Rate Distortion Lower Bounds for Video Sources and the HEVC Standard

Jing Hu, Malavika Bhaskaranand and Jerry Gibson
Department of Electrical and Computer Engineering
University of California, Santa Barbara
Santa Barbara, CA 93106, USA
Email: {jinghu, malavika, gibson}@ece.ucsb.edu

***Abstract*—Rate distortion bounds for video sources had eluded researchers for decades until our recent development of a new video source model. Our new model is composed of a five parameter spatial correlation model with the parameter selection dependent on texture information; and a series of temporal correlation coefficients that only depend on the video frame index offset. Using this new model and conditional rate distortion theory for the MSE distortion measure, we were able to obtain rate distortion functions that strictly lower bound the performance of the Advanced Video Coding (AVC/H.264) video codec. In this paper, we elaborate on the video source model and compare the performance of the newest high performance video codec, the High Efficiency Video Codec (HEVC/H.265), to our rate distortion curves.**

I. Introduction

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 studied extensively, the statistical models and their corresponding rate distortion theory are falling behind the rapidly 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 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 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 of image and video has been on setting up operational models for practical image and video compression systems to realize rate control [6]–[12] and to implement quality optimization algorithms [5], [13]–[16]. 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. For an in-depth discussion of the aforementioned related research, please refer to the Related Prior Work section of [17].

We address the difficult task of modeling the correlation in video sources by first proposing a new spatial correlation model for two close pixels in one digitized natural video frame that is conditional on the local texture. This new spatial correlation model is dependent upon five parameters whose optimal values are calculated for a specific image or specific video frames. The new spatial correlation model is simple, but it performs very well, as strong agreement is discovered between the approximate correlation coefficients and the correlation coefficients calculated by the new correlation model, with a mean absolute error (MAE) usually smaller than 0.05.

Further, we extend the correlation coefficient modeling from pixels within one video frame to pixels that are located in nearby video frames. We show that for two pixels located in nearby video frames, their spatial correlation and their temporal correlation are approximately independent. Therefore the correlation coefficient of two pixels in two nearby video frames, denoted by ρ, can be modeled as the product of ρ_s, the texture dependent spatial correlation coefficient of these two pixels, as if they were in the same frame, and ρ_t, a variable to quantify the temporal correlation between these two video frames. ρ_t does not depend on the textures of the blocks the two pixels are located in and is a function of the index offset of the two frames.

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 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; and the lowest rate that is theoretically achievable in *inter-frame* video compression can be as much as 0.7 bit per pixel lower than that without the side information.

The High Efficiency Video Coding (HEVC) standard is currently being developed by the ITU-T VCEG and the ISO/IEC MPEG organizations, who work together as the Joint Collaborative Team on Video Coding (JCT-VC) [18]. The first version of the HEVC standard has been finalized and approved by the ITU and MPEG in January 2013 [19], [20]. HEVC has been designed to address existing applications of H.264/AVC with particular focus on supporting increased video resolution and enabling increased use of parallel processing. It has been shown to achieve equivalent visual quality for 1280×720 high definition video at roughly half the bit rate required by the H.264/AVC standard [21].

In this paper we will show that our rate distortion bounds with local texture information taken into account while making no assumptions on coding, are indeed to be valid lower bounds with respect to the operational rate distortion curves of both *intra-frame* and *inter-frame* coding in AVC/H.264 and in the latest draft of High Efficiency Video Codec (HEVC/H.265).

II. A New Block-Based Conditional Correlation Model for Video

In this section we propose a new correlation model for a digitized natural video. We assume that all pixel values within one natural video form a three dimensional Gaussian random vector with memory, and each pixel value is of zero mean and the same variance σ^2. We first propose a new correlation model for a digitized natural image or an image frame in a digitized natural video, and then extend the spatial correlation model to pixels located in nearby frames of a video sequence.

A. The conditional correlation model in the spatial domain

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. A new coding technique, called "intra-frame prediction", in the video coding standard AVC/H.264, gave us inspirations on how to deal with the two aforementioned requirements.

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 an 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. 1, 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 though 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 [22], which should produce even better results in rate distortion modeling.

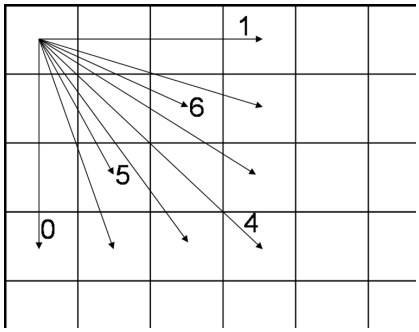


Fig. 1. The numbered arrows represent a few local textures that are defined as intra-prediction modes in AVC/H.264 and the unnumbered arrows represent a few local textures that are not defined as intra-prediction modes in AVC/H.264

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

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

for $-I \leq \Delta i \leq I$, $-J \leq \Delta j \leq J$. Please note that 1) $M \times N$ is not the size of a whole image, but the size of a 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. $\hat{\rho}_s(\Delta i, \Delta j)$ 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 $\hat{\rho}_s(\Delta i, \Delta j|y)$ where y denotes the local texture.

In Figs. 2, we plot $\hat{\rho}_s(\Delta i, \Delta j|y)$ (shown in the figure as the loose surfaces, i.e., the mesh surfaces that look lighter with fewer data points) for the first frame from paris.cif. 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 defined in AVC/H.264 standard for 4x4 blocks. 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. 2 shows that the average approximate correlation coefficient $\hat{\rho}_s(\Delta i, \Delta j|y)$ is very different for the blocks with different local textures. If we average $\hat{\rho}_s(\Delta i, \Delta j|y)$ across all the blocks in the picture, the important information about the local texture will be lost. Not surprisingly $\hat{\rho}_s(\Delta i, \Delta j|y)$ demonstrates certain shapes that agree with the orientation of the local textures. We also find out that although the average approximate correlation coefficients of the same local texture in different images demonstrate similar shapes their actual values are quite different.

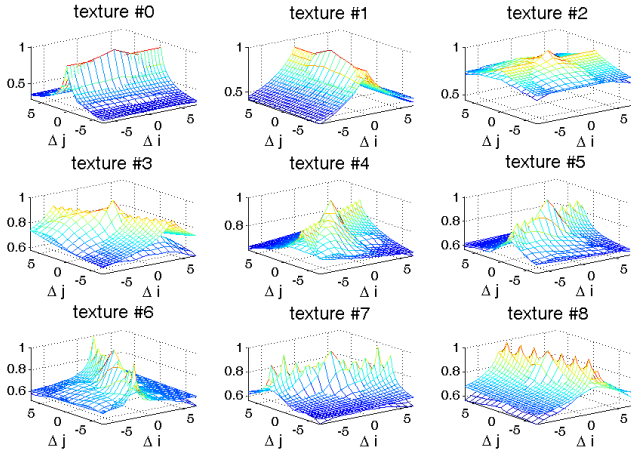


Fig. 2. The loose surfaces (the mesh surfaces that look lighter with less data points) are $\hat{\rho}_s(\Delta i, \Delta j|y)$, the approximate correlation coefficients of two pixel values in the first frame from paris.cif, averaged among the blocks that have the same local texture; the dense surfaces (the mesh surfaces that look darker with more data points) are $\rho_s(\Delta i, \Delta j|y)$, the correlation coefficients calculated using the proposed conditional correlation model, along with the optimal set of parameters

Motivated by these observations, in the following we present the formal definition of the new correlation coefficient model for a digitized natural image or an image frame in a digitized natural video that is dependent on the local texture.

Definition 2.1: 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

$$\rho_s(\Delta i, \Delta j|Y_1 = y_1, Y_2 = y_2) = \frac{\rho_s(\Delta i, \Delta j|y_1) + \rho_s(\Delta i, \Delta j|y_2)}{2} \quad \text{(II.2)}$$

where

$$\rho_s(\Delta i, \Delta j|y) = a(y) + b(y)e^{-|\alpha(y)\Delta i + \beta(y)\Delta j|^{\gamma(y)}}. \quad \text{(II.3)}$$

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 $\rho_s(\Delta i, \Delta j|Y_1 = y_1, Y_2 = y_2) = \rho_s(-\Delta i, -\Delta j|Y_1 = y_1, Y_2 = y_2)$. To satisfy the other restrictions for a function to be a correlation function: $\rho_s(\Delta i, \Delta j|Y_1 = y_1, Y_2 = y_2) \in [-1, 1]$ and $\rho_s(0, 0|Y_1 = y_1, Y_2 = y_2) = 1$, 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, $\rho_s(\Delta i, \Delta j|Y_1 = y_1, Y_2 = y_2)$ 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., $\hat{\rho}_s(\Delta i, \Delta j|y)$, and the correlation coefficients calculated using the new model, $\rho_s(\Delta i, \Delta j|y)$. 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 defined in AVC/H.264 standard for 4x4 blocks; 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. 2 we plot $\rho_s(\Delta i, \Delta j|y)$ of all the local textures for the same images from paris.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 $\hat{\rho}_s(\Delta i, \Delta j|y)$ 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 $\hat{\rho}_s(\Delta i, \Delta j|y)$ and $\rho_s(\Delta i, \Delta j|y)$ 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. 3 for four different videos. As we can see in Fig. 3 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

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

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

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.

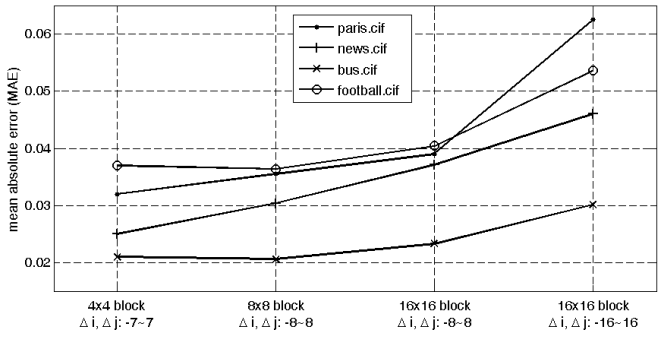


Fig. 3. The average MAE over all local textures, for different block sizes and spatial offsets of four videos

The new spatial 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. 4 we plot such MAE's for 90 frames of four CIF videos. We can see that for paris and news, which have low motion, the MAE's throughout the whole video sequences are almost the same as that of the first frame. This is not true for football, whose MAE's 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. 5. 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.

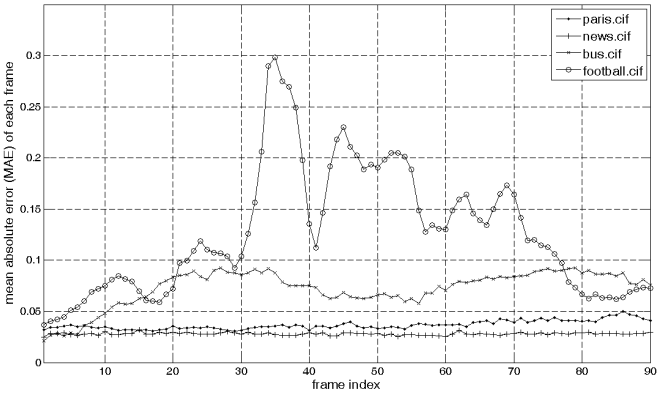
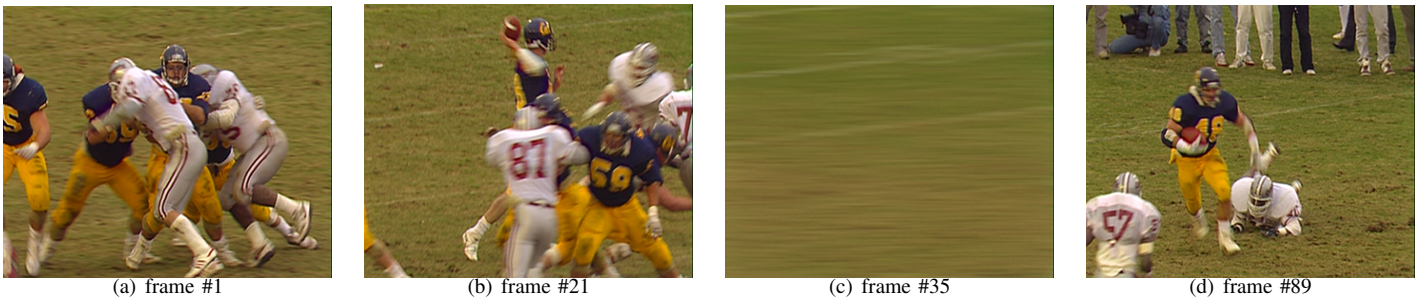


Fig. 4. The average MAE over all local textures, between the model-calculated correlation coefficients using the optimal parameters of the first frame in a video clip, and the average correlation coefficients of the following frames in the video clip

B. Correlation among pixels located in nearby frames

In this section we extend the correlation coefficient modeling from pixels within one video frame to pixels that are located in nearby video frames. Similar to the approach we take in deriving the spatial correlation model, we first study the approximate correlation coefficient between one block of size $M \times N$ in frame k_1 of a video, and another block of the same size, shifted by Δi vertically and Δj horizontally, in frame k_2 of the same video. Eq. (II.1) is used to calculate the approximate correlation coefficient of each pair of blocks, which is then averaged over all blocks with the same local texture. We denote this extended average approximate correlation coefficient as $\hat{\rho}_s(\Delta i, \Delta j, k_1, k_2|y)$. In Fig. 6 we plot $\hat{\rho}_s(\Delta i, \Delta j, k_1 = 1, k_2 = 16|y)$, with y being one of 9 local textures for video silent.cif. As shown in this figure, even though silent.cif is a video of a medium level



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

Fig. 5. Four frames in video clip football.cif

of motion, the pixels in the first frame and the pixels in the sixteenth frame have quite high correlation; and furthermore, the approximate correlation coefficients between these pixels show certain shapes that are similar to those modeled by the spatial correlation coefficient model.

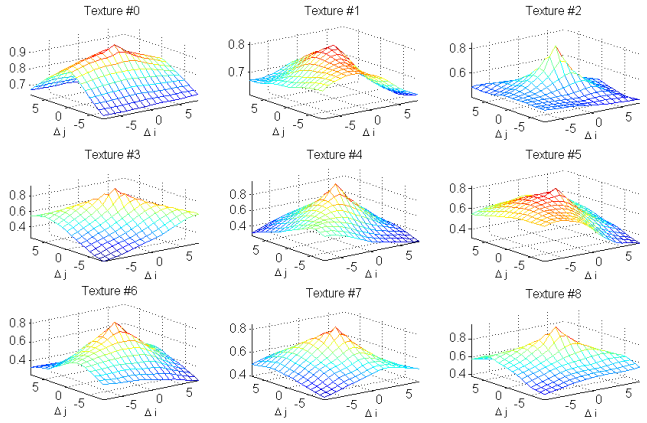


Fig. 6. $\hat{\rho}(\Delta i, \Delta j, k_1 = 1, k_2 = 16|y)$, the overall approximate correlation coefficients of two blocks, each in the 1^{st} and 16^{th} frames of silent.cif, respectively, averaged among the blocks that have the same local texture

To isolate the temporal correlation between two frames from the overall correlation, and to apply the spatial correlation coefficient model we already investigated, we first divide, element by element, the overall approximate correlation coefficients $\hat{\rho}(\Delta i, \Delta j, k_1 = 1, k_2 = 16|y)$, by the spatial approximate correlation coefficients $\hat{\rho}_s(\Delta i, \Delta j|y)$ of the first frame, i.e., $\hat{\rho}(\Delta i, \Delta j, k_1 = k_2 = 1|y)$. In [23] we show that it is adequate to use $\hat{\rho}_t(\Delta k)$, the average of $\hat{\rho}(\Delta i, \Delta j, k_1 = k_2 = 1|y)$ over all $\Delta i, \Delta j$, all k_1 and k_2 with the same shift $\Delta k = k_2 - k_1$, and over all local texture y's, to specify approximately the temporal correlation coefficient between two video frames with index difference Δk. This temporal correlation for paris.cif is plotted in Fig. 7.

We conclude this section with the following definition of the overall correlation coefficient model of natural videos that is dependent on the local texture.

Definition 2.2: The correlation coefficient of two pixel values within a digitized video, with spatial offsets Δi and

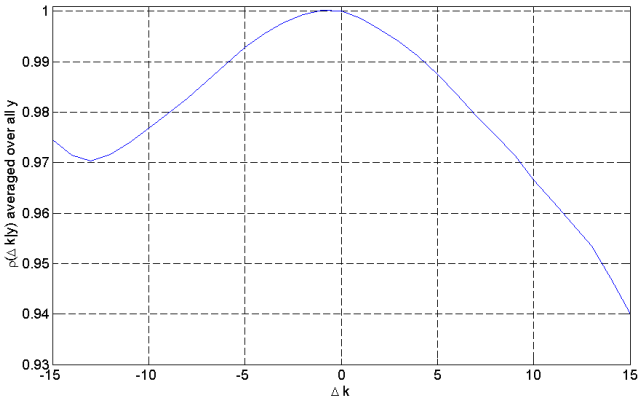


Fig. 7. $\hat{\rho}_t(\Delta k)$, the average of $\hat{\rho}(\Delta i, \Delta j, k_1 = k_2 = 1|y)$ over all $\Delta i, \Delta j$, all k_1 and k_2 with the same shift $\Delta k = k_2 - k_1$, and over all local texture y's, for paris.cif. This average is used to specify approximately the temporal correlation coefficient between two video frames with index difference Δk

Δj, and temporal offset Δk, is defined as

$$\begin{aligned} &\rho(\Delta i, \Delta j, \Delta k|Y_1 = y_1, Y_2 = y_2) \\ &= \rho_s(\Delta i, \Delta j|Y_1 = y_1, Y_2 = y_2)\rho_t(\Delta_k) \end{aligned} \quad \text{(II.4)}$$

where $\rho_s(\Delta i, \Delta j|Y_1 = y_1, Y_2 = y_2)$ is the spatial correlation coefficient as defined in Definition 2.1 and $\rho_t(\Delta_k)$ can be calculated by averaging the approximate temporal correlation coefficients $\hat{\rho}_t(\Delta_k|y)$, over all local texture y's.

In the following section, 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. 4 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 and often scene changes, such as bus and football, a new pixel value variance should be calculated based on the frames in each scene of the video.

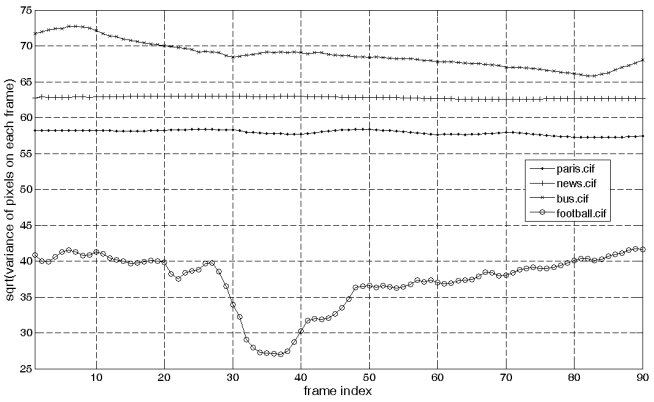


Fig. 8. Pixel value variance of 90 frames in four video clips

III. New theoretical rate distortion bounds of natural videos

In this section, we study the theoretical rate distortion bounds of videos based on the correlation coefficient model as defined in Definition 2.2. To facilitate the comparison with the operational rate distortion functions of AVC/H.264 and HEVC/H.265, we construct the video source in frame k by two parts: $\underline{X}_k$ as an M by N block (row scanned to form an $M \times N$ by 1 vector) and $\underline{S}_k$ as the surrounding $2M+N+1$ pixels ($2M$ on the top, N to the left and the one on the left top corner, forming a $2M+N+1$ by 1 vector). When we investigate the rate distortion bounds of a few frames k_1, k_2, ..., k_l, the video source across all these frames is defined as a long vector $\underline{V}$, where

$$\underline{V} = [\underline{X}_{k_1}^T, \underline{S}_{k_1}^T, \underline{X}_{k_2}^T, \underline{S}_{k_2}^T, \ldots, \underline{X}_{k_l}^T, \underline{S}_{k_l}^T]^T. \quad \text{(III.5)}$$

We assume that $\underline{V}$ is a Gaussian random vector with memory, and all entries of $\underline{V}$ are of zero mean and the same variance σ^2. The value of σ is different for different video sequences. The correlation coefficient between each two entries of $\underline{V}$ can be calculated using Definition 2.2.

We use Y to denote the information of local textures formulated from a collection of natural videos and Y is considered as universal side information available to both the encoder and the decoder. 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 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 rate distortion bound of $\underline{V}$ 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.

A. Formulation of rate distortion bound without local texture as 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

$$R_{\text{no texture}}(D) = \min_{p(\underline{\hat{v}}|\underline{v}): d(\underline{\hat{V}},\underline{V}) \le D} I(\underline{V};\underline{\hat{V}}), \quad \text{(III.6)}$$

which is the minimum mutual information between the source $\underline{V}$ and the reconstruction $\underline{\hat{V}}$, subject to a mean square distortion measure $d(\underline{\hat{v}},\underline{v}) = \frac{1}{|\underline{V}|}|\underline{\hat{v}}-\underline{v}|^T|\underline{\hat{v}}-\underline{v}|$. To facilitate the comparison with the case when side information Y is taken into account, we calculate the correlation matrix as

$$E\left[\underline{V}\underline{V}^T\right] = \sum_{y=0}^{|Y|-1} \sigma^2 \rho\left(\underline{V}|y\right) P[Y=y], \quad \text{(III.7)}$$

i.e., by taking the average of the texture dependent correlation coefficients $\rho\left(\underline{V}|y\right)$, as defined in Definition 2.2, over all local textures. To calculate $R_{\text{no texture}}(D)$, we first de-correlate the entries of the video source $\underline{V}$ by taking eigenvalue decomposition of the correlation matrix $E\left[\underline{V}\underline{V}^T\right]$. The reverse water-filling theorem [24] is then utilized to calculate the rate distortion bound of $\underline{V}$, whose entries are independent Gaussian random variables after de-correlation.

B. Formulation of 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. It is defined as [25, Sec. 6.1]

$$R_{\underline{V}|Y}(D) = \min_{p(\underline{\hat{v}}|\underline{v},y): d(\underline{V},\underline{\hat{V}}|Y) \le D} I(\underline{V};\underline{\hat{V}}|Y), \quad \text{(III.8)}$$

where

$$d(\underline{V},\underline{\hat{V}}|Y) = \textstyle\sum_{\underline{v},\underline{\hat{v}},y} p(\underline{v},\underline{\hat{v}},y) d(\underline{v},\underline{\hat{v}}|y), \quad \text{(III.9)}$$

and

$$I(\underline{V};\underline{\hat{V}}|Y) = \textstyle\sum_{\underline{v},\underline{\hat{v}},y} p(\underline{v},\underline{\hat{v}},y) log \frac{p(\underline{v},\underline{\hat{v}}|y)}{p(\underline{v}|y)p(\underline{\hat{v}}|y)}. \quad \text{(III.10)}$$

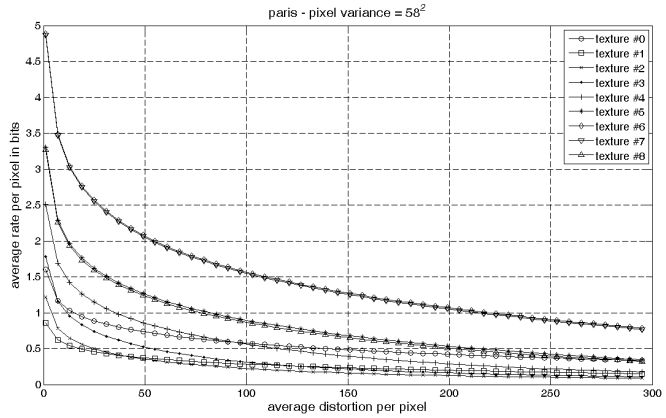
It can be proved [26] that the conditional rate distortion function in Eq. (III.8) can also be expressed as

$$R_{\underline{V}|Y}(D) = \min_{D_y' s: \sum_y D_y p(y) \le D} \sum_y R_{\underline{V}|y}(D_y) p(y), \quad \text{(III.11)}$$

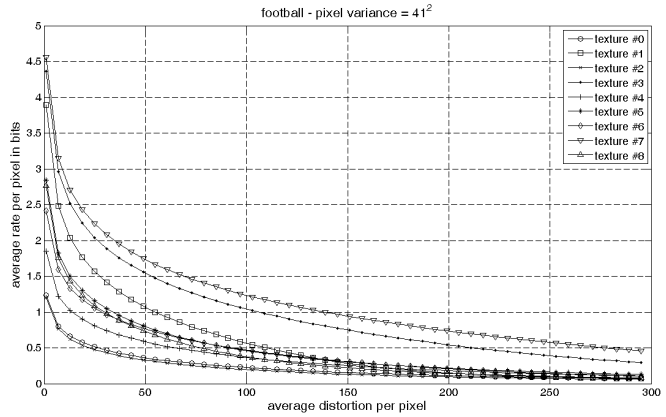
and the minimum is achieved by adding up $R_{\underline{V}|y}(D_y)$, the individual, also called marginal, rate-distortion functions, at points of equal slopes of the marginal rate distortion functions, i.e., when $\frac{\partial R_{\underline{V}|y}(D_y)}{\partial D_y}$ are equal for all y and $\sum_y D_y p(y) = D$. These marginal rate distortion bounds can also be calculated using the classic results on the rate distortion bound of a Gaussian vector source with memory and a mean square

error criterion, where the correlation matrix of the source is dependent on local texture y.

Because the proposed correlation model discriminates all the different local textures, we can calculate the marginal rate distortion functions for each local texture, $R_{\underline{V}|Y=y}(D_y)$, as plotted in Fig. 9 for one frame in paris.cif and football.cif, respectively. The local textures are calculated for each one of the 4 by 4 blocks, the available nine local textures are chosen to be those defined in AVC/H.264 standard for 4x4 blocks, 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.



(a) paris.cif



(b) football.cif

Fig. 9. Marginal rate distortion functions for different local textures, $R_{\underline{V}|Y=y}(D_y)$, for a frame in paris.cif and football.cif, respectively

In Fig. 10 we plot $R_{\underline{V}|Y}(D)$ and $R_{\text{no texture}}(D)$ as dashed and solid lines, respectively, for paris.cif and three different blocksizes. 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 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 but remains between $\frac{1}{4}$ bit per pixel and $\frac{1}{2}$ 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.

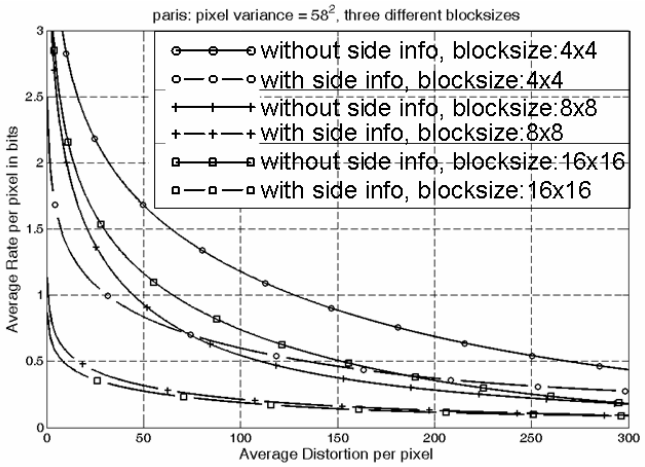


Fig. 10. Comparison of the theoretical rate distortion bounds for paris.cif and three different blocksizes: solid lines – $R_{no\ texture}(D)$ (Eq. (III.6)); dashed lines – $R_{with\ texture}(D)$ (Eq. ((III.8))

IV. Comparisons to the operational rate distortion curves of AVC/H.264 and HEVC/H.265

There are a few key features in HEVC that enable better compression performance and are different from those used in H.264/AVC [27]. The core processing unit (analogous to the macroblock in previous standards) in HEVC is called the coding tree unit (CTU) whose luma component can be of size 16×16, 32×32, or 64×64. The CTU is sub-divided into coding units (CUs) whose luma block can be as small as 8×8. The intra- or inter-picture prediction decision is made at the CU level which is further split into prediction units (PUs). For intra-picture prediction, the PUs are square and can have dyadic sizes ranging from 64×64 to 4×4; however, for inter-picture prediction, the PUs can be chosen as non-square. Additionally, each CU is partitioned into square transform blocks (TBs). Multiple integer transform bases that approximate the DCT are specified for dyadic sizes from 32×32 to 4×4. An integer transform approximating the 4×4 DST is also defined for the intra-picture prediction residuals. Intra-picture prediction supports 35 prediction modes: 33 directional, one planar, and one DC (flat). Inter-picture prediction uses 7-tap or 8-tap interpolation filters to achieve quarter-sample precision for the motion vectors (MVs) and

advanced motion vector prediction (AMVP) to reduce the number of bits required to code MVs. Two optional in-loop filters can be used within the inter-picture prediction loop: the deblocking filter (DBF) and the sample adaptive offset (SAO) filter. The deblocking filter is similar to the one in H.264/AVC but has simpler decision-making and filtering processes and is more suitable for parallel processing. The SAO filter is a non-linear amplitude mapping controlled using a few parameters determined by the encoder, with the goal to improve the reconstruction of the signal amplitudes. In order to enable parallel processing, high-level features such as slices, tiles, and wavefronts have been included in the HEVC standard.

In this section we compare our new theoretical rate distortion bounds to the *intra-frame* and *inter-frame* coding of AVC/H.264 and the recently approved HEVC/H.264. In AVC/H.264, for both intra-frame and inter-frame coding, 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. In HEVC, for both intra-frame and inter-frame coding, we choose CABAC and allow prediction unit sizes from 64x64 to 8x8 and transform block sizes from 32x32 to 4x4. We also allow the encoder to use two-level hierarchical B frames. For the rate distortion bounds, we choose the block size 16x16 and the spatial offsets as from -16 to 16.

A. Rate distortion bounds for one video frame

In Fig. 11 we plot the two rate distortion bounds $R_{\underline{V}|Y}(D)$ and $R_{\text{no texture}}(D)$ as dashed and solid lines, respectively, as well as the operational rate distortion functions of *intra-frame* coding in AVC/H.264 and in HEVC/H.265, for the first frame of paris.cif.

The rate distortion bound without local texture information, plotted as a black solid line, is higher than the actual operational rate distortion curves of H.264/AVC and HEVC. However, the rate distortion bounds with local texture information taken into account while making no assumptions in coding, plotted as a red dashed line, is indeed a lower bound with respect to the operational rate distortion curves of AVC/H.264 and HEVC/H.265.

B. Rate distortion bounds for a sequence of video frames

For multiple video frames we calculate $R_{\text{no texture}}(D)$ and two different conditional rate distortion bounds $R_{\underline{V}|Y}(D)$ depending on how the temporal correlation coefficient ρ_t is incorporated into the overall correlation among pixels in these video frames.

1) *With texture, one ρ_t for all textures:* this rate distortion bound is defined in the above Eq. (III.8) and correlation coefficients are exactly those defined in Definition 2.2.
2) *With texture, one ρ_t for each texture:* this rate distortion bound is also what is defined in Eq. (III.8), but when using Definition 2.2 to calculate the correlation coefficients among the entries of $\underline{V}$, we do not take the average of $\rho_t(\Delta k|y)$ over all textures but use $\rho_t(\Delta k|y)$ directly.

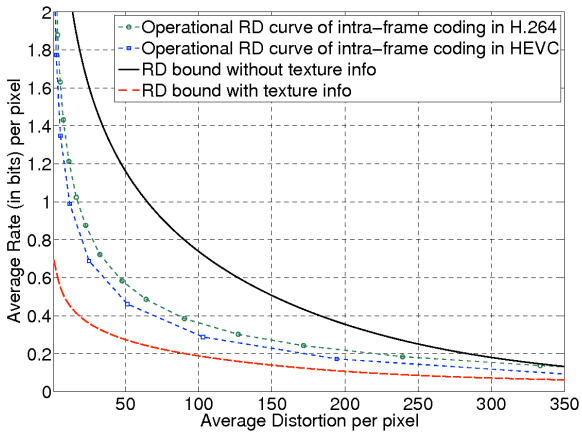


Fig. 11. Comparison of the rate distortion bounds and the operational rate distortion curves of paris.cif intra-coded in AVC/H.264 and in HEVC

The reason of studying these two different cases is that in Section II-B, we propose to use $\hat{\rho}_t(\Delta k)$, the average of $\hat{\rho}_t(k_1, k_2|y)$ over all k_1 and k_2 with the same shift $\Delta k = k_2 - k_1$ and over all local texture y's, to specify approximately the temporal correlation coefficient between two video frames with index difference Δk. For both cases, we first de-correlate the entries of the video source $\underline{V}$ by taking an eigenvalue decomposition of their respective correlation matrices. The reverse water-filling theorem [24] is then utilized to calculate the rate distortion bound of $\underline{V}$, whose entries are independent Gaussian random variables after de-correlation.

In Fig. 12 we plot these two conditional rate distortion bounds as well as $R_{\text{no texture}}(D)$ in Eq. (III.6) for paris.cif and the operational rate distortion curves for paris.cif, inter-coded in AVC/H.264. As shown in Fig. 12, the rate distortion bound without local texture information, plotted as solid lines, are higher than, or intersect with, the actual operational rate distortion curve of AVC/H.264. The rate distortion bounds with local texture information taken into account while making no assumptions in coding, both using one ρ_t for all textures and using one ρ_t for each texture, plotted as dotted lines and dashed lines respectively, are indeed lower bounds with respect to the operational rate distortion curves of AVC/H.264. The rate distortion bounds of using either temporal correlation definition agree with each other except at the very low distortion level, where the rate distortion bound of using one ρ_t for each texture is slightly higher than that of using one ρ_t for all textures.

Fig. 13 is similar to Fig. 12(d) but with the operational rate distortion function of HEVC/H.265 also included. As can be seen from Fig. 13, the theoretical rate distortion bound without the texture information is not a valid lower bound to the operational rate distortion function of HEVC/H.265 inter-frame coding with a group of pictures size of 5.

Comparing $R_{\text{no texture}}(D)$ (solid lines) and the conditional rate distortion bound *With texture, one ρ_t for all textures* (dotted lines) in Fig. 12(a) shows that by engaging the first-order statistics of the universal side information Y saves 0.5 bit per pixel at low distortion levels (distortion less than 25, PSNR higher than 35 dB), which corresponds to a reduction of about 50 Kbits per frame for the CIF videos and 750 Kbps if the videos have a group of picture size equal to 2 and are played at a medium frame rate of 15 frames per second. This difference decreases as the average distortion increases but remains 0.1 bit per pixel at high distortion level (distortion at 150, PSNR at about 26 dB), corresponding to about 150 Kbps in bit rate difference.

Another interesting observation of Fig. 12 is that as more video frames are coded, the actual operational rate distortion curves of inter-frame coding in AVC/H.264 become closer and closer to the theoretical rate distortion bound when no texture information is considered. This is because in AVC/H.264, only the intra-coded frames (i.e., only the 1^{st} frame in our simulation) take advantage of the local texture information through intra-frame prediction, while the inter-coded frames are blind to the local texture information. Therefore, when more frames are inter-coded, the bit rate saving achieved by intra-frame prediction in the 1^{st} frame is averaged over a larger number of coded frames. This suggests a possible coding efficiency improvement in video codec design by involving texture information even for inter-coded frames.

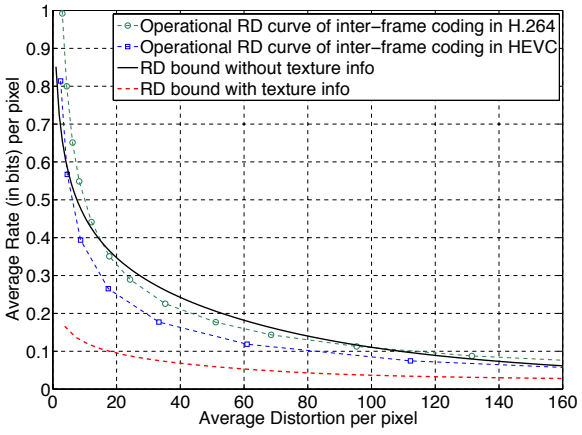


Fig. 13. Comparison of the rate distortion bounds and the operational rate distortion curves of paris.cif inter-coded in AVC/H.264 and in HEVC

V. Conclusion

We revisit the classic problem of developing a correlation model for natural videos and studying their rate distortion bounds. We addressed the difficult task of modeling the correlation in video sources by first proposing a new five parameter spatial correlation model for two close pixels in one frame of digitized natural video sequences that is conditional on the local texture. The correlation coefficient of two pixels in two nearby video frames, is then modeled as the product of the texture dependent spatial correlation coefficient of these two pixels, as if they were in the same frame, and a variable to quantify the temporal correlation between these two video frames. We also derive the conditional rate distortion function when the side information of local textures is available at both the encoder and decoder. We demonstrate that involving this side information can save as much as 1 bit per pixel for a single video frame and 0.5 bits per pixel for multiple video frames. 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 both intra-frame and inter-frame coding in both AVC/H.264 and HEVC/H.265.

Like all rate distortion bounds, these new bounds are valid only for the source models and distortion measures used in their calculation. For example, the theoretical calculations for the video sources are based on the assumption that the video sources are Gaussian. Gaussian sources are known to be the most pessimistic compared to other source distributions; that is, Gaussian source models yield rate distortion performance that upper bounds the rate distortion performance of sources with other distributions, such as Laplacian or Gamma. Additionally, the number of subsources (local textures), the size and shape of the local blocks are somewhat arbitrary. It is clear that these new bounds can be improved by more accurate composite source models and/or a more accurate method for quantifying local textures.

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.

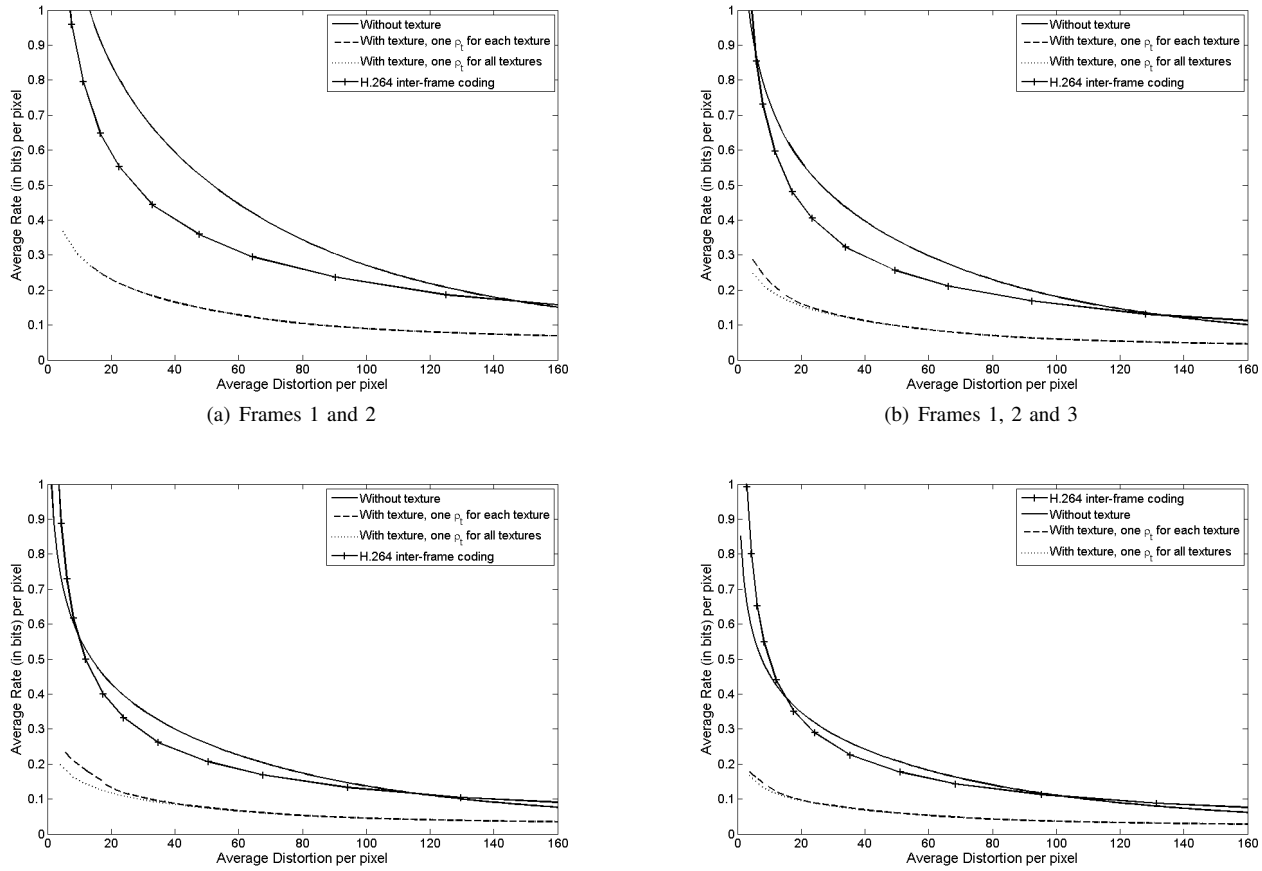
[8] J. Ribas-Corbera and S. Lei, "Rate control in DCT video coding for low-delay communications," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 9, no. 1, pp. 172–185, Feb. 1999.

[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.



(a) Frames 1 and 2

(b) Frames 1, 2 and 3

(c) Frames 1, 2, 3 and 4

(d) Frames 1, 2, 3, 4 and 5

Fig. 12. Theoretical rate distortion bounds and the rate distortion curves of inter-frame coding in AVC/H.264

[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] J. Hu and J. D. Gibson, "New rate distortion bounds for natural videos based on a texture dependent correlation model," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 19, Aug. 2009.

[18] J.-R. Ohm and G. J. Sullivan, "High efficiency video coding: The next frontier in video compression [standards in a nutshell]," *IEEE Signal Process. Mag.*, vol. 30, no. 1, pp. 152 –158, Jan. 2013.

[19] "ITU press release," Jan. 2013, (Retrieved on January 28, 2013). [Online]. Available: http://www.itu.int/net/pressoffice/press_releases/2013/01.aspx

[20] "MPEG press release," Jan. 2013, (Retrieved on January 28, 2013). [Online]. Available: http://mpeg.chiariglione.org/sites/default/files/files/meetings/docs/w13253_0.doc

[21] J.-R. Ohm, G. J. Sullivan, H. Schwarz, T. K. Tan, and T. Wiegand, "Comparison of the coding efficiency of video coding standards - including high efficiency video coding (HEVC)," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 22, no. 12, pp. 1669 –1684, Dec. 2012.

[22] 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.

[23] 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.

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

[25] T. Berger, *Rate Distortion Theory*. Prentice-Hall, 1971.

[26] 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.

[27] G. J. Sullivan, J.-R. Ohm, W.-J. Han, and T. Wiegand, "Overview of the high efficiency video coding (HEVC) standard," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 22, no. 12, pp. 1649 –1668, Dec. 2012.