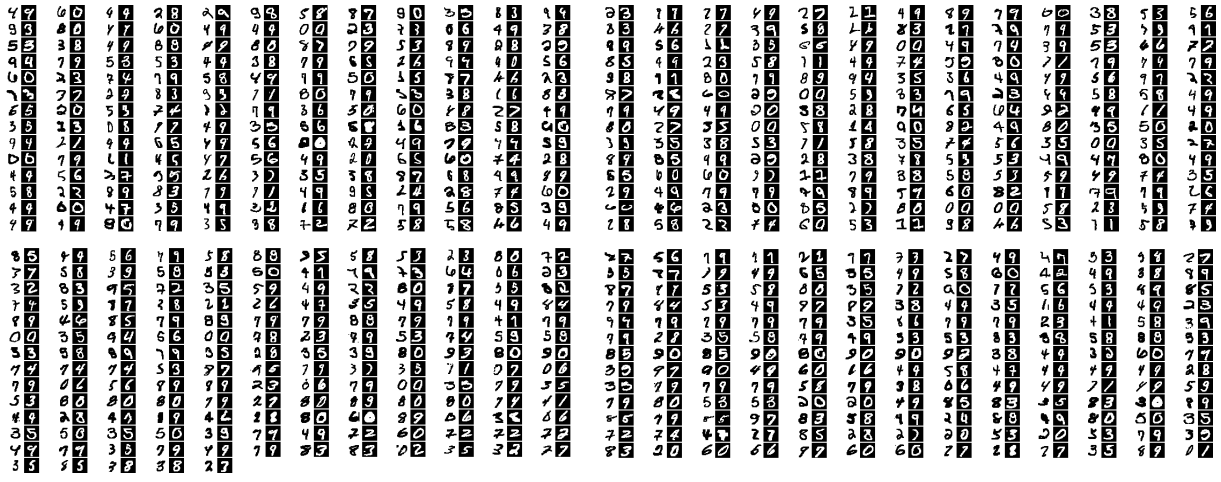


(b)

Figure C.4. Results of CW-SSIM Digits Recognition (a) Template; (b) Incorrectly Recognized Digits; Template set size is 50.



(a)



(b)

Figure C.5. Results of CW-SSIM Digits Recognition: (a) Template; (b) Incorrectly Recognized Digits; Template set size is 100.

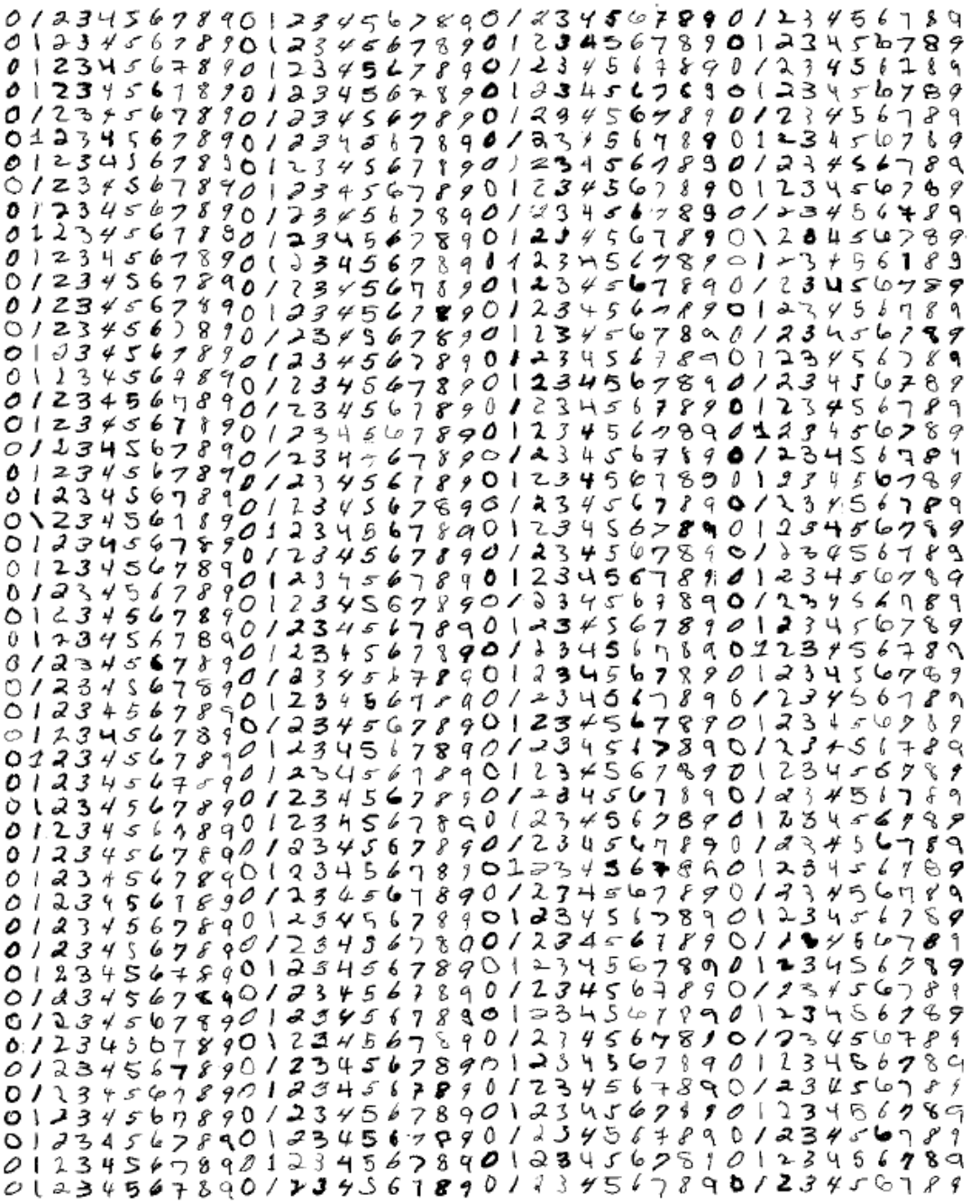


Figure C.6. Results of CW-SSIM Digits Recognition: Templates; Template set size is 200.

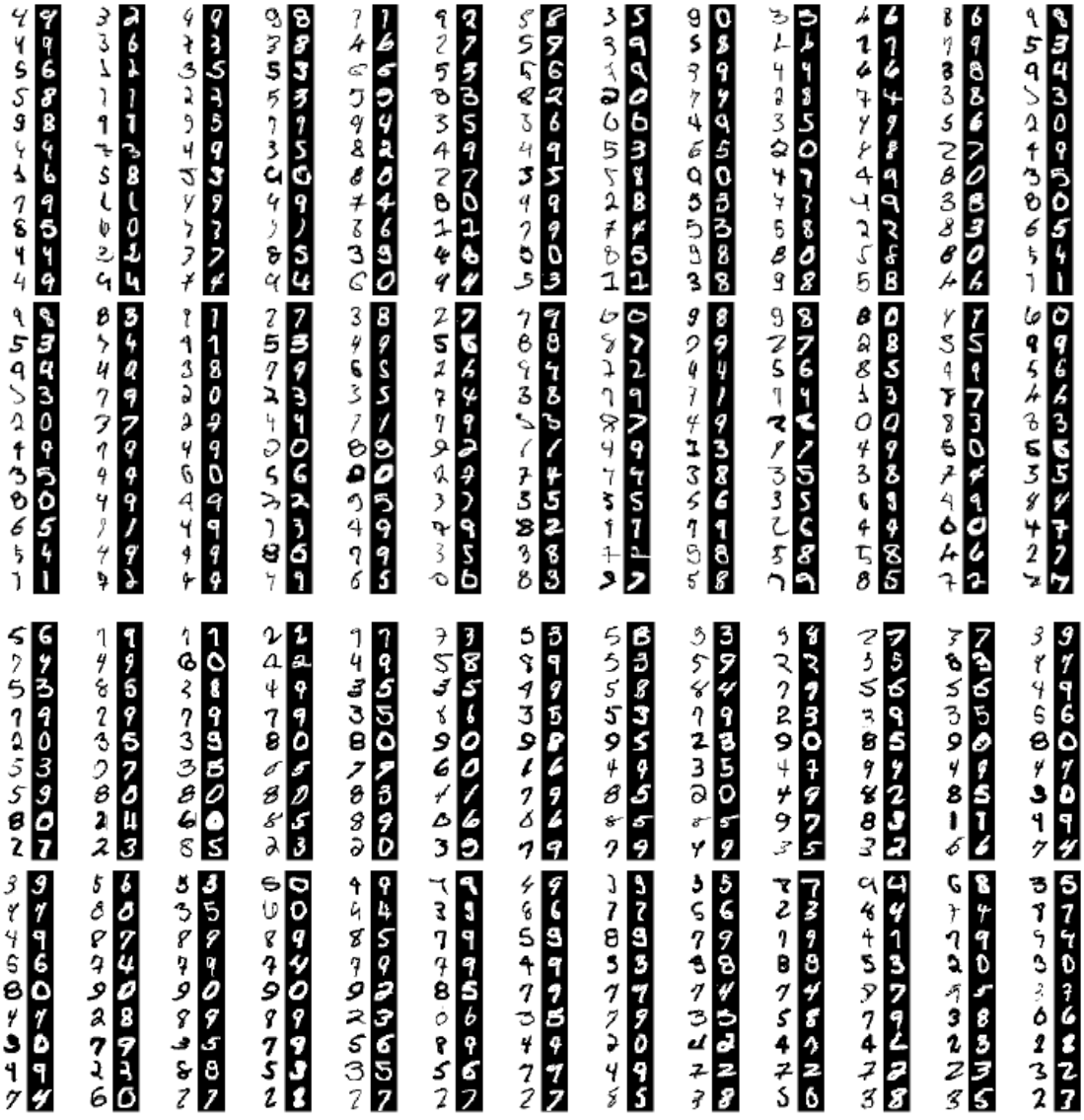


Figure C.7. Results of CW-SSIM Digits Recognition: Incorrectly Recognized Digits; Template set size is 200.

APPENDIX D

LIST OF ACRONYMS

Acronyms	Description
AI	Average Improvement
CSF	Contrast Sensitivity Function
CW-SSIM	Complex Wavelet Domain Structural Similarity Index Measurement
DCT	Discrete Cosine Transform
DWT	Discrete Wavelet Transform
EBCOT	Embedded Block Coding with Optimized Truncation
EZW	Embedded Zerotree Wavelet
FB	Filter Bank
FR	Full Reference
GQMF	Generalized Quadrature Mirror Filter
HVS	Human Visual System
ICT	Irreversible Color Transform
ICWP	Information Content-Weighted Pooling
IDWT	Inverse Discrete Wavelet Transform
JBIG	Joint Bi-level Image experts Group
JND	Just-Noticeable Distortion
JPEG	Joint Photographic Experts Group
LIVE	Laboratory of Image and Video Engineering at The University of Texas at Austin
LQDWP	Local Quality Distortion-Weighted Pooling

MAE	Mean Absolute Error
MNIST	Modified Database of the National Institute of Standards and Technology
MND	Minimally Noticeable Distortion
MOS	Mean Opinion Score
MP	Minkowski pooling
MSE	Mean Square Error
NR	No Reference
PCRD	Post-Compression Rate-Distortion
PR	Perfect-Reconstruction
PSNR	Peak Signal to Noise Ratio
PS	Pooling Strategy
QMF	Quadrature Mirror Filter
RCT	Reversible Color Transform
ROCC	Spearman Rank Order Correlation Coefficient
RR	Reduce Reference
SA	Spatial Average;
SPIHT	Set Partitioning in Hierarchical Trees
SSIM	Structural Similarity Index Measurement
UMD	Uniformly Maximally Decimated
VDP	Visible Difference Predictor
VQEG	Video Quality Experts Group

REFERENCES

[1] Y. LeCun, "The MNIST database of handwritten digits," *http://yann.lecun.com/exdb/mnist/*.

[2] Z. Wang and A. C. Bovik, *Modern Image Quality Assessment*. Morgan and Claypool, Mar. 2006.

[3] H. Wu and K. R. Rao, *Digital Video Image Quality and Perceptual Coding*. CRC press, 2006.

[4] J. L. Mannos and D. J. Sakrison, "The effects of a visual fidelity criterion on the encoding of images," *IEEE Trans. Information Theory*, vol. 4, pp. 525–536, 1974.

[5] A. B. Watson, *Digital Images and Human Vision*. Cambridge, Massachusets: The MIT Press, 1993.

[6] Z. Wang, A. C. Bovik, and E. P. Simoncelli, *Handbook of Image and Video Processing*. Academic Press, 2005.

[7] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Processing*, vol. 13, pp. 600–612, Apr. 2004.

[8] Z. Wang and E. P. Simoncelli, "Translation insensitive image similarity in complexwavelet domain," in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Processing*, Mar. 2005.

[9] "JPEG2000 verification model 7.0 software," *ISO/IEC/JTCI/SC29/WG1 N2415*.

[10] "JASPER JPEG2000 software," http://www.ece.uvic.ca/ mdadams/jasper/.

[11] "JJ2000 v. 4.2," *ISO/IEC/JTCI/SC29/WG1 N2136*.

[12] Z. Liu and L. J. Kurum, "JPEG2000 encoding with perceptual distortion control," *Proc. IEEE Int. Conf. Image Proc.*, vol. 1, pp. 637–640, 2003.

[13] S. Daly, "The visible difference predictor: An algorithm for the assessment of image fidelity," in *Proc. SPIE*, vol. 1616, 1992, pp. 2–15.

[14] J. Lubin, "The use of psychophysical data and models in the analysis of display system performance," in *Digital images and human vision*, A. B. Watson, Ed. Cambridge, Massachusetts: The MIT Press, 1993, pp. 163–178.

[15] T. N. Pappas and R. J. Safranek, "Perceptual criteria for image quality evaluation," in *Handbook of Image and Video Proc.*, A. Bovik, Ed. Academic Press, 2000.

[16] P. C. Teo and D. J. Heeger, "Perceptual image distortion," in *Proc. IEEE Int. Conf. Image Proc.*, 1994, pp. 982–986.

[17] A. B. Watson, "DCT quantization matrices visually optimized for individual images," *Proc. SPIE*, vol. 1913, pp. 202–216, 1993.

[18] A. B. Watson, G.Y.Yang, J.A.Solomon, and J.Villasenor, "Visibility of wavelet quantization noise," *IEEE Trans. Image Processing*, vol. 6, pp. 1164–1175, 1997.

[19] I.Hontsch and L.Karam, "APIC: Adaptive perceptual image coding based on subband decomposition with locally adaptive perceptual weighting," *Proc. IEEE Int. Conf. Image Proc.*, vol. 1, pp. 37–34, 1997.

[20] W. Osberger, N. Bergmann, and A. Maeder, "An automatic image quality assessment technique incorporating high level perceptual factors," *Proc. IEEE Int. Conf. Image Proc.*, pp. 414–418, 1998.

[21] Z. Wang, L. Lu, and A. C. Bovik, "Foveation scalable video coding with automatic fixation selection," *IEEE Trans. Image Processing*, vol. 12, no. 2, pp. 243–254, Feb. 2003.

[22] S. Saha, "Image Compression - from DCT to Wavelets: An Overview," in *ACM Crossroads*, vol. 6, no. 3, June 2000, http://www.acm.org/crossroads/xrds6-3/sahaimgcoding.html.

[23] J. M. Shapiro, "Embedded image coding using zerotrees of wavelets coefficients," *IEEE Trans. Signal Processing*, vol. 41, pp. 3445–3462, December 1993.

[24] A. Said and W. A. Pearlman, "A new, fast, and efficient image codec based on set partitioning in hierarchical trees," *IEEE Trans. Circuits and Systems for Video Tech.*, vol. 6, no. 3, pp. 243–250, June 1996.

[25] D. S. Taubman and M. W. Marcellin, *JPEG2000 Image Compression Fundamentals, Standards, And Practice*. Kluwer Academic Publishers, 2004.

[26] A. S. Lewis and G. Knowles, "Image compression using the 2-D wavelet transform," *IEEE Trans. Image Processing*, vol. 1, no. 2, pp. 244–250, Apr. 1992.

[27] P. M. M. Antonini, M. Barlaud and I. Daubechies, "Image coding using wavelet transformyear," *IEEE Trans. Image Processing*, vol. 1, no. 2, pp. 205–220, Apr. 1992.

[28] I. H. Witten, R. M. Neal, and J. G. Cleary, "Arithmetic coding for data compression," *Comm. of the ACM*, vol. 30, no. 6, pp. 520–540, 1987.

[29] A. Mazzarri and R.Leonardi, "Perceptual embedded image coding using wavelet transforms," *Proc. IEEE Int. Conf. Image Proc.*, vol. 1, pp. 586–589, 1995.

[30] D. Taubman, "High Performance Scalable Image Compression with EBCOT," *IEEE Trans. Image Processing*, vol. 9, no. 7, July 2000.

[31] P. Y. Simard, Y. A. LeCun, J. S. Denker, and B. Victorri, "Transformation invariance in pattern recognition – tangent distance and tangent propagation," *International Journal of Imaging System and Technology*, vol. 11, no. 3, pp. 181–194, 2001.

[32] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, pp. 2278–2344, 1998.

[33] M. Hoque and M. Fairhurst, "A moving window classifier for off-line character recognition," *Proceedings of the 7-th International Workshop on Frontiers in Handwriting Recognition*, pp. 595–600, 2000.

[34] R. Cox, B. Haskell, Y. LeCun, B. Shahraray, and L. Rabiner, "On the application of multimedia processing to telecommunications," *Image Processing, 1997. Proceedings., International Conference on*, vol. 1, pp. 5–8, 1997.

[35] P. Y. Simard, R. Szeliski, and J. Benaloh, "Using character recognition and segmentation to tell computers from humans," *International Conference on Document Analysis and Recogntion*, pp. 60–65, 2003.

[36] H. R. Sheikh, Zhou.Wang, A. C. Bovik, and L. K. Cormack, "Image and video quality assessment research at LIVE," *http://live.ece.utexas.edu/research/quality/.*

[37] Z. Wang and E. P. Simoncelli, "Stimulus synthesis for efficient evaluation and refinement of perceptual image quality metrics," *"Human Vision and Electronic Imaging III, Proc. SPIE"*, vol. 5292, pp. 99–108, Jan. 2004.

[38] H. R. Sheikh, A. C. Bovik, and G. de Veciana, "An information fidelity criterion for image quality assessment using natural scene statistics," *IEEE Trans. Image Processing*, vol. 14, pp. 2117–2128, Dec. 2005.

[39] W. Zeng, S. Daly, and S. Lei, "An overview of the visual optimization tools in JPEG2000," *Signal Processing: Image Communication*, pp. 85–104, 2002.

[40] M. Rabbani and P. W. Jones, "Digital image compression techniques," *SPIE Opt. Eng. Press*, 1991.

[41] Y. M. Yeung and O. C. Au, "Efficient Rate Control for JPEG2000 Image Coding," *IEEE Trans. Circuits and Systems for Video Tech.*, vol. 15, no. 3, pp. 335–344, 2005.

[42] F.W.Campbell and J.G.Robson, "Application of Fourier analysis to the visibility of gratings," *Journal of Physiology (London)*, no. 197, pp. 551–566, 1968.

[43] A. R. Calderbank, I. Daubechies, W. Sweldens, and B.-L. Yeo, "Wavelet transforms that map integers to integers," *Applied and Computational Harmonic Analysis*, vol. 5, no. 3, pp. 332–369, 1998.

[44] D. L. Gall and A. Tabatabai, "Sub-band coding of digital images using symmetric short kernel filters and arithmetic coding techniques," *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. 2, pp. 761–764, 1988.

[45] C. M. Brislawn, J. N. Bradley, R. J. Onyshczak, and T. Hopper, "The FBI compression standard for digitized fingerprint images," *Proc. SPIE*, vol. 2847, pp. 344–355, Aug. 1996.

[46] J. Portilla and E. P. Simoncelli, "A parametric texture model based on joint statistics of complex wavelet coefficients," *International Journal of Computer Vision*, vol. 40, pp. 49–71, 2000.

BIOGRAPHICAL STATEMENT

Xinli Shang was born in Henan, China, in 1974. He received his B.S. and M.S. degrees in Chemistry Engineering and Electrical Engineering from University of Electronic Science and Technology of China in 1998 and 2003, respectively. From 1998 to 2000, he was an engineer at Sobey Inc. He was a software engineer with Lucent Technology Bell Labs, China, from 2003 to 2005. His current research interests include image processing, wireless communications, and data communications.