USC-SIPI REPORT #406

Advanced Intra Prediction Techniques for Image and Video Coding

by

Yunyang Dai

August 2010

Signal and Image Processing Institute
UNIVERSITY OF SOUTHERN CALIFORNIA
Viterbi School of Engineering
Department of Electrical Engineering-Systems
3740 McClintock Avenue, Suite 400
Los Angeles, CA 90089-2564 U.S.A.

ADVANCED INTRA PREDICTION TECHNIQUES FOR IMAGE AND VIDEO CODING

by

Yunyang Dai

A Dissertation Presented to the
FACULTY OF THE USC GRADUATE SCHOOL
UNIVERSITY OF SOUTHERN CALIFORNIA
In Partial Fulfillment of the
Requirements for the Degree
DOCTOR OF PHILOSOPHY
(ELECTRICAL ENGINEERING)

August 2010

Table of Contents

List Of Tables

List Of Figures

Abstract

Intra prediction has been used in the H.264/AVC video coding standard to improve the coding efficiency of the intra frame. We present different intra prediction techniques that outperform the existing ones adopted by H.264/AVC and JPEG-LS in this research:

1. joint block/line-based intra prediction (JBLIP),

2. hierarchical (or multi-resolution) intra prediction (HIP), and

3. context-based hierarchical intra prediction (CHIP).

We consider two image/video coding scenarios: *lossy* compression and *lossless* compression. For lossy compression, we conduct a comprhensive study and show that the existing line-based prediction (LIP) technique adopted by the H.264/AVC standard can only be effective in smooth and simple edge regions. However, it is not as useful in predicting complex regions that contain texture patterns. To overcome this difficulty, we propose a JBLIP scheme with 2D geometrical manipulation to improve coding efficiency. The complexity of the JBLIP scheme is however quite hight due to the need to search the best matched block for the prediction purpose. Thus, we propose a fast search algorithm to reduce the coding complexity. The proposed JBLIP scheme outperforms the LIP scheme in H.264/AVC by up to 1.68dB in the PSNR improvement at the same bit rate.

Next, for lossless compression, we present an advanced intra frame coding using a hierarchical (or multi-resolution) approach called HIP. The objective is to support lossless image/video compression with spatial scalability. We analyze the characteristics of the underlying input signal characteristics and previously proposed signal modeling algorithms and show that most of the existing signal models cannot capture the dynamic signal characteristics through one fixed model. Hence, we propose a spatially scalable intra-prediction scheme that decompose signals according to their characteristics in the frequency domain. A block-based linear combination with edge detection and training set optimization is used to improve coding efficiency for complex textured areas in the EL. It is shown by experimental results the proposed lossless HIP scheme outperforms the lossless LIP scheme of H.264/AVC and JPEG-LS by a bit rate saving of 10%.

Finally, we analyze the inefficiency of the proposed lossless HIP scheme and present an enhanced hierarchical intra prediction coding called the context-based hierarchical intra prediction (CHIP). To save bits for the coding of modes, we propose a mode estimation scheme. To improve prediction accuracy, we employ the principal components analysis (PCA) to extract dominant features from the coarse representation of the base layer. The extracted features are clustered using a k-means clustering algorithm. Then, the context-based interlayer prediction (CIP) scheme is used to select the best prediction candidate without any side information. To enhance coding efficiency furthermore, an adaptive precoding process is performed by analyzing the characteristics of the prediction residual signal and a more accurate approach is proposed to estimate the context model. Experimental results show that the proposed lossless CHIP scheme outperforms the lossless LIP scheme of H.264/AVC and JPEG-LS by 16% in the bit rate saving.

Chapter 1

Introduction

1.1 Significance of the Research

Modern video coding standards such as H.264/AVC exploit different types of redundancy in the video content to allow efficient coding such as temporal redundancy and spatial redundancy. Temporal redundancy refers to the correlation between adjacent frames in a sequence of frames while spatial redundancy refers to the correlation between pixels of the same frame. To exploit them, H.264/AVC adopts inter-frame prediction and intra-frame prediction, respectively. Successive inter frames and intra frames are arranged into a group of pictures (GOP). A GOP always begins with an intra-frame, which is followed by a number of inter-frames.

Since the intra frame uses only spatial redundancy in the same frame, it is independent of other frames in the sequence. The intra frame is located in the beginning of a GOP to serve as the main reference frame and, as a result, it has a significant performance impact on the entire GOP. The inter frame exploits both temporal and spatial redundancy via sophisticated motion compensated prediction (MCP). A number of reference frames are searched against the current frame to find the most similar block which is called the predictive block. The prediction residual is then coded. Generally speaking, inter frame prediction often offers better coding efficiency than intra frame prediction. On the other hand, since MCP demands one or multiple reference frames in the memory buffer for compensation in inter frame coding, its complexity and memory cost is significantly higher.

In this research, we focus on effective intra frame prediction and coding techniques for both lossy and lossless coding requirements. These techniques can be applied to image and video coding applications. In the following, we examine the advantages of intra prediction and hierarchical intra prediction, and the need for their further improvement.

1.1.1 Intra Prediction

Intra prediction is important due to the following considerations.

- Complexity

  Although inter prediction has been studied more extensively than intra prediction due to its coding efficiency, recent research on high definition (HD) digital cinema and digital camcorder has brought renewed interest to intra frame coding. For HD video capturing, the large image size together with a high frame rate places an enormous burden on the encoder. In such a case, inter frame coding could be challenging due to its high memory and complexity requirements. All-intra-frame coding can be adopted as an alternative to meet both coding efficiency and the cost constraint.

- Reliability

  Intra frame coding provides greater reliability than inter frame coding since an error will only impact one frame instead of the entire GOP. A professional camcorder records at a higher speed than the normal human visual speed (*i.e.*, 60-90 frames a second vs. 30 frames a second), the loss of one frame is tolerable. On the other hand, the loss of one entire GOP is far less tolerable.

- Random Access

  After the digital content is created, it may have to be edited and post-processed before it can be released. This defines a complete "work flow" in the movie industry. Although the inter frame has higher coding efficiency, the display of a single frame demands the decoding of the entire GOP due to dependency between the current frame and frames before and after. For the normal display purpose, this is not an issue since the GOP has to be decoded in order. In the context of post-editing, frames have to be accessed in an arbitrary order, the need to decode extra frames becomes a burden. In addition, if a frame within a GOP is modified, the entire GOP will be affected since the error tends to propagate throughout the GOP.

An all-intra-frame sequence provides an advantage since each frame can be independently accessed and edited. Despite its many obvious advantages, the performance of intra-frame's coding still have room for further improvement. In addition, the camera of resolution 2K-by-2K was considered to be of professional quality just a few years ago, a camera of resolution 3K-by-3K under USD$3000 is available for advanced amateurs. Such a dramatic increase in the image capturing speed demands a more efficient intra-frame coding algorithm.

1.1.2 Scalabe Intra Prediction

As digital contents become more popular, the need to deliver them in a scalable fashion becomes an important issue. The need arises from two fronts: 1) content preview and editing for the on-board camcorder environment and studio enviorment and 2) content delivery over a a heterogeneous network. They are detailed below.

- Preview and Editing

  In the context of on-board preview, the limited LCD size (typically of 3-inch diagonally), decoder complexity and power restriction makes the on-board content preview at the original resolution impractical. Even in the studio setting, the large size of today's digital negative of up to 28K resolution per frame cannot be searched and indexed efficiently. Thus, an efficient and scalable video format is needed.

- Content Delivery

  For content delivery over a heterogeneous network, there is a need to tailor its format to different networking and display requirement, which demands an efficient scalable solution. Although high quality video is desired, video contents often have to be searched, indexed and previewed over the network. Performing such a task on the full resolution video stream encounters some issues such as complexity, power consumption, processing time and communication delay. In addition, the increased number of end-devices with different hardware and resolutions also imposes the need of scalable video.

Being similar to the single-layer counterpart, intra prediction in SVC utilities line-based spatial correlation within each enhancement layer. Sophisticated intra prediction in H.264/SVC also utilities inter-layer correlation by providing inter-layer prediction for the enhancement layer. However, even with inter-layer prediction, the performance of intra-frame coding in H.264/SVC is still substantially worse than its single-layer counterpart. This drawback has a negative impact on its coding gain. This deficiency shows the need and an opportunity for a more efficient scalable intra coding algorithm.

1.1.3 Lossless Image/Video Compression

Digital storage capability has increased by more than 1000 folds and the cost of per storage capacity decreased significantly in recent years. Thus, more attention has been shifted from the low-bit-rate coding to the high-bit-rate coding. To take medical imaging as an example, it often generates sequences of strongly correlated images as in computerized axial tomography (CAT), magnetic resonance imaging (MRI) and positron emission tomography (PET). These images need to be stored losslessly. For other applications such as scientific geographic information, and professional digital photography, lossless coding is also preferred due to the requirement to store and access original contents without any distortion. The importance of lossless coding has been recognized so that even the consumer/amateur video processing systems have already have the lossless compression encoder/decoder unit built on board.

1.2 Review of Previous Work

In this section, we will review some of the previous work that is closely related to our current research.

1.2.1 Pixel-based Intra Prediction (PIP)

The JPEG Committee of the International Standards Organization (ISO) issued a call to solicit proposals for lossless and near-lossless compression of continuous tone pictures in 1994. This call for proposal resulted in an increased research interest on lossless image compression schemes, especially in the area of lossless differential pulse code modulation (DPCM) or lossless predictive coding. The low complexity lossless coder (LOCO-I)[49] and the context-based adaptive lossless image codec (CALIC) [52] offered the most superior performance [30]. Both of them are based on the context modeling of image pixels. Context switching is conducted with an heuristically tuned switching function under a stationary image model. As most of natural images contain abrupt changes in local statistics, a large prediction error could occur around edges and textures. The linear prediction with least-squares optimization can adapt to the local statistics well and, hence, deliver noticeable improvement over JPEG-LS and CALIC.

1.2.2 Line-based Intra Prediction (LIP)

H.264/AVC intra prediction was developed for the coding of intra frame in video coding. Several latest intra prediction algorithms were compared in [34], which shows that H.264/AVC intra prediction delivers comparable or, in many cases, better coding performance than other solutions using the line-based prediction (LIP). That is, pixel values from neighbors of the current block are extrapolated to construct a prediction block and different directional predictions are defined by modes. The LIP scheme exploits the spatial correlation that may exist between the predicted and actual pixels in a target block. There has been a large amount of effort on improving the efficiency of H.264/AVC intra prediction. Conklin [17] proposed to extend the coded pixel line to include more pixel samples for prediction.

1.2.3 Displacement-based Intra Prediction (DIP)

Yu and Chrysafis [45] extended this idea further by introducing an idea similar to motion compensation into intra prediction, where a displacement vector that measures the distance between the reference and the target blocks in the same frame is recorded. This method is referred to as the displaced intra prediction (DIP) here. As revealed by experimental results in [45], the DIP is scene/complexity-dependent. As spatial correlation allows better intra prediction results, Shiodera *et al.,* [48] proposed to change the coding order of subblocks, *i.e.,* to encode the sub-block on the right bottom position first. A template matching technique was introduced in [4, 47] to improve coding efficiency of a textured surface.

1.2.4 Inter-Layer Intra Prediction (ILIP)

Due to the layered structure in the H.264/SVC, its intra prediction is done differently for the base and the enhancement layers. The LIP in H.264/AVC is still adopted as the base layer prediction for H.264/SVC. For the enhancement layer, additional modes are used to improve coding efficiency via inter-layer correlation. The I_BL mode, which copies the co-located block in the upsampled based layer (using an efficient upsampling filter [39]) as one of the prediction to the target block in the enhancement layer, is called the inter-layer intra prediction in H.264/SVC. Park *et al.* [32] observed a global/local phase shift between the base and the enhancement layers and claimed that the upsampling filter could produce an output sample that is situated at a different location as compared to the input sample. Hence, a frame-based (or a block-based) shift parameter can be used to improve coding efficiency. Wong *et al.* [50] proposed to add the right and down pixel lines in the base layer to improve the prediction performance.

1.2.5 Multi-hypothesis Motion Compensated Prediction (MH-MCP)

Multi-hypothesis motion compensated prediction (MH-MCP) was proposed in [42] to further exploit the long term prediction and the multiple reference frame structure. Theoretical investigation conducted for MH-MCP in [18] has shown good potential for further improvement in inter prediction. MH-MCP uses the temporal correlation between frames for prediction. However, due to the high complexity in optimizing coefficients and the prediction signal simultaneously, coefficients are often fixed as a constant (average) in each prediction [15].

1.3 Contribution of the research

Several contributions have been made in this research. First, we consider an enhancement of the single-layer intra prediction for high-bit-rate coding in Chapter 3. The specific contributions of this chapter are summarized below.

- Borrowing techniques from fractal image coding, we propose a block-based intra prediction (BIP) scheme with 2D geometrical manipulations. Previous line-based intra prediction is only effective in homogeneous areas. For frames with complex, repeating patterns such as architecture and natural life video streams, our algorithm is highly effective.

- We propose a rate-distortion optimization (RDO) procedure that jointly selects the best mode among LIP and BIP modes. A block classification method is used to identify blocks that can be well predicted by LIP alone.

- We develop a zone-based fast search method to reduce complexity, which results in little rate-distortion loss as compared with the full search. A self-correlation analysis of the local image region is conducted to identify regions of different properties. A translational vector prediction scheme is proposed to first establish a near optimal search position. A candidate block pruning method is further proposed to assist in fast convergence to the optimal position.

- The proposed joint block/line-based intra prediction (JBLIP) scheme can achieve a significant coding gain of up to 1.68dB in PSNR with the same bit rate for images with various resolutions. The proposed fast search scheme can obtain an average 10 times speedup compared to the full search method with very little performance degradation.

Then, we consider an intra prediction scheme with built-in spatial and quality scalability for lossless coding in Chapter 4. The contributions of this chapter are given below.

- We conduct an in-depth analysis on previous signal correlation modeling research and show that most of the models cannot provide a good approximation to the underlying signal characteristics and, hence, the prediction performance degrades.

- To address the shortcoming of previous signal modeling studies, we adopt a novel hierarchical structure to achieve scalability with a spatial domain frequency decomposition scheme. Furthermore, due to different signal statistics existing in the base and the enhancement layers, different predictors are used to maximize the decorrelation among layers.

- Due to the existence of high frequency components in the enhancement layer, the LIP is ineffective in textured areas. Block-based solutions such as the BIP demands a repetative pattern to be effective. We propose a block-based linear prediction scheme, which produces a small prediction error in high textured regions while demanding no side information.

- To optimize prediction and reduce complexity, we propose a layer-based prediction with edge detection. The proposed frequency decomposition scheme would produce an enhancement layer that contains a large percentage of flat areas and complex textured areas in the remaining areas. As a result, a novel edge detection mechanism is used to identify edges of the enhancement layer and the linear prediction can be applied selectively.

- To improve the performance of the linear prediction scheme furthermore, we optimize the training set. That is, instead of using a fixed training window or selecting predictions with explicitly specified displacement vectors, the training set is optimized using the base layer statistics. This training set selection scheme allows coefficients to be optimized according to the target block features. In addition, no extra side information is needed. The decoder can obtain the same training information and calculate the coefficients accordingly.

- The proposed hierarchical intra prediction (HIP) cuts down the overhead information significantly and allows a small block size to be employed without a rate penalty. The proposed solution can acheive a 12-14% bit rate saving as compared with the H.264 lossless intra prediction and JPEG-LS. At the same time, the proposed solution enables spatial/quality scalability.

Finally, we consider several enhancements to the hierarchical intra prediction (HIP) scheme for lossless coding in Chapter 5. The contributions of this chapter are given below.

- We first investigate issues associated with the HIP in Chapter 4 and propose a model-free prediction scheme to eliminate the side information.

- We analyze the HIP prediction residual and show that, for some textured areas, the linear prediction introduced in Chapter 4 does not offer good results if the training set is not chosen properly. To address this issue, we extract features from the BL using the Principal Component Analysis (PCA). A k-means clustering algorithm is then used to cluster these features to form the context.

- A context-based interlayer prediction (CIP) is proposed based on available prediction samples within the same context. To improve prediction accuracy, their N-th casual neighbors on the EL were used to evaluate the best prediction sample. This CIP method can be integrated with the training set optimization technique so that no side information is needed for mode selection.

- We conduct an analysis on the probability of DCT coefficients for the traditional and the HIP coding environments and observe some unique characteristics with the prediction residual generated by the CIP, which can be exploited to improve its coding efficiency.
- The context-based hierarchical intra prediction (CHIP) is proposed for lossless image coding without any side information. The proposed CHIP scheme outperforms the H.264 LS and JPEG-LS with a bit rate saving of 16-24% with built-in spatial/quality scalability.

1.4 Organization of the Dissertation

The rest of this dissertation is organized as follows. Related research background on predictive coding for a single layered structure and scalable video is reviewed in Chapter 2. The joint block/ line-based intra prediction (JBLIP) scheme is presented in Chapter 3. The hierarchical intra prediction (HIP) for lossless image/video coding is proposed in Chapter 4. The context-based hierarchical intra prediction (CHIP) for lossless image/video coding is proposed in Chapter 5. Finally, concluding remarks and future work are given in Chapter 6.

Chapter 2

Research Background

Predictive coding is one of the fundamental principles behind existing image and video coding techniques. It is a statistical estimation procedure where future random variables are predicted from past and present observable random variables [20]. For an input signal sample of x_i, the prediction is carried out based on a finite set of past samples $x_1, x_2, ..., x_{i-1}$ as $\hat{x_i} = f(X) = f(x_1, x_2, ..., x_i)$. The optimal predictor of $\hat{x_i}$ is based on minimizing the error between x_i and $\hat{x_i}$ as

$$\epsilon\{(x_i - \hat{x_i})^2\} \quad = \quad \epsilon\{[x_i - f(x_1, x_2, ..., x_i)]^2\}. \tag{2.1}$$

As there usually exists a very strong spatial and temporal correlation in an image or video frames, good predictive coding techniques become extremely powerful in minimizing the redundancy to achieve efficient coding. Predictive coding can be classified into two broad categories: prediction *without side information* and prediction *with side information.* In the following sections, we will introduce some of the most influential prediction algorithms that have been proposed.

2.1 Pixel-based DPCM

Differential Pulse Code Modulation (DPCM) is one of the most typical predictive coding schemes that generate prediction $\hat{x_i}$ based on neighboring data samples without extra side information as shown in Fig. 2.1. Most of existing DPCM methods such as JPEG-LS [49], CALIC [52] and least squares prediction [51] are performed on a pixel-to-pixel level to maximize the effect of inter-pixel correlation.

2.1.1 JPEG-LS Prediction

Weinberger *et al.* [49] proposed a pixel-based lossless compression algorithm for continuous-tone images, which is known as the low complexity lossless compression for Images (LOCO-I). It was later accepted by ISO/ITU as a standard for lossless and near-lossless compression as JPEG-LS. The prediction/modeling scheme used in JPEG-LS is simple yet effective. As shown in Fig. 2.1, the input data sample to be predicted is labeled by x. The prediction is conducted based on the casual set $\{a, b, c, d\}$.

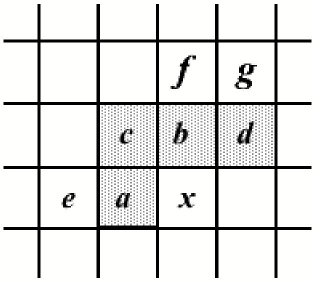


Figure 2.1: Neighboring pixel samples used in JPEG-LS and CALIC prediction schemes.

The prediction is done by adaptively learning a model conditioned on the local edge condition. To limit the complexity, a simplified edge detector called the median edge detector (MED) was adopted, which was developed based on the median adaptive predictor (MAP) [37]. The MED attempts to detect a vertical or horizontal edge given the casual set data. If an edge is not detected, the prediction uses $a + b - c$ as the prediction result for x if the current pixel belongs to the plane defined by three neighboring pixels with values a, b and c. This process tries to capture the expected smoothness of the region in the absence of an edge. The prediction rule can be written mathematically as

$$\hat{x} = \begin{cases} \min(a,b) & \text{if } c \geq \max(a,b), \\ \max(a,b) & \text{if } c \leq \min(a,b), \\ a+b-c & \text{otherwise.} \end{cases} \tag{2.2}$$

In the context determination stage, the context that conditions the encoding of the current prediction error $\epsilon = x - \hat{x}$ is derived based on the local activities from the pre-set casual template as $g_1 = d - a, \quad g_2 = a - c, \quad g_3 = c - b,$ and $g_4 = b - e$. These gradient estimators are used to describe the local smoothness or edginess around the current pixel to be encoded.

2.1.2 CALIC

The context-based, adaptive, lossless image codec (CALIC) [52] was proposed in response to the call-for-proposals by ISO/IEC JTC 1/SC 29/WG 1 (JPEG) in an attempt to establish a new international standard for lossless compression of continuous-tone image.

CALIC was actually a competing proposal to JPEG-LS. It operates in two modes: binary and continuous tones. Here, we consider the continuous-tone mode only. The codec consists of four major stages: gradient-adjusted prediction (GAP), context-selection and quantization, context modeling of prediction errors, and entropy coding.

As compared with JPEG-LS, GAP utilizes more neighbor pixel samples to form its prediction and the scheme adapts to the intensity gradients around the pixel to be coded. It is a non-linear prediction scheme in the sense that it assigns weights to neighboring pixels according to the estimated gradient of the area in horizontal and vertical directions, respectively. The horizontal and vertical gradients of the intensity function at pixel x can be estimated as

$$d_h = |a - e| + |b - c| + |b - d|, \tag{2.3}$$

$$d_v = |a - c| + |b - f| + |b - g|, \tag{2.4}$$

respectively. The values of d_h and d_v are used to detect the magnitude and orientation of edges in the area, and they allow necessary adjustments in the prediction to improve the prediction error in the presence of local activities.

To model the correlation between the prediction error and the local activity, an error estimator is defined as

$$\Delta = a \cdot d_h + b \cdot d_v + c \cdot |e_w|, \tag{2.5}$$

where d_h and d_v are given above and $e_w = x_{i-1} - \hat{x}_{i-1}$ is the previous prediction error. The e_w value is included in Eq. (2.5) since it is observed that large prediction errors tend to occur consecutively. By conditioning the prediction error distribution based on Δ, prediction errors can be classified into different classes. Then, the entropy coder in the following stage will estimate conditional probability $p(\epsilon|\Delta)$ to improve coding efficiency over using $p(\epsilon)$ alone.

2.1.3 Least Squares Prediction

The idea of least-squares (LS) prediction dates back to [7]. By adaptively optimizing prediction coefficients based on local statistics, LS prediction offers improved performance over JPEG-LS or CALIC [51, 53, 28, 46]. Lossless LS prediction was proposed using the linear combination of a small number N of casual neighbor pixels x_i of the current pixel (so-called "context") to form its prediction. An example is shown in Fig. 2.2 (b) with $N = 10$.

$$p = \sum_{i=1}^{N} c_i x_i, \tag{2.6}$$

where c_i are coefficients to form the linear combination of the prediction. A training window, usually takes the form of a double rectangular window, is built from its previously coded pixels to derive optimal coefficients as shown in Fig. 2.2(b). LS methods have been proven to be the most robust prediction method used in lossless coding due to its efficient adaptation to the spatially changing features in an image.

2.2 Block-based Predictive Coding

As it is found that the prediction error requires more bits while side information requires much less, later developed prediction algorithms for lossy coding used in image

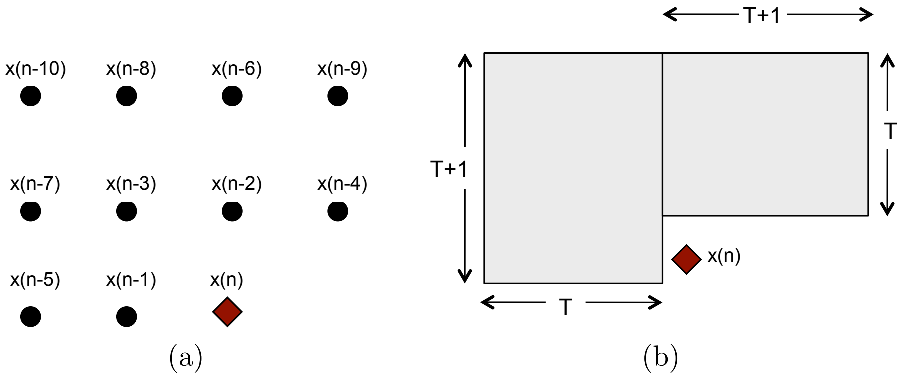


Figure 2.2: Least squares (LS) prediction using (a) 10 nearest casual neighbors and (b) the training window used to optimize prediction coefficients.

and video compression mainly fall in the category of *prediction with side information*. It improves the overall rate-distortion (RD0 performance by trading residual bits with side information bits.

2.2.1 H.264/AVC Intra Prediction

In contrast to previous video coding standards such as H.263 and MPEG-4, intra prediction in H.264/AVC is conducted in the spatial domain by referring to neighboring samples of previously coded blocks which are to the left and/or above the block. H.264/AVC offers a rich set of prediction patterns for Intra prediction to reduce the spatial redundancy of target video data. Comparison of various image coding methods reported that H.264/AVC intra prediction provides the best coding performance in most test images.

In H.264/AVC, a line-based intra prediction (LIP) is used. Pixel values from the neighbors of the current block are extrapolated to construct a prediction block. Different directional predictions are defined by modes and used to exploit the spatial correlation that may exist between the predicted and actual pixels in a target block. H.264/AVC intra prediction is carried out independently for both Luminance (green component) and Chrominance (red and blue) channels. The neighboring pixel samples from the coded blocks are used to form prediction blocks and a total of nine different prediction directions are specified as in Fig. 2.3.

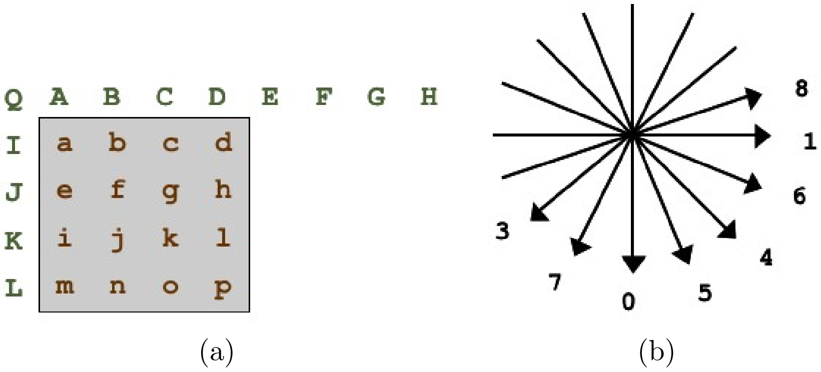


Figure 2.3: H.264/AVC intra prediction samples and directions.

Each 16x16 marcoblock can be further divided into sub-marcoblocks of 4x4 partitions and predict separately as Intra_16x16 and Intra_4x4. Each partition is predicted separately and a mode parameter is sent with the prediction residual to the decoder for reconstruction. In the latest *fidelity range extension* (FRExt) [43], Intra_8x8 block prediction modes have been added. Therefore, for the luma components, a marcoblock can use 16x16, 8x8 and 4x4 block prediction modes. For chroma components, a marcoblock can use either 8x8 or 4x4 prediction modes. There are four prediction modes in Intra_16x16 and nine prediction modes in Intra_8x8 and Intra_4x4 modes respectively. 16x16, 8x8 and

4x4 modes. Using the horizontal prediction mode as an example, the prediction can be expressed as

$$r_0 = p_0 - q_0, \tag{2.7}$$

$$r_1 = p_1 - q_0, \tag{2.8}$$

$$r_2 = p_2 - q_0, \tag{2.9}$$

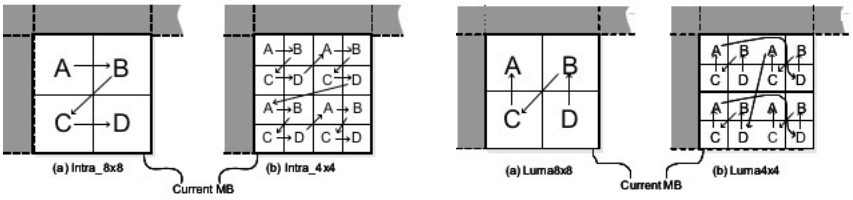
$$r_3 = p_3 - q_0. \tag{2.10}$$

It can be seen that the residual difference of r_0 through r_3 are predicted from the previous block line boundary samples, and the residual difference are sent to the decoder together with the mode information for correct reconstruction of the block.

2.2.2 Other Improvements on H.264/AVC Intra Prediction

To further improve the efficiency of H.264 intra prediction, different schemes are proposed. Due to the limited prediction samples available from the neighboring coded blocks, Conklin [17] proposed to extend the coded pixel line to include more pixel samples for prediction. Yu and Chrysafis [45] extended this idea further by introducing a displacement vector (DIP) which measures the distance between the reference and the target blocks in the same frame. As revealed by experimental results in [45, 25], DIP is scene/complexity-dependent.

As spatial correlation allows better intra prediction results, Shiodera *et al.* [48] proposed to change the coding order of subblocks, *i.e.*, to encode the sub-block on the right



(a) (b)

Figure 2.4: Sub-marcoblock partitions and coding order: (a) the original and (b) the one proposed by Shiodera *et al.* [48].

bottom position first. By doing so, the other three sub-blocks can utilize the spatial correlation from the right bottom sub-block and, thus, improve coding efficiency. As shown in Fig. 2.4, by coding the D blocks first, C blocks and B block would have more coded block prediction samples from D. Experimental results show that this approach can achieve around 0.2 dB PSNR improvement. A template matching technique was introduced in [4, 47] to improve coding efficiency of a textured surface.

2.2.3 H.264/SVC Intra Prediction

H.264/SVC is an extension of the current H.264/AVC. It offers spatial and quality scalability through a layered coding structure. Hence, the intra prediction for SVC is done differently for the base and the enhancement layers. Due to the compatibility requirement, the base layer intra coding remains the same as in H.264/AVC. The H.264 standard exploits the spatial correlation between adjacent macroblocks or blocks for intra prediction as described in Sec. 2.2.1.

For the coding of the enhancement layer, it introduces additional prediction method. Because of the layered structure, its corresponding frame in the base layer has already been coded and reconstructed, the information available from the base layer frame can

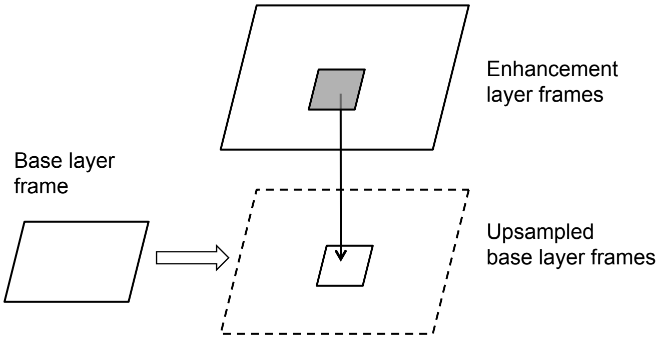


Figure 2.5: Inter-layer intra prediction in H.264/SVC.

be used for prediction in enhancement layer coding. Hence, H.264/SVC intra prediction introduced a I_BL mode, which copies the co-located block in the upsampled based layer (using an efficient upsampling filter [39]) as one of the prediction to the target block in the enhancement layer as shown in Fig. 2.5.

This I_BL mode offers good performance improvement to the coding of the enhancement layer. There have been many additional improvement schemes proposed for intra prediction in SVC to further explore the correlation between the base and the enhancement layer. For example, Park *et al.* [32] considered a global/local phase shift between the base and the enhancement layers. They observed that the upsampling filter could produce an output sample that is situated at a different location compared to the input sample and, therefore, proposed a frame-based (or block-based) shift parameter to improve coding efficiency. Wong *et al.* in [50] proposed to add the right and down pixel lines in the base layer to improve prediction.

2.2.4 JPEG-XR Intra Prediction

As compared with existing intra prediction schemes, JPEG-XR [2] (formerly known as HD Photo) has a different take on the prediction problem. Instead of predicting based on spatial correlation in intra blocks, intra prediction in JPEG-XR uses a LIP-based prediction in the frequency domain that predict frequency coefficients directly.

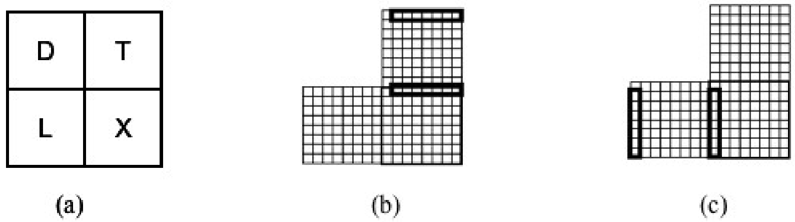
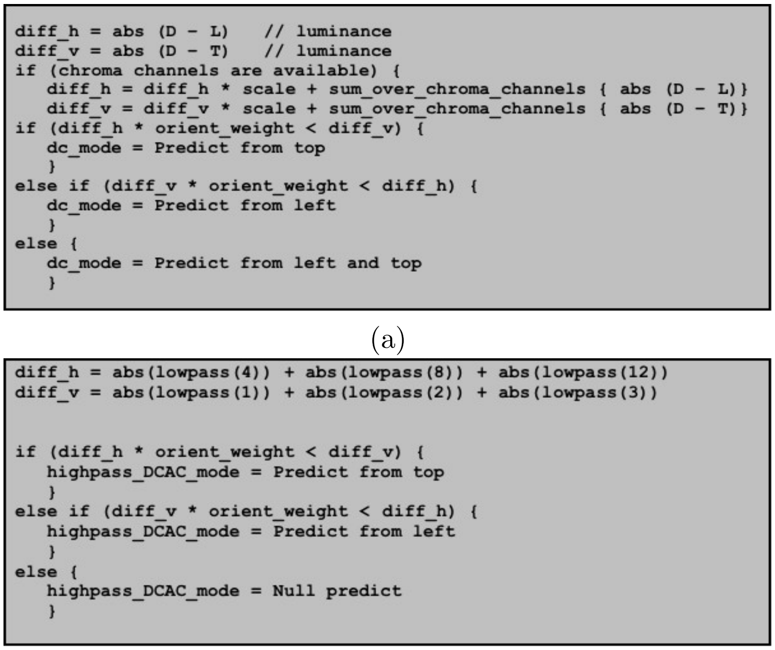


Figure 2.6: JPEG-XR prediction modes: (a) DC values to the left (L), top-left (D) and top (T) of the current DC value (X); (b) prediction from top for AC values and (c) prediction from left for AC values.

The DC and AC coefficients are predicted separately. There are four modes for DC prediction: prediction from left ($p = DC[L]$), from top ($p = DC[T]$), from left and top ($p = 1/2(DC[L] + DC[L])$) and no prediction. The left, top neighboring blocks are indicated in Fig. 2.6(a). The pseudo coding to perform the DC prediction is given in Fig. 2.7(a). The prediction of AC coefficients is determined by a lowpass_AC_mode, and a highpass_AC_mode. The lowpass_AC_mode is similar to the DC mode except that the prediction direction from left and top ($p = 1/2(DC[L] + DC[L])$) is not available. The highpass_AC_mode is shown in Fig. 2.6(b) and the pseudo code is shown in Fig. 2.7(b).

In summary, most recent intra prediction schemes adopt LIP in the pixel or the frequency domain from neighboring coded blocks. One advantage with LIP is that it is relatively simple to compute while providing good prediction for blocks in a homogeneous



```
diff_h = abs (D - L)    // luminance
diff_v = abs (D - T)    // luminance
if (chroma channels are available) {
   diff_h = diff_h * scale + sum_over_chroma_channels { abs (D - L)}
   diff_v = diff_v * scale + sum_over_chroma_channels { abs (D - T)}
if (diff_h * orient_weight < diff_v) {
   dc_mode = Predict from top
   }
else if (diff_v * orient_weight < diff_h) {
   dc_mode = Predict from left
   }
else {
   dc_mode = Predict from left and top
   }
```

(a)

```
diff_h = abs(lowpass(4)) + abs(lowpass(8)) + abs(lowpass(12))
diff_v = abs(lowpass(1)) + abs(lowpass(2)) + abs(lowpass(3))


if (diff_h * orient_weight < diff_v) {
   highpass_DCAC_mode = Predict from top
   }
else if (diff_v * orient_weight < diff_h) {
   highpass_DCAC_mode = Predict from left
   }
else {
   highpass_DCAC_mode = Null predict
   }
```

(b)

Figure 2.7: The pseudo code for (a) DC prediction and (b) AC high pass prediction in JPEG-XR.

region. However, for blocks with gradient changes, such as edges and textures, LIP may not have the ability to provide good prediction.

2.2.5 Multi-hypothesis Prediction

Sullivan [42] proposed a "multi-hypothesis motion compensation" (MH-MCP). Although it is an inter-frame prediction algorithm, we included it here since it is based on the same principle of previously mentioned least squares (LS) prediction.

MH-MCP extends the idea of pixel-based LS prediction to the block level. A total number of n blocks of $h_1, h_2, ..., hn$ from previous reference frames were used to form a single prediction to the target block s in the current frame. These blocks h_v are termed as *hypothesis*. The prediction block $\hat{s}$ is determined by a linear combination of the available hypothesis h_v as

$$\hat{s} \;=\; \sum_{v=1}^{N} c_v h_v, \tag{2.11}$$

where c_i are coefficients that determine the weight of each components for the prediction block. The prediction difference of $s - \hat{s}$ is coded together with the displacement vector of each hypothesis h_v and their coefficients h_v.

In order to find h_i, an exhaustive search combined with the rate-distortion optimization (RDO) is used. Theoretical investigation conducted by Girod [18] shows that a linear combination of multiple predictors can improve the motion compensation performance. However, there are a few drawbacks with the MH-MCP. Most of all, the search method used in finding the optimal set of h_i is computationally prohibitive. In the practical application, a compromise can be achieved by assigning the same coefficients to each of the hypotheses instead of finding the optimal coefficients individually [15]:

$$\hat{s} = \frac{1}{N} \sum_{v=1}^{N} h_i. \tag{2.12}$$

Second, although coefficients are constant for each hypothesis, the MH-MCP still has to send out the spatio-temporal displacement vectors for each optimal hypothesis $(\Delta_x, \Delta_y, \Delta_t)$ as side information to the decoder. The increase in the amount of side information could potentially mask the efficiency provided by the MH-MCP.

2.3 Conclusion

In this chapter, we reviewed some of key developments in the area of predictive coding. The challenges and requirements for efficient intra prediction in both classic single layered coding environment and in scalable coding environment were discussed. In Chapter 3, we will present an advanced intra prediction scheme that can deliver higher coding efficiency and better image quality. In Chapter 4, we will propose a scalable intra prediction with an alternative hierarchical structure and a more efficient prediction scheme that can deliver improved coding performance as compared to the existing SVC intra prediction.

Chapter 3

Joint Block/Line-based Intra Prediction

3.1 Introduction

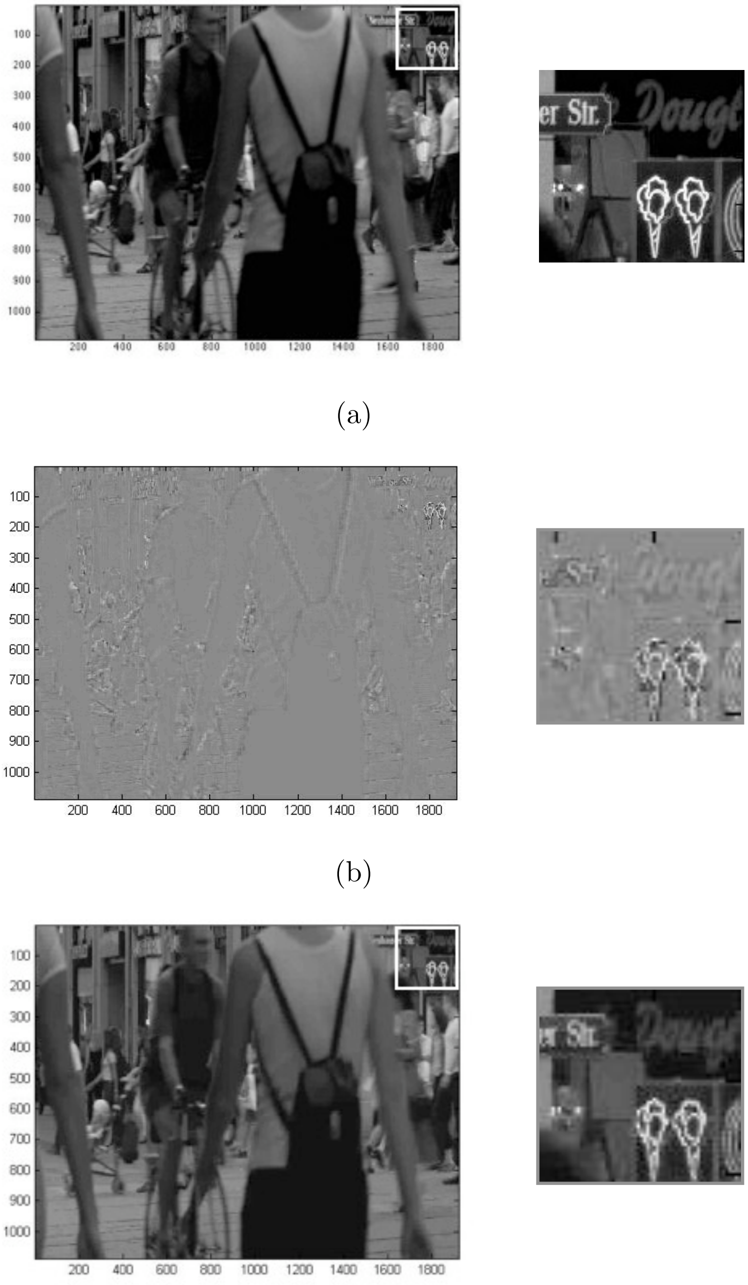
Current video coding standards such as H.264/AVC utilizes both temporal and spatial redundancy to improve coding performance. Temporal redundancy refers to correlations between adjacent frames within a sequence of a video along the timeline while spatial redundancy refers to the correlation between pixels within the same frame. Inter frame and intra frame are developed to exploit these correlations. In actual coding algorithms such as MPEG-2, MPEG-4 and H.264/AVC, inter frame and intra frame are arranged into a group of pictures (GOP). A GOP typically contains one intra-frame in the beginning of a GOP, serving as the key frame for the sequence followed by multiple inter-frame. The Intra frame exploits only spatial redundancy while the inter frame exploits both spatial and temporal redundancies. Due to the strong correlation in the temporal domain, inter frame prediction is typically more efficient than intra frame prediction. However, the inter frame is difficult to access in a random fashion for the editing purpose due to its dependence on other frames within the same GOP. As digital video production becomes popular in the movie industry as well as the consumer electronics market, an easily accessible and editable coding format is desirable. H.264/AVC intra prediction is one of the leading choices due to its random accessibility and overall excellent coding efficiency as compared with other image-based compression algorithms such as JPEG, JPEG-2000 and JPEG XR.

The JPEG standard offers one of the most influential image compression algorithms. It revolutionized image compression with the use of the block-based Discrete Cosine Transform (DCT), which allows good frequency decomposition. JPEG-2000 was standardized by the JPEG committee in 2000 and later incorporated in the digital cinema initiative (DCI) as a way to achieve improved coding efficiency and individual frame accessibility. JPEG XR, formerly known as HD photo and developed by Microsoft, is the latest attempt in still image coding. Comparison of various image coding methods was recently conducted in [34], which reported that H.264/AVC intra prediction provides the best coding performance in most test images.

Intra prediction in H.264/AVC is a line-based approach, where pixel values from neighbors of the target block are extrapolated to construct a prediction block. Different directional predictions are defined by modes and used to exploit the spatial correlation that may exist between the predicted and the actual pixels in a target block. For the luminance component, three types of intra prediction are available, Intra16x16, Intra8x8 and Intra4x4. Different directional predictions are defined by modes and used to exploit the spatial correlation that may exist between the predicted and the actual pixels in a target block. For example, nine prediction modes are adopted for Intra4x4. Similarly, nine and four prediction modes are used for Intra8x8 and Intra16x16, respectively. In this work, this prediction method used in H.264/AVC intra prediction is referred to as line-based intra prediction (LIP).

One advantage associated with the LIP is that it is simple to compute and efficient to predict blocks in a homogeneous region. However, the LIP has some performance limitation in the coding of textured blocks and images with high PSNR requirements. For blocks with gradient change or edges, the LIP may not be able to capture details, which often results in a clear contour of the residual image after prediction. An example is shown in Fig. 3.1. Fig. 3.1 (a) shows a frame taken from Pedestrian sequence of resolution 1920x1080 and a highlighted region in the upper-right corner of the original image. Fig. 3.1 (b) gives the H.264/AVC intra prediction error frame and the prediction residual in the highlighted region. Fig. 3.1 (c) shows the reconstructed frame and the highlighted region. We see a clear contour of prediction errors in Fig. 3.1 (b). This would result in a higher bit rate since high frequency components are not completely removed and, therefore, coding efficiency is negatively affected. Hence, for latest high definition image coding with high PSNR requirements, there is an urgent need to develop an intra coding algorithm that can deliver a high compression ratio while maintaining a desired PSNR level.

Due to the limited prediction samples available from the neighboring coded blocks in the LIP, Conklin [17] proposed to extend the coded pixel line to include more pixel samples for prediction. Yu and Chrysafis [45] extended this idea further by introducing a displacement vector (DIP) which measures the distance between the reference and the target blocks in the same frame. As revealed by experimental results in [45], the DIP is scene/complexity-dependent. As spatial correlation allows better intra prediction results,



(a)

(b)

(c)

Figure 3.1: Visual comparison for full and partially enlarged Pedestrian (a) original frame; (b) residual after H.264/AVC intra prediction; and (c) reconstructed frame.

Shiodera *et al.* [48] proposed to change the coding order of subblocks, *i.e.*, to encode the sub-block in the right bottom position first. By doing so, the other three sub-blocks can utilize the spatial correlation from the right bottom sub-block and, thus, improve coding efficiency. Experimental results show that this approach can achieve around 0.2 dB PSNR improvement. A template matching technique was introduced in [47], [4] to improve coding efficiency of a textured surface.

The rest of the chapter is organized as follows. A novel block-based intra prediction (BIP) coding technique is proposed to improve DIP in Sec. 3.2 by applying 2D geometrical manipulations to the reference block. An advanced joint BIP and LIP scheme (JBLIP) with Rate-Distortion optimization is proposed in Sec. 3.3. A fast mode selection scheme for JBLIP is described in Sec. 3.4. Experiment results are provided to show the effectiveness of the proposed intra prediction scheme in Sec. 3.5. Finally, concluding remarks are given in Sec. 3.6.

3.2 Block-based Intra Prediction (BIP)

The idea of the DIP can be illustrated in Fig. 3.2, where blocks B and P are the target and the reference blocks, respectively. For target block B to encode, there could exist another coded block P that exhibits visual similarity to B. To explore this similarity, a displacement based intra prediction method is adopted in the DIP. This idea is similar to motion prediction in inter-frame coding except that blocks P and B belong to the same frame (rather two different frames) now. However, the pool of reference blocks is relatively small, which affects the coding efficiency of the DIP. Being motivated by fractal image coding, one way to enrich the pool of the predictive blocks is to apply 2D geometrical manipulations to all reference blocks, such as scaling, rotation and reflection [11]. They will be examined in this section in detail. This block-based intra prediction scheme with 2D geometrical manipulations is denoted by the BIP for the rest of this chapter.

3.2.1 Translation

2D Translation enables the search for a reference block with the strongest similarity within a specified space. As shown in Fig. 3.2, block B is the target block, shaded region R_c is the previously coded region, and region R_f is the region to be coded. Clearly, the position vector (x', y') of reference block P is limited to region R_c. The prediction block P is obtained through 2D translation via

$$P(x', y') \quad = \quad B(x + \delta x, y + \delta y), \tag{3.1}$$

where $(\delta x, \delta y)$ is the translational vector (or the displacement). 2D Translation allows the encoder to search for a similar copy within the coded region by a spatial displacement. Once a good prediction block is found via 2D translation, additional transformation on the prediction block would introduce more matching options, which could allow exploitation of further reduce the prediction residual. Therefore, we would like to consider a larger pool of candidate blocks obtained by 2D geometrical manipulations such as scaling, rotation and reflection as explained in the next subsections.

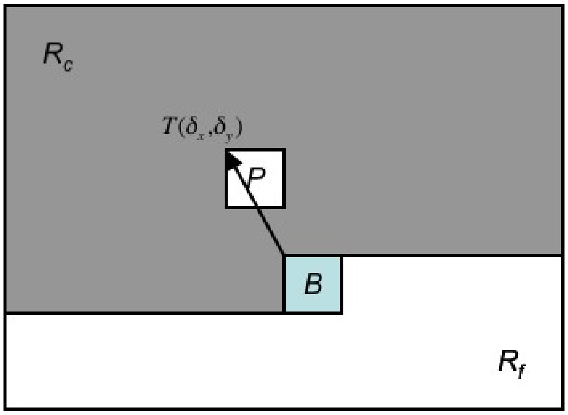


Figure 3.2: Block-based intra prediction with 2D translation.

3.2.2 Scaling

To achieve the 2D scaling of a prediction block, a upsampling interpolation scheme can be applied. As shown in Fig. 3.3, pixels A, B, C and D are existing input pixels in the coded region. If we intend to get a scaling of 1:2 along both vertical and horizontal directions, we need to interpolate values of half-pixels denoted by shaded circles; namely, *e, f, g, h, i*. A straighforward bilinear interpolation scheme can be applied to obtain the half-pixel position as:

$$e = [A + B + 1] >> 1, \tag{3.2}$$

$$f = [A + C + 1] >> 1, \tag{3.3}$$

$$g = [B + C + 1] >> 1, \tag{3.4}$$

$$h = [C + D + 1] >> 1, \tag{3.5}$$

$$i = [A + B + C + D + 2] >> 2. \tag{3.6}$$

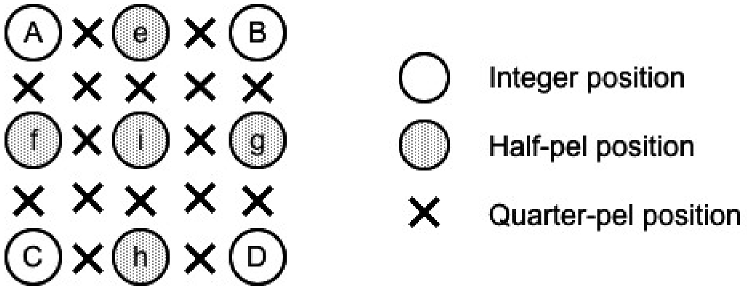


Figure 3.3: 2D scaling using bilinear interpolation.

Similarly, for 1:4 scaling, the pixel samples as indicated by the crosses on Fig. 3.3 can be obtained using the same formula as shown in Eq. (3.6). In this work, the 1:4 horizontal and vertical interpolations are implemented by the cascade of two 1:2 simple horizontal and vertical bilinear interpolations. A more sophisticated interpolation scheme such as the six-tap filter employed for subpixel interpolation in H.264/AVC, can also be used to provide better scaling results. After the prediction block is scaled, current block will perform a search similar to the process described in subsection 3.2.1. Thus, the position of the prediction block can be obtained as

$$P(x', y') \quad = \quad S(S_x, S_y). \tag{3.7}$$

It is important to point out the difference between the search of 2D scaled blocks as presented here and the subpixel search scheme used in H.264/AVC motion compensation. In H.264/AVC motion search, sub-pixel search is commonly viewed as a local refinement technique. That is, we refine the motion vector to allow it takes the half- or quarter-pixel positions as shown in Fig. 3.4(a). Basically, the size of the region covered by the reference and target blocks remains the same.

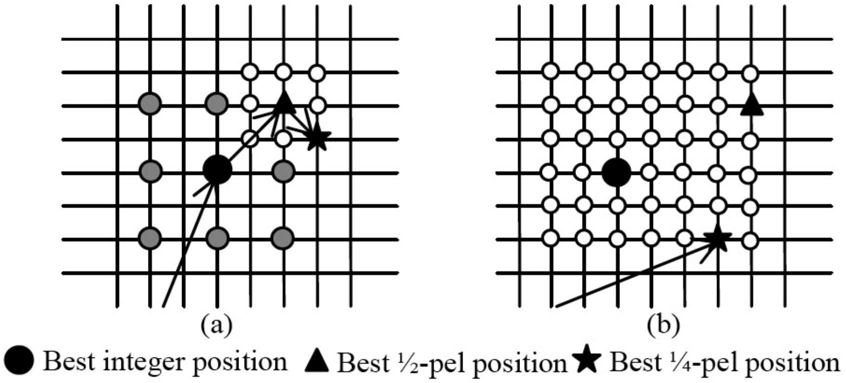


Figure 3.4: (a) H.264/AVC subpel motion search scheme v.s. (b) 2D scaling search scheme.

In contrast, with the proposed 2D scaling method, we can increase the size of the reference blocks 4 or 16 times and there is no definitive correlation between the best matched integer position and the best matched sub-pixel position in our current context. An example is given in Fig. 3.4 4(b), where we show the best matched positions under the integer-, half- and quarter-pixel interpolation results. Hence, each position bares the same weight and needs to be treated equally.

3.2.3 Rotation

Besides translation and scaling, the rotation of reference blocks in coded region R_c is also included to enhance prediction results. The rotation angle, θ, is chosen to be 90, 180, and 270 in our work. The use of only 90-degree interval rotation avoids the problem of block cropping, and efficient implementation of 90-degree rotation is widely available in both software and hardware. The rotation operation can create more prediction blocks with different orientations. This allows blocks with lines and edges to be better aligned. The prediction block from 2D rotation is thus obtained as

$$P(x', y') \quad = \quad P(x', y') \times R(\theta), \tag{3.8}$$

where

$$R(\theta) \quad = \quad \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}.$$

3.2.4 Reflection

The last 2D geometrical manipulation is block reflection with respect to central horizontal and vertical lines and diagonal lines. However, since the reflection with respect to the diagonal line does not provide significant performance enhancement in our experiments, we only include the reflection with respect to the x- or the y-axis. The prediction block can be obtained as

$$P(x', y') \quad = \quad P(x', y') \times F(2\theta), \tag{3.9}$$

where

$$F(2\theta) \quad = \quad \begin{bmatrix} \cos 2\theta & \sin 2\theta \\ \sin 2\theta & -\cos 2\theta \end{bmatrix}.$$

3.2.5 BIP with Enhanced Mode

In most cases, the BIP provides much lower prediction error compared to the LIP scheme. However, for some highly complexed textured area, the performance gain from the previously proposed 4x4 is still limited. To further improve the prediction error of these textured area, we incorporate a 2x2 block decomposition in the existing BIP based on 4x4 blocks. For complex blocks that do not have a good prediction result in the 4x4 BIP, we allow them to be further partitioned into 2x2 blocks and perform 2x2 block translation. This is considered an enhancement mode for the BIP. As 2x2 enhancement mode is designed to target the texture prediction, the 1:2 scaling in the previous geometrical manipulations are eliminated as 1:2 scaling to the prediction blocks to reduce the block variance and hence would not fit the search requirements in the 2x2 enhancement mode.

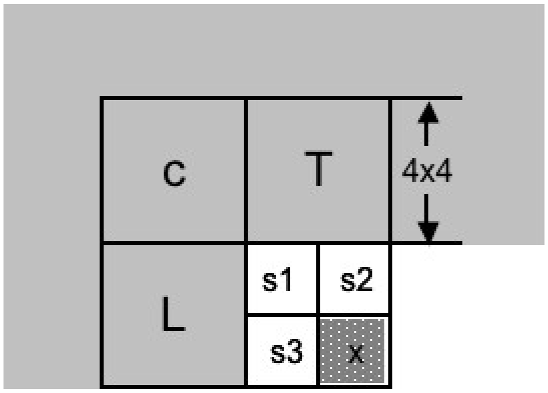


Figure 3.5: Valid search candidate blocks for target 2x2 block x.

One design concern is that, to avoid unaccounted quantization errors, all predictions much be kept in-loop. That is, for a 2x2 block to perform the BIP, all the search candidates much be based on reconstructed and decoded copies rather than original blocks. Since the minimal transform size used here is 4x4 DCT, sub-blocks $s1$, $s2$, and $s3$ cannot be reconstructed until target block x finishes its prediction as shown in Fig. 3.5. Hence, subblocks $s1$, $s2$, and $s3$ cannot be used as the prediction candidate blocks for sub-block x. The same principle applies to subblocks $s1$, $s2$, and $s3$ as well. In other words, only the shaded region in Fig. 3.5 can be used for prediction for sub-block x.

3.3 Joint Block/Line-based Intra Prediction with Rate and Distortion Optimization

The BIP with an enhancement mode has shown to have the ability to effectively reduce prediction errors for textured blocks using 2D geometrical manipulations including translation, rotation, scaling and reflection. However, common images or video sequences are usually a mixture of homogeneous regions and textured regions. As it pointed out earlier that when coding homogeneous regions, the LIP has the advantage of low mode overhead and simpler computation. Furthermore, if either LIP or BIP cannot produce a good prediction error, the LIP still has the advantage of lower overheads. Therefore, for these blocks, it is sometimes desirable to use the LIP modes in place of the BIP modes for an improved performance gain.

The prediction error in terms of the rate obtained from the LIP and the BIP are plotted in Fig. 3.6. The center dotted line of $y = x$ indicates the prediction error in terms of rate produced by both the LIP and the BIP are the same. It is obvious that although the BIP provides much lower prediction errors in most blocks, there still exist some blocks that have a lower prediction error from the LIP. Hence, in this section, we propose a joint block/line-based intra prediction scheme (JBLIP). In JBLIP, the LIP and

BIP modes are jointly selected via a rate-distortion optimization (RDO)-based scheme. The mode that delivers the best RD performance will be used as the final coding mode.

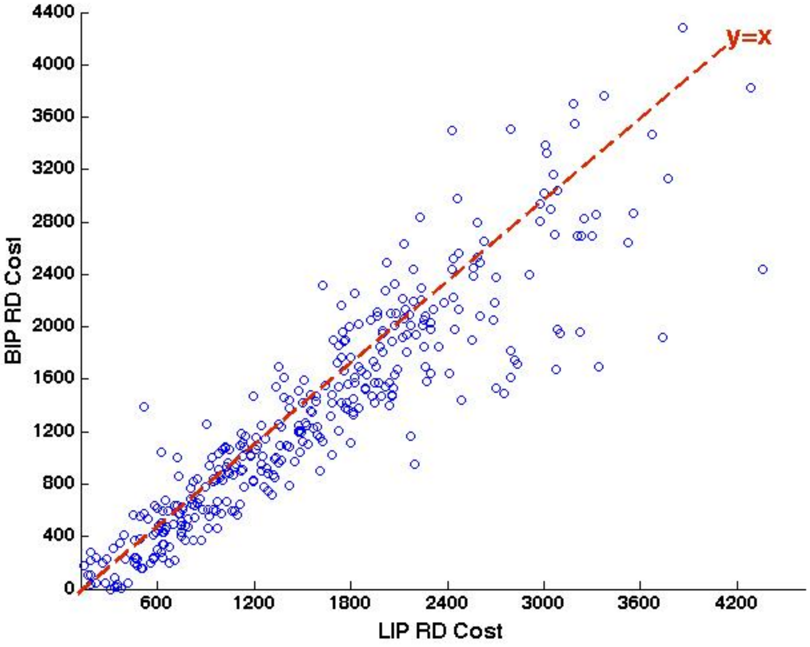


Figure 3.6: The rate-distortion cost comparison between LIP and BIP.

We describe the process of selecting the best intra prediction mode among all these modes. Being similar to the mode decision process in H.264/AVC, we adopt the Lagrangian Rate-Distortion optimization (RDO) procedure for the optimal mode decision. The RDO process evaluates all available modes simultaneously and chooses the one that minimizes the Lagrangian cost function. Mathematically, this can be written as

$$\min\{J(blk_i|QP,m)\},$$

$$J(blk_i|QP,m) = D(blk_i|QP,m) + \lambda_m \cdot R(blk_i|QP,m), \tag{3.10}$$

where $D(blk_i|QP,m)$ and $R(blk_i|QP,m)$ are the distortion and the bit rate of target block blk_i for a given coding mode of m and quantization parameter (QP), respectively. The minimization of the RD cost function in Eq. (3.10) implies that the RDO process decided the best mode which minimizes the distortion $D(blk_i|QP,m)$ and also satisfies the rate constraint $R_{st,i}$, i.e. $R(blk_i|QP,m) \leq R_{st,i}$. the Lagrangian multiplier λ_m in (3.10) is used to balance the distortion and bit rate. It is usually set as $\lambda_m = c \cdot 2^{(QP-12)/3}$ as in H.264/AVC. Therefore by selecting a larger Lagrangian multiplier λ_m, the Lagrangian cost function will be more biased towards the mode that produces a smaller rate. The Lagrangian multiplier λ_m can be obtained by the formula as suggested by [44].

The rate model for $R(blk_i|QP,m)$ is primarily consisted of the rate produced by headers R_{hdr} and the rate produced by the prediction error/residual R_{res}, $R(blk_i|QP,m) = R_{hdr} + R_{res}$. The rate model for header bits R_{hdr} will be different from LIP modes and BIP modes. As the translational vectors (TV) used in the BIP is similar that of the motion vectors (MV) in inter frames we could adopt the rate model in [9]. Thus R_{hdr} can be modeled as a function of number of TVs N_{tv} and number of non-zero TV elements N_{nzTVe} which is written as

$$R_{hdr,BIP} = \gamma \cdot (N_{nzTVe} + N_{tv}), \tag{3.11}$$

where γ is a model parameter andω is a weighting factor depending on the number of reference frames used in an inter frame. As the BIP does not use multiple reference frames, ω will be set as 0.3. For all LIP modes, the header bits can be modeled as

$$R_{hdr,LIP} = N_{LIP} \cdot b_{LIP}, \tag{3.12}$$

where N_{LIP} is the number of LIP modes and b_{LIP} is the average number of mode bits used for LIP. The rate model for prediction residual bit can be modeled as a function of quantization parameter (QP) as

$$R_{res} = \alpha \cdot \frac{SATD}{QP^{\tau_r}}, \tag{3.13}$$

where α is a model parameter and τ_r is set as 0.8. A Hadamard Transform (HT) is used to find the sum of absolute transformed difference (SATD).

Distortion $D(blk_i|QP, m)$ can be modeled as

$$D = \beta \cdot SATD \cdot QP^{\tau_d}, \tag{3.14}$$

where β is a model parameter and τ_d is 1.2 for LIP modes and 1.0 for BIP modes. Note that a 4x4 Hadamard transform is used in the RDO process for LIP and BIP modes except the BIP enhancement modes. Thus, to facilitate the 2x2 block matching process, the SATD used in both rate and distortion models used in Eqs. (3.11)-(3.14) is replaced by the sum of absolute difference (SAD) for simpler computation.

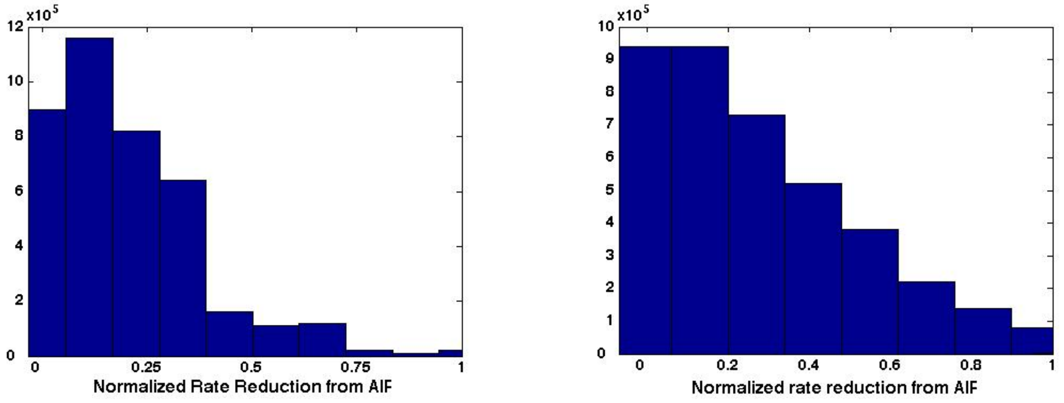
Typically, the BIP and the LIP are applied simultaneously, and we can choose the best mode among all possible BIP and LIP modes to be the ultimate intra prediction mode. It was shown in [10] that JBLIP scheme achieves a significantly improved performance gain than the application of the LIP alone. However, due to the exhaustive search method used to obtain the best prediction block, the computation complexity of JBLIP is very high. Hence, it is important to reduce the complexity of the BIP mode search while maintaining its good RD gain. In the next subsection, we will introduce a block classification method that could pre-filter those blocks that could be well predicted by the LIP alone.

3.4 Fast JBLIP Mode Selection Scheme

Being inspired by fractal coding, a block-based advanced intra prediction scheme that utilizes both the simplicity of LIP and the good prediction quality from prediction with 2D geometrical manipulations that includes block translation, scaling, rotation and reflection was introduced in Sec. 3.3. The purpose of geometrical manipulations is to increase the number of match candidates so as to provide a better rate-distortion (R-D) tradeoff. However, the overall computational complexity of JBLIP is extremely high when the RDO scheme evaluate all the LIP and BIP modes simultaneously to select the best mode. Therefore, it is desirable to develop a fast model selection scheme to speed up the prediction process while maintaining the good R-D performance.

3.4.1 Block Classification

The BIP was designed to provide better prediction to blocks with medium to high surface textures where the LIP fails. Hence, if the target block does not contain much surface

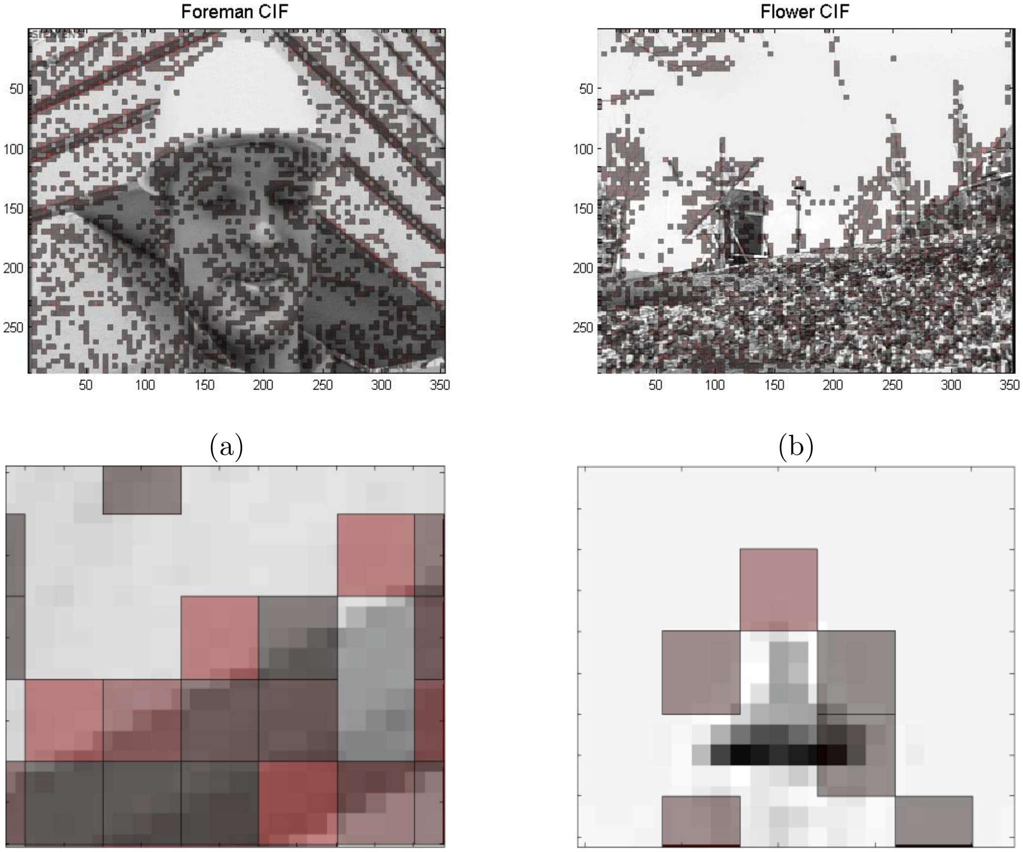


(a) (b)

Figure 3.7: JBLIP rate reduction histogram for a total number of 10 frames from (a) Mobile and (b) Foreman sequences.

variation, the JBLIP scheme would select LIP mode as its best prediction mode due to the RDO process employed in Sec. 3.3. Fig. 3.7 provides two histograms of the rate reduction using the BIP for frames with different complexity. It shows that the performance gain only comes from 50 to 60 percent of blocks that benefit from the BIP significantly. For those blocks clustered around the origin, there is no performance gain, and we can use the simple LIP method. Consequently, by properly categorizing the target block, the overall computational complexity can be reduced dramatically. Therefore, for homogenous blocks, we classify them as *"L"* blocks and others *"D"* blocks.

Intuitively speaking, block classification can be done based on its surface property. For example, if the block is relatively homogeneous, it could be well predicted by the LIP only. This idea can be illustrated in Fig. 3.8, where highlighted blocks indicate those benefit from the BIP substantially. All blocks in flat regions are unhighlighted, as the BIP has no clear advantage there. On the other hand, the effectiveness of the BIP is not simply proportional to the complexity of surface texture. Actually, the block that benefit



(a) (b) (c) (d)

Figure 3.8: Highlighted blocks in frames (a)Foreman and (b)Flower garden are those that have joint RD gain from BIP; (c) and (d) are the partial enlarged areas for Foreman and Flower garden respectively.

the most from the BIP are those that are relatively homogeneous but with slight texture variations as shown in Fig. 3.8(c) and (d).

We employ a parameter to capture the surface complexity of a block. If it is too small, the RDO process will evaluate LIP modes only and would not consider BIP modes.

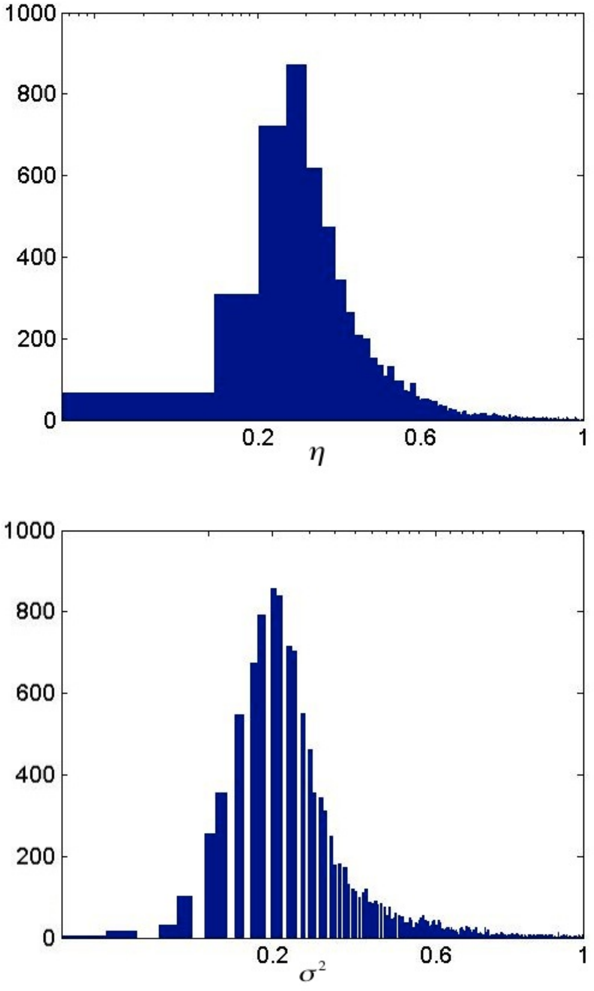


Figure 3.9: Histograms for normalized η and variance σ^2 with x-axis in Logarithmic scale.

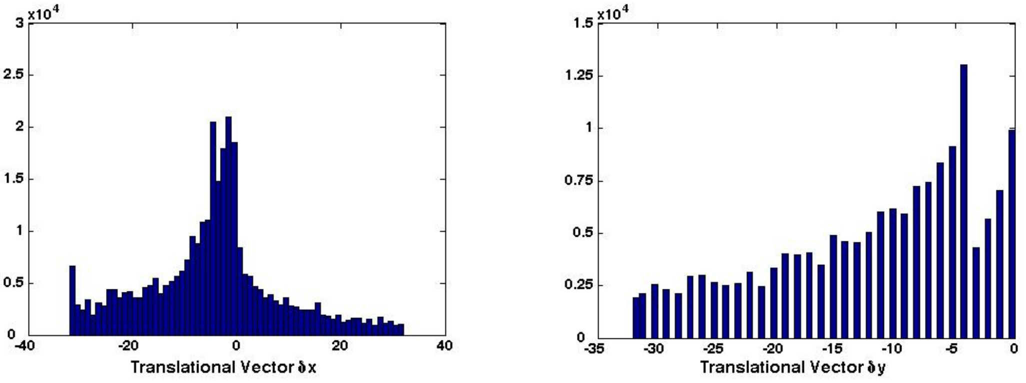
Because this is a pre-processing step for every block, the simpler, the better. Hence, we compute the following feature

$$\eta \quad = \quad \max(X) - \bar{X}, \tag{3.15}$$

where $\bar{X}$ is the average intensity level of the target block. We normalize the variance and η of the target block within [0, 1] and plot them in Fig. 3.9 (a) and (b), respectively, for visual comparison. We see that they follow a similar trend while Eq. (3.15) has a lower complexity as compared to the variance computation. If η is greater than a predefined threshold, we classify these blocks as *"L"* blocks and they will be predicted with LIP modes only. For all other *"D"* blocks, the RDO in AIP will evaluate all LIP and BIP modes.

3.4.2 JBLIP Search Window

The block classification method introduced in Sec. 3.4.1 can effectively identify those *"L"* blocks that can be well predicted by LIP alone. It reduces the overall computation complexity as the RDO process will no longer evaluate BIP modes for those blocks that are classified as *"D"* blocks. However, for those *"D"* block, the complexity is still very high as shown in Fig. 3.2, when using the BIP modes, the area allowed for coded region R_c becomes larger as block B moves towards to the lower-right corner. It is desirable to decide a proper size of the search range within coded region R_c to avoid the growing number of computations. To meet this goal, some statistically data were collected from different test images and plotted in Fig. 3.10, where we show the histogram of the horizontal and vertical displacements of the optimal translational vectors.



(a) (b)

Figure 3.10: Translational vector histogram for a combination of HD frames.

As shown in the figure, the optimal horizontal translational vectors are bounded about by +40 and -40 pixels while the vertical translational vectors are bounded by 0 and -35 pixels. Therefore, we could set the translational vector search window to be a stride of w pixels upwards, left-wards and rightwards of the current block as shown in Fig. 3.11. This limited search window can reduce the complexity of the spatial domain search for each block greatly while providing good performance.

Although the specified search window helps to limit the search complexity, the full search (FS) method used in this window still constitutes a great percentage of the computation time and a different treatment is proposed in the next subsections to further reduce the overall complexity.

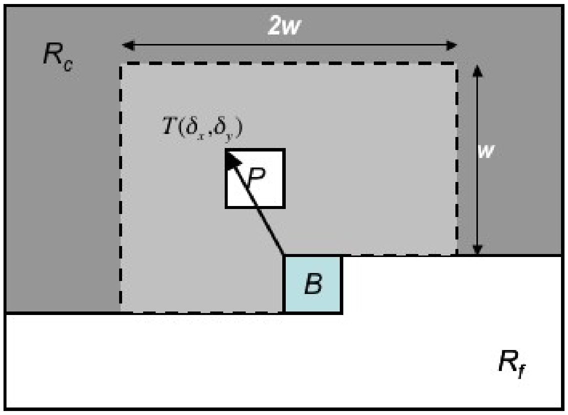
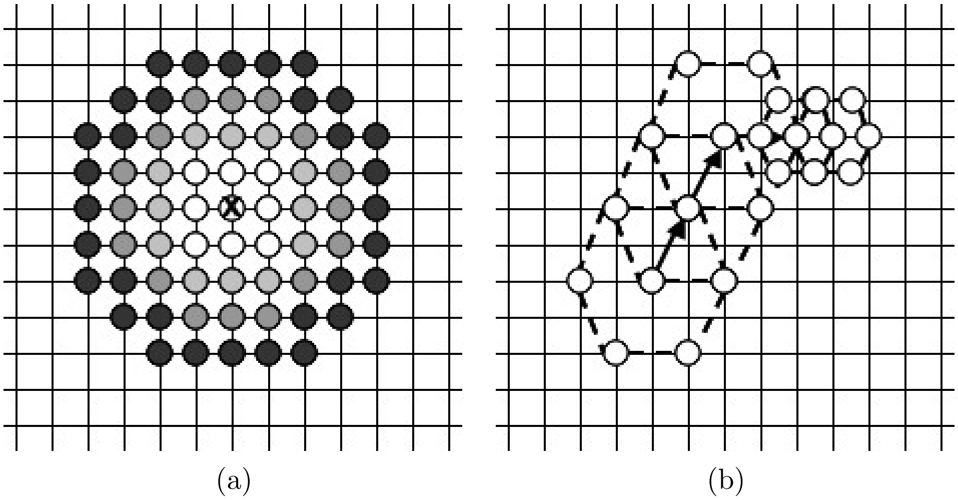


Figure 3.11: Search window setup for the BIP.

3.4.3 Self-Correlation Analysis of Local Image Region

Fast search of similar blocks has been extensively studied in the context of inter-frame prediction, known as the fast motion vector (MV) search. It consists of two major steps [21]: 1) find a good initial MV position and 2) local MV refinement. For Step 1, the temporal and spatial correlation of the MV field can be exploited. For Step 2, there exist several well known techniques in the literature, including the diamond search (DS) and the hexagonal search (HS) [8], etc.

Both DS and HS adopt two search patterns, *i.e.,* a large one and a small one as shown in Fig. 3.12. The large pattern is used first for faster convergence. Once the optimum is found at the center, it switches to the small pattern to allow finer refinement. As compared with DS, HS allows a better approximation of the circular shape yet with a small number of search points to speed up the search. However, DS and HS work under the assumption that the distortion function near the global minimum is an unimodal function. The closer to the global minimum, the smaller the distortion.


(a)

(b)

Figure 3.12: Search patterns for circular zonal search (CZS) with each zone specified by the same intensity. The position marked by "x" is target block position.

In contrast, CZS is more relaxed on this assumption. It only demands a center-biased property [36]. The zone is constructed so that each point within the same zone has the same distance from center as shown in Fig. 3.12. Each zone can be described by

$$round\left(\sqrt{MV_x^2 + MV_y^2}\right) = d - 1, \tag{3.16}$$

where d is the zone index and (MV_x, MV_y) is the displacement vector of a search block from the center (*i.e.* the position of the target block). When evaluating the best search position, CZS considers both distortion and rate. As compared to FS, CZS offers a good speed-up factor yet with very little R-D degradation.

The local MV refinement schemes all work under one assumption; namely, the inter-prediction error surface is unimodal within a small neighborhood of the global minimum, where the prediction error decreases monotonically as the search moves closer to the global minimum [8]. For the BIP, temporal correlation cannot be exploited, and we need to study the spatial correlation between neighboring blocks. To achieve this, we perform the self-correlation analysis in the spatial domain, which will reveal the error surface structure around the global miminum. A stronger correlation implies a smaller error.

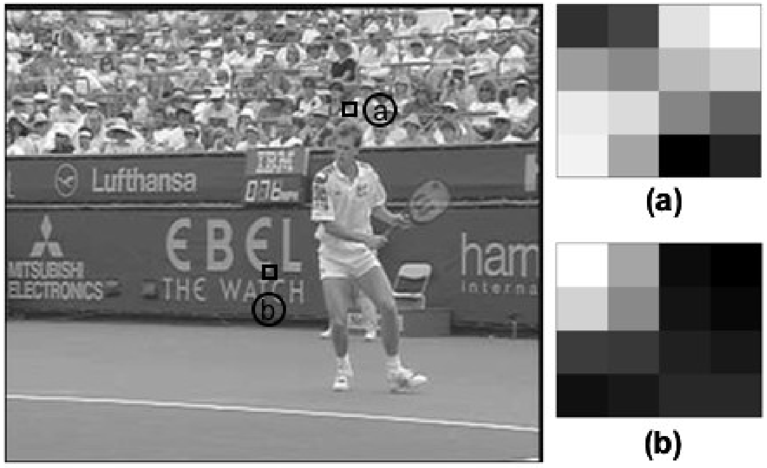


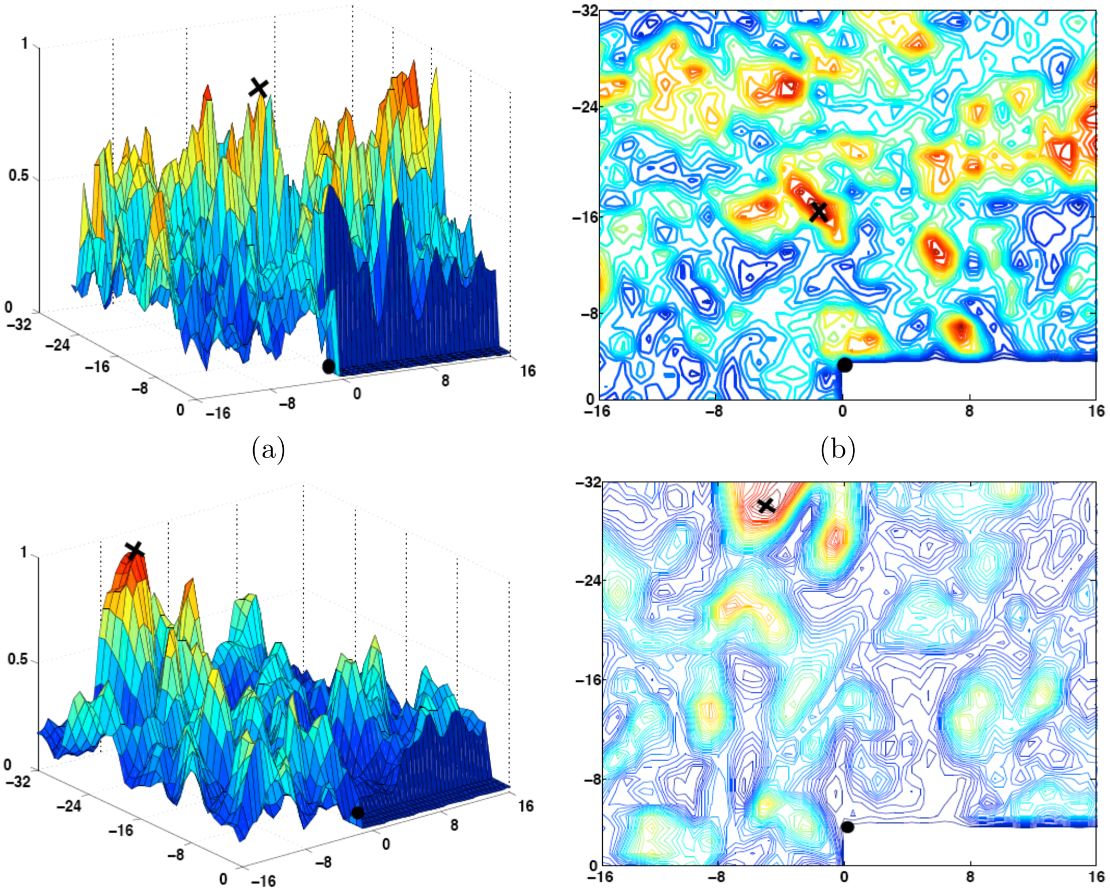
Figure 3.13: A frame from Stefan CIF sequence with two sample blocks using the BIP.

Since the BIP is more efficient in predicting textured blocks, the spatial correlation is conducted for textured block only. Two blocks from a frame of the Stefan CIF sequence are chosen as test blocks as shown in Fig. 3.13. Block (a) is a texture block taken from the background audience. It has many similar texture blocks in its neighborhood. Block (b) is another block taken from the advertisement board whose neighboring blocks have different surface properties.

The self-correlation of the target block (marked with black dot) with its spatial neighbor region is shown in Fig. 3.14 with the 3D and the 2D contour plots. The position with the highest correlation is marked with "x". The correlation surfaces are depicted in different colors in 2D contour plots. The warmer the color, the higher the correlation is. Since a causality constraint is required by the BIP search, we only show the correlation surfaces in the coded region. As shown in Fig. 3.14, there are too many local maxima in these correlation surfaces. Consequently, we are not able to apply the DS or the HS directly for local prediction refinement since they may miss local (or global) minima easily. The circular zonal search (CZS) [3] appears to be a more suitable choice. However, for the case of block (b), the CZS will exhaust almost the entire search window to reach the global maxima, if the search starts from the origin.

We can draw a simple conclusion from the above discussion. That is, the error surface of BIP is more irregular than that of the MV search in inter prediction. The correlation surface may have many local minima in the neighborhood of the best match. If the DS/HS is applied, the search could be trapped to a local minimum easily and the overall RD performance can be degraded significantly as compared with that of the FS. On the other hand, if the CZS is applied, its complexity could be close to that of the FS. Thus, the speed-up of the BIP search is a more challenging problem than the fast MV search problem in inter prediction.

In the next section, we propose a fast BIP search scheme that can maintain good RD performance as provided by the CZS yet with much reduced computational complexity. First, a translational vector prediction scheme is introduced in Sec. 3.4.4 to establish a close approximation to the global minimum. A zone-based fast search algorithm is proposed in Sec. 3.4.5. In addition, a candidate block pruning method is presented to limit the search complexity by eliminating those reference blocks that do not provide a good match in advance in Sec. 3.4.6.



(a)

(b)

(c)

(d)

Figure 3.14: Self-correlation surfaces for blocks within the same frame: (a) the 3D plot and (b) the 2D contour plot of a correlation surface obtained for block (a) in Fig. 3.13; and (c) the 3D plot and (b) the 2D contour plot of a self-correlation surface for block (b) in Fig. 3.13.

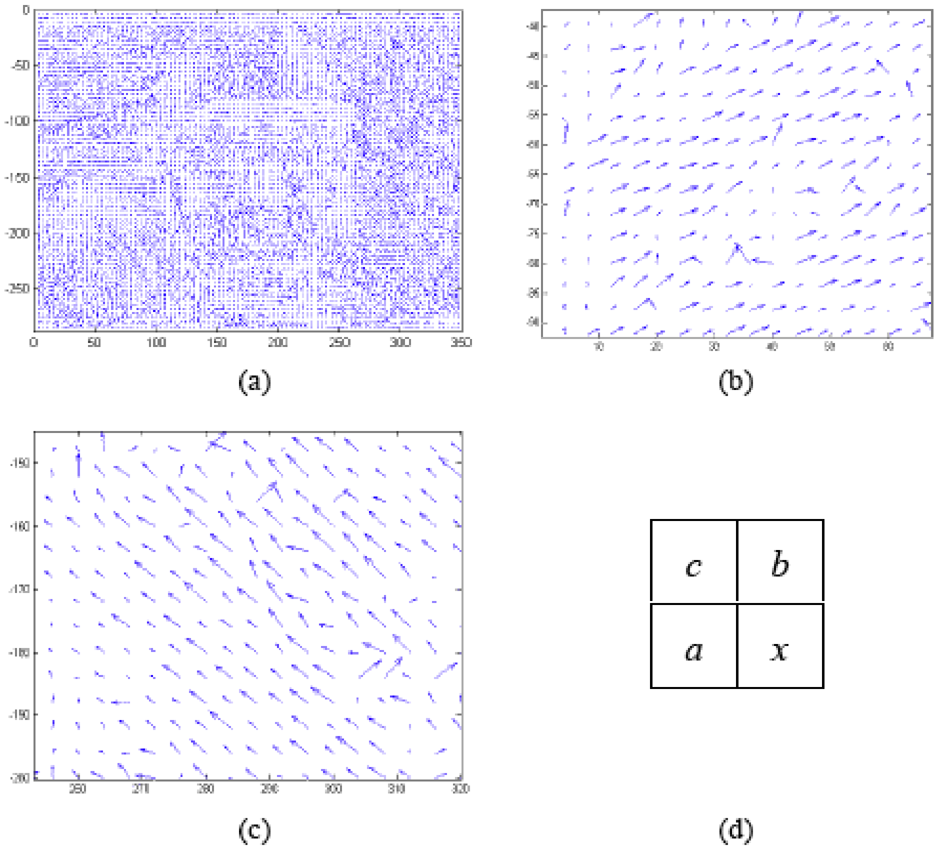


Figure 3.15: The BIP displace map for (a) a full frame in the CIF Foreman sequence, (b) and (c) contain partially cropped close-up maps from (a), and (d) displacement prediction.

3.4.4 Search Initialization via Prediction

One important component of a fast search algorithm is to have a good initial search position which is close to the optimal one. This not only speeds up the search process but also reduces the possibility of being trapped to a local minimum. Classic MV prediction is conducted based on the MV correlation in the spatial or the temporal domain. For the BIP, we observe that the neighboring translational vectors exhibit some degree of correlation spatially. Fig. 3.15 gives an example of a translational vector map for a typical frame. Two close-up views are provided in Figs. 3.15 (b) and (c).

As shown in these figures, many translational vectors exhibit some degree of similarity. Thus, by borrowing the concept of MV prediction in the spatial domain, we can predict an initial translational vector via

$$\hat{x} \quad = \quad \text{median}\{a, b, c\}, \tag{3.17}$$

where $\hat{x}$ is the estimated translational vector of the current block while a, b and c are the translational vectors of its neighboring blocks as shown in Fig. 3.15 (d).

Median prediction is simple and effective at predicting translational vectors that are highly correlated, thus it is very powerful in predicting MVs that has a very strong correlation in temporal domain. However, a closer examination of the translational vector map in Fig. 3.15 shows that the translational vectors form a map that roughly represents the actual frame content, which is the head of foreman in this frame. Therefore, the correlation between translational vectors are not as high as that of the MV in inter frames. In the translational vector map, the edges of the contour are less correlated compared to the flat regions of the frame. Hence, it is less effective to use median predictor for translational vector prediction. On the other hand, it inspired us to treat this translational vector map as a content frame and predict accordingly. A median edge detector (MED) as given in Eq. (3.18) is proven to provide a good edge prediction and thus can be adopted in our translational vector prediction without incurring additional mode overhead.

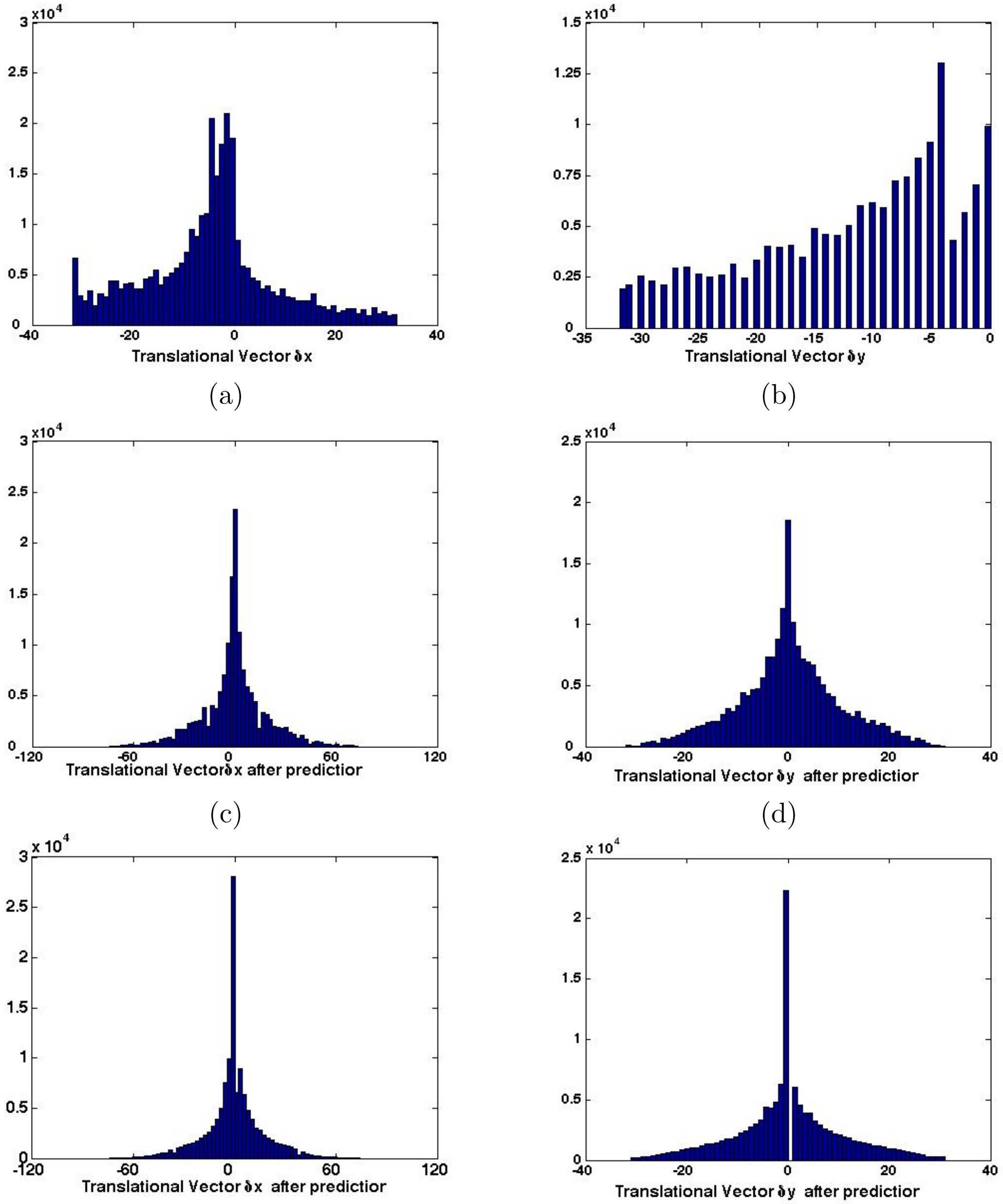
$$\hat{x} = \begin{cases} \min(a,b) & \text{if } c \geq \max(a,b), \\ \max(a,b) & \text{if } c \leq \min(a,b), \\ a+b-c & \text{otherwise.} \end{cases} \tag{3.18}$$

Fig. 3.16 shows the prediction results using both median predictor and median edge detector. It can be seen that median edge detector provides a better prediction results than median predictor.

3.4.5 Zone-based Fast Search

The fast search algorithm starts from the predicted displacement position as given in Eq. (3.18) and a zone called *Zone 1* with search window of w_1 is constructed around this initial search position, where the CZS is performed. *Zone 1* is set to be relatively small since translational vectors are spatially correlated and the prediction is expected to be close to the actual global minimum. If the search in *Zone 1* yields a sufficiently good result (*e.g.,* less than a pre-set threshold γ), the search will terminate immediately.

However, if the search in *Zone 1* does not yield a satisfying result, it is most likely that the global minimum is not close to the position pointed by the predicted translational vector. Hence, we need to re-initialize the start position and prepare the second round of search. In a local texture region, the global minimum is often close to the target block itself. Thus, the block immediately to the left of the target block is selected as the center of *Zone 2* with search window of w_2 at this new location and the CZS is performed.



(a)

(b)

(c)

(d)

(e)

(f)

Figure 3.16: Translational vector histograms for x- and y-axis respectively: (a)and(b) without prediction; (c) and (d) with MAP predictor; (e) and (f) with MED predictor.

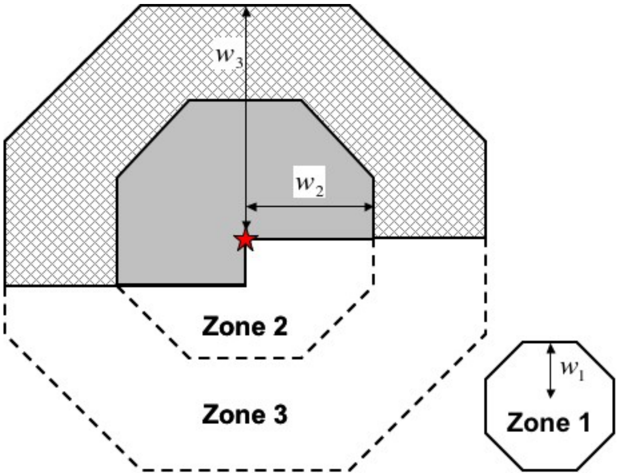


Figure 3.17: The proposed adaptive search zones and patterns, where non-searchable areas in Zones 2 and 3 are marked out for ease of visualization.

Again, w_2 is set to a small value since the likelihood to have a good displacement around this position is high.

However, if results from *Zone 1* or *Zone 2* are poor, the global maxima could be outside of these zones and we need to enlarge the search window. That is, we adopt the same center of *Zone 2* but extend the search window to w_3. The above process is illustrated in Fig. 3.17.

3.4.6 Candidate Block Pruning

When the CZS is applied, the search complexity will be close to that of the FS if the optimal point is located in the boundary of a zone as the example shown in Fig. 3.14(d). Thus, we need some mechanism to trim unlikely positions to save the computational cost. A block pruning method based on the difference of the variances between target block t and candidate block c is often adopted in the context of fractal image coding:

$$|\sigma_t^2 - \sigma_c^2| \quad > \quad T, \tag{3.19}$$

where T can be adaptively selected as $T = \theta \cdot \sigma_t^2$ and where θ is another parameter. However, the variance computation is still costly, and we would like to look for a simpler way for block pruning.

In the CZS scenario, candidate blocks are often offset by one pixel either horizontally or vertically. Hence, we for a block of size $n \times n$, could compute average line-based variances μ_v and μ_h horizontally or vertically as:

$$\mu_v \quad = \quad \frac{1}{n}\sum_{i=1}^{n} \sigma_v^2(x = i), \tag{3.20}$$

$$\mu_h \quad = \quad \frac{1}{n}\sum_{i=1}^{n} \sigma_h^2(y = i), \tag{3.21}$$

where $\sigma_v^2(x = i)$ and $\sigma_h^2(y = i)$ are horizontal or vertical line-variances for line segments with index i. Thus, instead of re-computing the variance of the entire block at each now position, we only need to compute the new line variance as the search windows moves by one pixel. Then, we can perform the average of line-based variances μ_v and μ_h and obtain

$$\mu_{v,h} \quad = \quad 0.5(\mu_v + \mu_h). \tag{3.22}$$

for each block. Then, we can use $\mu_{v,h}$ in Eq. (3.22) to replace σ^2 in Eq. (3.19).

To further reduce complexity, we can further replace variances in Eqs. (3.20) and (3.21) by sums of absolute mean differences as:

$$\begin{aligned} \mu'_v &= \frac{1}{n}\sum_{i=1}^{n} diff_v(x=i), & (3.23)\\ \mu'_h &= \frac{1}{n}\sum_{i=1}^{n} diff_h(y=i), & (3.24) \end{aligned}$$

where $diff_v(x = i)$ and $diff_h(y = i)$ are sums of absolute mean differences for line segments with index i horizontally or vertically. Eqs. (3.23) and (3.24) are implemented in the experimental section for the block pruning test:

$$|\mu'_{(v,h)_t} - \mu'_{(v,h)_c}| \quad > \quad T', \tag{3.25}$$

where T' is an adaptively selected threshold. If the test is met, the candidate block is pruned so that we can move to the next candidate block. This simplification allows the CZS to move faster towards the true global minimum. The search performed in *Zone 3* may also terminate early if a search result meets the predefined threshold.

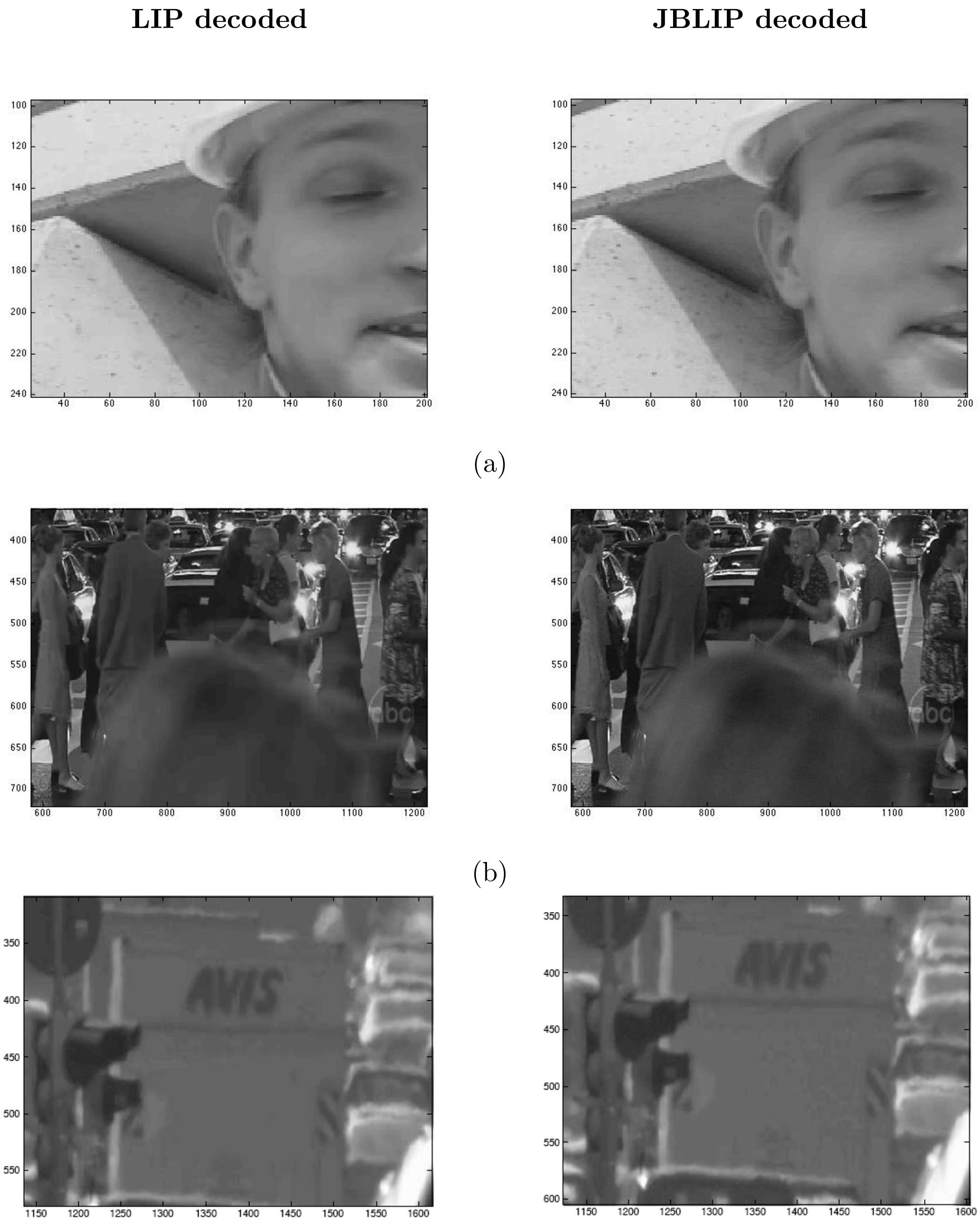
In the full BIP search, 2D geometrical manipulations such as rotation and reflection will be performed for every translation vector. This can be implemented by applying the rotation and reflection to the target block to generate 8 blocks. Then, we will perform the matching between these 8 target blocks and a candidate block. Therefore, as the target block is being encoded one at a time, it can first go through the 2D manipulation and each manipulation can be saved without incurring much overhead. As a result, the mode selection process will happen between the unaltered reference blocks and the manipulated target blocks. A global flag must be set for the entire sequence to indicate that the mode information must be interpreted on the decoder side as the manipulation on the target block rather than on the reference block. 2D scaling is disabled in this paper since the target block size will change under the fast search approach.

3.5 Experimental Results

Experiments were conducted with the H.264/AVC reference codec JM12.1 [1] on a PC with Intel core 2 duo@1.8GHz. The search window of the BIP full search was $w = 16$. A total of twelve YUV 4:2:0 sequences with three different resolutions were tested. For CIF resolution, we used Flowergarden, Foreman, Mobile and Coastguard. For HD sequences at 1280 x 720 resolution, we tested Night, City corridor, Harbor and Sheriff. For HD sequences at 1920x1080 resolution, we tested Vintage car, Pedestrian, Rush hour and Sunflower.

We first compare the average coding efficiency between 1) H.264/AVC using LIP only, and 2) proposed joint block/line-based intra prediction with full BIP search. The average PSNR gain and bit rate reduction results are obtained based on [16] and collectively shown in Table. 3.1. The decoded images from each scheme are shown in Fig. 3.18.

The second set of experiments are conducted between the following three schemes: 1) H.264/AVC using LIP only, and 2) proposed JBLIP with joint LIP/BIP with full



LIP decoded

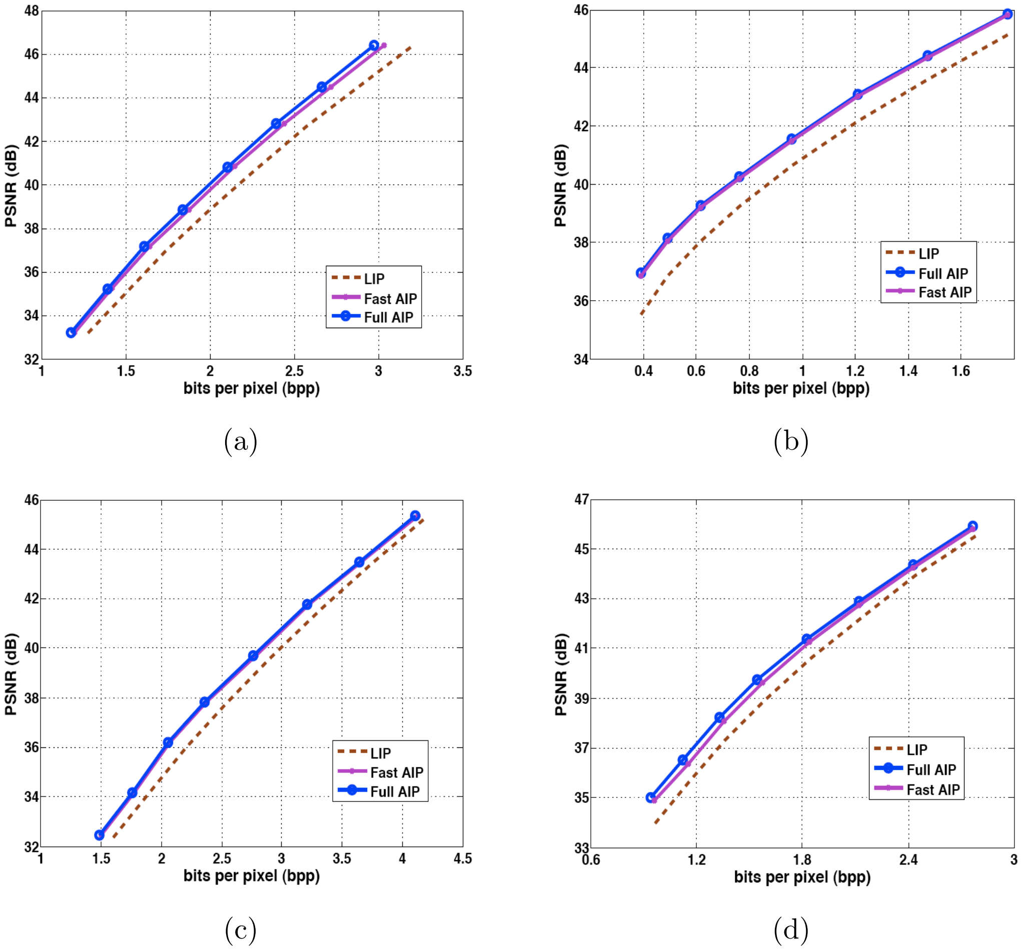
JBLIP decoded

(a)

(b)

(c)

Figure 3.18: Decoded partially enlarged image for (a) Foreman (b) Night, and (c) Rush hour.



(a) (b)

(c) (d)

Figure 3.19: Rate-Distortion curves for (a) Flowergarden, (b) Foreman, (c) Mobile and (d) Stefan.

search BIP; and 3) joint LIP/BIP with the proposed fast BIP search. for each resolution in Figs.3.19, 3.20, and 3.21,: The proposed fast search algorithm was set up with the following parameters: $w_1 = w_2 = 8$, $w_3 = 16$, and $\gamma = 32$. We define a speed-up factor as

$$f_s \quad = \quad 1 - \frac{C_{fast}}{C_{full}}, \tag{3.26}$$

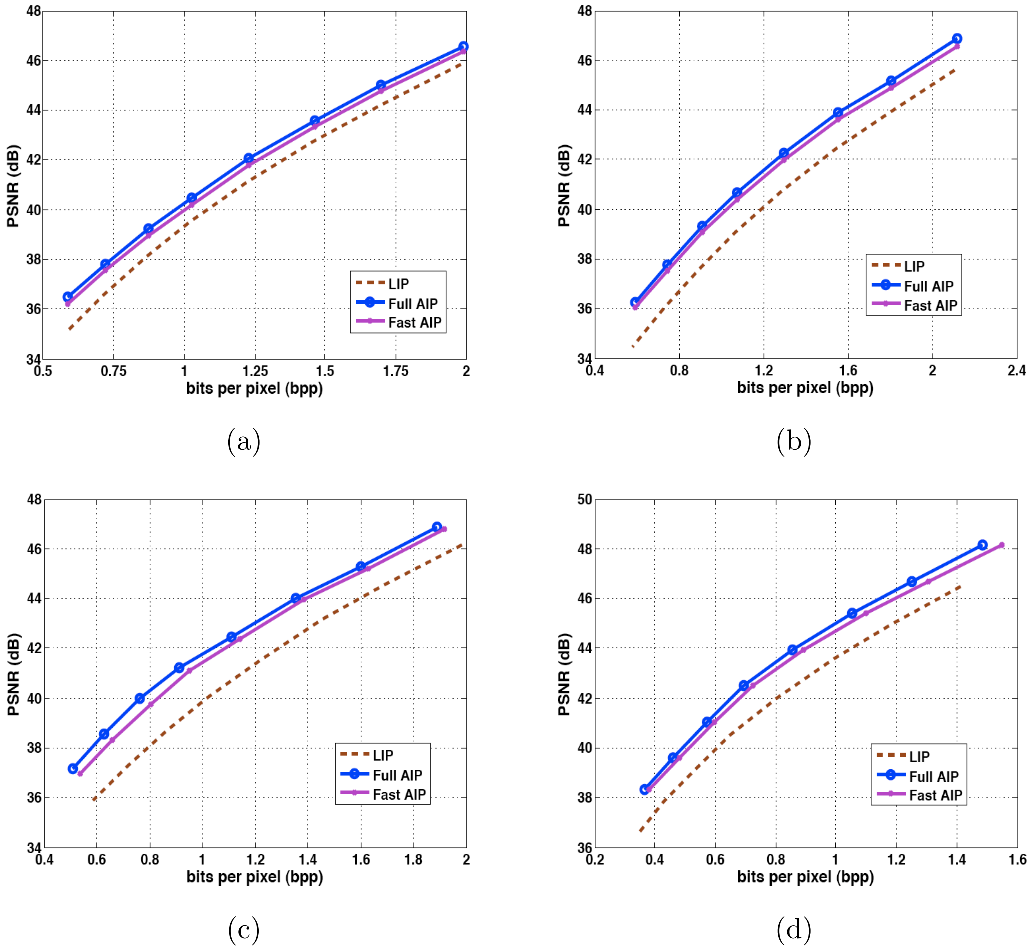
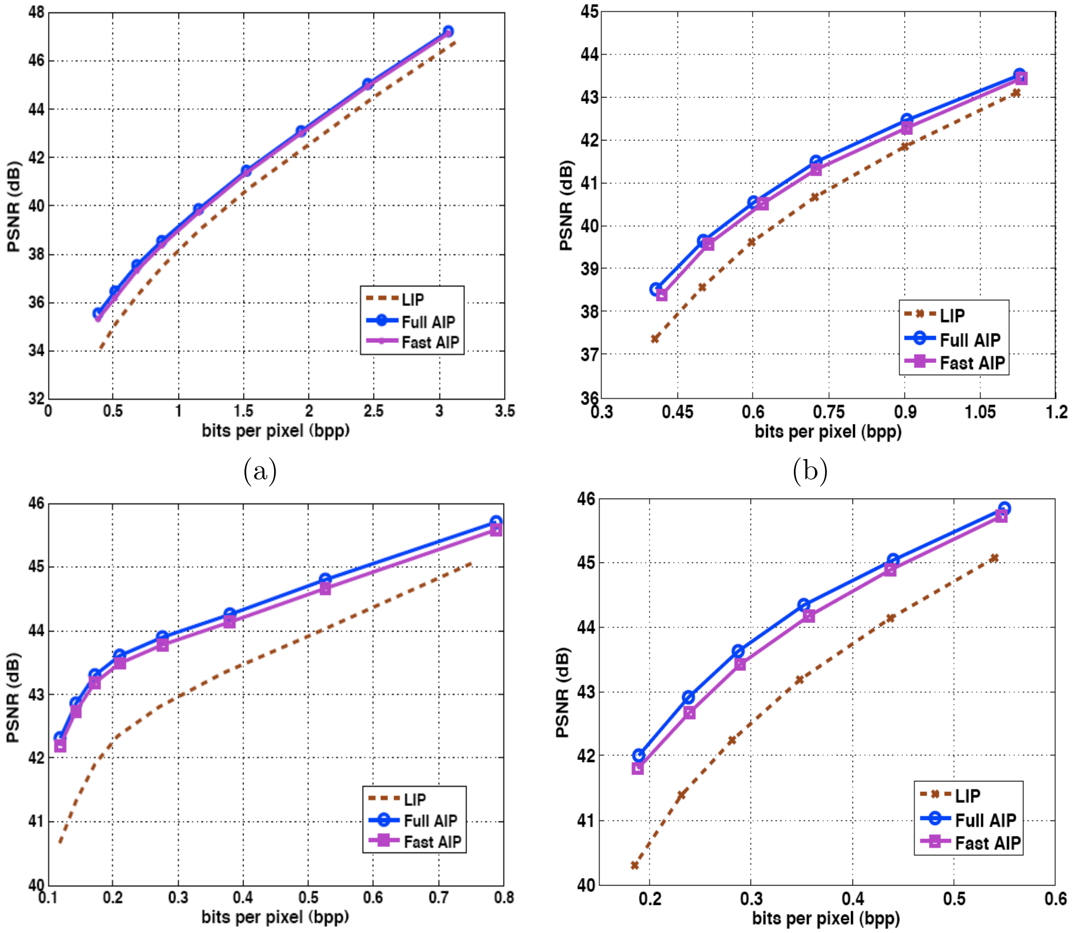


Figure 3.20: Rate-Distortion curves for (a) Harbor, (b) City Corridor, (c) Night and (d) Sheriff.



(a) (b)

(c) (d)

Figure 3.21: Rate-Distortion curves for (a) Vintage car, (b) Pedestrian, (c) Rush Hour and (d) Sunflower.

Table 3.1: Coding Efficiency Comparison between JBLIP and LIP.

CIF Sequences

Sequence	$\Delta Bitrate(\%)$	$\Delta PSNR(dB)$
Flowergarden	-6.75	0.46
Foreman	-10.34	0.69
Mobile	-3.53	0.34
Coastguard	-10.73	0.76

HD Sequences @ 1080x720

Sequence	$\Delta Bitrate(\%)$	$\Delta PSNR(dB)$
Night	-11.02	0.76
City corridor	-11.01	0.69
Harbor	-12.67	0.78
Sheriff	-15.37	0.82

HD Sequences @ 1920x1080

Sequence	$\Delta Bitrate(\%)$	$\Delta PSNR(dB)$
Vintage car	-10.29	0.88
Pedestrian	-30.11	1.36
Rush Hour	-28.45	1.43
Sunflower	-36.04	1.68

where C_{full} is the computational time demaned by the full BIP search and C_{fast} is the computational time using the proposed fast BIP search. The speedup factors are shown in Table. 3.2.

We see that the performance of the proposed fast search method lies between the full BIP search and the LIP-based H.264/AVC. The proposed fast BIP search reduces the BIP FS complexity by an average of 67 to 72% at the cost of small RD degradation.

3.6 Conclusion

In this work we proposed a novel BIP scheme for efficient image (or intra-frame) coding, where we apply various 2D geometrical manipulations to reference image blocks to enrich the pool of prediction blocks for a given target block. This BIP is further enhanced and

Table 3.2: The speed-up of the proposed fast search with respect to the full search BIP.

CIF Sequences

QP	Flowergarden	Foreman	Mobile	Coastguard
20	0.673	0.702	0.556	0.743
22	0.722	0.722	0.560	0.746
24	0.738	0.715	0.568	0.751
26	0.731	0.737	0.586	0.758
28	0.745	0.750	0.571	0.753
30	0.750	0.758	0.612	0.766

HD Sequences @ 1080x720

QP	Night	City corridor	Harbor	Sheriff
20	0.745	0.683	0.766	0.803
22	0.771	0.695	0.768	0.833
24	0.779	0.722	0.810	0.850
26	0.792	0.738	0.784	0.858
28	0.790	0.759	0.801	0.850
30	0.811	0.770	0.819	0.887

HD Sequences @ 1920x1080

QP	Vintage car	Pedestrian	Rush Hour	Sunflower
20	0.775	0.770	0.776	0.754
22	0.782	0.748	0.754	0.756
24	0.789	0.812	0.769	0.783
26	0.775	0.837	0.828	0.788
28	0.782	0.835	0.851	0.805
30	0.786	0.856	0.826	0.829

a joint block/line-based intra prediction scheme (JBLIP) is proposed to jointly evaluate LIP and BIP based on a rate-distortion optimization method. As compared with the traditional line-based intra prediction in H.264/AVC, the new JBLIP scheme offers a significant coding gain (about 0.68-1.68dB in the PSNR value with the same bit rate). Furthermore, a fast JBLIP mode decision scheme is proposed to reduce encoder complexity. A block classification method is used to identify those blocks that can be well predicted with LIP only. In addition, self correlation analysis is provided to address the weakness with existing fast search algorithms due to the unique characteristics in BIP.

Translational vector prediction is used for near optimal search initialization and candidate block pruning is adopted to assist in faster convergence. Results show that the average speedup is around 10 times faster compared to the full search employed in JBLIP and the average PSNR degradation is 0.14dB.

Chapter 4

Hierarchical Intra Prediction for Lossless Image Coding

4.1 Introduction

Earlier image/video coding algorithms focused on low bit rate coding due to the limited storage space and transmission bandwidth. However, due to the increased popularity of high definition video and the availability of the broadband networking infrastructure, recent research interests have gradually shifted from low-bit-rate to high-bit-rate (or high fidelity) video coding. For applications in medical imaging, scientific geographic information, professional digital photography, and digital videography, lossless coding is often preferred due to the requirement to store and access original contents without any distortion.

More digital contents are delivered over the Internet nowadays, and there is another urgent need to be addressed. That is, these contents are delivered in a scalable fashion. This need arises from two fronts: *1) content preview and editing* for on-board playback of a camcorder and *2) content delivery* over a a heterogeneous network. For on-board preview, the limited LCD size (typically of 3-inch diagonally), decoder complexity and power restriction make the preview at the original resolution impractical. For content delivery over networks, there is a need to tailor its format to different networking and display requirements. Both of them demand an efficient scalable solution. H.264/SVC offers a powerful scalable video solution. Being similar to its single-layer counterpart in H.264/AVC, intra prediction in H.264/SVC adopts line-based spatial correlation within each enhancement layer. Besides, H.264/SVC has sophisticated intra prediction that exploits inter-layer correlation to perform inter-layer prediction for the enhancement layer. However, even with inter-layer prediction, the coding performance of intra-frame coding in H.264/SVC is still substantially worse than its single-layer counterpart H.264/AVC. This deficiency motivates us to develop a more efficient scalable intra coding algorithm.

Since lossless coding and scalable coding are presented as solutions targeting at different market requirements, their codec designs are generally proposed in parallel and developed independently by different standard committees. As a result, the existing scalable coding techniques put major emphasis on quality and spatial flexibility for low bit rate streaming environments, while all lossless compression schemes are designed with coding efficiency as its main design goal. Therefore, current scalable solutions not only have no capability to deliver contents in a lossless fashion, the coding efficiency will also suffer a substantial degradation in order to accommodate the added scalability. Experimental results in actual coding environments reported in [38, 31] have shown that even the current leading scalable coding solutions, such as H.264 scalable video coding (H.264/SVC) suffers a bit rate increase of 50 - 100% compared to the single-layered approach. This is mainly due to the repeated encoding of redundant information amongst different layers and the inefficiency to fully explore these correlation between layers. On the other hand, current state-of-the-art lossless compression solutions, such as H.264 intra lossless (H.264-LS) [27] and JPEG-LS [49], do not have any scalability built into the codec. More importantly, the compression performance delivered by these codec are quite limited. For images with medium complexity, the average compression ratio for lossless codec is around 50%. That means the compressed file size is only reduced to about half of that of its original copy.

In summary, it has been a major research challenge to support scalability while delivering acceptable compression performance. Consequently, none of these aforementioned coding solutions can provide such compression architecture with built-in scalability ranging from lossy to lossless quality requirements while at the same time delivers better compression performance compared to the single layered approach. Thus, in this paper, we propose a novel lossless hierarchical intra prediction scheme to try to address these new requirements and provide an unified solution that can give a cohesive and efficient treatment to these issues.

The rest of the chapter is organized as follows. we will first examine existing issues with current codec, in Sec. 4.2 we propose an new coding architecture with a pyramidal image decomposition scheme to establish a novel hierarchical coding structure that is different from the existing scalable codec by separating signals of different correlation characteristics. Based on this decomposition method, a hierarchical coding system with multiple predictors is proposed in Sec. 4.3. To further improve prediction accuracy, a joint inter-layer prediction with training set optimization is proposed by using the gradient measured on the BL blocks to adaptively select the best suited training sets without sending the displacement vectors to the decoder explicitly. Experimental results for the proposed lossless HIP scheme are given in Sec. 4.4 to demonstrate the effectiveness of the proposed scheme. Finally, concluding remarks are given in Sec. 4.5.

4.2 Lossless Hierarchical Intra Prediction

Predictive coding is one of the fundamental principles behind existing image and video coding techniques. It is a statistical estimation procedure where future random variables are predicted from past and present observable random variables [20]. For an input signal sample of x_i, the prediction is carried out based on a finite set of past samples $x_1, x_2, ..., x_{i-1}$ as $\hat{x_i} = f(X) = f(x_1, x_2, ..., x_i)$. The optimal predictor of $\hat{x_i}$ is based on minimizing the error between x_i and $\hat{x_i}$ as

$$\epsilon\{(x_i - \hat{x_i})^2\} \quad = \quad \epsilon\{[x_i - f(x_1, x_2, ..., x_i)]^2\}. \tag{4.1}$$

As there usually exists a very strong spatial and temporal correlation in an image or video frames, good predictive coding techniques become extremely powerful in minimizing the redundancy to achieve efficient coding. Next, we will examine the issues with existtng signal modeling and decorrelation methods.

4.2.1 Signal Modeling and Decorrelation

Signal modeling is one of the main building blocks in signal compression since it is highly related to entropy coding. In previous research, the Gaussian distribution is often used to describe the distribution of AC coefficients [33]. However, it was soon found that the Laplacian distribution is more suitable to describe the signal statistics when the

Kologorov-Smirnov goodness-of-fit test is used [35]. Many other distributions of DCT coefficients have been proposed in the past, including the generalized Gaussian, Cauchy, etc. [26, 35, 41, 54].

In the actual coding system, natural images generally consist of nonstationary signals, which cannot be well represented by a single distribution. A complex mixed Gaussian distribution was proposed in [14] to allow better approximation. Although there has been some debate on which model offers the best approximation, it is generally agreed that there exists some discrepancy between the actual input signal and its model due to the non-stationary property. Besides, researchers in the field of texture synthesis found that natural images can be modeled as two-dimensional random fields (RF), which consiste of both long- and short-range correlations. The long-range correlation generally refers to the textures with most low frequency components and little variation in pixel intensity while the short-range correlation usually refers to the areas with more complex textures and variations.

For texture synthesis (or sometimes image compression), some random field models are used to represent a wide variety of textures with a small number of parameters and estimate the image activity based on spatial interaction of pixels in a local neighborhood. For these natural texture with long-range dependency is generally known as the Hurst phenomenon [19]. Therefore, earlier work on long correlation (LC) models [22] was based on fractional differencing of low-order autoregressive processes. However, these LC models lack the capability to represent short correlation (SC) characteristics. Thus, another LC model [5] was introduced as fractionally differenced short correlation models. Unfortunately, the two LC models mentioned above are not efficient in modeling long correlation structures in typical natural images since the autocorrelation model decays too rapidly to model the long correlation within the image accurately. To address this problem, generalized long correlation (GLC) models were proposed to explore these correlations. It turns out that the difference between the three models are small. They are not well suited for images with structured elements and, therefore, cannot be directly used for efficient prediction.

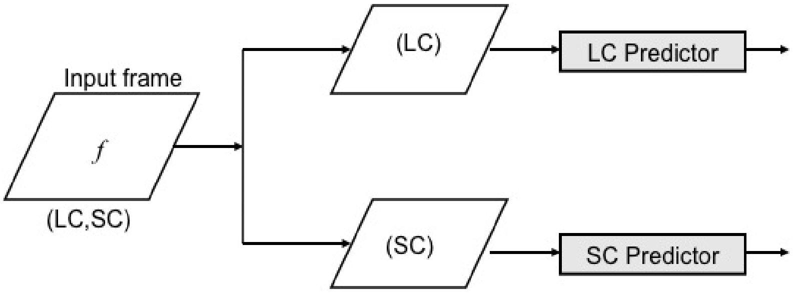


Figure 4.1: The prediction system based on the divide-and-conquer (D&C) strategy.

In summary, natural images are non-stationary and they contain both long and short correlations that cannot be well represented with a single model. However, to explore long- and short-range correlation within an image serves as a fundamental building block in our image decomposition scheme. In other words, we want to design a system that extracts the short-range correlation to one layer while simultaneously shorten the long-range correlation to a well defined span within the neighbor set. This would allow both long- and short-range correlations to be addressed independently. Therefore, in the proposed image decomposition scheme, we introduce an additional frequency decomposition phase before the prediction phase. This phase allows the system to identify signal components of different correlation so as to perform a different prediction scheme based on signal's characteristics.

To achieve this design goal, we adopt a divide-and-conquer (D&C) strategy by dividing a problem into two or more smaller subproblems. Each of them is solved independently, and their solutions are integrated to yield the final solution to the original complex problem. For practical image compression application, one intuitive D&C algorithm is to perform signal separation. By separating a signal into components of different characteristics, we divide a complex problem into multiple subproblems as shown in Fig. 4.1.

In the next section, we present a spatial-domain signal decomposition scheme that decomposes signals into subsignals according to their signal correlation characteristics.

4.2.2 Pyramidal Image Decomposition

A common image decomposition scheme can be achieved by performing a spatial-to-frequency domain transform on the input image. The Discrete Cosine Transform (DCT) is widely adopted in image and video coding standards such as JPEG, MPEG-1, MPEG-2, MPEG-4, H.261 and H.264/AVC codec. It allows good energy compaction with relatively simple computation as compared with the optimal Karhunen-Loeve Transform (KLT). However, the coding system is set up in such a way that DCT usually works in the second phase of decorrelating the prediction error after the prediction phase has removed majority of the signal redundancy in the spatial domain. This scheme has one problem. That is, if the prediction phase cannot provide good prediction for high frequency components of the image in high bit rate coding, high frequency components will exist after image decomposition, which will affect the overall coding efficiency. To achieve signal decorrelation in the spatial domain, we propose an pyramidal image decomposition

(PID) method to separate signals with LC and SC, where the down-sampling process is used as a low-pass filter. After the decorrelation process, low and medium frequency components in the form of LC signal components will reside in the downsampled upper layers, leaving a small amount of residuals of SC signal components in the form of high frequency components on the lower layer.

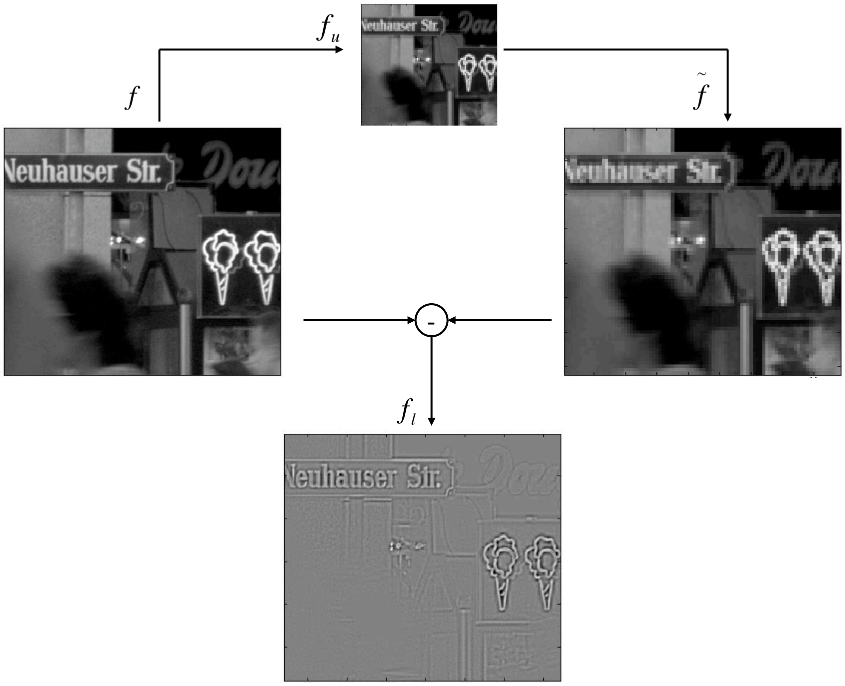


Figure 4.2: Pyramidal image decomposition using a downsampling lowpass filter.

A pyramidal image decomposition scheme is shown in Fig. 4.2. The input sample, f, is first downsampled into an image of lower resolution denoted by f_u. This low resolution image is up-sampled back into $\tilde{f}$ of the original resolution. The difference image between the upsampled f_u and the original input f is denoted by f_l. The image consisting of the low to medium frequency components is extracted to the upper layer of f_u while that of higher frequency components such as sharp edges and fine textures are filtered into the lower layer of f_l. Although the dynamic range of input pixel values does not change much in $\tilde{f}$, $\tilde{f}$ is however a blurred version of the original input, f, with most of its high frequencies removed. These high frequency components are well preserved in lower layer f_l.

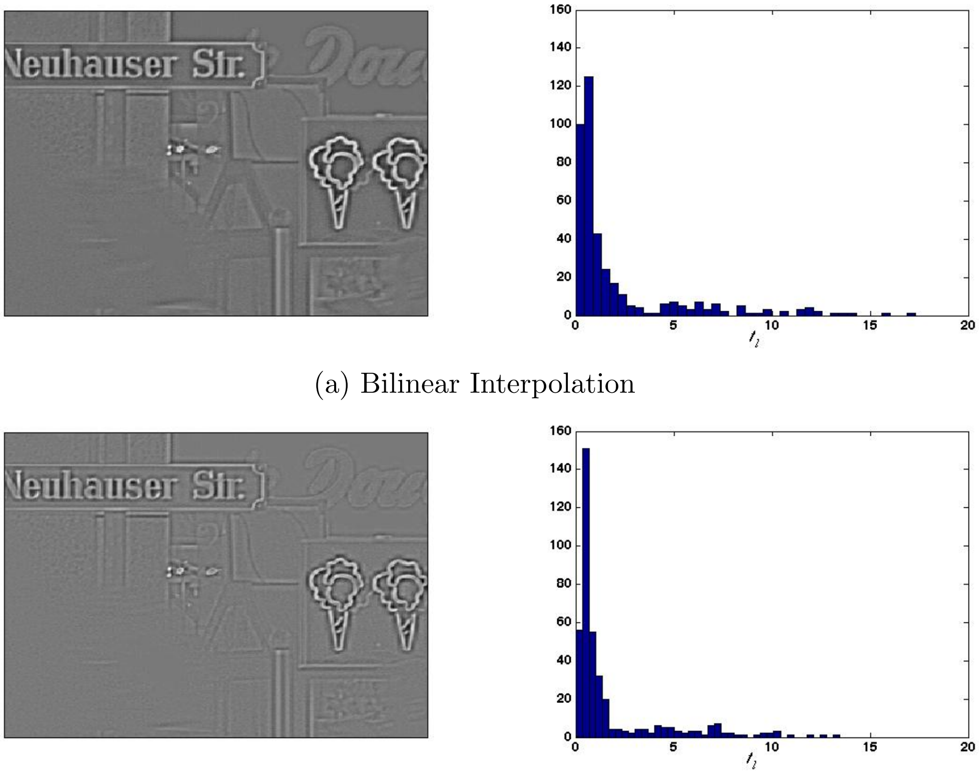
One important design aspect with the pyramidal image decomposition system is clean signal decorrelation. We need to find a method that retains as much low and medium frequency components in the lower layer as possible while keeping most high frequency components in the upper layer. Because this PID scheme has the similar scalable characteristics as SVC, we will retain the conventional phrasing of base layer (BL) and enhancement layer (EL) in our proposed system as well.

As the image of medium and low frequency components are now extracted into the BL with its long correlation captured within the span, they can be well predicted using the highly effective scheme proposed in [11] and coded by a small number of bits. In the next subsection, we will examine the impact of different interpolation schemes.

4.2.3 Interpolation Scheme

Image interpolation techniques have been well studied in the past. For bilinear interpolation, the resampled pixel value is obtained by the weighted average of 4 closest pixels. It tends to yield a blurred version of the original image since pixel values are altered during the process. Bicubic interpolation determines the output using the weighted average of 16 closest pixels. Its result is slightly sharper than that produced by bilinear interpolation.

The influence of two interpolation schemes on the EL statistics is compared in Fig. 4.3, where the image in display is a subimage taken from from the Pedestrian sequence of



(a) Bilinear Interpolation

(b) Bicubic Interpolation

Figure 4.3: The impact of interpolation schemes on the low-layer image: (a) bilinear interpolation and (b) bicubic interpolation, where the left-hand-side is the lower-layer difference image while the right-hand side is its histogram.

resolution 1920 by 1080. Since the test image contains both homogeneous and textured regions, it serves as a good test case. The left-hand-side image in Fig. 4.3 is the lower-layer difference image while the right-hand side figure is the corresponding histogram. The bicubic interpolation scheme provides a smaller dynamic range for the EL, which implies a better prediction result. Hence, it is used in our work.

4.2.4 Proposed Hierarchical Intra Prediction (HIP)

This pyramidal image decomposition serves as a good platform for hierarchical intra prediction (HIP) for the following reasons. First, the low and medium frequencies can be efficiently compacted into the BL which is smaller than the original size. Second, although the EL is of the original size, most of low frequency components are removed. This allows us to focus on high frequency components only. Third, as the BL is predicted independently from the EL, the decoder can decode the BL individually without waiting for the arrival of the EL. Even if part of the EL is lost or damaged during the transmission, the decoder can still decode the BL to produce a coarser representation of the block instead of an entirely corrupted block. This offers robustness to the decoding as the impact of corrupted blocks will not affect the decoding of other blocks in the upper layer.

This last feature is even more important in video coding. Since P or B frames in an image sequence rely on the availability of blocks of the I frame. If some blocks of the I frame are lost or corrupted, those blocks in P and B frames will not be reconstructed properly. However, this can be remedied by the HIP. As long as the BL is received properly, even if the EL is lost, the P or B frames can still reconstruct based on the coarser representation in the BL. In general, this decomposition-based HIP can achieve the same benefit offered by the traditional scalable structure with added features.

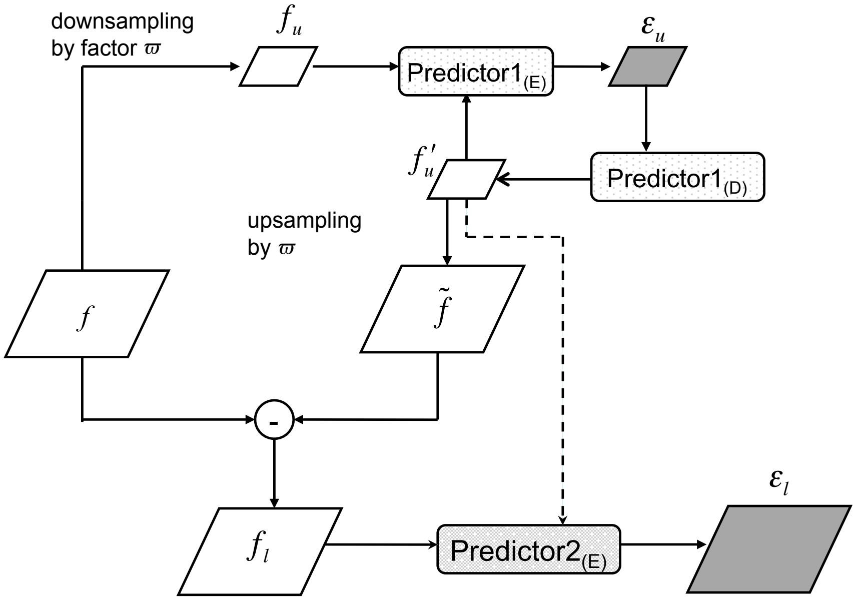


Figure 4.4: The block diagram of the hierarchical intra prediction (HIP) scheme.

The block diagram of the hierarchical intra prediction (HIP) scheme is shown in Fig. 4.4. For an input image/frame, f, a pyramidal decomposition scheme with the bicubic interpolation is used to downsample the original frame into a BL denoted by f_u, one predictor (called Predictor 1) is applied to f_u to yield the prediction error of the BL, denoted by ε_u. The BL prediction residual goes through a lossless entropy coder to produce the final bit rate for f_u. For correction block reconstruction, the prediction must be kept within the coding loop. That is, the prediction must be performed based on the reconstructed block instead of the original block. The module, $Predictor1_{(D)}$, indicates the decoding process to reconstruct $f_u^{'}$. The same principle must be applied to the process of obtaining the EL, f_l. In this figure, $\tilde{f}$ is an upsampled frame based on reconstructed $f_u^{'}$ instead of original f_u.

The EL f_l is further fed to a different prediction scheme called Predictor 2 to allow the high frequency components to be compensated. The proposed HIP scheme produces two layers of residuals with different resolutions, upper layer residual ε_u and lower layer residual ε_l. Both of them need to be entropy coded. However, the decoder does not need to receive both layers to decode the image. As the BL can be independently predicted from the EL, the decoder can decode the BL first without waiting for the EL to arrive. The decoded EL is added back to the BL at a later time for visual enhancement or discarded if the time stamp expires in the case of video compression.

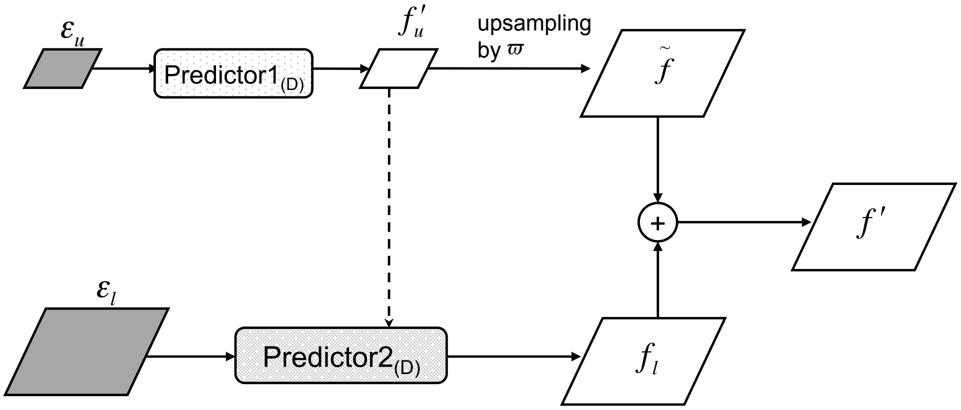


Figure 4.5: The block diagram of the hierarchical intra prediction in the decoder.

Once the HIP decoder receives the BL prediction residual, ε_u, it can decode ε_u. However, in order to properly decode ε_l, the decoder must possess the correctly decoded ε_u to reconstruct $f_u^{'}$. As the decoder relies on the information in $f_u^{'}$ to correctly reconstruct f_l. Fig. 4.5 shows the block diagram of the hierarchical intra prediction in the decoder.

The reconstructed and upsampled BL, $\tilde{f}$, can be added back to the decoded EL, f_l, to obtain the final reconstructed copy of f'.

In our scheme, the EL has the same resolution as the original input and consists of high frequency components, which poses a serious challenge on the existing prediction scheme. To address this problem, we propose a block-based joint linear prediction method that is designed to provide efficient prediction to the complex edge textures presented in the EL. Since the lossless line-based prediction (LIP-LS) in H.264/AVC is designed for homogeneous blocks with little texture variations, the prediction error is high in textured regions as the LIP cannot deal with complex texture efficiently. In the next section, we consider a different type of predictor that is designed to predict edge efficiently.

4.3 Predictive Coding with Linear Combinations

Differential Pulse Code Modulation (DPCM) is a typical predictive coding scheme that generates prediction $\hat{x_i}$ based on the past data samples without side information. Most existing DPCM methods, such as JPEG-LS [49] and CALIC [52], are pixel-based to maximize inter-pixel correlation. However, for nontrivially oriented edges, DPCM performs poorly. Lossless least squares prediction was proposed in form of [28, 46, 51, 53].

$$p = \sum_{i=1}^{N} c_i x_i, \tag{4.2}$$

where c_i are coefficients to form the linear combination of the prediction. A training window, usually in form of a double rectangular window, is built from its previously coded pixels to derive the optimal coefficients. Least squares methods are proven to be effective in adaptation to spatially changing features in an image. However, it is limited to lossless image compression as the spatial correlation between the target block and blocks in the training window is too low to derive good prediction.

Recent prediction algorithms used in image and video compression mainly fall in the category of prediction with side information. As it is observed that prediction error requires more bits while the side information demands less. Hence, it actually improves the overall rate-distortion (RD) performance by trading the side information bits for the residual bits. H.264/AVC provides an example that uses a large percentage of the side information to indicate the best mode to minimize the prediction error. Study in [9] shows that H.264's side information could occupy from 20-60% of the overall coding bits. This prediction with the side information was pushed to another level by Sullivan [42] in the so-called "multi-hypothesis motion compensation" (MH-MCP). It is based on the similar concept of linear prediction, n blocks of $c_1, c_2, ..., cn$ from previous frames were used to predict a block s in the current frame. Each of these predictive blocks is called a *hypothesis*. Block $\hat{s}$ is determined by a linear combination of multiple hypotheses:

$$\hat{s} \quad = \quad \sum_{v=1}^{N} c_i h_i, \qquad (4.3)$$

where c_i is a weight. To find c_i and h_i, an exhaustive search combined with the rate-distortion optimization (RDO) method is used. Theoretical investigation conducted in [18] shows that a linear combination of multiple predictors can improve the motion compensation performance. However, there is one severe drawback with this MH-MCP scheme; namely, the complexity of finding the optimal set of c_i and h_i is very high.

In actual coding, a compromise was achieved in [15] by assigning the same coefficient to all hypotheses as

$$\hat{s} \;=\; \frac{1}{N}\sum_{v=1}^{N} h_i. \tag{4.4}$$

Even if coefficients are constant for all hypotheses, MH-MCP still has to determine the hypothesis displacement vector $(\Delta_x, \Delta_y, \Delta_t)$ as the side information. The increase of the side information could potentially mask the efficiency provided by the MH-MCP. If we can obtain superior quality by exploiting the prediction with the side information while inferring the side information based on available samples, this will enhance the coding performance. Such a scheme will be presented in the next subsection.

It is important to choose a predictor that is most suited to image characteristics in different layers in the pyramidal image decomposition. The EL consists of mainly high frequency components and its features are presented as contours of objects in the lower layer. Here, we propose a linear prediction scheme. Being different from the MH-MCP, the prediction demands no side information such as the displacement vectors. The decoder can follow the exact optimization process to obtain coefficients and reconstruct the predicted block.

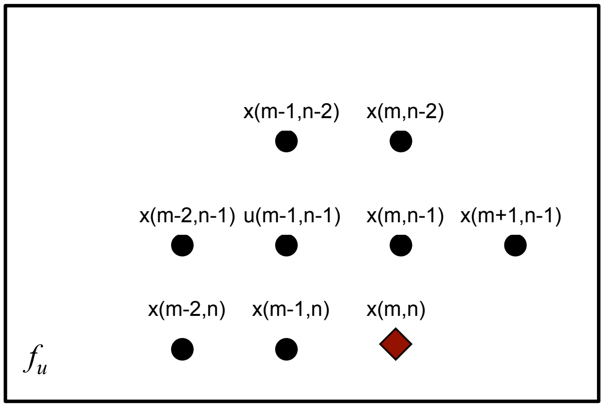


Figure 4.6: Linear prediction with an N-th order casual neighborhood where $N = 8$.

4.3.1 Linear Predictive Coding for Image Pixel

Under the assumption of the N-th order Markov propery, we can construct an N-th order linear prediction based on its N nearest neighbors as shown in Fig. 4.6:

$$\hat{x}(i) = \sum_{k=1}^{N} \theta(k) x(i-k), \tag{4.5}$$

where $\theta(k)$ are predictive coefficients. The prediction error can be written as

$$\begin{aligned}\hat{e}(i) &= x(i) - \sum_{k=1}^{N} \theta(k) x(i-k) \\ &= \sum_{k=1}^{N} \theta(k) \Big[x(i) - x(i-k) \Big] \\ &= \sum_{k=1}^{N} \theta(k) e(i-k). \end{aligned} \tag{4.6}$$

To determine coefficients, we use the previously coded block to construct a training window of M casual neighbors of $x(n)$ as shown in Fig. 4.7. For simplicity, the training window is fixed here. The optimization of the training set will be discussed later.

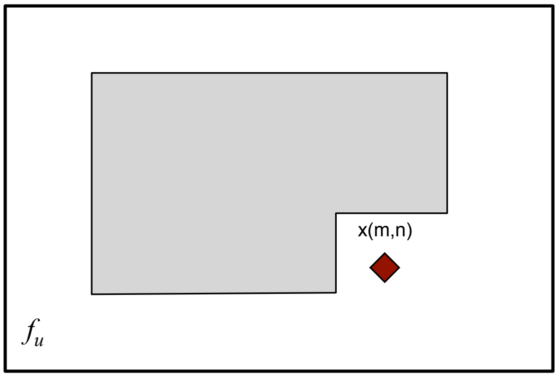


Figure 4.7: Linear predictive coding with a fixed training window.

Each sample is denoted by $x(L), x(L+2), \cdots, x(L+M)$. This training window can be expressed as a $M \times 1$ column vector in form of $\overrightarrow{y} = [x(L+1), x(L+2), ...x(L+M)]^T$. Then, the prediction neighbors of $\overrightarrow{y}$ can be arranged in into an $M \times N$ matrix as

$$C = \begin{bmatrix} x(L) & \cdots & x(L-N) \\ \vdots & & \vdots \\ x(L-M) & \cdots & x(L-M-N) \end{bmatrix}, \tag{4.7}$$

where $x(L-j-k)$ is the kth prediction neighbor of $x(L-j)$. To solve for optimal coefficients $\theta(k)$ is to minimize the sum of squared differences in the training window:

$$\min \sum_{j=1}^{M} \left[\overrightarrow{y} - C\Theta \right]^2 \tag{4.8}$$

The least squares optimization has a close form solution; namely,

$$\hat{\Theta} = (C^T C)^{-1} C^T \overrightarrow{y}, \tag{4.9}$$

where

$$\hat{\Theta} = \Big[\theta(1), \theta(2), ..., \theta(N)\Big]^T. \tag{4.10}$$

The above optimization process allows prediction coefficients to be adaptive to edge neighbors without explicitly detecting the edge orientation as it is done in JPEG-LS or CALIC.

4.3.2 Joint Inter-layer Linear Prediction

In the HIP, the pyramidal image decomposition extracts high frequency components to the EL while leaving low and medium frequency components to the BL. As a result, the EL consists of flat regions or highly complex regions. One problem with the linear prediction is that matrix C may not be of full rank in flat regions. This could produce a prediction value that is worse than an intuitive prediction as the linear prediction may not produce an unique value. Besides, as the resolution of the image/intra frame increases, the computational complexity to calculate the covariance matrix C^TC increases significantly. Different blending algorithms based on the context of previously coded samples were employed in [23]. However, the context switching method could be heuristic and inaccurate without the exact knowledge of the current sample to be predicted. Here, we propose a different edge detection mechanism to apply linear prediction for complex regions with strong edges and patterns adaptively, while leaving flat regions to be predicted by the LIP. This avoids the potential large prediction error arising from linear prediction and reduces the overall complexity. In addition, we would like to perform a predictor switch scheme without incurring the overhead bits for each block to indicate which predictor is used for the target block. To achieve the predictor selection adaptively, we utilize the information present in the BL.

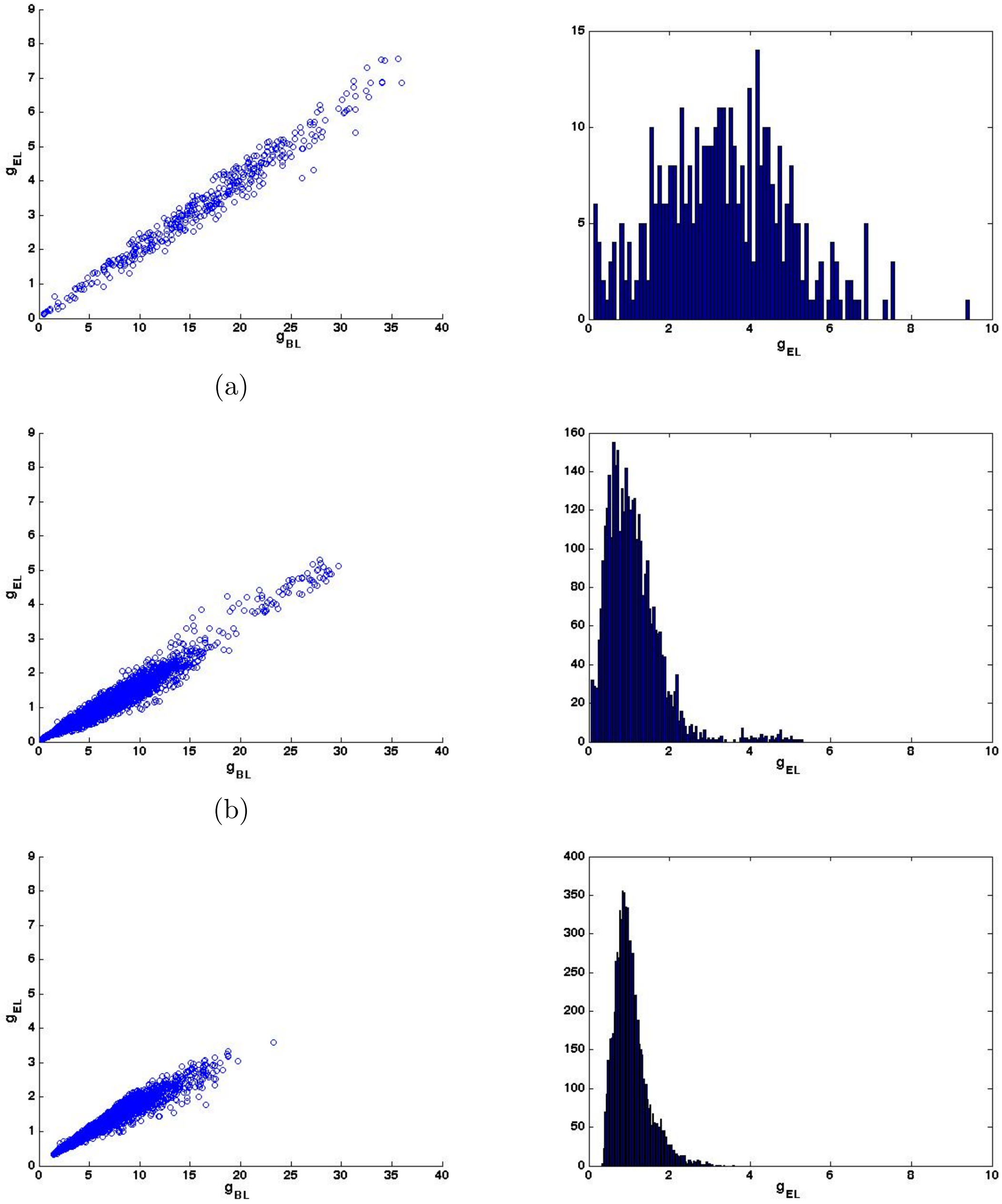
In the proposed lossless HIP scheme as depicted in Fig. 4.4, the BL, $f_u^{'}$, is available to the EL, f_l. Thus, if there exists a correlation between $f_u^{'}$ and f_l, the block statistics obtained in $f_u^{'}$ can be utilized for the EL to determine which prediction method is more suited for that block without searching all possible prediction schemes. To show this, we conduct a correlation study on the block texture between BL $f_u^{'}$ and EL f_l. To get the one-to-one correspondence, the BL is first re-upsampled to the same resolution as the EL. Block texture is measured using the gradient method along vertical and horizontal direction as

$$\begin{aligned} g_v &= \frac{1}{M}\sum_{j=1}^{M}\sum_{i=1}^{M}|x(i,j) - x(i+1,j)|, && (4.11) \\ g_h &= \frac{1}{M}\sum_{j=1}^{M}\sum_{i=1}^{M}|x(i,j) - x(i,j+1)|, && (4.12) \end{aligned}$$

with

$$g = \frac{1}{2}(g_v + g_h) \qquad (4.13)$$

is used as an average gradient for each block. The relationship between the gradient values of the BL and the EL is plotted in the left column of Fig. 4.8 while the gradient histogram in the EL is plotted in the right column of Fig. 4.8.



(a)

(b)

(c)

Figure 4.8: The relationship between the gradient values of the BL and the EL and the histogram is texture histogram for EL for (a) Mobile (CIF), (b) City corridor (1280x720), and (c) Vintage car (1920x1080).

We see from Fig. 4.8 that the relationship between the texture of co-located blocks in the BL and the EL are highly correlated. Besides, this linear relationship is consistent from frame to frame with different block characteristics, distributions and different resolutions. Thus, an edge detection mechanism based on the gradient measure of co-located blocks on the BL can be derived to identify the type of target blocks. We apply the linear prediction for target block k_{EL} in the EL, if its co-located counterpart k_{BL} in the BL contains high frequency components

$$g(k)_{BL} \quad \geq \quad \gamma_1, \tag{4.14}$$

where γ_1 is a threshold value.

There are a few advantages with the proposed scheme exploiting the inter-layer correlation. First, the co-located block in the BL is available to the decoder and hence can be used to perform accurate edge detection before starting to decode the target block. Second, the switch between the LIP and linear prediction does not demand any side information except for the value of γ_1. Basically, the decoder can use γ_1 to identify a switch in the predictor type. Finally, the gradient measure is done on the BL of smaller resolution, which can reduce the computational complexity on edge detection.

4.3.3 Training Set Optimization

For a linear prediction scheme, the prediction accuracy is highly dependent on the correlation between samples in the training window and the sample to be predicted because the algorithm assumes the N-th order Markovian property. The question is the choice of the window size N. If the samples in the training window are highly uncorrelated from the target sample, coefficients obtained from this training set would produce a large prediction error. Our correlation analysis has shown that the co-located blocks can be a reliable indicator to identify the texture of the target block in the EL. Here, we can resort to the BL block gradient to eliminate those blocks from the training set that have different characteristics from the target block.

We define a training window as specified in Fig. 4.7 in the upsampled BL. Beginning from the position of the co-located block in BL, we use the gradient of the co-located block as our target $g(k)_{BL}$ and adopt the circular search and remove those blocks that do not qualify:

$$g(k)_{BL} - g(k-i)_{BL} \quad \leq \quad \gamma_2. \tag{4.15}$$

Based on the size of the training set, we record the positions of the first M blocks that satisfy Eq. (4.15). The EL layer will then use the blocks with these marked positions in the EL as the desired training set. This will allow us to select the best training set adaptively without extra side information.

4.4 Experimental Results

In this section, experiments were conducted to compare the performance of the following coding schemes: lossless H.264/AVC intra prediction (H.264 LS) [1], JPEG-LS, and the proposed lossless hierarchical intra prediction (HIP-LS). For the HIP-LS, we adopted the two-layered setup with the BL of a quarter size of the EL, and quantization parameter as

QP=20. The enhancement layer is coded losslessly to achieve the final goal of a lossless coding system with the added spatial/quality scalability. In addition, the joint block and line-based intra prediction proposed in [11] is chosen as Predictor 1 for the BL prediction. However, we disabled the 2D geometrical manipulations and the 2x2 block enhancement to reduce the coding complexity. For predictor 2 in the EL, we choose among the H.264-LS and the linear prediction based on the detected edge mechanism. We consider a casual neighbor of size $N = 3$ and a training set of size $M = 6$.

Table 4.1: Coding efficiency comparison between H.264-LS, JPEG-LS and HIP-LS.

Resolution @ 1280x720			
Sequence	Method	Compression Ratio (%)	Bit rate saving w.r.t H.264-LS(%)
Sheriff	H.264-LS	50.64	0
	JPEG-LS	47.75	5.69
	HIP-LS	45.81	9.54
Sailorman	H.264-LS	63.31	0
	JPEG-LS	60.64	4.22
	HIP-LS	55.17	12.85
Harbor	H.264-LS	59.85	0
	JPEG-LS	55.97	6.49
	HIP-LS	52.76	11.84
Preakness	H.264-LS	84.03	0
	JPEG-LS	80.32	4.42
	HIP-LS	76.19	9.33
Average	H.264-LS	**64.46**	**0**
	JPEG-LS	**61.17**	**5.20**
	HIP-LS	**57.48**	**10.89**

Eight HD YUV sequences of two different resolutions were used in the experiments. They were: Sheriff (@1280x720), Night (@1280x720), Harbor (@1280x720), City corridor (@1280x720), Station2 (@1920x1080), Blue sky (@1920x1080), Riverbed (@1920x1080) and Vintage car (@1920x1080). The entropy coding method was CAVLC. All values were

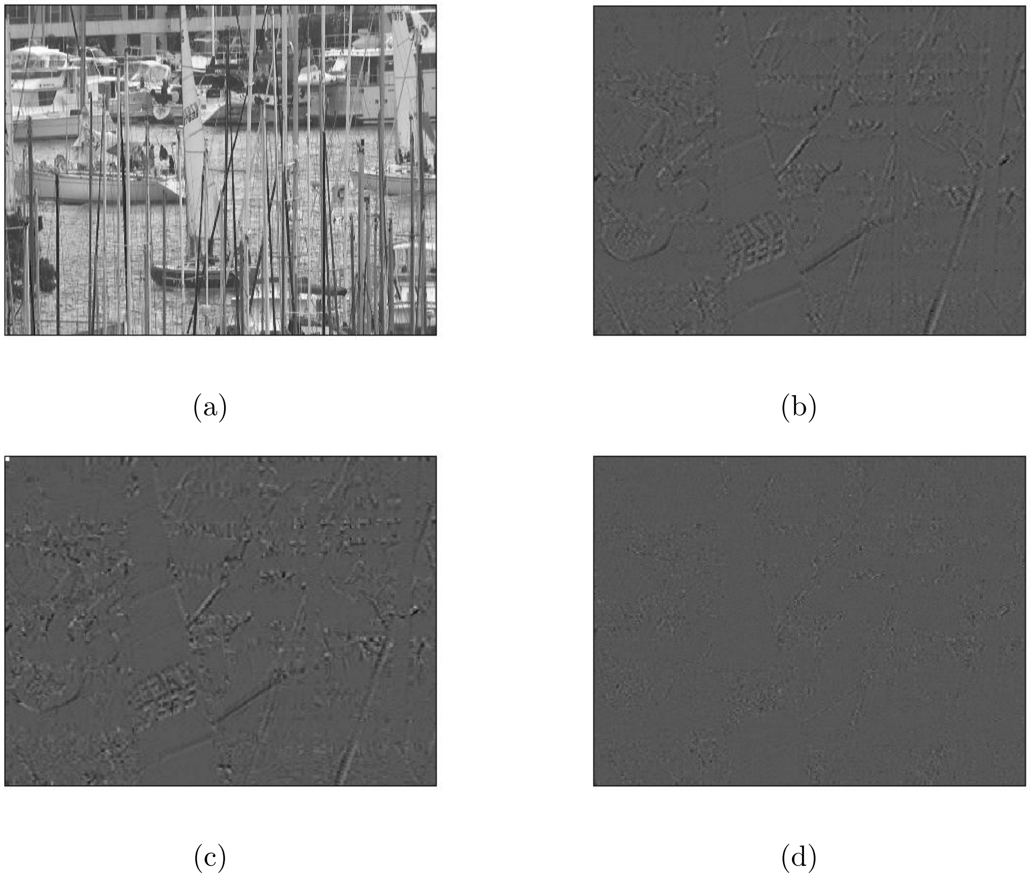
Table 4.2: Coding efficiency comparison between H.264-LS, JPEG-LS and HIP-LS.

Resolution @ 1920x1080			
Sequence	Method	Compression Ratio (%)	Bit rate saving w.r.t H.264-LS(%)
Rush Hour	H.264-LS	43.37	0
	JPEG-LS	42.42	2.20
	HIP-LS	38.91	10.28
Blue Sky	H.264-LS	46.37	0
	JPEG-LS	43.65	3.88
	HIP-LS	41.89	9.65
Sunflower	H.264-LS	39.95	0
	JPEG-LS	38.01	4.86
	HIP-LS	35.67	10.72
Vintage Car	H.264-LS	71.26	0
	JPEG-LS	67.14	5.78
	HIP-LS	63.32	11.14
Average	H.264-LS	**50.24**	**0**
	JPEG-LS	**47.78**	**4.18**
	HIP-LS	**44.95**	**10.45**

averaged over 20 pictures of equi-spaced frame indices from respective sequences. The compression ratio is defined as

$$R = \frac{\text{output file size}}{\text{original file size}}.$$

The results are presented in Table 4.1 and Table 4.2. We see that the proposed lossless HIP can achieve a consistent improvement of an average of 14% in coding efficiency compared to the single layered H.264-LS and JPEG-LS. The proposed HIP is especially effective for contents with complex textures in all resolutions such as Sailorman and Vintage car. For better visual comparison, we also draw the prediction error maps from all three methods as shown in Fig. 4.9 and 4.10. We can see that the proposed lossless HIP scheme generates the smallest prediction errors for both sample frames.



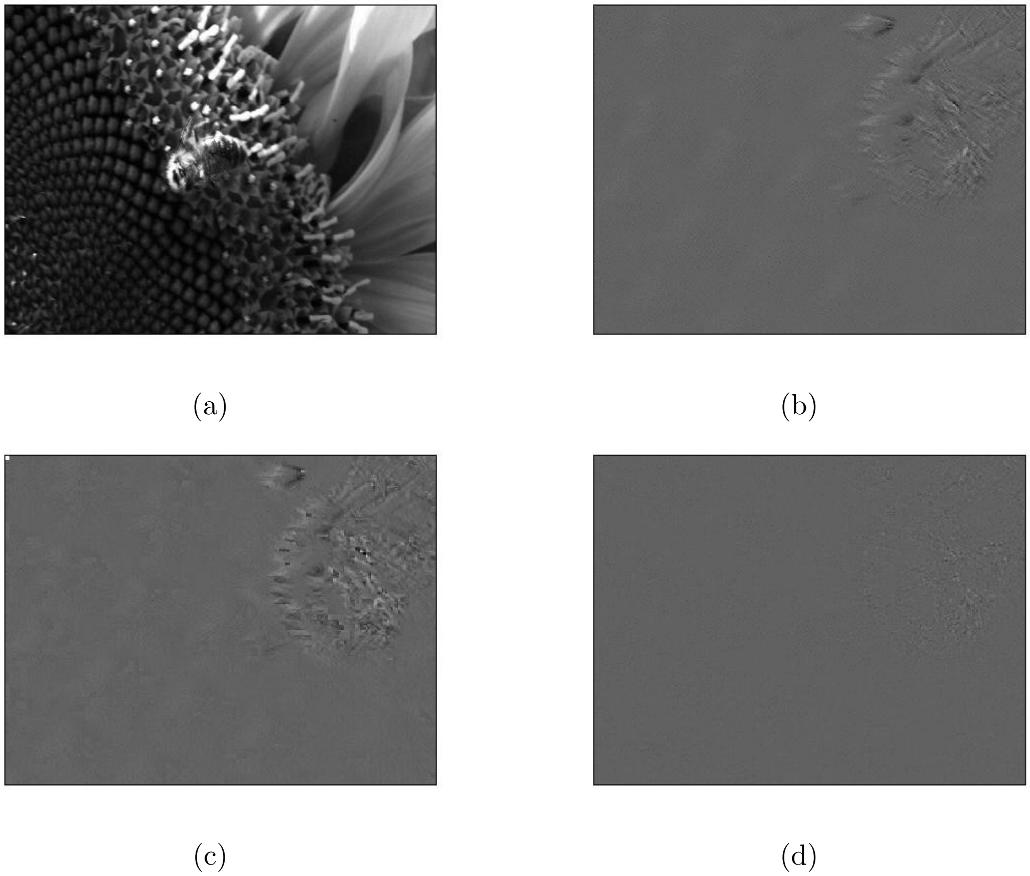
(a)

(b)

(c)

(d)

Figure 4.9: (a) Original Harbor frame and its prediction error maps using (b) JPEG-LS, (c) H.264 LS intra prediction and (d) proposed lossless HIP scheme.



(a)

(b)

(c)

(d)

Figure 4.10: (a) Original City Corridor frame and its prediction error maps using (b) JPEG-LS, (c) H.264 lossless intra prediction and (d) proposed lossless HIP scheme.

4.5 Conclusion

In this paper, we proposed a novel lossless hierarchical intra prediction scheme that delivers much improved coding efficiency for lossless coding with added spatial/quality scalability based on the proposed image pyramidal decomposition scheme in spatial domain. A novel linear prediction scheme with edge detection was designed specifically for the EL coding. A training set optimization using the gradient measured on the BL blocks was used to select the best suited training sets adaptively without overhead. Experimental results were given to demonstrate that the proposed lossless HIP scheme can deliver an average of 14% improvement on compression efficiency while deliver the added flexibility of an additional spatial/quality scalable layer.

Chapter 5

Context-based Hierarchical Intra Prediction for Lossless Image Coding

5.1 Introduction

The importance of efficient lossless image coding has been widely recognized over years due to its application to many professional fields such as medical imaging, professional digital photography, scientific geographic information systems and etc. Furthermore, as the image size become increasingly larger, the importance of a scalable lossless solution become more evident, especially with the steadily growing popularity of delivering digital contents over a heterogeneous network. Scalable coding standards use a layered approach by decomposing the content into one low resolution fundamental presentation, known as the base layer (BL), and other higher resolution enhancement layers (EL). When the content is to be previewed, indexed or searched on the camera, in the studio or over the network, BL can be used to allow speedy access at lower hardware cost and network bandwidth requirement. When the content is to be edited or viewed at its full resolution, additional enhancement layers can be further decoded to recover the content at the full resolution.

Generally speaking, the entire process of digital content production, post-production and on-line delivery can benefit since the scalable image/video format will reduce the hardware cost, improve production efficency and facilitate content delivery. However, today's image/video codecs typically sacrifice the coding performance to offer additional flexibility. Since they result in overall coding efficiency degradation, scalable image/video coding formats are not widely used. A coding solution that offers additional flexibility such as spatial scalability and improve the overall coding performance at the same time is highly desirable.

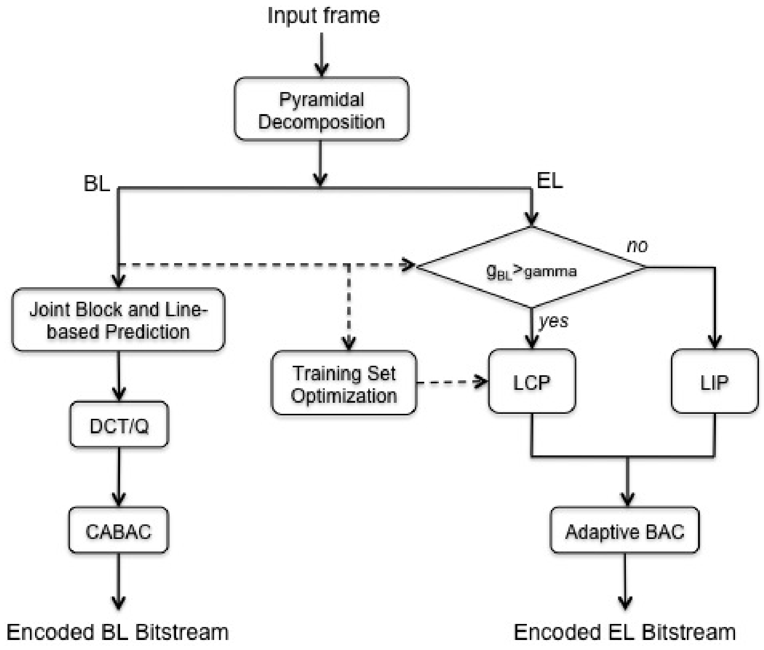


Figure 5.1: The proposed lossless HIP scheme.

A lossless hierarchical intra prediction (HIP) scheme was proposed in the previous chapter for lossless image/video coding. As shown in Fig. 5.1, the HIP scheme was designed to provide a unified scalable solution to the tradition lossless image/video coding while allows more efficient compression. Being different from the traditional scalable solutions, the proposed HIP employs a pyramidal image decomposition scheme to obtain the BL and ELs. The hierarchical structure serves as an image decorrelation scheme in the spatial domain, where the down-sampling operation is used to extract long correlation to the self-contained BL. Based on this decomposition method, a hierarchical coding system with multiple predictors was proposed by exploiting the correlation characteristics of each layer. Despite its excellent coding performance, there is still room for further improvement, which is the main focus of this chapter.

The rest of this chapter is organized as follows. We study the construction of the H.264/AVC LIP modes and propose an approximation scheme to estimate the best LIP mode that is mostly likely to be used for the target block in Sec. 5.2. This scheme eliminates the need of the side information. To further improve prediction accuracy, we use the Singular Value Decomposition (SVD) method to extract dominant features from the coarse representation of BL in Sec. 5.3. The extracted features are clustered using the k-means algorithm, and the context-based interlayer prediction (CIP) scheme is adopted to select the best prediction candidate. To achieve more efficient coding, an adaptive CABAC entropy coder is developed in Sec. 5.4 by analyzing the characteristics of prediction residuals. Context modeling is also enhanced in the CABAC coding process. Experimental results are give in Sec. 5.5 to demonstrate the effects of proposed improvements. Finally concluding remarks are given in Sec. 5.6.

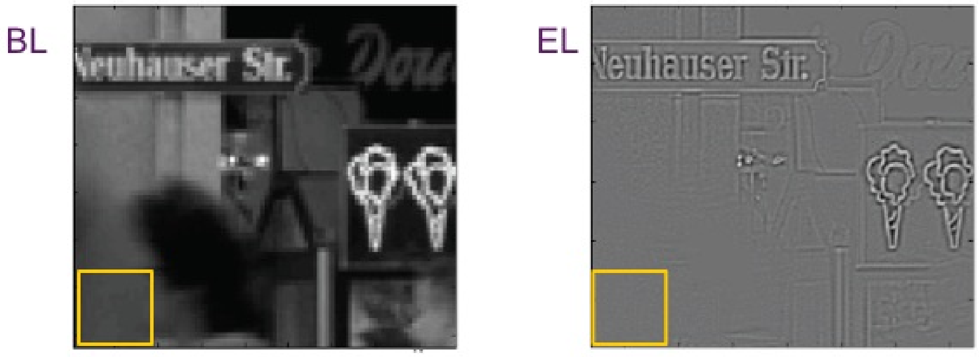


Figure 5.2: The co-located blocks in the upsampled BL and its EL.

5.2 Modeless Prediction

For the lossless HIP scheme proposed in Chapter 4, we studied the correlation between the co-located blocks in the BL and the EL and showed that the complexity of a block in the EL is linearly proportional to its upsampled co-located blocks in the BL. Thus, we proposed a modeless method that performs edge detection to capture the complexity of the upsampled co-located block in the BL. Then, we apply the linear combination prediction (LCP) to target block k_{EL} in the EL, if the high frequency components of its co-located counterpart k_{BL} in the BL are higher than a certain threshold. Otherwise, other blocks will be handled by regular LIP methods. In Fig. 5.2, we show the visual appearance of a block that can be handled with the regular LIP. Although the above scheme eliminates the need of bits to differentiate the LIP and the LCP modes, it still needs bits to encode the suitable LIP mode for that block. For example, we have to encode the LIP mode information such as Intra4x4 vertical (mode 0), horizontal (mode 1), DC mode (mode 2), etc. Consequently, it is still not a modeless scheme.

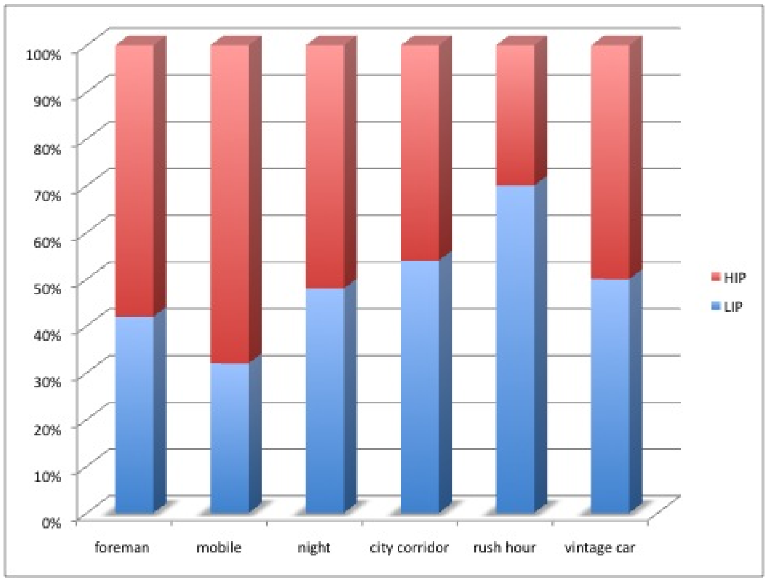
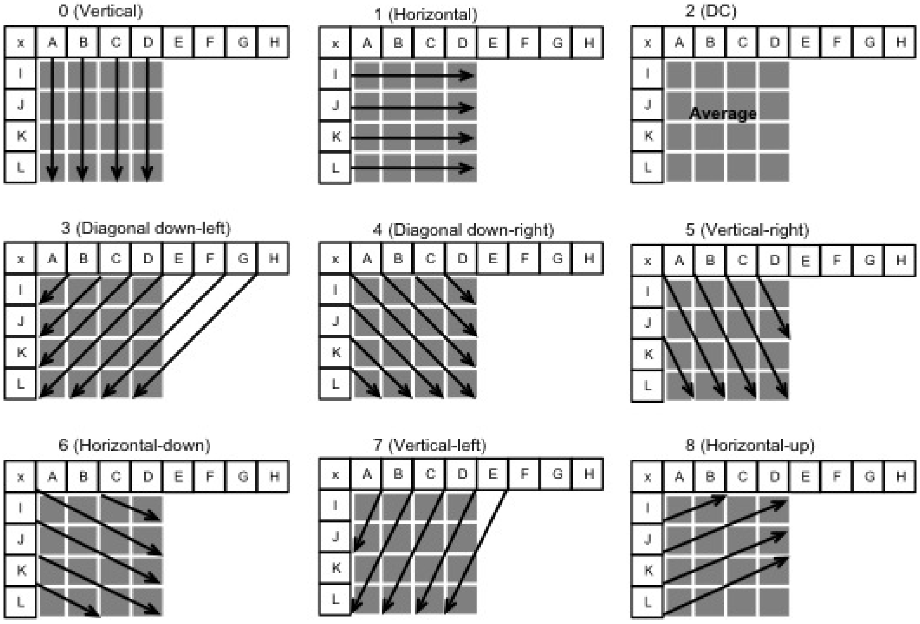


Figure 5.3: The percentages of blocks coded using HIP and LIP modes for six test vide sequences.

The percentages of blocks encoded using HIP and LIP modes for six test vide sequences are shown in Fig.5.3. We see that all sequences except Mobile have more than 50% LIP modes. This is because that the EL often contains a large amount of regions that are relatively flat and can be well encoded using the LIP. The number of bits required to encode the mode information can go up to 5% or more of the overall bits needed to encode the entire frame. If we can find an scheme to estimate which one of the nine LIP modes would be chosen as the LIP mode, the overhead bits of mode coding can be saved.

In H.264/AVC, LIP modes are constructed by extrapolating the decoded pixel values from the neighbors of the current block to form a prediction block. Different directional predictions are defined by modes and used to exploit the spatial correlation that may exist between the predicted and actual pixels in a target block. A total of nine different


0 (Vertical)

1 (Horizontal)

2 (DC)

3 (Diagonal down-left)

4 (Diagonal down-right)

5 (Vertical-right)

6 (Horizontal-down)

7 (Vertical-left)

8 (Horizontal-up)

Figure 5.4: Nine prediction modes for 4 × 4 blocks in H.264/AVC intra prediction.

prediction directions for target block S of size 4×4 are shown in Fig. 5.4. Pixel samples A, B, C, $\cdots$, H in Fig. 5.4 are denoted by $p(0,-1)$, $p(1,-1)$, $\cdots$, $p(7,-1)$ while pixel samples I, J, K and L are denoted by $q(-1,0)$, $q(-1,1)$, $q(-1,2)$ and $q(-1,3)$, respectively.

We use mode 0 (vertical prediction) and mode 1 (horizontal prediction) as examples, whose prediction residuals can be expressed as

$$r_{m0}(i,j) = s(i,j) - p(i,-1), \tag{5.1}$$

$$r_{m1}(i,j) = s(i,j) - q(-1,j), \tag{5.2}$$

where i, j are the relative position of pixels in the target block. The formation of the LIP prediction blocks indicates that the prediction attempts to capture the intensity continuity from neighboring blocks to the current target block along different directions. If this intensity continuity can be revealed by some existing information in the coded bit stream, the bits required for mode coding can be saved.

Recall that we adopted the gradient based edge detection scheme to identify which blocks should use LIP modes (rather than the HIP mode) in the HIP scheme. The gradient values used in the mode selection process is an average of the gradient values computed in both horizontal and vertical directions of that block as shown in Eqs. (4.12) and (4.12).

In the following, we will generalize the edge detection mechanism by including the decoded neighboring block pixel lines in the gradient computation. Being similar to the previous edge detection scheme, the new mode estimation mechanism is derived based on the information available from the BL. For example, to estimate the prediction accuracy of mode 0 and mode 1 in the LIP, we modify the gradient measurements in Eqs. (4.12) and (4.12) by taking neighboring pixels into account as

$$\begin{aligned} g_{m0} &= \sum_{i,j=-1}^{M-1} \Big| s(i,j) - s(i,j+1) \Big|, && (5.3) \\ g_{m1} &= \sum_{i,j=-1}^{M-1} \Big| s(i,j) - s(i+1,j) \Big|, && (5.4) \end{aligned}$$

where decoded pixel lines $p(i,-1)$ and $q(-1,j)$ in Eqs. (5.1) and (5.2) are re-written as $s(i,-1)$ and $s(-1,j)$ for convenience in the mathematical formulation to be shown later. For mode 2, which is the DC mode, we can update the average gradient computation in Eq. (4.13) as

$$g_{m2} = \frac{1}{2}\big(g_{m0} + g_{m1}\big). \tag{5.5}$$

By following the same idea, we can generalize the gradient method for other modes. For example, the prediction residuals for mode 3 (diagonal down-left) and mode 4 (diagonal down-right) can be written as

$$g_{m3} \quad = \quad \sum_{i,j=-1}^{M-1} \Big| s(i,j) - s(i-1,j+1) \Big|, \tag{5.6}$$

$$g_{m4} \quad = \quad \sum_{i,j=-1}^{M-1} \Big| s(i,j) - s(i+1,j+1) \Big|, \tag{5.7}$$

To avoid redundant computation and unify the proposed mode estimation scheme, we replace (4.12) and (4.12) with (5.3) and (5.4), respectively.

One advantage of this LIP mode estimation mechanism is that we can have all LIP modes tested on the BL first without the knowledge of the target block statistics of the EL. Then, we can select the one with the minimum gradient values as the best mode for the colocated block in the EL. That is,

$$m_{opt} = \min \Big\{ g_{m1}, g_{m2}, g_{m3}, g_{m4}, ..., g_{m8} \Big\}. \tag{5.8}$$

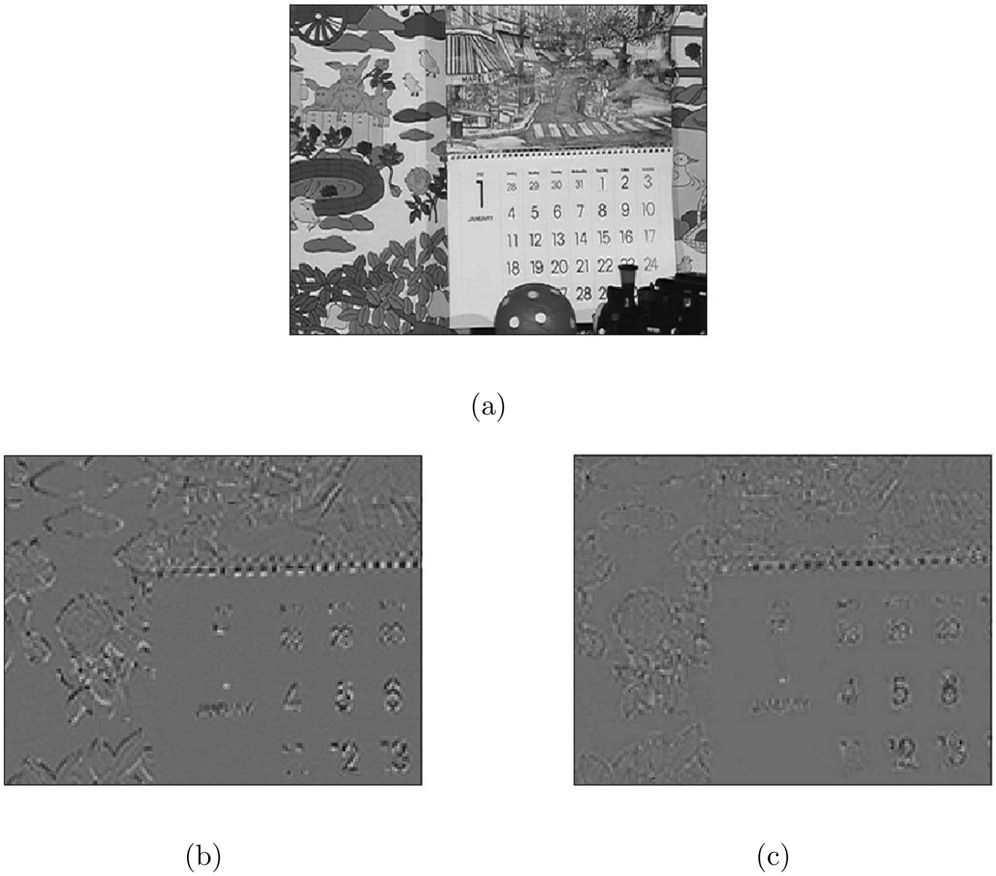
This LIP mode estimation mechanism allows mode selection done in BL blocks only. The prediction residual of EL blocks can be coded without the mode information since the decoder can follow the same estimation rules and derive the correct LIP mode.

5.3 Context-based Hierarchical Intra Prediction (CHIP)

5.3.1 Research Motivation

In Chapter 4, we mentioned that the performance of the LCP scheme is influenced by the similarity between the training set and the target sample since the algorithm assumes the N-th order Markovian property. If the samples in the training window and the target sample are uncorrelated, coefficients obtained from the training set would yield a large prediction error. Hence, to address this problem, a training set optimization technique was proposed to select the best suited training set for the coefficient training in the LCP scheme. However, after a fixed number of iterations, the technique has to re-use a limited number of choices to control the computation complexity. Alternative methods such as template matching [47] method proposed by T.K.Tan *et al.,* or texture synthesis inspired solution[4] propsed by Balle and Wien would not be able to sufficiently solve this problem as they are fundamentally still based on the valid assumption of the Nth order Markovian property. As a result, the performance improvement could be limited for images that consist of many heterogeneous subregions.

In Fig. 5.11, we show an example where the LCP scheme cannot produce a good prediction since the training set samples are very different from the target samples. We see that the numbers on the calendar in Fig. 5.11(a) are separated by flat regions which will not serve as good training sets. Furthermore, the training set optimization technique would not be effective since it would demand a large memormy for training set optimization. As a result, although the prediction scheme in the HIP scheme has prediction errors smaller than those of H.264/AVC as shown in Fig. 5.11(c), the tips of these



(a)

(b)

(c)

Figure 5.5: (a) One frame of the Mobile sequence with a resolution of 352x288; (b) the enlarged prediction residual using the LIP prediction and (c) the enlarged prediction residual using the HIP prediction.

numbers still contain large prediction errors. This means that a more effective prediction scheme is in need. One intuitive way to achieve this is to incorporate the JBLIP scheme as proposed in Chapter 3. However, in the hierarchical architecture, the frequent use of translational vectors might impact the coding performance negatively due to the coding cost of translational vectors. Hence, it is desirable to have a prediction scheme that has a good prediction result that can target these isolated textured areas but without the use of large translational vectors.

5.3.2 Block Classification via Feature Extraction and Clustering

To address this problem described above, we propose a method that identifies similarities between target samples and samples within the prediction frame that might not fall within the range of the training window. Simply speaking, similar training samples define a context by utilizing the information available from the decoded and upsampled BL. Then, we develop a context-based HIP prediction scheme, which allows us to identify the best fit prediction samples based on the context.

The Principal Component Analysis (PCA) [6] is widely used to reduce a complex data set to a lower dimension feature set that captures its underlying characteristics. The PCA is obtained by performing the singular value decomposition (SVD) on the data matrix [12]. In the current context, we compute the gradient of image f at pixel i as

$$\nabla f(i) \quad = \quad \begin{bmatrix} \frac{\partial f(i)}{\partial x} & \frac{\partial f(i)}{\partial y} \end{bmatrix}^T . \tag{5.9}$$

To estimate the local orientation of a block of size $m \times m = M$ centered at the target sample, we group their gradient values into an $M \times 2$ as

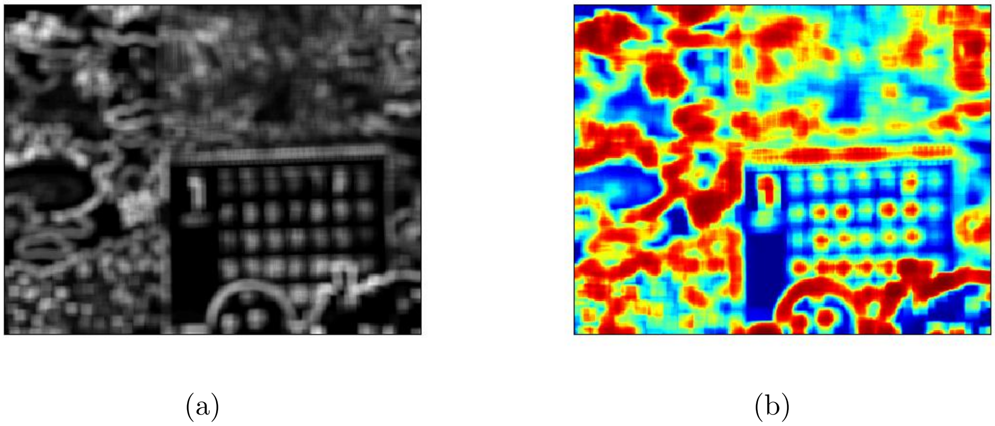
$$G = \begin{bmatrix} \nabla f(1)^T \\ \nabla f(2)^T \\ \vdots \\ \nabla f(M)^T \end{bmatrix}, \tag{5.10}$$

The SVD of matrix G can be expressed as

$$G = USV^T, \tag{5.11}$$

where U and V are orthogonal matrices of dimension $M \times M$ and 2×2, respectively, and matrix S of dimension $M \times 2$ contains two singular values that described the energy in the dominant orientation of the gradient field at pixel i.

Given the dominant SVD feature of each sample, we partition them into k clusters using the k-means algorithm [29]. An example is given in Fig.5.6. We show the magnitude of the singular value $S[0, 0]$ for the Mobile frame in Fig. 5.6 (a) and the clustered result in Fig. 5.6 (b), where the cluster number is equal to $k = 64$.



(a)

(b)

Figure 5.6: (a) The plot of the magnitude of the 2-D singular value vector for a frame in the Mobile sequence and (b) its corresponding k clustered regions with $k = 64$.

5.3.3 Context-based Inter-layer Prediction (CIP)

After applying the SVD computation and clustering to the upsampled BL, we obtain f clusters denoted by k_f as shown in Fig. 5.8(a). Typically, we have multiple samples associated with each cluster, where each sample has its own prediction rule. To illustrate the above idea, we give an example in Fig.5.8. Samples x_i and x_j belong to the same context as shown in Fig. 5.7.

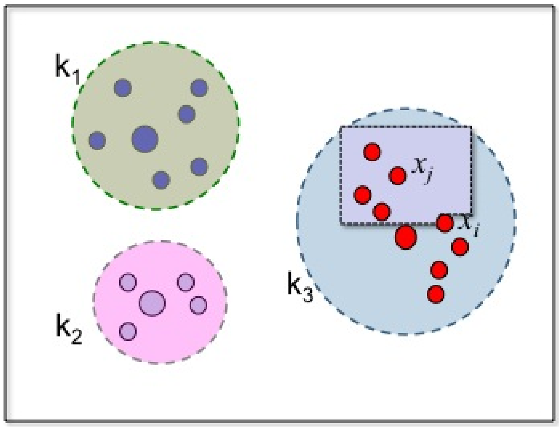


Figure 5.7: Graphical illustration of multiple clusters after the k-means clustering method, where the bigger dots indicate the centroid of that cluster.

If the prediction and selection is directly performed between x_i and x_j, additional side information must be recorded and sent to the decoder for correct reconstruction as the decoder has no knowledge of x_i. However, we want to avoid the use of side information. Therefore, we define the term of "context". Each context consists of the $N$the causal neighbors of the sample. Therefore, the prediction is processed between the context C_i of the target sample x_i and those C_j of the prediction samples x_j as shown in Fig. 5.8. The highlighted blue regions are the contexts.

This SSD computation will be conducted for all samples within the same context k_f and the one that yields the smallest d_j will be used as the final prediction rule for the target sample.

$$\begin{aligned} & \min\{d_j\}, \\ d_j(x_j|k_f) &= \sum \Big[c_i(x_i|k_f) - c_j(x_j|k_f)\Big]^2. \end{aligned} \tag{5.12}$$

In case of a tie, an Euclidean distance criterion can be used as a tie-breaker.

Hence, the best prediction can be obtained as

$$\hat{x_i} = x_j, with \min_j\{d_j|c_k\}. \tag{5.13}$$

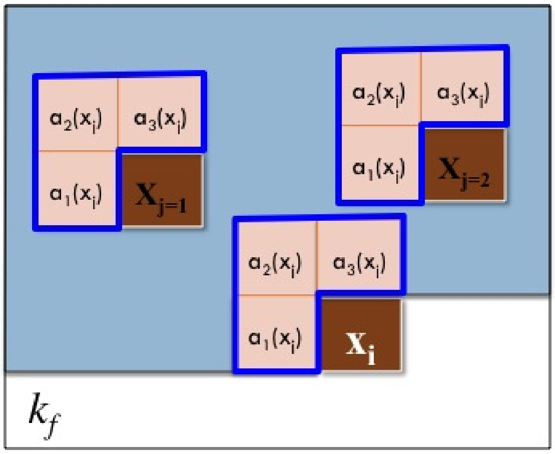


Figure 5.8: Graphical illustration of contexts of the target sample and that of the prediction samples.

As the clustering is applied to the entire frame, it is possible that some samples within the same context fall in the range of future samples R_c as shown in Fig. 3.2. Those samples should be eliminated from the SSD evaluation to allow accurate reconstruction of the target sample. Only the samples within the rectangular shaded region as shown in Fig.5.8(a) can be used for prediction.

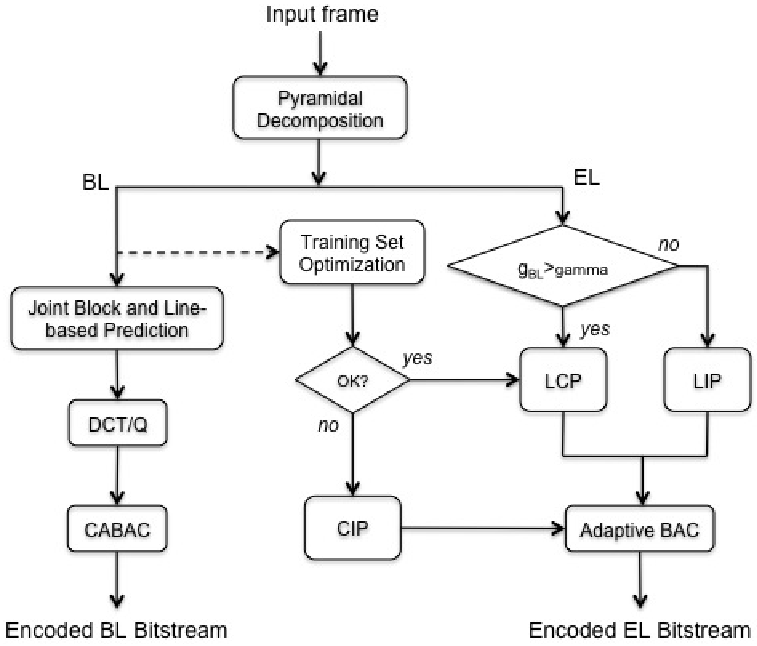


Figure 5.9: System architecture of proposed lossless context based HIP scheme.

To incorporate the context-based inter-layer prediction (CIP) scheme in the HIP scheme, we could use the classic rate-distortion optimization (RDO) technique to determine if the sample should use the LCP mode or the CILP mode. However, this would need header bits to indicate which mode is selected. Besides, a large amount of computation is needed for the RDO optimization. For the simplification purpose, we adopt the CIP only if the previously proposed training set optimization scheme cannot converge after a fixed number of iterations. The failure to converge is an indication that samples in the training set are too different from the target sample and, hence, cannot provide a good prediction. As a result, the prediction mode will switch to the CIP. The entire context-based hierachical intra prediction (CHIP) coding system is illustrated in Fig. 5.9. In the next section, we will examine the last building block in the compression system; namely, entropy coding.

5.4 Enhanced Context-based Adaptive Binary Arithematic Coding

Earlier coding standards such as MPEG-2, H.263 and MPEG-4 used the variable length coders (VLCs) as their entropy coders. The context adaptive binary arithmetic coder (CABAC) was adopted by H.264/AVC [13], which provides a bit-rate saving of 10-20% as compared with the context based VLCs [24]. The design of CABAC consists of four key elements: 1) binarization, 2) context modeling, 3) binary arithmetic coding and 4) probability estimation as shown in Fig. 5.10. It achieves good performance by selecting probability models according to the context. In addition, it allows probability estimation to be adaptive to the local statistics. However, CABAC was initially designed for DCT transformed coefficients after quantization in a lossy low-bit-rate video coding environment. In the current context of lossless coding, the prediction residuals have a very different characteristics since the DCT and quantization process are bypassed. In this section, we examine the unique characteristics of residual signals produced by the proposed CHIP system and propose several enhancements for further bit rate saving.

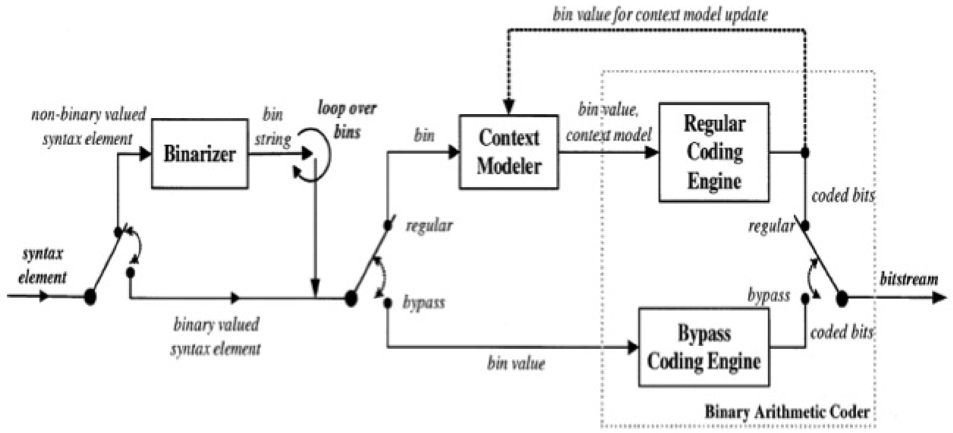


Figure 5.10: The flow chart of the CABAC.

5.4.1 Pre-processing of DCT Coefficients

The CABAC adopts several optimization techniques to facilitate the coding of a binary symbol set with a highly skewed probability distribution. After the quantization of DCT coefficients, for each block that has at least one nonzero coefficient, a sequence of binary maps were introduced; namely, the *Sign* map, the *Significance flag* map and the *last flag* map. The *Sign* map extracts the negative sign from each coefficient and leave only the absolute levels of coefficients. This is designed to further explore the redundancy between coefficients if their levels are the same. The *Significance flag* is used to indicate the position of each nonzero coefficient on the scanning path. If the *Significance flag* for that coefficient location is equal to 1, an additional *last flag* map is generated to signal to the decoder that if it is the position of the last nonzero coefficient in the scanning path. If the *last flag* indicates that the current residual sample value is indeed the last one in a block, the *Significance flag* coding will be terminated. The detailed process is shown in Table 5.1 for an DCT transformed and quantized coefficients block.

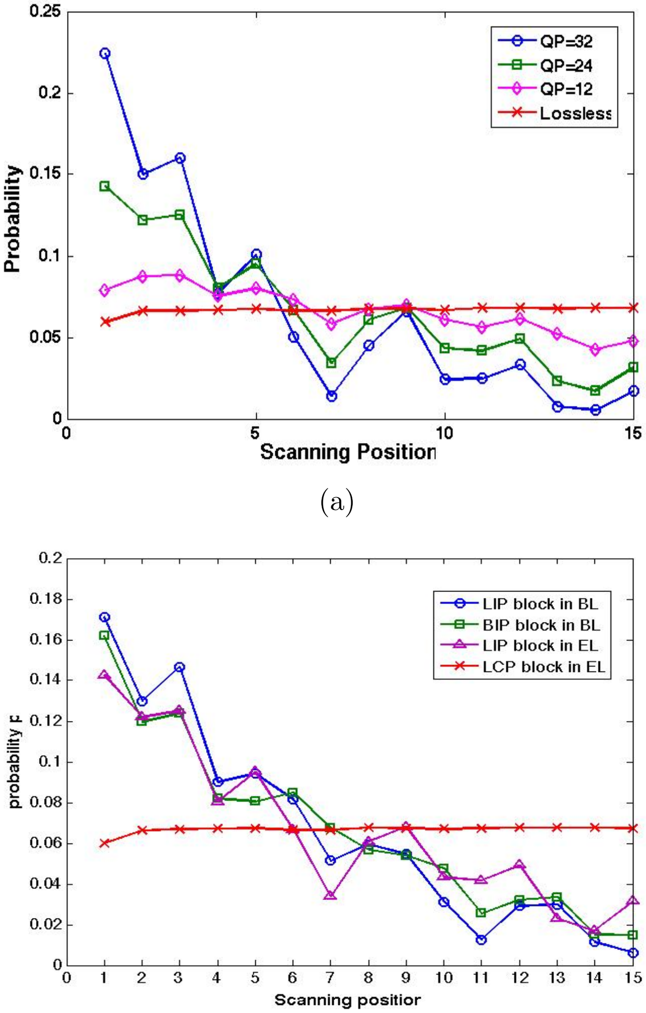
Table 5.1: Statistics of a DCT transformed and quantization block with $QP = 24$.

(a)																
Scanning Position	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Coefficient	8	-2	0	1	-2	1	1	0	0	0	0	0	0	0	0	0

(b)																
Scanning Position	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Abs level	8	2	0	1	2	1	1	0	0	0	0	0	0	0	0	0
Sign map	0	1	0	0	1	0	0									
Significance flag	1	1	0	1	1	1	1									
Last flag	0	0		0	0	0	1									

As mentioned before, this pre-processing works well if the block contains very few nonzero coefficients such as the case in Table. 5.1. In Fig. 5.11 (a), we show an example of the normalized probability distribution of nonzero coefficients in each scanning position [55]. Note that the probability of the scanning position having a nonzero coefficient decreases significantly when the scanning position is higher than 5 for the low-bit-rate coding case (say, $QP = 32$). That means for a large percentage of residual blocks, only 5 or less nonzero coefficients need to be encoded. However, as the quantization step-size becomes smaller, (say $QP < 12$), we notice that the probability distribution of nonzero coefficients is close to the uniform distribution. Consequently, the efficiency of the CABAC degrades for the high-bit-rate or lossless coding applications.

Similarly, we can analyze the prediction residual of the lossless CHIP scheme by plotting the probability distribution of nonzero coefficients for the LIP modes (for the BL) and the LCP/CIP modes (for the EL) in Fig. 5.11(b). The distribution of the BL blocks has a more skewered distribution since the BL is encoded with a lossy scheme in the CHIP. The LIP blocks in the EL has more zero coefficients than the LCP/CIP blocks in the EL since only homogeneous blocks are selected for the LIP and their residuals are usually small.



(a)

(b)

Figure 5.11: The probability of a zero coefficient along the scanning path for a frame of the Mobile sequence with resolution of 352x288: (a) quantized DCT coefficients and (b) the prediction residual of the CHIP scheme.

By examining an individual residual block after the LCP/CIP prediction as given in Table 5.2(a), we see that almost all scanning positions have nonzero coefficients, which is very different from the statistics in Table 5.1(a). The direct result of this type of blocks is shown in Table 5.2(b). Note that many 1's appear in the *significance* map to indicate the nonzero coefficient position and the *Last flag* map is filled almost to the end with the last nonzero coefficient being located at position 14.

Table 5.2: Statistics of a residual block after lossless LCP/CIP prediction

(a)																
Scanning Position	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Coefficient	9	-5	1	4	-3	7	2	0	-5	0	3	-2	1	0	-2	0

(b)																
Scanning Position	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Abs level	9	5	1	4	3	7	2	0	5	0	3	2	1	0	2	0
Sign map	0	1	0	0	1	0	0	0	1	0	0	1	0	0	1	0
Significance flag	1	1	0	1	1	1	1	0	1	0	1	1	1	0	1	0
Last flag	0	0		0	0	0	0		0		0	0	0		1	

In the proposed lossless CHIP scheme, LIP modes are most effective for relatively homogeneous blocks in the EL. Thus, the prediction residual of these blocks will be relatively low. In other words, the residual block after transform and quantization is likely to have a much skewered zero probability distribution and, therefore, the use of the *Significance flag* map and *last flag* map is helpful for further data compression for this type of blocks. However, if the LCP/CIP prediction is invoked for the target block, this block may have heterogeneous properties and the prediction might generate larger residuals. Hence, the addition of the *Significance flag* map and *Last flag* map may incur code expansion [40].

This observation motivates us to use the pre-processing step selectively based on the inferred block characteristics. Thus, we have the following rule:

$$\text{Last Flag Map, Significance Flag Map} \;=\; \begin{cases} \textit{null} & \text{if LCP=1 or CIP=1,} \\ \textit{inuse} & \text{otherwise.} \end{cases} \tag{5.14}$$

As the mode information for each block can be determined based on the statistic of the BL block, no additional bits are needed to signal the selection result.

5.4.2 Enhanced Context Modeling

Context modeling is one of the key elements in CABAC. It identifies a probability model for the syntax of binary sequences to be coded. As the accuracy of the model impacts the final coding efficiency, it is of critical importance that the model is adaptive and the actual data statistics can be well captured. In CABAC, the context index is expressed in form of

$$\gamma = \Gamma_s + \Delta_s(ctx\_cat) + \chi_s, \tag{5.15}$$

where Γ_s is the context index offset defined as the initialized value (lower bound) of the context range of a syntax element S, Δ_s refers to the context category and χ_s is the context index increment for a given syntax element S. The context index offset and the context category offset are determined by the corresponding syntax element, the block type, etc. Context indices are pre-computed and mapped to a state. Intuitively, the larger the context index, the lower the probability of the least probable symbol (LPS).

Since CABAC was initially designed for low bit rate coding, the context index increment method reflects the statistics of the corresponding residual data as well. Therefore, the coding engine considers that the probability of having a nonzero coefficient reduces as the the scanning position SP increases. Hence, the context index increment is implemented by setting $\chi_s = SP$. For the *Coefficient Level*, the level of significant coefficients also depends on the scanning position and the reverse scanning of the level information is adopted. As the reverse scanning path approaches the end, the occurrence of successive significant coefficients is likely to be observed. In addition, the absolute level will become larger as the scanning position decreases. Consequently, the context index increment for the *Coefficient Level* can be determined using the accumulated encoded trailing 1 and the accumulated encoded levels with absolute values greater than one.

In the CHIP system, we proposed several techniques to save the bits for the EL mode coding. Therefore, we only have to focus on the residual coding for the EL. However, the statistics of the residual sample values obtained from LCP/CIP prediction in lossless coding are different from the statistics of the transformed and quantized coefficients obtained in lossy coding. Actually, the probability of prediction residual from LCP/CIP is close to the uniform distribution in all scanning positions. To accommodate the difference, we have made some adjustment to the context index:

$$\gamma = \begin{cases} \Gamma_s + \Delta_s(ctx\_cat) & \text{if LCP=1 or CIP=1,} \\ \Gamma_s + \Delta_s(ctx\_cat) + SP & \text{otherwise.} \end{cases} \tag{5.16}$$

We no longer consider the impact of the scanning position on the context index increment, which offers a more accurate probability model for the residual data in the CHIP system.

5.5 Experimental Results

In this section, experiments were conducted to compare the performance of the following coding schemes: lossless H.264/AVC intra prediction (H.264 LS) [1], JPEG-LS, the lossless hierarchical intra prediction (HIP-LS) presented in Chapter 4 and the lossless context based hierarchical intra prediction (CHIP-LS) proposed in this chapter.

For the HIP-LS and the CHIP-LS, we adopted the two-layered structure with the BL of a quarter size of the EL with quantization parameter QP=20. The enhancement layer is coded losslessly to achieve the ultimate lossless coding goal with spatial/quality scalability. The joint block and line-based intra prediction proposed in [11] was chosen as Predictor 1 for the BL prediction, where the options of 2D geometrical manipulations and 2x2 block enhancement were turned off to reduce the coding complexity. For predictor 2 in the EL, we choose between H.264-LS and the linear prediction with the detected edge information. We consider a casual neighbor of size $N = 3$ and a training set of size $M = 6$. For the CHIP-LS, the k-means clustering process adopts $k = 512$ contexts. The iteration in the training set optimization is set to 6. The CABAC is used for the BL coding while the enhanced CABAC presented in the last section is adopted for the EL coding.

Eight HD YUV sequences of two different resolutions were used in the experiments. They were: Sheriff (@1280x720), Sailman (@1280x720), Harbor (@1280x720), Preakness

Table 5.3: Coding efficiency comparison between H.264-LS, JPEG-LS, HIP-LS and CHIP-LS.

Resolution @ 1280x720			
Sequence	Method	Compression Ratio (%)	Bit rate saving w.r.t H.264-LS(%)
Sheriff	H.264-LS	50.64	0
	JPEG-LS	47.75	5.69
	HIP-LS	45.81	9.54
	CHIP-LS	42.98	15.12
Sailorman	H.264-LS	63.31	0
	JPEG-LS	60.64	4.22
	HIP-LS	55.17	12.85
	CHIP-LS	51.64	18.43
Harbor	H.264-LS	59.85	0
	JPEG-LS	55.97	6.49
	HIP-LS	52.76	11.84
	CHIP-LS	49.73	16.91
Preakness	H.264-LS	84.03	0
	JPEG-LS	80.32	4.42
	HIP-LS	76.19	9.33
	CHIP-LS	70.57	16.02
Average	H.264-LS	**64.46**	**0**
	JPEG-LS	**61.17**	**5.20**
	HIP-LS	**57.48**	**10.89**
	CHIP-LS	**53.73**	**16.62**

(@1280x720), Rush Hour (@1920x1080), Blue sky (@1920x1080), Sunflower (@1920x1080) and Vintage car (@1920x1080). All values were averaged over 15 pictures of equi-spaced frame indices of respective sequences. The compression ratio is defined as

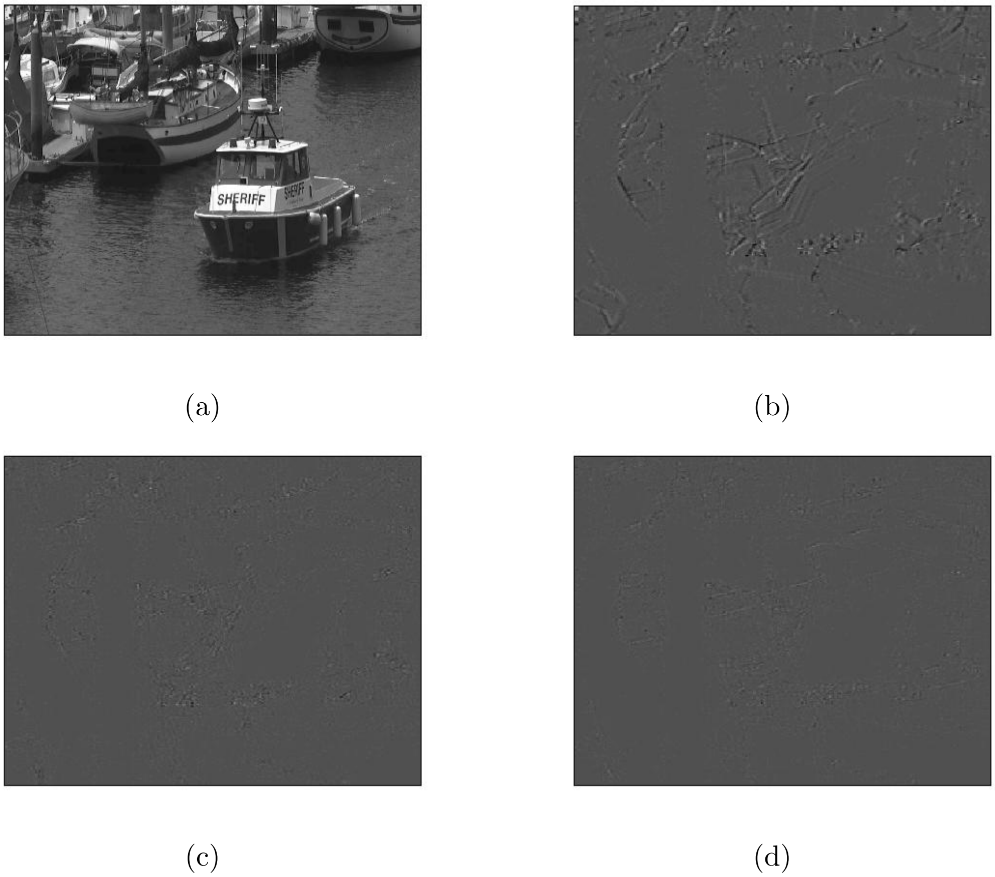
$$R = \frac{\text{output file size}}{\text{original file size}}.$$

The results are shown in Table 5.3 and Table 5.4. We see that the CHIP-LS achieves a bit rate saving of 18-22% with respect to H.264-LS and JPEG-LS, which are based on

Table 5.4: Coding efficiency comparison between H.264-LS, JPEG-LS, HIP-LS and CHIP-LS.

Resolution @ 1920x1080			
Sequence	Method	Compression Ratio (%)	Bit rate saving w.r.t H.264-LS(%)
Rush Hour	H.264-LS	43.37	0
	JPEG-LS	42.42	2.20
	HIP-LS	38.91	10.28
	CHIP-LS	36.15	16.63
Blue Sky	H.264-LS	46.37	0
	JPEG-LS	44.57	3.88
	HIP-LS	41.92	9.65
	CHIP-LS	38.88	16.15
Sunflower	H.264-LS	39.95	0
	JPEG-LS	38.01	4.86
	HIP-LS	35.67	10.72
	CHIP-LS	33.55	16.02
Vintage Car	H.264-LS	71.26	0
	JPEG-LS	67.14	5.78
	HIP-LS	63.32	11.14
	CHIP-LS	58.55	17.83
Average	H.264-LS	**50.24**	**0**
	JPEG-LS	**47.78**	**4.18**
	HIP-LS	**44.95**	**10.45**
	CHIP-LS	**41.78**	**16.66**

the single layer prediction. The CHIP-LS also outperforms the HIP-LS by 11-14% in the bit rate saving.

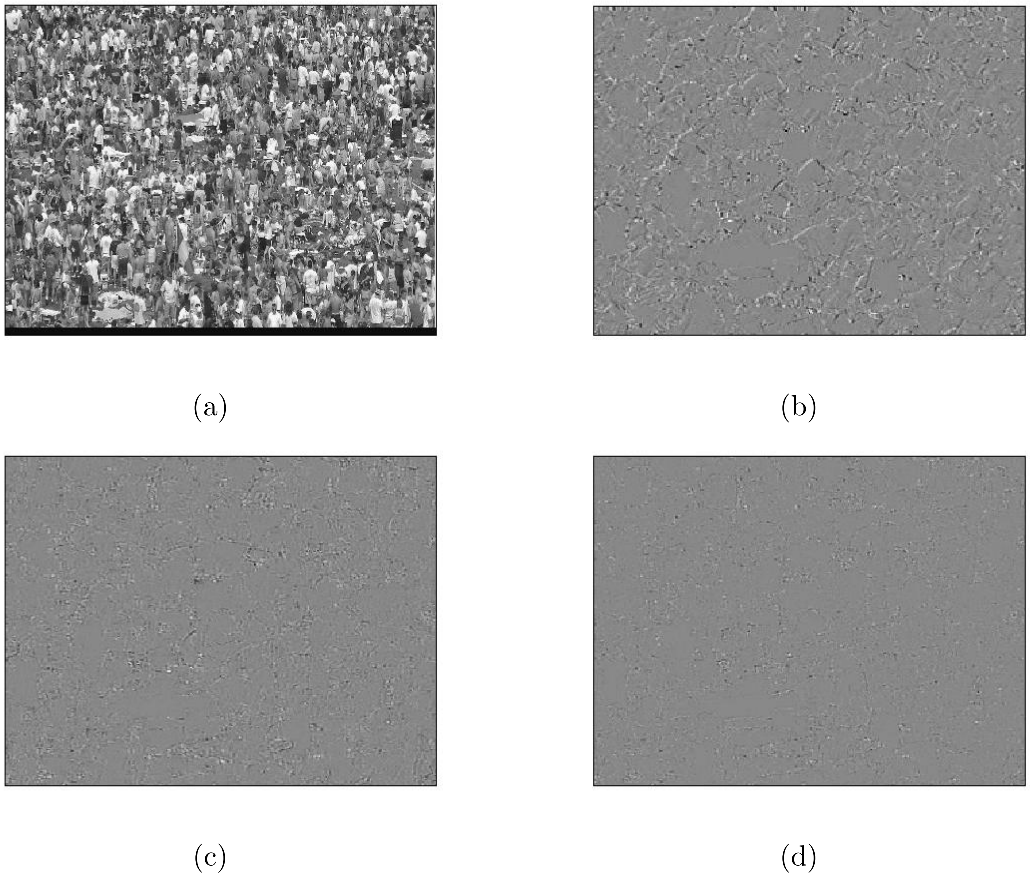


(a) (b)

(c) (d)

Figure 5.12: (a) One frame of the Sheriff sequence and the enlarged prediction residuals using prediction schemes in (b) the H.264-LS, (c) the HIP-LS and (d) the CHIP-LS.

It is worthwhile to point out that the CHIP-LS scheme offers smaller prediction residuals for contents with complex textures; e.g., the Preakness and Vintage Car sequences. In Figs. 5.12 and 5.13, we show the side-by-side comparison of prediction residuals generated by each prediction scheme. It is apparent that the proposed CHIP scheme has the lowest prediction error for each input frame consistently. In general, the proposed



(a) (b)

(c) (d)

Figure 5.13: (a) One frame of the Preakness sequence and the enlarged prediction residuals using prediction schemes in (b) the H.264-LS, (c) the HIP-LS and (d) the CHIP-LS.

CHIP-LS scheme provide the best coding performance with respect to the other three benchmarking schemes.

5.6 Conclusion

A novel lossless context-based hierarchical intra prediction scheme based on the image pyramidal decomposition was proposed and called the CHIP-LS scheme. The context of the BL is used to improve the prediction and coding for the EL in the CHIP-LS with spatial/quality scalability. It was demonstrated by experimental results that the proposed CHIP-LS scheme outperforms other three state-of-the-art lossless image coding schemes (*i.e.* JPEG-LS, H.264-LS and HIP-LS) by a significant margin in terms of bit rate saving.

Chapter 6

Conclusion and Future Work

6.1 Summary of the Research

In this research, we studied advanced intra prediction techniques for more effective Intra frame coding in the contexts of high-bit-rate and lossless image/video coding. The proposed schemes include: the joint block/line-based intra prediction (JBLIP) scheme, the hierarchical intra prediction (HIP) scheme and the context-based hierarchical intra prediction (CHIP) scheme. These techniques can be applied to the high definition digital cinema and digital camcording. For high definition digital camcording, the large image size together with a high frame rate puts an enormous burden on the encoder. In such a case, inter-frame prediction may not be desirable due to its high memory and complexity requirements, and an effective all-intra-frame coding scheme can offer a good alternative.

The JBLIP scheme was proposed in Chapter 3 for high-bit-rate coding. First, we conducted an in-depth analysis on the line-based intra prediction (LIP) schemes to reveal their inefficiency in prediction surfaces with complex textures. Then, we presented the BIP scheme with 2D geometrical manipulation to address this issue. Furthermore, we proposed additional enhancement to the BIP scheme and developed a rate-distortion optimization scheme (JBLIP) that jointly selects the best mode among the LIP and the BIP modes. A block classification method was proposed to identify blocks that can be well predicted by the LIP alone. To reduce the coding complexity, a zone-based fast search method was proposed. A translational vector prediction scheme was used to first establish a near optimal search position. A candidate block pruning method was further employed to achieve fast convergence to the optimal position. The proposed JBLIP scheme can achieve a significant coding gain of up to 1.68dB in PSNR with the same bit rate for images with various resolutions. The fast search scheme can obtain an average 10 times speed-up as compared to the full search with very little performance degradation.

The HIP scheme was presented in Chapter 4 for lossless compression. We analyzed the problem associated with intra prediction in H.264/SVC and then proposed a different approach to achieve the same scalability with frequency decomposition in the spatial domain. Specifically, we applied different predictors to different layers. For the enhancement layer, we proposed a block-based linear combination prediction to effectively estimate edges, which yields a small prediction error in high textured regions without sending the side/mode information to the decoder. The proposed HIP scheme cuts down the overhead cost significantly. The proposed lossless HIP scheme can achieve a 10% better compression efficiency as compared with the intra prediction schemes in H.264/AVC LS and JPEG-LS.

The CHIP scheme was presented in Chapter 5 to improve the coding performance of the HIP scheme furthermore. We analyzed the coding inefficiency associated with the HIP scheme. A mode estimation scheme was derived based on the block statistics of the

BL to achieve the modeless prediction. A PCA-based feature extraction method with the k-means clustering algorithm was first applied. Then, the context-based interlayer prediction scheme (CIP) was proposed to select the best prediction from the sample context. The CIP scheme determined the predicted value without the help of any side information. To improve the coding efficiency of the entropy coder, an enhanced precoding process for CABAC was introduced by exploiting the characteristics of the LCP/CIP prediction residuals. A enhanced context model was proposed to estimate the context more accurately. The proposed lossless CHIP scheme outperforms the H.264/AVC LS and the JPEG-LS schemes by a bit rate saving of 16%.

6.2 Future Work

Several possible extensions of the current research are stated below.

- Advanced interpolation scheme

  In both the HIP and the CHIP schemes, the best prediction mode is obtained from the statistics gathered from the BL so that the side information can be saved. A simple interpolation scheme was adopted to relate the BL and the EL in the current work. An advanced interpolation scheme may reduce the prediction error and result in a better mode selection scheme for better coding performance.

- Advanced feature extraction and clustering algorithm

  In the CIP scheme, the efficiency and adaptivity of feature extraction and clustering play an important role in finding a good prediction sample. It is desirable to identify the best features to be extracted and study whether the clustering algorithm can be further improved.

- Enhanced entropy coding scheme

  Since CABAC was initially designed for low-bit-rate coding, this technique has been finetuned for the coding of DCT coefficients after coarse quantization. However, as the need for image/video compression is moving towards high-bit-rate and even lossless coding, we may explore other entropy coders to meet the new requirements.

Bibliography

[1] "H.264/AVC Reference software JM12.1," in *[Online] Available: http://iphome.hhi.de/suehring/tml/*, July 2005.

[2] "HD Photo Specification V1.0," in *[Online] Available: www.microsoft.com_whdc_xps_wmphoto.mspx*, November 2006.

[3] A.M.Tourapis, O.C.Au, and M.L.Liu, "Fast motion estimation using modified circular zonal search," *IEEE Intl. Symposium on Circuits and Systems*, vol. 4, pp. 231–234, July 1999.

[4] J. Balle and M.Wien, "Extended texture prediction for H.264 intra coding," in *Joint Video Team (JVT) of ISO/IEC MPEG & ITU-T VCEG (ISO/IEC JTC1/SC29/WG11 and ITU-T SG16 SG16 Q.6) VCEG-AE11.doc*, January 2007.

[5] J. Bennett and A. Khotanzad, "Modeling textured images using generalized long-correlation models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-6, pp. 800–809, 1998.

[6] J. Bigun, G. H. Granlund, and J. Wiklund, "Multidimensional orientation estimation with applications to texture analysis and optical flow," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 15, pp. 775–791, August 1991.

[7] Y.-S. Chung and M. Kanefsky, "On 2-D recursive LMS algorithms using ARMA prediction for ADPCM encoding of images," *IEEE Transactions on Image Processing*, vol. 1, pp. 416–422, 1992.

[8] C.Zhu, X.Lin, and L.Chau, "Hexagon-base search pattern for fast block motion estimation," *IEEE Trans. on Circuits and Video Technologies*, vol. 12, January 2007.

[9] D.-K.Kwon, M.-Y. Shen, and C.-C. J. Kuo, "Rate control for h.264 video with enhanced rate and distortion models," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 17, pp. 517–529, May 2007.

[10] Y. Dai, Q. Zhang, and C.Kuo, "2d enhanced intra prediction (2DEIP) for high definition image/video coding," *IEEE Intl Symposium on Circuits and Systems*, May 2009.

[11] Y. Dai, Q. Zhang, A. Tourapis, and C.-C.J.Kuo, "Efficient block-based intra prediction for image coding with 2D geometrical manipulations," *IEEE Intl Conf on Image Processing*, October 2008.

[12] E. F. Deprettere, "SVD and signal processing: algorithms, applications and architectures," in *Elsevier Science Publishing Co.*, 1988.

[13] D.Marpe, H. Schwarz, and T. Wiegand, "Context-based adaptive binary arithmetic coding in the H.264/AVC video compression," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 13, pp. 620–636, July 2003.

[14] T. Eude, R. Grisel, H. Cherifi, and R. Debrie, "On the distribution of the DCT coefficients," *Proc. 1994 IEEE Int. Conf. Acoustics, Speech, Signal Processing*, vol. 5, pp. 365–368, April 1994.

[15] M. Flierl and T. Wiegand, "A video codec incorporating block-based multi-hypothesis motion compensated prediction," *Proc. of SPIE Conf. on Visual Communications and Image Processing*, June 2000.

[16] G.Bjontegaard, "Calculation of average PSNR difference between RD-curves," in *Joint Video Team (JVT) of ISO/IEC MPEG & ITU-T VCEG (ISO/IEC JTC1/SC29/WG11 and ITU-T SG16 SG16 Q.6) VCEG-M33.doc*, April 2001.

[17] G.Conklin, "New intra prediction modes," in *Joint Video Team (JVT) of ISO/IEC MPEG & ITU-T VCEG (ISO/IEC JTC1/SC29/WG11 and ITU-T SG16 SG16 Q.6) VCEG-N54.doc*, September 2001.

[18] B. Girod, "Efficiency analysis of multihypothesis motion-compensated prediction for video coding," *IEEE Trans. on Image Processing*, vol. 9, pp. 173–183, February 2000.

[19] H. Hurst, "Long-term storage capacity of reservoirs," *Trans. Amer. Soc. Civil Eng.*, vol. 116, pp. 770–799, 1951.

[20] N. S. Jayant and P. Noll, "Digital coding of waveforms principles and applications to speech and video," pp. 16–17, 1984.

[21] J.Chalidabhongse and C.-C. J. Kuo, "Fast motion vector estimation using multiresolution-spatio-temporal correlations," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 7, pp. 477–488, 1997.

[22] R. Kashyap and A. Khotanzad, "Synthesis and estimation of random fields using long-correlation models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-6, pp. 800–809, July 1984.

[23] L.-J. Kau and Y.-P. Lin, "Adaptive lossless image coding using least squares optimization with edge-look-ahead," *IEEE Trans. on Circuits and Systems-II: Express briefs*, vol. 52, pp. 751–755, November 2005.

[24] S.-H. Kim and Y.-S. Ho, "Improved CABAC for H.264 lossless intra coding," *IEEE Trans. on Image Processing*, 2009.

[25] S. Kondo, H. Sasai, and S. Kadono, "Tree structured hybrid intra prediction," October 2004, pp. 473–476.

[26] E. Y. Lam and J. W. Goodman, “A mathematical analysis of the DCT coefficient distributions for images,” *IEEE Trans. on Image Processing*, vol. 9, pp. 1661–1666, October 2000.

[27] Y.-L. Lee, K. H. Han, and G. J. Sullivan, “Improved lossless intra prediction for H.264/MPEG-4 AVC,” *IEEE Trans. on Image Processing*, vol. 15, pp. 2610–2615, September 2006.

[28] X. Li and M. T. Orchard, “Edge-directed prediction for lossless compression of natual images,” *IEEE Trans. on Image Processing*, vol. 10, pp. 813–817, June 2001.

[29] S. P. Lloyd, “Least squares quantization in PCM,” *IEEE Trans. on Information Theory*, vol. 28, pp. 129–137, 1982.

[30] N. Memon and X. Wu, “Recent developments in context-based predictive techniques for lossless image compression,” *The Computer Journal*, vol. 40, pp. 127–135, 1997.

[31] J.-R. Ohm, “Advances in scalable video coding,” *IEEE Proceedings*, vol. 93, pp. 42–56, January 2005.

[32] S.-W. Park, D. H. Yoon, J.-H. Park, and B.-M. Jeon, “Intra-bl prediction considering phase shift,” in *Joint Video Team (JVT) of ISO/IEC MPEG & ITU-T VCEG JVT-o023.doc*, April 2005.

[33] W. K. Pratt, “Digital image processing,” 1978.

[34] P.Topiwala, “Comparative study of JPEG2000 and H.264/AVC FRExt I-frame coding on high definition video sequences,” *Proceedings of SPIE, Optical Information Systems III*, September 2005.

[35] R. Reininger and J. Gibson, “Distributions of the two-dimensional DCT coefficients for images,” *IEEE Trans. on Communication*, vol. COM-31, pp. 835–839, June 1983.

[36] R.Li, B.Zeng, and M.L.Liou, “A new three-step search algorithm for block motion estimation,” *IEEE Trans. on Circuits and Video Technologies*, vol. 4, August 1994.

[37] S.A.Martucci, “Reversible compression of HDTV images using median adaptive prediction and arithmetic coding,” vol. 2, pp. 1310–1313, April 1990.

[38] H. Schwarz, D. Marpe, and T. Wiegand, “Overview of the scalable video coding extension of the H.264/AVC standard,” *IEEE Trans. on Circuits and Systems*, vol. 17, pp. 1103–1120, September 2007.

[39] A. Segall and S. Lei, “Adaptive upsampling for spatially scalable coding,” in *Joint Video Team (JVT) of ISO/IEC MPEG & ITU-T VCEG JVT-o010.doc*, April 2005.

[40] S.H.Kim and Y.-S. Ho., “Improved CABAC for H.264 lossless intra coding,” *IEEE Trans. on Image Processing.*

[41] S. R. Smoot and L. A. Rowe, “Study of DCT coefficient distributions,” *Proc. SPIE*, pp. 403–411, January 1996.

[42] G. J. Sullivan, “Multi-hypothesis motion compensation for low bit-rate video coding,” *Proc. of IEEE Intl. Conf. on Acoustics, Speech and Signal Processing*, vol. 5, pp. 437–440, April 2000.

[43] G. J. Sullivan, P. Topiwala, and A. Luthra, “The H.264/AVC advanced video coding standard: overview and introduction to the fidelity range extensions,” vol. 5558, pp. 454–474, August 2004.

[44] G. J. Sullivan and T. Wiegand, “Rate-distortion optimization for video compression,” *IEEE Signal Processing Magazine*, vol. 15, pp. 74–90, November 1998.

[45] S.Yu and C.Chrysafis, “New intra prediction using intra-macroblock motion compensation,” in *Joint Video Team (JVT) of ISO/IEC MPEG & ITU-T VCEG (ISO/IEC JTC1/SC29/WG11 and ITU-T SG16 SG16 Q.6) JVT-c151r1.doc*, April 2004.

[46] S. Takamura and M. Takagi, “A hybrid lossless compression of still images using Markov model and linear prediction,” vol. 974, pp. 203–208, 19.

[47] T.K.Tan, C.S.Boon, and Y.Suzuki, “Intra prediction by template matching,” *IEEE Intl Conf on Image Processing*, October 2006.

[48] T.Shiodera, A.Tanizawa, and T.Chujoh, “Bidirectional intra prediction,” in *Joint Video Team (JVT) of ISO/IEC MPEG & ITU-T VCEG (ISO/IEC JTC1/SC29/WG11 and ITU-T SG16 SG16 Q.6) VCEG-AE14.doc*, January 2007.

[49] M. Weinberger, G. Seroussi, and G.Sapiro, “The LOCO-I lossless image compression algorithm: priniciples and standarization into JPEG-LS,” *IEEE Trans. on Image Processing*, May 2000.

[50] Z. Wong, J. Liu, Y. Tan, and J. Tian, “Inter layer intra prediction using lower layer information for spatial scalability,” vol. 345/2006, pp. 303–311, October 2006.

[51] X. .Wu and K. Barthel, “Piesewise 2D autoregression for predictive image coding,” *Proc. Int. Conf. Image Processing*, vol. 3, pp. 901–905, October 1998.

[52] X. Wu and N. Memon, “Context-based, adaptive, lossless image coding,” *IEEE Trans. on Communications*, Apirl 1997.

[53] H. Ye, G. Deng, and J. C. Devlin, “Least squares approach for lossless image coding,” *Fifth Intl. Symposium on Signal Processing and its Applications*, pp. 63–67, August 1999.

[54] G. S. Yovanof and S. Liu, “Statistical analysis of the DCT coefficients and their quantization error,” *Conf. Rec. 30th Asilomar Conf. Signals, Systems, Computers*, vol. 1, pp. 601–605, 1997.

[55] Q. Zhang, S.H.Kim, Y. Dai, and C.-C. J. Kuo, “Multi-order-residual (MOR) video coding: framework, analysis and performance,” *SPIE Visual Communications and Image Processing*, July 2010.