New Rate Distortion Bounds for Natural Videos Based on a Texture Dependent Correlation Model

Jing Hu
Digital Signal Processing Group
Cisco Systems
jinghu@cisco.com

Jerry D. Gibson
Department of Electrical and Computer Engineering
University of California, Santa Barbara
gibson@ece.ucsb.edu

***Abstract*—We revisit the classic problem of developing a spatial correlation model for natural images and videos by proposing a conditional correlation model for relatively nearby pixels that is dependent upon five parameters. The conditioning is on local texture and the optimal parameters can be calculated for a specific image or video with a mean absolute error (MAE) usually smaller than 5%. We use this conditional correlation model to calculate the conditional rate distortion function when universal side information on local texture is available at both the encoder and the decoder. We demonstrate that this side information, when available, can save as much as 1 bit per pixel for selected videos at low distortions. We further study the scenario when the video frame is processed in macroblocks (MBs) or smaller blocks and calculate the rate distortion bound when the texture information is coded losslessly and optimal predictive coding is utilized to partially incorporate the correlation between the neighboring MBs or blocks. These rate distortion bounds are compared to the operational rate distortion functions generated in intra-frame coding using the AVC/H.264 video coding standard.**

I. Introduction

Parsimonious statistical models of natural images and videos can be used to calculate the rate distortion functions of these sources as well as to optimize particular image and video compression methods. Although they have been studied extensively, the statistical models and their corresponding rate distortion theories are falling behind the fast advancing image and video compression schemes.

The research on statistically modeling the pixel values within one image goes back to the 1970s when two correlation functions were studied [1], [2]. Both correlation functions assume a Gaussian distribution of zero mean and a constant variance for the pixel values and treat the correlation between two pixels within an image as dependent only on their spatial offsets. These two correlation models for natural images were effective in providing insights into image coding and analysis. However they are so simple that, as shown later in this paper, the rate distortion bounds calculated based on them are actually much higher than the operational rate distortion curves of the current intra-frame video coding schemes. For the same reason, more recent rate distortion theory work on video coding such as [3], [4] that adopt these two spatial correlation models have limited applicability.

Due to the difficulty of modeling the correlation among the pixel values in natural image and video sources, studying their rate distortion bounds is often considered infeasible [5]. As a result, in the past two decades, the emphasis of rate distortion analysis has been on setting up operational models for practical image/video compression systems to realize rate control [6]–[12] and to implement quality optimization algorithms [5], [13]–[16]. For example, a very popular such model treats the discrete cosine transform (DCT) coefficients in the predicted frames of a video sequence as uncorrelated Laplacian random variables [17], [18] so that the coding bit rate R and reconstruction distortion D can be expressed as simple functions of the quantization parameter q. Other popular operational rate and distortion models include those proposed in [10]–[12], [15], [19]–[22] that do not consider packet loss over communication networks and those proposed in [16], [23]–[27] that do take into account possible packet loss over the networks. These operational rate and distortion models are derived for specific coding schemes, and therefore, they cannot be utilized to derive the rate distortion bound of videos.

In this paper we address the difficult task of modeling the correlation in video sources by proposing a new correlation model for two close pixels in one frame of digitized natural video sequences that is conditional on the local texture. This new correlation model is dependent upon five parameters whose optimal values are calculated for a specific image or video. The new correlation model is simple, but it performs very well, as strong agreement is discovered between the approximate correlation coefficients (as defined in Eq. (III.4)) and the correlation coefficients calculated by the new correlation model, with a mean absolute error (MAE) usually smaller than 5%.

With the new block-based local-texture-dependent correlation model, we first study the marginal rate distortion functions of the different local textures. These marginal rate distortion functions are shown to be quite distinct from each other. Classical results in information theory are utilized to derive the conditional rate distortion function when the universal side information of local textures is available at both the encoder

 This work was supported by the California Micro Program, Applied Signal Technology, Cisco, Inc., Dolby Labs, Inc., Marvell, Inc. and Qualcomm, Inc., by NSF Grant Nos. CCF-0429884 and CNS-0435527, and by the UC Discovery Grant Program and Nokia, Inc..

and the decoder. We demonstrate that by involving this side information, the lowest rate that is theoretically achievable in intra-frame video compression can be as much as 1 bit per pixel lower than that without the side information. This rate distortion bound with local texture information taken into account while making no assumptions on coding, is shown indeed to be a valid lower bound with respect to the operational rate distortion curves of intra-frame coding in AVC/H.264.

The incorporation of the new correlation model into existing models of practical image/video compression systems is also promising. We demonstrate this by studying the common "blocking" scheme used in most video compression standards [28]–[31], which divides a video frame into 16×16 macroblocks (MB) or smaller blocks before processing. With the block based nature of the new correlation model, we study the penalty paid in average rate when the correlation among the neighboring MBs or blocks is disregarded completely or is incorporated partially through predictive coding. A rate distortion bound is calculated for the scenario when the texture information is coded losslessly and optimal predictive coding is employed. This lower bound is shown to be reasonably tight with respect to the operational rate distortion curves of intra-frame coding in AVC/H.264. Furthermore, it is near linear in terms of average bit rate per pixel versus PSNR of a video frame and therefore can easily be utilized in future video codec designs.

The correlation model and the rate distortion bounds proposed in this paper only deal with the pixels within one frame of a video. The model needs to be expanded to modeling the correlation of the pixels that are located in different video frames. This is currently under investigation and recent results show promise when a single temporal correlation coefficient is introduced for every two frames [32]. This local texture dependent correlation model and its corresponding rate distortion bounds are a significant step toward obtaining rate distortion bounds for video compression, which has seen few new results in the last twenty years. In the meantime, the intra-frame coding modes in video compression and some applications that only use intra-coded frames, such as digital cinema and low frame rate surveillance cameras, can exploit these new results directly.

The remainder of this paper is organized as follows. In Section II we review the existing statistical models of natural images and videos, as well as the rate and distortion analysis of practical video compression systems in the literature. In Section III we propose the novel new correlation model based on local texture. In Section IV we study the marginal rate distortion bounds of different local textures and derive the theoretical rate distortion bound with the local texture as the side information. In Section V we derive the rate distortion bounds for the "blocking" scheme that is commonly used in video coding, with or without prediction across the blocks. These various rate distortion bounds are compared to the operational rate distortion curves of intra-frame coding in AVC/H.264 in Section VI. We conclude this paper and provide insights into future research in Section VII.

II. Existing statistical models

1) Statistical models of images and videos: The research on statistically modeling the pixel values within one image goes back to the 1970s when two correlation functions were studied. Both assume a Gaussian distribution of zero mean and a constant variance for the pixel values.

The first correlation model is

$$\rho(\Delta i, \Delta j) = e^{(-\alpha|\Delta i| - \beta|\Delta j|)}, \tag{II.1}$$

with Δi and Δj denoting offsets in horizontal and vertical coordinates. The parameters α and β control the correlation in the horizontal and vertical directions, respectively, and their values can be chosen for different images [1]. The separability in spatial coordinates in this correlation model facilitates the analysis of the two-dimensional rate distortion behavior of images using the one-dimensional Karhunen Lòeve transform (KLT).

The second correlation model is an isotropic function

$$\rho(\Delta i, \Delta j) = e^{-\alpha\sqrt{\Delta i^2 + \Delta j^2}}. \tag{II.2}$$

This model implies that the correlation between two pixels within an image depends only on the Euclidean distance between them [2]. The major advantage of this model is that it has a closed-form two-dimensional Fourier transform and therefore leads to a closed-form rate function and distortion function on a common parameter.

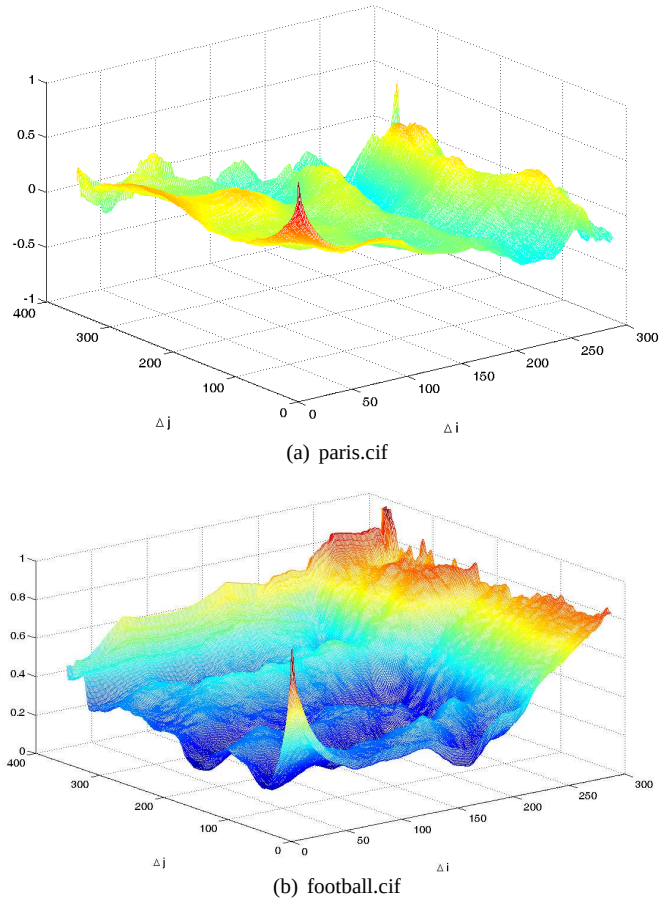
These two correlation models for natural images are simple yet effective in providing insights into image coding and analysis. However image and video coding schemes have advanced significantly and a rate distortion theory that is relevant to these more sophisticated methods is needed.

Let $X(i, j)$ denote the pixel value at the i^{th} row and the j^{th} column of a digitized image, and let M and N denote the numbers of rows and columns in the image. The approximate correlation coefficient $\hat{\rho}(\Delta i, \Delta j)$ of this image can be expressed as

$$\hat{\rho}(\Delta i, \Delta j) = \frac{\frac{\sum[X(i,j)X(i+\Delta i, j+\Delta j)]}{\sqrt{\sum[X^2(i,j)]\sum[X^2(i+\Delta i, j+\Delta j)]}}}{(M - \Delta i)(N - \Delta j)}, \tag{II.3}$$

for $0 \leq \Delta i \leq M - 1$, $0 \leq \Delta j \leq N - 1$. The summations in (II.3) are taken over all pixels whose coordinates satisfy $0 \leq i \leq M - 1 - \Delta i$, $0 \leq j \leq N - 1 - \Delta j$. Fig. 1 plots the approximate correlation coefficients $\hat{\rho}(\Delta i, \Delta j)$ of two digitized natural images, selected from two digitized natural video sequences, paris.cif and football.cif, respectively. We can see in Fig. 1 that when Δi and Δj are larger than 50, which is still much smaller than the image size we encounter in present applications, for example 352×288 in this figure, the approximate correlation coefficients $\hat{\rho}(\Delta i, \Delta j)$ are rather and neither of the two correlation functions can model this behavior. Correspondingly, the rate distortion analysis of natural images based on these two correlation functions will be inaccurate. This is confirmed later in this paper as the rate distortion bounds calculated based on these two correlation

functions are shown to be actually much higher than the operational rate distortion curves of the current intra-frame video coding schemes.



(a) paris.cif

(b) football.cif

Fig. 1. The approximate correlation coefficient $\hat{\rho}(\Delta i, \Delta j)$ of two digitized natural images

For the same reason, more recent rate distortion theory work for videos, such as [3], [4], [33] that adopt these two spatial correlation models, is limited in scope. For example, In [4], [33], distortion-rate performance is analyzed by deriving the power spectral density of the prediction error with respect to the probability density function of the displacement error. This is shown, however, to be incapable of describing, with sufficient accuracy, the measured distortion-rate performance of a typical video encoder [23].

2) Statistical models of practical video compression systems: Researchers working on video compression also have developed statistical models of images in the transformed domain. The most popular among them treats the discrete cosine transform (DCT) coefficients in the predicted frames of a video sequence as uncorrelated Laplacian random variables [17], [18]. If we use the absolute magnitude distortion measure $d(x, \hat{x}) = |x - \hat{x}|$, there is a closed form rate distortion function for the memoryless Laplacian source that can be expanded into a Taylor series and approximated by $R(D) \sim= aq^{-1} + bq^{-2}$. In this formula, the distortion is measured by the average quantization scale q used in the frame.

This quadratic rate distortion function is the foundation of the rate control schemes [6]–[8] that are adopted by the international video coding standards, such as ISO MPEG-2/4 [28], [29] and ITU-T H.263 [30]. In these rate control schemes, the quantization stepsizes, which are indexed by the quantization parameters (QPs), are chosen optimally based on the quadratic rate distortion function, number of bits left to consume and the approximate coding complexity. The bits spent coding the other syntax elements, considered to be mainly the motion vectors, are monitored and predicted through simple linear or nonlinear functions.

The memoryless Laplacian model for DCT coefficients becomes less appropriate, even for practical video compression system design purposes, since the emergence of new coding standards such as AVC/H.264. The new schemes and refinements in AVC/H.264 [34] reduce the applicability of the memoryless Laplacian model of the DCT coefficients for at least two reasons. First, with all the options offered in the codecs and the very small processed block sizes, the majority of the bandwidth is very likely to be allocated to transmit the coding parameters and the motion vectors of each block rather than the DCT coefficients, especially in the low to medium bit rate applications. Since the Laplacian model only treats the DCT coefficients, it becomes insufficient to represent the information in the video source. Second and more importantly, these coding options and parameters are to be chosen, in an optimal way if possible, before the DCT or DCT-like transforms can be applied to the residue block. This is considered as a rate distortion optimization problem and the most popular solution to this problem is to conduct the optimization with a fixed quantization parameter. However, from the perspective of rate control, the quantization parameter is to be optimally chosen based on the residue data after the rate distortion optimization is performed. Therefore there is a "chicken and egg" dilemma artificially caused by modeling the statistics in the transformed domain that has prevented a global optimum from being obtained, even for a specific codec [9], [12], [35].

Two recently proposed schemes following in the same vein [9], [35] try to tackle this dilemma by either engaging a "two pass scheme" or defining a "basic unit". This is an ongoing research direction and for more recent activities please refer to [12]. Another recent work on rate distortion modeling for H.264 [15] treats the residue blocks after intra/inter prediction in the spatial domain as Laplacian random vectors with separable correlation coefficients that depend only on one *a priori* parameter. The statistics in the spatial domain are then used to calculate rate distortion models in the transformed domain. Even though this work also studies the statistics in the spatial domain of videos, it relies on a very simple model of the residue block, and therefore does not address the interdependence between the rate control and rate distortion optimization.

In summary, a new statistical correlation model for digitized

natural videos is much needed in both theory and application. This correlation model should be independent of any coding schemes, rather than modeling the processed values, such as the DCT coefficients, in a coding scheme, so that the theoretical rate distortion bounds can be derived to predict the fundamental limit on the number of bits (per pixel) needed to represent a video at a given distortion level. This correlation model should also be more sophisticated than the old correlation models in Eqs. (II.1) and (II.2) so that the derived theoretical rate distortion bounds are valid. It will be a plus if this correlation model has a simple form with parameters that can be calculated for a specific video, which makes the incorporation of the correlation model into a practical video codec design and evaluation possible. In the next section we propose such a correlation model.

III. Definition of Block-Based Conditional Correlation Model

In this section we propose a new correlation model for a digitized natural image or an image frame in a digitized natural video. We assume that all pixel values within one natural image form a two dimensional Gaussian random vector with memory, and each pixel value is of zero mean and the same variance σ^2. From the discussion in Section II-1, we know that to study the correlation between two pixel values within one natural image, these two pixels should be located close to each other compared to the size of the image. Also for a sophisticated correlation model, the correlation between two pixel values should not only depend on the spatial offsets between these two pixels but also on the other pixels surrounding them.

Intra-frame prediction is a new feature in AVC/H.264 which removes, to a certain extent, the spatial redundancy in neighboring 4×4 blocks or 16×16 macroblocks (MBs). If a block or MB is encoded in intra-mode, a prediction block is formed based on previously encoded and reconstructed surrounding pixels. The prediction block P is then subtracted from the current block prior to encoding. For the luminance samples, P may be formed for a 4×4 block or for a 16×16 MB. There are a total of nine optional prediction modes for each 4×4 luminance block as shown in Fig. 2 and four optional prediction modes (mode 0 to 3 in Fig. 2) for each 16×16 luminance MB.

To quantify the effect of the surrounding pixels on the correlation between pixels of interest, we utilize the concept of local texture, which is simplified as local orientation, i.e., the axis along which the luminance values of all pixels in a local neighborhood have the minimum variance. The local texture is similar to the intra-prediction modes in AVC/H.264, but with a generalized block size and a arbitrary number of total textures. To calculate the local texture of a block, we also employ the pixels on the top and to the left of this block as surrounding pixels. However we use the original values of these surrounding pixels rather than the previously encoded and reconstructed values used in intra-frame prediction of AVC/H.264. The block can have any rectangular shape as long as its size is small compared to the size of the image. The local textures need not to be restricted to those defined in AVC/H.264. For example, in Fig. 3, the numbered arrows represent a few local textures that are defined as intra-prediction modes in AVC/H.264 and the unnumbered arrows represent a few local textures that are not defined as intra-prediction modes in AVC/H.264. Once the block size and the available local textures are fixed, the local texture of the current block is chosen as the one that minimizes the mean absolute error (MAE) between the original block and the prediction block constructed based on the surrounding pixels and the available local textures. It is important to point out that even through we choose a very simple and computationally inexpensive way to calculate the local texture, there are other, more sophisticated schemes of doing so, as summarized in [36], which should produce even better results in rate distortion modeling.

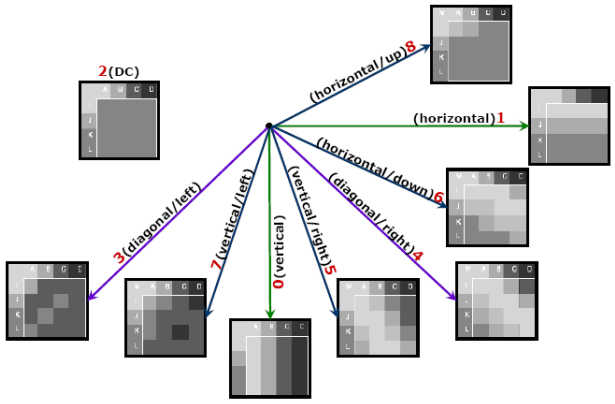


Fig. 2. The intra-prediction modes for 4×4 blocks in AVC/H.264 [34]

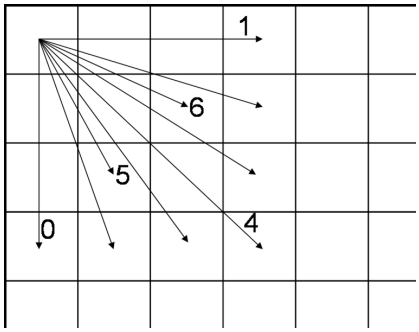


Fig. 3. The numbered arrows represent a few local textures that are defined as intra-prediction modes in AVC/H.264 and the unnumbered arrows represent a few local textures that are not defined as intra-prediction modes in AVC/H.264

The local texture reveals which one, out of the different available local textures, is the most similar to the texture of the current block. It is reasonable to conjecture that the difference in local texture also affects the correlation between two close pixels within one video frame. To confirm this we first calculate the approximate correlation coefficient between one block of size $M \times N$, and another nearby block of the same

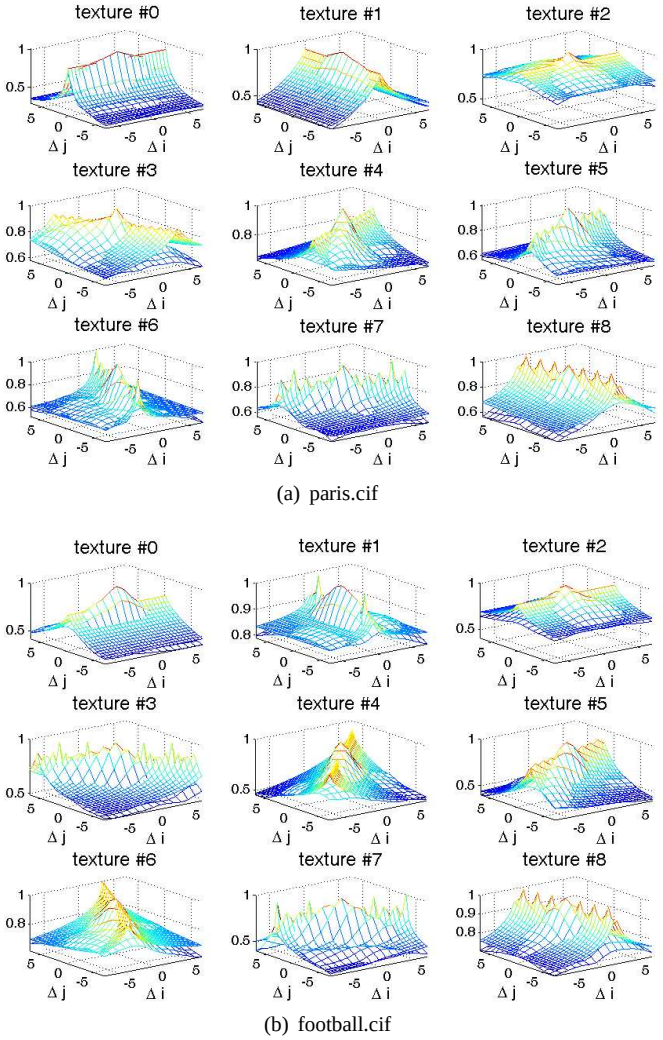
size, shifted by Δi vertically and Δj horizontally, according to the following formula

$$\hat{\rho}(\Delta i, \Delta j) = \frac{1}{MN} \frac{\sum[X(i,j)X(i+\Delta i, j+\Delta j)]}{\sqrt{\sum[X^2(i,j)]\sum[X^2(i+\Delta i, j+\Delta j)]}}, \tag{III.4}$$

for $-I \leq \Delta i \leq I$, $-J \leq \Delta j \leq J$. This formula is similar to Eq. (II.3), except that 1) $M \times N$ is not the size of a whole image, but the size of block, usually much smaller than the image size; 2) the ranges for Δi and Δj are different and need not be smaller than M and N. $\hat{\rho}(\Delta i, \Delta j)$ is first calculated for each $M \times N$ block in an image frame. Then they are averaged among the blocks that have the same local texture. We denote this average approximate correlation coefficient for each local texture as $\hat{\rho}(\Delta i, \Delta j|y)$ where y denotes the local texture.

In Figs. 4(a) and 4(b), we plot $\hat{\rho}(\Delta i, \Delta j|y)$ (shown in the figures as the loose surfaces, i.e., the mesh surfaces that look lighter with fewer data points) for the first frames from paris.cif and football.cif, respectively. The dense surfaces, i.e., the mesh surfaces that look darker with more data points, are the correlation coefficients calculated using the proposed conditional correlation model, which will be discussed later in this section. The block size is $M = N = 4$. The available nine local textures are chosen to be those plotted in Fig. 2. We set Δi and Δj to be very small, ranging from -7 to 7, to concentrate on the dependence of the statistics on local texture in an image frame. Fig. 4 shows that the average approximate correlation coefficient $\hat{\rho}(\Delta i, \Delta j|y)$ is very different for the blocks with different local textures. If we average $\hat{\rho}(\Delta i, \Delta j|y)$ across all the blocks in the picture, we get what is shown in Fig. 1 in the corresponding region of Δi and Δj, but the important information about the local texture is lost. Not surprisingly $\hat{\rho}(\Delta i, \Delta j|y)$ demonstrates certain shapes that agree with the orientation of the local textures. It is also interesting that although the average approximate correlation coefficients of the same local texture in both images demonstrate similar shapes their actual values are quite different.

Motivated by these observations, in the following we present the formal definition of the new correlation coefficient model that is dependent on the local texture.



(a) paris.cif

(b) football.cif

Fig. 4. The loose surfaces (the mesh surfaces that look lighter with less data points) are $\hat{\rho}(\Delta i, \Delta j|y)$, the approximate correlation coefficients of two pixel values in the first frame from paris.cif and football.cif respectively, averaged among the blocks that have the same local texture; the dense surfaces (the mesh surfaces that look darker with more data points) are $\rho(\Delta i, \Delta j|y)$, the correlation coefficients calculated using the proposed conditional correlation model, along with the optimal set of parameters

Definition 3.1: The correlation coefficient of two pixel values with spatial offsets Δi and Δj within a digitized natural image or an image frame in a digitized natural video is defined as

$$\rho(\Delta i, \Delta j|Y_1 = y_1, Y_2 = y_2) = \frac{\rho(\Delta i, \Delta j|y_1) + \rho(\Delta i, \Delta j|y_2)}{2}, \tag{III.5}$$

where

$$\rho(\Delta i, \Delta j|y) = a(y) + b(y)e^{-|\alpha(y)\Delta i + \beta(y)\Delta j|^{\gamma(y)}}. \tag{III.6}$$

Y_1 and Y_2 are the local textures of the blocks the two pixels are located in, and the parameters a, b, α, β and γ are functions of the local texture Y. Furthermore we restrict $b(y) \geq 0$ and $a(y) + b(y) \leq 1$.

This definition satisfies $\rho(\Delta i, \Delta j|Y_1 = y_1, Y_2 = y_2) = \rho(-\Delta i, -\Delta j|Y_1 = y_1, Y_2 = y_2)$. To satisfy the other restrictions for a function to be a correlation function: $\rho(\Delta i, \Delta j|Y_1 = y_1, Y_2 = y_2) \in [-1, 1]$ and $\rho(0, 0|Y_1 = y_1, Y_2 = y_2) = 1$, we need $a(y) + b(y) = 1$ and $a(y) \geq -1$. In order for the correlation model to approximate as closely as possible the average correlation coefficients in an video, we loosen the requirement $a(y) + b(y) = 1$ to $b(y) \geq 0$ and $a(y) + b(y) \leq 1$.

This new correlation model discriminates different local textures. As the spatial offsets between the two pixels, Δi and Δj, increase, $\rho(\Delta i, \Delta j|Y_1 = y_1, Y_2 = y_2)$ decreases at a different speed depending on the five parameters a, b, α, β and γ, which will be shown to be quite different for different local

textures. For each local texture, we choose the combination of the five parameters that jointly minimizes the MAE between the approximate correlation coefficients, averaged among all the blocks in a video frame that have the same local texture, i.e., $\hat{\rho}(\Delta i, \Delta j|y)$, and the correlation coefficients calculated using the new model, $\rho(\Delta i, \Delta j|y)$. These optimal parameters for one frame in Paris.cif and Football.cif and their corresponding MAEs are presented in Table I. (The local textures are calculated for each one of the 4 by 4 blocks; the available nine local textures are chosen to be those plotted in Fig. 2; Δi and Δj range from -7 to 7.) We can see from this table that the parameters associated with the new model are quite distinct for different local textures while the MAE is always less than 0.05. The values of all five parameters are also different for the two videos. In Fig. 4 we plot $\rho(\Delta i, \Delta j|y)$ of all the local textures for the same images from paris.cif and football.cif using these optimal parameters (as the dense surfaces, i.e., the mesh surface with more data points). We can see that the new spatial correlation model does capture the dependence of the correlation on the local texture and fits the average approximate correlation coefficients $\hat{\rho}(\Delta i, \Delta j|y)$ very well.

The parameters a, b, α, β and γ should have different optimal values when the block size used to calculate the local texture is different. Generally speaking, when the available local textures are fixed, the larger the block size, the less the actual average correlation coefficients should agree with the shape designated by the local texture. What also matters are the ranges of spatial offsets Δi and Δj over which the MAE between $\hat{\rho}(\Delta i, \Delta j|y)$ and $\rho(\Delta i, \Delta j|y)$ is calculated. The larger the range of spatial offsets, the more average correlation coefficients the model needs to approximate which will normally yield a larger MAE. These two aspects are shown in Fig. 5 for four different videos. As we can see in Fig. 5 the average MAE over all local textures increase, when the block size and/or the ranges of Δi and Δj increase. Therefore, when we employ the proposed correlation model and its corresponding optimal parameters in applications such as rate distortion analysis, we need to choose the block size and spatial offsets that yield a small MAE, chosen here to be 0.05.

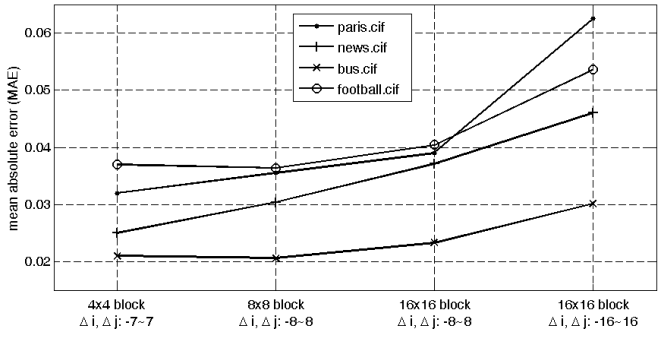


Fig. 5. The average MAE over all local textures, for different block sizes and spatial offsets of four videos

The new correlation model with its optimal parameters a, b, α, β and γ is expected to capture the characteristics of the content of the frames of a video scene. Therefore, the change of the optimal parameters a, b, α, β and γ from one frame to another in a video clip with the same scene is of great interest. To study this dependence, instead of calculating the optimal parameters of each local texture for each frame in a video clip and look at their variations, we use the optimal parameters calculated based on the average correlation coefficients of the first frame, and then study the average MAE over all local textures between the model-calculated correlation coefficients using these parameters and the average correlation coefficients of the following frames in the video clip. In Fig. 6 we plot such MAEs for 90 frames of four CIF videos. We can see that for paris and news, which have low motion, the MAEs throughout the whole video sequences are almost the same as that of the first frame. This is not true for football, whose MAEs quickly reach beyond 0.1 at frame # 21 and jump to 0.3 at frame # 35. However, this becomes less surprising when we look at the video frames of this clip presented in Fig. 7. With the high motion in the football video, the frames in this video do not have the same scene any more. For example, frame # 35 looks completely different than the first frame. Therefore, the optimal parameters generated based on one frame can be used in the other frames of the same scene. Different optimal parameters need to be calculated for different scenes even though the frames might reside in the same video.

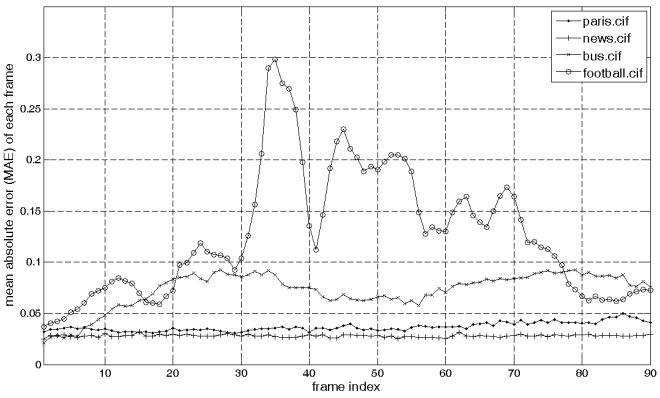


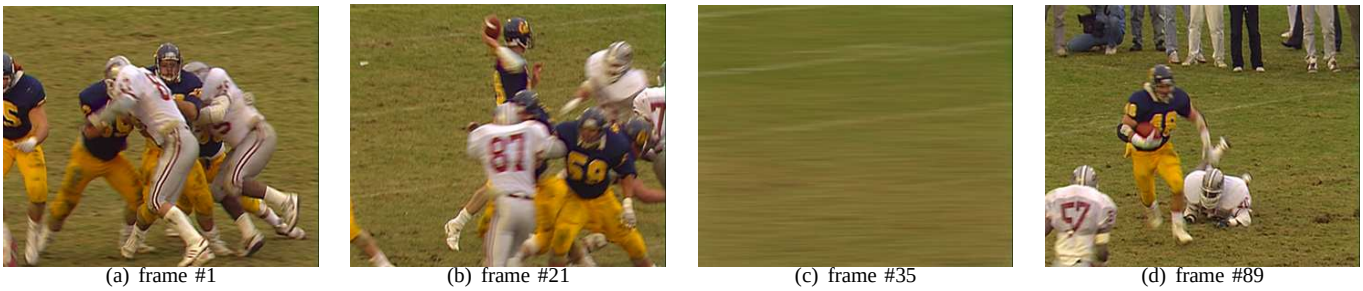
Fig. 6. The average MAE over all local textures, between the model-calculated correlation coefficients using the optimal parameters of the first frame in a video clip, and the average correlation coefficients of the following frames in the video clip

In the following sections, we study the rate distortion bounds of digitized natural videos which depend not only on the correlation model, but also on the pixel variance. Therefore we discuss briefly here the change in pixel variance from one frame to another in a video clip as plotted in Fig. 8. The results in Fig. 8 agree with those in Fig. 6 very well: for videos paris and news which have low motion and therefore can be considered as having only one scene in the entire clips, the change in pixel variance throughout the video clip is almost negligible; for videos with higher motion, such as bus and football, a new pixel value variance should be calculated based

TABLE I
THE OPTIMAL PARAMETERS FOR ONE FRAME IN PARIS.CIF AND FOOTBALL.CIF AND THEIR CORRESPONDING MEAN ABSOLUTE ERRORS (MAES)

Paris.cif	a	b	γ	α	β	MAE
texture #0	0.3	0.6	0.7	0.0	0.6	0.022
texture #1	0.3	0.6	0.9	-0.2	0.0	0.024
texture #2	0.6	0.3	0.9	0.0	-0.1	0.035
texture #3	0.6	0.3	0.9	-0.2	-0.1	0.043
texture #4	0.6	0.3	0.7	0.1	-0.2	0.034
texture #5	0.6	0.3	0.7	0.2	-0.6	0.028
texture #6	0.6	0.4	0.5	-1.3	0.4	0.026
texture #7	0.6	0.4	0.5	0.4	1.1	0.030
texture #8	0.6	0.4	0.6	0.4	0.1	0.046

Football.cif	a	b	γ	α	β	MAE
texture #0	0.2	0.6	0.8	0.0	-0.1	0.045
texture #1	0.8	0.2	0.3	-1.0	0.1	0.017
texture #2	0.6	0.3	0.8	0.0	-0.2	0.043
texture #3	0.5	0.5	0.5	0.4	0.5	0.048
texture #4	0.3	0.6	0.7	-0.1	0.1	0.040
texture #5	0.4	0.5	0.9	0.1	-0.3	0.034
texture #6	0.6	0.4	0.5	-0.2	0.1	0.031
texture #7	0.4	0.6	0.5	-0.3	-0.7	0.044
texture #8	0.7	0.3	0.6	0.4	0.1	0.029



(a) frame #1 (b) frame #21 (c) frame #35 (d) frame #89

Fig. 7. Four frames in video clip football.cif

on the frames in each scene of the video.

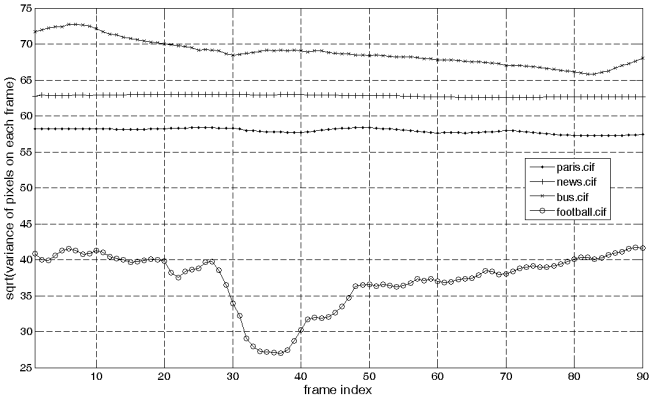


Fig. 8. Pixel value variance of 90 frames in four video clips

IV. THEORETICAL RATE DISTORTION BOUND WITH LOCAL TEXTURE AS UNIVERSAL SIDE INFORMATION

With the new block-based local-texture-dependent correlation model, we study the rate distortion bound of the video source where no compression scheme is assumed. To facilitate the comparison with other rate distortion bounds involving certain compression schemes derived later in this paper and the operational rate distortion functions, the video source is constructed by two parts: $\underline{X}$ as an M by N block and $\underline{S}$ as the surrounding $2M+N+1$ pixels ($2M$ on the top, N to the left and the one on the left top corner). Y denotes the information of local textures formulated from a collection of natural images and is considered as universal side information available to both the encoder and the decoder. The number of available local textures is denoted by $|Y|$. We only employ the first order statistics of Y, $P[Y=y]$, i.e., the frequency of occurrence of each local texture in the natural images and videos. In simulations, when available, $P[Y=y]$ is calculated as the average over a number of natural video sequences commonly used as examples in video coding studies.

In the following we first investigate briefly the joint coding of $\underline{S}$ and $\underline{X}$ without the universal side information Y, the case normally studied in information theory; we then focus on the case when Y is taken into account in the rate distortion analysis, where interesting new results lie.

Two different distortion constraints are considered in this paper, denoted by "avgD" and "sepD" respectively:

Average distortion constraint (avgD):

$$\frac{1}{|\underline{S}|+|\underline{X}|}\left\{E[||\underline{S}-\hat{\underline{S}}||^2]+E[||\underline{X}-\hat{\underline{X}}||^2]\right\}\le D. \quad \text{(IV.7)}$$

Separate distortion constraint (sepD):

$$\frac{1}{|\underline{S}|}E[||\underline{S}-\hat{\underline{S}}||^2]\le D \text{ and } \frac{1}{|\underline{X}|}E[||\underline{X}-\hat{\underline{X}}||^2]\le D. \quad \text{(IV.8)}$$

The average distortion constraint is used dominantly in image and video compression, while recent research in perceptual quality measurement of videos has suggested the importance of the separate distortion constraint on maintaining perceptual video quality, because the variation in video quality from frame to frame or from one region to another in the same

frame induces an unpleasant viewing experience of the human users. In this section the lowest rate that can be achieved by coding $\underline{X}$ and $\underline{S}$ together is studied; therefore, we only use the average distortion constraint.

A. Rate distortion bound without taking into account side information

The rate distortion bound without taking into account the texture as side information is a straightforward rate distortion problem of a source with memory which has been studied extensively. It can be expressed as

$$R_{\underline{S},\underline{X} jointly-without Y}(D) = \min_{p(\underline{\hat{x}},\underline{\hat{s}}|\underline{x},\underline{s}):\ avgD\ in\ Eq.\ (IV.7)} \frac{I(\underline{X},\underline{S};\underline{\hat{X}},\underline{\hat{S}})}{|\underline{S}|+|\underline{X}|}, \tag{IV.9}$$

which is the minimum mutual information between the source $\underline{X},\underline{S}$ and the reconstruction $\underline{\hat{X}},\underline{\hat{S}}$, subject to the average distortion measure, avgD, as defined in Eq. (IV.7). To facilitate the comparison with the case when side information Y is taken into account, we calculate the correlation matrix as

$$E\left[\begin{pmatrix}\underline{X}\\ \underline{S}\end{pmatrix}(\underline{X}^T\ \underline{S}^T)\right] = \sum_{y=0}^{|Y|-1}\sigma^2\rho\left(\begin{pmatrix}\underline{X}\\ \underline{S}\end{pmatrix}|y\right)P[Y=y], \tag{IV.10}$$

where the conditional correlation coefficients are exactly what the new model defines.

B. Rate distortion bound with local texture as side information

The rate distortion bound with the local texture as side information is a conditional rate distortion problem of a source with memory.

The conditional rate distortion function of a source $\underline{X}$ with side information Y is defined as [37, Sec. 6.1]

$$R_{\underline{X}|Y}(D) = \min_{p(\underline{\hat{x}}|\underline{x},y):D(\underline{X},\underline{\hat{X}}|Y)\le D} I(\underline{X};\underline{\hat{X}}|Y), \tag{IV.11}$$

where

$$\begin{aligned} D(\underline{X},\underline{\hat{X}}|Y) &= \textstyle\sum_{\underline{x},\underline{\hat{x}},y} p(\underline{x},\underline{\hat{x}},y)D(\underline{x},\underline{\hat{x}}|y),\\ I(\underline{X};\underline{\hat{X}}|Y) &= \textstyle\sum_{\underline{x},\underline{\hat{x}},y} p(\underline{x},\underline{\hat{x}},y)log\frac{p(\underline{x},\underline{\hat{x}}|y)}{p(\underline{x}|y)p(\underline{\hat{x}}|y)}. \end{aligned} \tag{IV.12}$$

It can be proved [38] that the conditional rate distortion function in Eq. (IV.11) can also be expressed as

$$R_{\underline{X}|Y}(D) = \min_{D'_y s:D(\underline{X},\underline{\hat{X}}|Y)=\sum_y D_y p(y)\le D} \sum_y R_{\underline{X}|y}(D_y)p(y), \tag{IV.13}$$

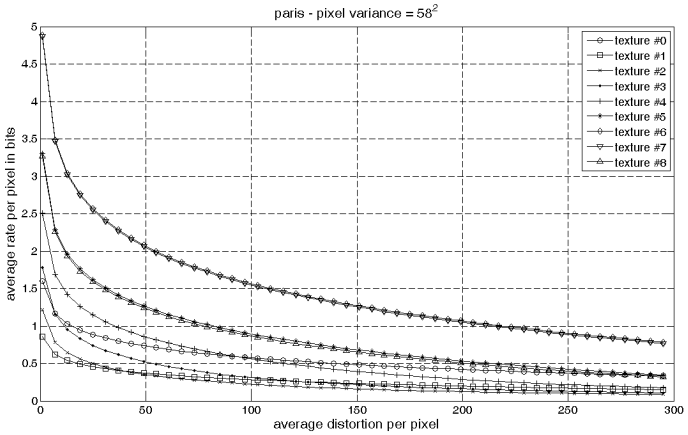
and the minimum is achieved by adding up the individual, also called marginal, rate-distortion functions at points of equal slopes of the marginal rate distortion functions.

Following the above classic results of conditional rate distortion theory, the rate distortion bound based on the new correlation model with the local texture as universal side information, is

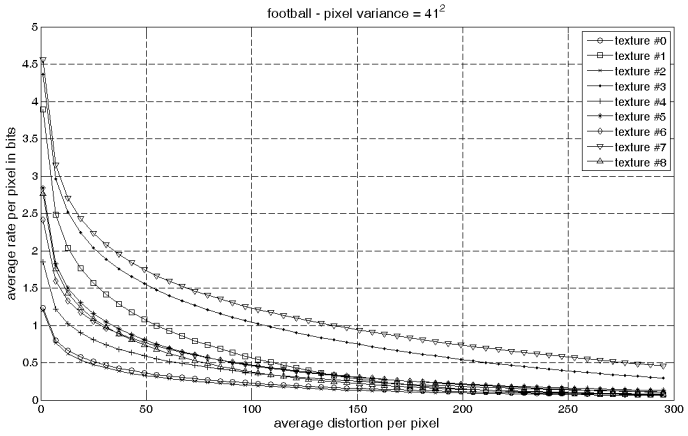
$$\begin{aligned} &R_{\underline{S},\underline{X} jointly-with Y}(D)\\ &= \min_{p(\underline{\hat{x}},\underline{\hat{s}}|\underline{x},\underline{s},y):\ avgD\ in\ Eq.\ (IV.7)} \frac{I(\underline{X},\underline{S};\underline{\hat{X}},\underline{\hat{S}}|Y)}{|\underline{S}|+|\underline{X}|}\\ &= \min_{D_y:\sum_y D_y P[Y=y]\le D} \textstyle\sum_y R_{\underline{X},\underline{S}|Y=y}(D_y)P[Y=y]. \end{aligned} \tag{IV.14}$$

Because the proposed correlation model discriminates all the different local textures, we can calculate the marginal rate distortion functions for each local texture, $R_{\underline{X},\underline{S}|Y=y}(D_y)$, as plotted in Fig. 9 for paris.cif and football.cif. The local textures are calculated for each one of the 4 by 4 blocks, the available nine local textures are chosen to be those plotted in Fig. 2, and the spatial offsets Δi and Δj are set to range from -7 to 7. The two plots in Figs. 9(a) and 9(b) show that the rate distortion curves of the blocks with different local textures are very different. Without the conditional correlation coefficient model proposed in this paper, this difference could not be calculated explicitly. The relative order of the nine local textures in terms of the average rate per pixel depends not only on the texture but also on the parameters associated with the correlation coefficient model for each local texture. For example, texture # 1, which is horizontal prediction, by intuition should consume less rate compared to other more complicated textures (# 3 through #8), which is the case for paris.cif. However for football.cif, texture # 1 consumes higher rate for some of the more complicated textures. This can be explained by looking at Fig. 4. In Fig. 4(b) both the approximate correlation coefficients and the model-calculated correlation coefficients of texture #1 are above 0.8, which is very high compared to those of the other textures. This means that the marginal rate distortion functions depend not only on the local texture, but also on the characteristics of a specific video. The latter dependence is captured by the five parameters a,b,α,β,γ in the new correlation model.

Utilizing the classical results for conditional rate distortion functions in Eq. (IV.13), the minimum in Eq. (IV.14) is achieved at $D'_y s$ where the slopes $\frac{\partial R_{\underline{X},\underline{S}|Y=y}(D_y)}{\partial D_y}$ are equal for all y and $\sum_y D_y P[Y=y] = D$. In Fig. 10 we plot this minimum $R_{\underline{S},\underline{X} jointly-with Y}(D)$ as well as $R_{\underline{S},\underline{X}\ jointly-without Y}(D)$ as dashed and solid lines, respectively, for two videos and three different blocksizes. In order to have a better idea of the region of interest for the average distortion levels, we plot in Fig. 11 the correspondence between peak signal to noise ratio (PSNR) and the average distortion when the maximum pixel value is 255. Comparing each pair of curves (solid line - without side information; dashed line - with side information, the same markers for the same blocksize) for paris.cif in Fig.10(a) shows that engaging the first-order statistics of the universal side information Y saves at least 1 bit per pixel at low distortion levels (distortion less than 25, PSNR higher than 35 dB), which corresponds to a reduction of about 100 Kbits per frame for the CIF videos and 1.5 Mbps if the videos only have intra-coded frames and are played at a medium frame rate of 15 frames per second. This difference decreases as the average distortion increases

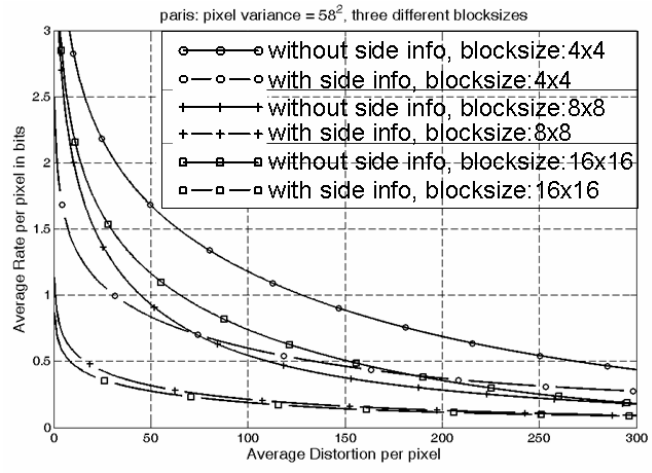


(a) paris.cif

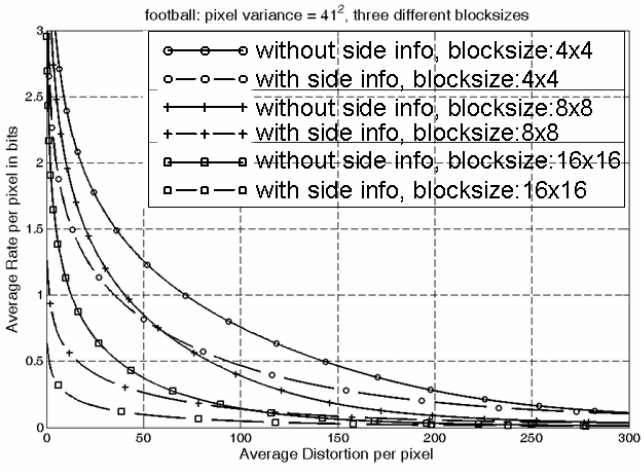


(b) football.cif

Fig. 9. Marginal rate distortion functions for different local textures, $R_{\underline{X},\underline{S}|Y=y}(D_y)$



(a) paris.cif



(b) football.cif

Fig. 10. Comparison of the theoretical rate distortion bounds in Section IV for two videos and three different blocksizes: solid lines – $R_{\underline{S},\underline{X}\ jointly-without Y}(D)$ (Eq. (IV.9)); dashed lines – $R_{\underline{S},\underline{X}\ jointly-with Y}(D)$ (Eq. ((IV.14))

but remains between quarter a bit and half a bit per pixel at high distortion level (distortion at 150, PSNR at about 26 dB), corresponding to about 375 Kbps to 700 Kbps in bit rate difference.

The rate distortion curves of paris.cif are generally higher than those of football.cif due to the higher pixel variance in paris.cif. For both videos, the higher the blocksizes, the lower the rate distortion curves. This is reasonable because when correlation among a larger set of pixels is explored the average rate per pixel should be lower. The difference between each pair of curves (solid line - without side information; dashed line - with side information, the same markers for the same blocksize) in Figs. 10(a) and 10(b), however, does not have a monotonic relationship with the block size at any distortion level. For example, at distortion 50, for paris.cif, this difference for blocksize 8x8 is lower than those of the other two blocksizes; but for football.cif, this difference for blocksize 8x8 is higher than those of the other two blocksizes.

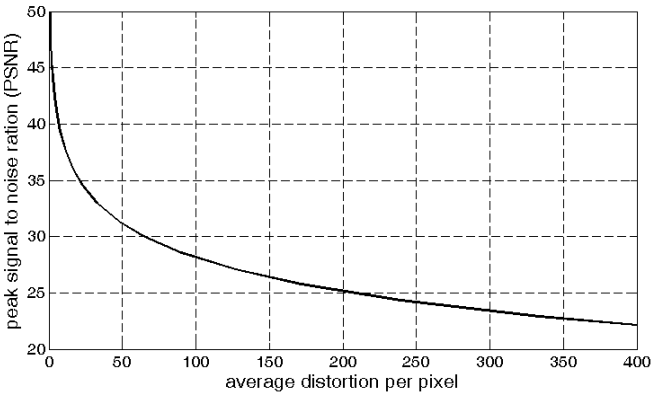


Fig. 11. The correspondence between peak signal to noise ration (PSNR) in dB and the average distortion when the maximum pixel value is 255 for CIF video frames

V. Rate distortion bounds for blocking and prediction across neighboring blocks

Breaking an image frame into 16×16 pixel MBs and processing one MB at a time, commonly known as the

"blocking" scheme, has been employed in the most popular image coding standards such as JPEG and almost all video coding standards such as MPEG-2/4 and the H.26x series [28]–[31]. In AVC/H.264 intra-frame prediction is utilized to reduce the spatial redundancy in the intra-coded frames, as discussed in Section III. With the new block-based local-texture-dependent correlation model, an explicit study of the rate distortion behavior of these key schemes, such as blocking and intra-prediction, is feasible. The basic set up can be summarized in the block diagram in Fig. 12. $\underline{X}$ denotes the M by N block currently being processed. The surrounding $2M+N+1$ pixels ($2M$ on the top, N to the left and the one on the left top corner), denoted by $\underline{S}$, are used to form a prediction block for each one of the available local textures, as

$$\underline{Z} = \underline{X} - P_d^{(A)}\underline{S}, \tag{V.15}$$

where $P_d^{(a)}$ is a $M \times N$ by $2M+N+1$ matrix, different for each local texture. A is the local texture chosen for the current block which yields the smallest prediction error. $\underline{Z}$ and A are further coded and transmitted to the decoder, where the predicted value is added in to obtain

$$\underline{\hat{X}} = \underline{\hat{Z}} + P_d^{(\hat{A})}\underline{\hat{S}}. \tag{V.16}$$

In the block diagram in Fig. 12, Y denotes the information of local textures formulated from a collection of natural images and is considered as universal side information available to both the encoder and the decoder. The number of available local textures is denoted by $|Y|$.

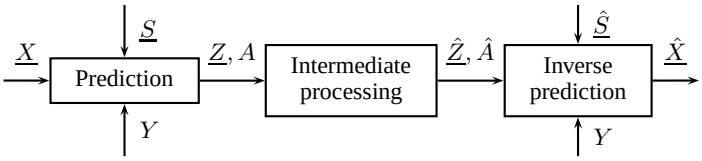


Fig. 12. Coding of one M by N block $\underline{X}$ and the surrounding $2M+N+1$ pixels $\underline{S}$

With the block based nature of the new correlation model, we study the penalty paid in average rate when the correlation among the neighboring MBs or blocks is disregarded completely (blocking, Section V-A) or is incorporated partially through the predictive coding (blocking and intra-frame prediction, Section V-B). In this Section we use the separate distortion measure, sepD as defined in Eq. (IV.8) since in video coding each MB is processes sequentially and only local distortion is considered. The rate distortion bounds calculated using sepD should be slightly higher than those when avgD is used.

A. Rate distortion bound for blocking

Since in this subsection we are interested in the penalty paid in average rate when the correlation among the neighboring MBs or blocks are disregarded completely, $\underline{S}$ and $\underline{X}$ are coded separately with the separate distortion constraint sepD in Eq. (IV.8). The total rate can be calculated as

$$R_{\underline{S},\underline{X} separately-without Y}(D) = \frac{R_{\underline{X}}(D)|\underline{X}| + R_{\underline{S}}(D)|\underline{S}|}{|\underline{S}|+|\underline{X}|}, \tag{V.17}$$

which is the average of the rate distortion functions of $\underline{X}$ and $\underline{S}$. We plot $R_{\underline{S},\underline{X}\ separately-without Y}(D)$ as dotted lines in Fig. 13 for two videos and three different block sizes. Not surprisingly for both videos and all three block sizes, coding $\underline{S}$ and $\underline{X}$ separately costs more bits than coding them jointly. The difference in bit rate decreases as the block size increases, since for smaller block sizes information on stronger correlation across the blocks is disregarded. With the new correlation coefficient model and the corresponding rate distortion curves, we can calculate explicitly the bit rate increase caused by blocking. For example, this penalty is one sixth bit per pixel in this plot at all distortion levels in Fig. 13(a), which is quite significant.

B. Rate distortion bound for blocking and optimal prediction

In the following we focus on the scenario when the video frames are processed block by block sequentially but the correlation among the blocks is utilized through predictive coding. We restrict ourselves to the separate distortion measure sepD in Eq. (IV.8) and therefore $\underline{S}$ is coded with no consideration of $\underline{X}$, after which $\underline{Z}$ and A are calculated by using intra-prediction in Eq. (V.15). The rate distortion function for this scenario is

$$R_{\underline{S},\underline{Z},A\ separately-without Y}(D) = \Big(\min_{p(\underline{\hat{s}}|\underline{s}):\frac{E[||\underline{S}-\underline{\hat{S}}||^2]}{|\underline{S}|}\le D} I(\underline{S};\underline{\hat{S}}) + \min_{p(\underline{\hat{z}},\hat{a}|\underline{z},a,\underline{s},\underline{\hat{s}}):\frac{E[||\underline{X}-\underline{\hat{X}}||^2]}{|\underline{X}|}\le D} I(\underline{Z},A;\underline{\hat{Z}},\hat{A})\Big)/(|\underline{S}|+|\underline{X}|) \tag{V.18}$$

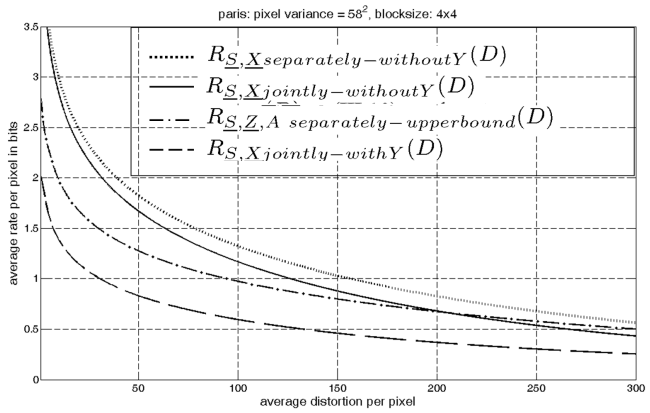
If we restrict that $A=\hat{A}$, i.e., we code the local texture A losslessly, the second part in Eq. (V.18) becomes

$$\min_{p(\underline{\hat{z}},\hat{a}|\underline{z},a,\underline{s},\underline{\hat{s}}):\frac{1}{|\underline{X}|}E[||\underline{X}-\underline{\hat{X}}||^2]\le D} I(\underline{Z},A;\underline{\hat{Z}},\hat{A}) = \min_{p(\underline{\hat{z}}|\underline{z},a,\underline{s},\underline{\hat{s}}):\frac{1}{|\underline{X}|}E[||\underline{X}-\underline{\hat{X}}||^2]\le D} I(\underline{Z};\underline{\hat{Z}}|A) + H(A), \tag{V.19}$$

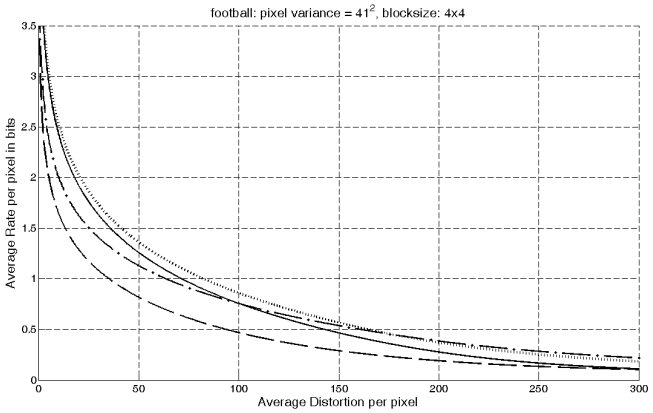
which forms an upper bound for all the scenarios when A is coded either losslessly or subject to a fidelity criterion. Also when $A=\hat{A}$, we have

$$\begin{aligned} &E[||\underline{X}-\underline{\hat{X}}||^2] = \textstyle\sum_a Pr(a)E[||(\underline{Z}+P_d^{(a)}\underline{S}) - (\underline{\hat{Z}}+P_d^{(a)}\underline{\hat{S}})||^2|a] \\ &= \textstyle\sum_a Pr(a)\int_{\underline{s}}\int_{\underline{\hat{s}}}\int_{\underline{z}}\int_{\underline{\hat{z}}} p(\underline{z},\underline{\hat{z}},\underline{s},\underline{\hat{s}}|a)(\underline{\hat{z}}-\underline{z})^T(\underline{\hat{z}}-\underline{z}) + \\ &(\underline{\hat{s}}-\underline{s})^T P_d^{(a)T}P_d^{(a)}(\underline{\hat{s}}-\underline{s}) + 2(\underline{\hat{s}}-\underline{s})^T P_d^{(a)T}(\underline{\hat{z}}-\underline{z})d\underline{s}d\underline{\hat{s}}d\underline{z}d\underline{\hat{z}}. \end{aligned} \tag{V.20}$$

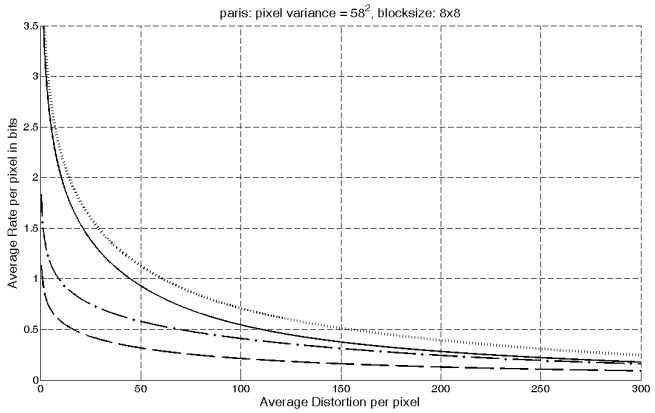
In order to investigate the lowest rate when predictive coding is employed, we use the optimal linear predictor $P_{opt}^{(a)} = E[\underline{X}\underline{S}^T|a](E[\underline{S}\underline{S}^T|a])^{-1}$ assuming that $E(\underline{S}\underline{S}^T|a)$ is non-singular. Since the source is assumed to be zero-mean Gaussian, the optimal linear predictor is also the optimal conditional mean predictor. The optimality is in the sense of minimizing MSE of $\underline{X}$. When the optimal linear predictor $P_{opt}^{(A)}$ is used, the cross-product term in Eq. (V.20) disappears.



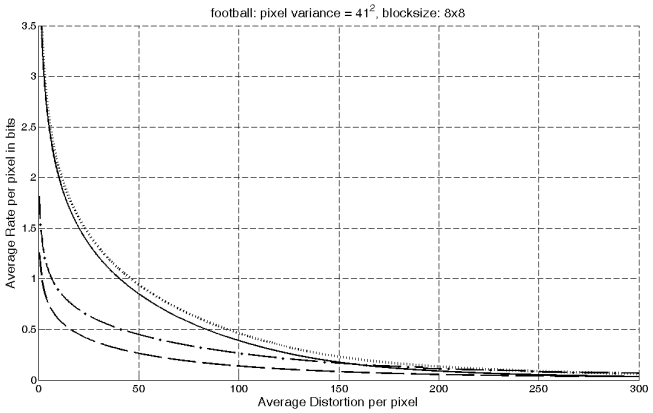
(a) paris, block size 4x4



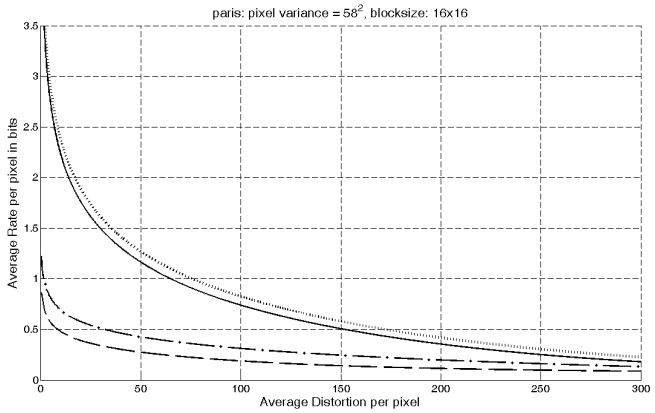
(b) football, block size 4x4



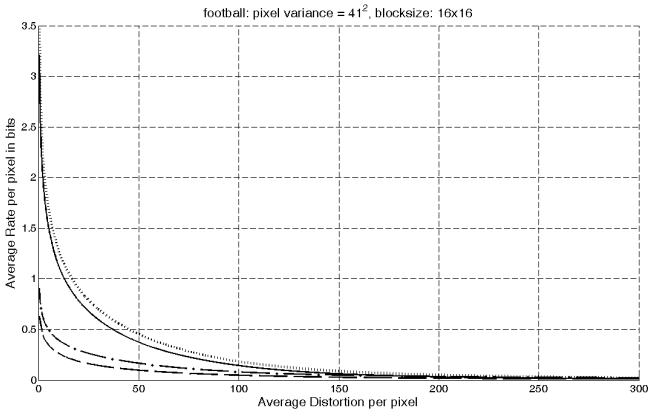
(c) paris, block size 8x8



(d) football, block size 8x8



(e) paris, block size 16x16



(f) football, block size 16x16

Fig. 13. Comparison of rate distortion bounds in Sections IV and V for two videos and three blocksizes: solid lines – $R_{\underline{S},\underline{X}\ jointly-withoutY}(D)$ in Eq. (IV.9); dashed lines – $R_{\underline{S},\underline{X}\ jointly-withY}(D)$ in Eq. (IV.14); dotted lines – $R_{\underline{S},\underline{X}\ separately-withoutY}(D)$ in Eq. (V.17) ; dash dot lines – $R_{\underline{S},\underline{Z},A\ separately-sep-upperbound}(D)$ in Eq. (V.24)

Let

$$D'_{\underline{S}} = \sum_a Pr(a) \int_{\underline{s}} \int_{\hat{\underline{s}}} p(\underline{s}, \hat{\underline{s}}|a)(\hat{\underline{s}}-\underline{s})^T P_{opt}^{(a)T} P_{opt}^{(a)}(\hat{\underline{s}}-\underline{s}) d\underline{s} d\hat{\underline{s}}. \tag{V.21}$$

Eq. (V.20) becomes

$$E[||\underline{X} - \hat{\underline{X}}||^2] = |\underline{Z}| D_{\underline{Z}} + D'_{\underline{S}}. \tag{V.22}$$

Since $\underline{S}$ is optimally coded without consideration of $\underline{X}$ as in the first part of Eq. (V.18), $D'_{\underline{S}}$ is fixed as well. The constraint on the distortion of $\underline{Z}$ becomes

$$D_{\underline{Z}} \leq (|\underline{X}| D - D'_{\underline{S}})/|\underline{Z}|. \tag{V.23}$$

An upper bound for Eq. (V.18) is thus

$$R_{\underline{S},\underline{Z},A\ separately-upperbound}(D) = \frac{1}{|\underline{S}|+|\underline{X}|} \left(|\underline{S}| R_{\underline{S}}(D) + |\underline{Z}| R_{\underline{Z}|A}(\frac{|\underline{X}|D - D'_{\underline{S}}}{|\underline{Z}|}) + H(A) \right) \tag{V.24}$$

The conditional rate distortion function $R_{\underline{Z}|A}(D_{\underline{Z}})$ in Eq. (V.24) is again calculated based on the "equal slope" theorem of the marginal rate distortion functions $R_{\underline{Z}|A=a}(D_a)$ [38]. In this case since the actual local texture A is coded without any loss, the exact statistics of A are available at both the encoder and the decoder; therefore, whether the universal side information Y is available or not becomes insignificant. The only complexity in computation is caused because $E(\underline{S}\underline{S}^T|a)$ is usually singular when the direction of the local texture is DC, horizontal, vertical, or too close to horizontal/vertical. In these cases we use the pseudo-inverse matrix of $E(\underline{S}\underline{S}^T|a)$ in the calculation.

The bit rate decrease from the dotted lines (coding $\underline{S}$ and $\underline{X}$ separately, Eq. (V.17)) to the dash-dotted lines (the upper bound of coding $\underline{S}$, $\underline{Z}$ and A separately with optimal prediction, Eq. (V.24)) is truly phenomenal in all the plots in Fig. 13 at low distortion levels, corresponding to about 1 bit per pixel for paris and between half a bit to 1 bit per pixel for for football at distortion 25 (corresponding to PSNR 35 dB). This bit rate saving decreases as the distortion increases, and interestingly, it vanishes for football at certain distortions. This is because spending bits coding the local texture A losslessly becomes unjustifiable at high distortion levels. This is especially true when the bit rate is low and the processing block size is small. We can see that in Fig. 13(b) the dash-dotted line and the dotted line intersect at a distortion of about 180, corresponding to an average rate of 0.4 bits per pixel. The average bit rate spent on coding the local texture A losslessly is simply the entropy of A, divided by the number of pixels per block, which is 16 in Fig. 13(b) since 4×4 blocks are investigated. This average rate is about 0.2 bits per pixel, or 50% of the total average rate. This is to say that for this particular video football.cif, processed in 4×4 blocks, 0.4 bits per pixel is the threshold in average rate that depicts when incorporating the correlation among the neighboring blocks through optimal predictive coding and coding the local texture A losslessly, becomes worse than discarding the correlation among the neighboring blocks. This crossover average rate is different for different videos and different processing blocksizes, as can be seen in Fig. 13. It can be calculated along with the rate distortion bounds we derive in this paper and be utilized in real video codecs. More discussions about $R_{\underline{S},\underline{Z},A\ separately-upperbound}(D)$ are presented in Section VI when compared to the operational rate distortion curves of AVC/H.264.

VI. Comparison to the operational rate distortion curves of AVC/H.264

Among all the rate distortion functions we investigate in the previous sections, engaging prediction and coding $\underline{S}$, $\underline{Z}$ and A separately with the separate distortion constraint, as in Section V-B, is the most similar to intra-fame coding in state-of-the-art codecs such as AVC/H.264. The upper bound $R_{\underline{S},\underline{Z},A\ separately-upperbound}(D)$ in Eq. (V.24) is achieved when the local texture A is losslessly coded and optimal prediction is employed. Since in AVC/H.264, for intra-coded frames, the intra-modes are always coded losslessly, $R_{\underline{S},\underline{Z},A\ separately-upperbound}(D)$ should be a lower bound for the operational rate distortion function of intra-frame coding in AVC/H.264. If we remove all the assumptions on coding, the rate distortion bound of a video frame is $R_{\underline{S},\underline{X}jointly-withY}(D)$ in Eq. (IV.14). It is the theoretical rate distortion bound that is solely based on the proposed correlation model of the video source and takes advantage of the universal side information on the local texture. $R_{\underline{S},\underline{X}jointly-withY}(D)$ should always be lower than $R_{\underline{S},\underline{Z},A\ separately-upperbound}(D)$ according to the data processing theorem [39]. A third rate distortion bound is $R_{\underline{S},\underline{X}jointly-withoutY}(D)$ as calculated in Eq. (IV.9). Without taking into account the texture information this rate distortion bound should perform similarly to those based on the old correlation models as discussed in Section II-1.

In Fig. 14 we plot these three rate distortion bounds for paris.cif and the operational rate distortion functions for paris.cif intra-coded in AVC/H.264. In AVC/H.264 we choose the main profile with context-adaptive binary arithmetic coding (CABAC), which is designed to generate the lowest bit rate among all profiles. Rate distortion optimized mode decision and a full hierarchy of flexible block sizes from MBs to 4x4 blocks are used to maximize the compression gain. For the rate distortion bounds, we choose the block size 16x16 and the spatial offsets as from -16 to 16.

As shown in Fig. 14, the rate distortion bound without local texture information, $R_{\underline{S},\underline{X}jointly-withoutY}(D)$ as in Eq. (IV.9), plotted as a solid line, is higher than the actual operational rate distortion curve of AVC/H.264 at all distortion levels. The rate distortion bound with local texture information taken into account while making no assumptions in coding, i.e., $R_{\underline{S},\underline{X}jointly-withY}(D)$ as in Eq. (IV.14), plotted as a dashed line, is indeed a lower bound with respect to the operational rate distortion curves of AVC/H.264. The rate distortion bound calculated based on the new texture dependent correlation model for the scenario where optimal predictive coding is engaged to code $\underline{S}$, $\underline{Z}$ and A separately with separate distortion constraint, i.e., $R_{\underline{S},\underline{Z},A\ separately-upperbound}(D)$ as

in Eq. (V.24), plotted as a dash dotted line, is a reasonably tight lower bound, especially at medium to high distortion levels. In Fig. 15(a) we plot this lower bound $R_{\underline{S,Z},A\ separately-upperbound}(D)$ (Eq. (V.24)) and the operational rate distortion function using AVC/H.264 for two other videos. We can see that although the lower bounds are calculated based on only five parameters generated from each video, they do agree with the operational rate distortion curves of the corresponding video reasonably well. If we further plot these lower bounds as average rate per pixel versus PSNR of a video frame as in Fig. 15(b), the lower bounds appear to be nearly, linear which shows promises in codec design.

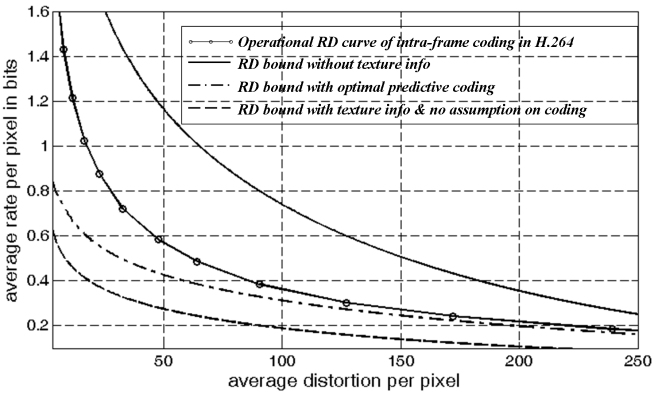
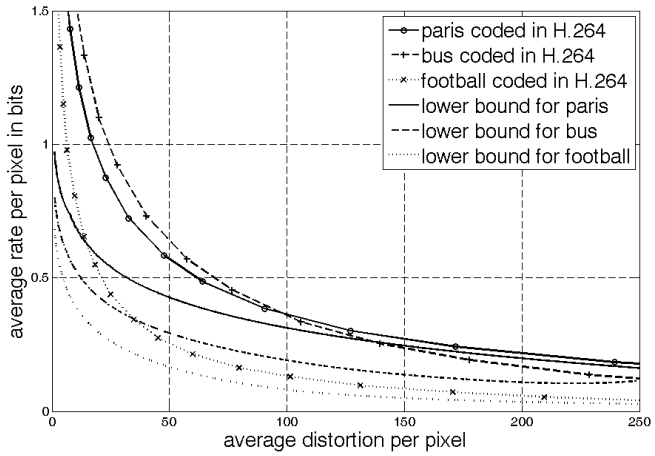


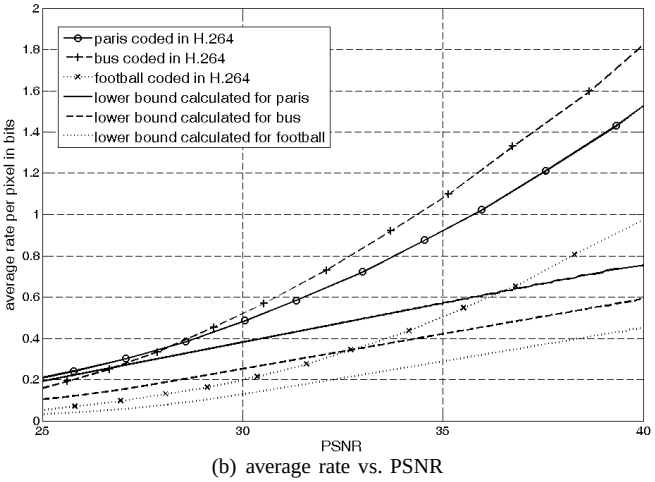
Fig. 14. Comparison of the rate distortion bounds and the operational rate distortion curves of paris.cif intra-coded in AVC/H.264

VII. Conclusions

We propose a conditional correlation model for two close pixels in one frame of digitized natural video sequences, with the conditioning being on the texture of the blocks where the two pixels are located. This new correlation model is dependent upon five parameters whose optimal values can be calculated for a specific image or video with a mean absolute error (MAE) usually smaller than 5%. Classical results in information theory are utilized to derive the conditional rate distortion function when the universal side information of local textures is available at both the encoder and the decoder, which is shown to save as much as 1 bit per pixel for selected videos at low distortions. We further study the common "blocking" scheme which divides a video frame into 16×16 macroblocks or smaller blocks before processing. With the block based nature of the new correlation model, we find the penalty paid in average rate when the correlation among the neighboring MBs or blocks is disregarded completely or is incorporated partially through predictive coding. The three rate distortion bounds investigated are compared to the operational rate distortion functions generated in intra-frame coding using AVC/H.264 video coding standard. The rate distortion bound without local texture information is shown to be much higher than the actual operational rate distortion curve of AVC/H.264. The rate distortion bound with local texture information taken into account while making no assumptions in coding, is indeed a lower bound with respect to the operational rate distortion curves of AVC/H.264. The rate distortion bound involving lossless coding of texture information and optimal prediction, is a reasonably tight lower bound and can be utilized in video codec design.



(a) average rate vs. average distortion



(b) average rate vs. PSNR

Fig. 15. The lower bounds calculated based on the new correlation coefficient model and its corresponding optimal parameters for three videos, compared to the operational rate distortion curves of these videos coded in AVC/H.264

References

[1] A. Habibi and P. A. Wintz, "Image coding by linear transformation and block quantization," *IEEE Transactions on Communication Technology*, vol. Com-19, no. 1, pp. 50–62, Feb. 1971.

[2] J. B. O'neal Jr. and T. R. Natarajan, "Coding isotropic images," *IEEE Transactions on Information Theory*, vol. IT-23, no. 6, pp. 697–707, Nov. 1977.

[3] G. Tziritas, "Rate distortion theory for image and video coding," *International Conference on Digital Signal Processing, Cyprus*, 1995.

[4] B. Girod, "The efficiency of motion-compensating prediction for hybrid coding of video sequences," *IEEE Journal on selected areas in communications*, vol. SAC-5, no. 7, pp. 1140–1154, Aug. 1987.

[5] A. Ortega and K. Ramchandran, "Rate-distortion methods for image and video compression," *IEEE Signal Processing Magazine*, vol. 15, no. 6, p. 2350, Nov. 1998.

[6] T. Chiang and Y.-Q. Zhang, "A new rate control scheme using quadratic rate distortion model," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 7, no. 1, pp. 246–251, Feb. 1997.

[7] H.-J. Lee, T. Chiang, and Y.-Q. Zhang, "Scalable rate control for MPEG-4 video," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 10, no. 6, pp. 878–894, Sep. 2000.

[8] J. Ribas-Corbera and S. Lei, "Rate control in DCT video coding for low-delay communications," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 9, no. 1, pp. 172–185, Feb. 1999.

[9] S. Ma, W. Gao, and Y. Lu, "Rate control on JVT standard," *Joint Video Team (JVT) of ISO/IEC MPEG & ITU-T VCEG, JVT-D030*, Jul. 2002.

[10] Z. G. Li, F. Pan K. P. Lim, X. Lin and S. Rahardj, "Adaptive rate control for h.264," *IEEE International Conference on Image Processing*, pp. 745–748, Oct. 2004.

[11] Y. Wu et al., "Optimum bit allocation and rate control for H.264/AVC," *Joint Video Team of ISO/IEC MPEG & ITU-T VCEG Document*, vol. JVT-O016, Apr. 2005.

[12] D.-K. Kwon, M.-Y. Shen and C.-C. J. Kuo, "Rate control for H.264 video with enhanced rate and distortion models," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 17, no. 5, pp. 517–529, May 2007.

[13] G. J. Sullivan and T. Wiegand, "rate-distortion optimization for video compression," *IEEE Signal Processing Magazine*, vol. 15, no. 6, pp. 74–90, Nov. 1998.

[14] Z. He and S. K. Mitra, "From rate-distortion analysis to resource-distortion analysis," *IEEE Circuits and Systems Magazine*, vol. 5, no. 3, pp. 6–18, Third quarter 2005.

[15] Y. K. Tu, J.-F. Yang and M.-T. Sun, "Rate-distortion modeling for efficient H.264/AVC encoding," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 17, no. 5, pp. 530–543, May 2007.

[16] R. Zhang, S. L. Regunathan, and K. Rose, "Video coding with optimal inter/intra-mode switching for packet loss resilience," *IEEE Journal on Selected Areas in Communications*, vol. 18, no. 6, pp. 966–976, 2000.

[17] R. C. Reininger and J. D. Gibson, "Distributions of the two-dimensional DCT coefficients for images," *IEEE Transactions on Communications*, vol. 31, pp. 835–839, Jun. 1983.

[18] S. R. Smoot and L. A. Rowe, "Study of DCT coefficient distributions," *SPIE Symposium on Electronic Imaging, San Jose, CA*, vol. 2657, Jan. 1996.

[19] W. Ding and B. Liu, "Rate control of MPEG video coding and recording by rate-quantization modeling," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 6, no. 1, pp. 12–20, Feb. 1996.

[20] H. M. Hang and J. J. Chen, "Source model for transform video coder and its application part (I): Fundamental theory," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 7, p. 1997, Apr. 287298.

[21] Z. He and S. K. Mitra, "A unified rate-distortion analysis framework for transform coding," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 11, pp. 1221–1236, Dec. 2001.

[22] L.-J. Lin and A. Ortega, "Bit-rate control using piecewise approximated rate-distortion characteristics," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 8, no. 4, pp. 446–459, Aug. 1998.

[23] K. Stuhlmuller, N. Farber, M. Link, and B. Girod, "Analysis of video transmission over lossy channels," *IEEE Journal on Selected Areas in Communications*, vol. 18, no. 6, Jun. 2000.

[24] M. van der Schaar, S. Krishnamachari, S. Choi, and X. Xu, "Adaptive cross-layer protection strategies for robust scalable video transmission over 802.11 WLANs," *IEEE Journal on Selected Areas in Communications*, vol. 21, no. 10, pp. 1752–1763, Dec. 2003.

[25] M. Wang and M. van der Schaar, "Model-based joint source channel coding for subband video," *IEEE Signal Processing Letters*, vol. 13, no. 6, Jun. 2006.

[26] ——, "Operational rate-distortion modeling for wavelet video coders," *IEEE Transactions on Signal Processing*, vol. 54, no. 9, Sep. 2006.

[27] C. Hsu, A. Ortega, and M. Khansari, "Rate control for robust video transmission over burst-error wireless channels," *IEEE Journal on Selected Areas in Communications, Special Issue on Multimedia Network Radios*, vol. 17, no. 5, pp. 756–773, May 1999.

[28] ISO/IEC 13818-1:2000, "Information technology – generic coding of moving pictures and associated audio information: Systems," 2000.

[29] ISO/IEC 14496-1:2001, "Information technology – coding of audio-visual objects – part 1: Systems," 2001.

[30] ITU Recommendations, "Video coding for low bit rate communication," *ITU-T rec. H.263*, Jan. 2005.

[31] ITU-T and ISO/IEC JTC 1, "Advanced video coding for generic audio-visual services," 2003.

[32] J. Hu and J. D. Gibson, "New rate distortion bounds for natural videos based on a texture dependent correlation model in the spatial-temporal domain," *Forty-Sixth Annual Allerton Conference on Communication, Control, and Computing*, Sep. 2008.

[33] B. Girod, "Motion-compensating prediction with fractional-pel accuracy," *IEEE Transactions on Communications*, vol. 41, pp. 604–612, Apr. 1993.

[34] T. Wiegand, G. J. Sullivan, G. Bjontegaard, and A. Luthra, "Overview of the H.264/AVC video coding standard," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 13, pp. 560–576, Jul. 2003.

[35] Q. Li and M. van der Schaar, "Providing adaptive QoS to layered video over wireless local area networks through real-time retry limit adaptation," *IEEE Transactions on Multimedia*, vol. 6, no. 2, pp. 278–290, Apr. 2004.

[36] T. Aach, C. Mota, I. Stuke, M. Mhlich, and E. Barth, "Analysis of superimposed oriented patterns," *IEEE Transactions on Image Processing*, vol. 15, no. 12, pp. 3690–3700, Dec. 2006.

[37] T. Berger, *Rate distortion theory*. New York: Wiley, 1971.

[38] R. M. Gray, "A new class of lower bounds to information rates of stationary sources via conditional rate-distortion functions," *IEEE Tran. Inform. Theory*, vol. IT-19, no. 4, pp. 480–489, Jul. 1973.

[39] T. M. Cover and J. A. Thomas, *Elements of information theory*. Wiley-Interscience, 1991.