New Rate Distortion Bounds for Natural Videos Based on a Texture Dependent Correlation Model in the Spatial-Temporal Domain

Jing Hu and Jerry D. Gibson
Department of Electrical and Computer Engineering
University of California, Santa Barbara, California 93106-9560
Email: jinghu and gibson@ece.ucsb.edu

***Abstract*— We revisit the classic problem of developing a correlation model for natural videos and studying their theoretical rate distortion bounds. We propose the correlation coefficient of two pixels in two nearby video frames as the product of the 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. The spatial correlation model for pixels within one video frame is a conditional correlation model. The conditioning is on local texture and the optimal parameters can be calculated for a specific 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 a single video frame and 0.7 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 AVC/H.264.**

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 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 in [3], [4], 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 [5], [6] 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 [7]. 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 [8]–[14] and to implement quality optimization algorithms [7], [15]–[18]. 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 [19], [20] 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 [12]–[14], [17], [21]–[24] that do not consider packet loss over communication networks and those proposed in [18], [25]–[29] 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 our previous work [3], [4] we addressed the difficult task of modeling the correlation in video sources by proposing a new spatial correlation model for two close pixels in one frame of digitized natural video sequences that is conditional on the local texture. This new spatial correlation model is dependent upon five parameters whose optimal values can be calculated for a specific image or video. 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 5%. With the new block-based local-texture-dependent spatial correlation model, we first studied the marginal rate distortion functions of the different local textures. These marginal rate distortion functions were shown to be quite distinct from each other. Classical results in information theory were 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 demonstrated that by involving this side information, the lowest rate that is theoretically

This research has been supported by the California Micro Program, Applied Signal Technology, Cisco, Dolby Labs, Inc., Sony-Ericsson, and Qualcomm, Inc., and by NSF Grant Nos. CCF-0429884, CNS-0435527, and CCF-0728646.

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, was shown indeed to be a valid lower bound with respect to the operational rate distortion curves of *intra-frame* coding in AVC/H.264.

In this paper we extend the correlation coefficient modeling and rate distortion analysis from pixels within one video frame to pixels that are located in nearby video frames. The correlation coefficient of two pixels in two nearby video frames, denoted by ρ, is 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. We show that for two pixels located in nearby video frames, their spatial correlation and their temporal correlation are approximately independent. Therefore ρ_t does not depend on the textures of the blocks the two pixels are located in and is a function of the indices of the two frames. With ρ_t calculated for the nearby frames of a video, we again derive the conditional rate distortion function when the side information of local textures is available at both the encoder and decoder. We demonstrate that by involving this side information, 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. 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 *inter-frame* coding in AVC/H.264.

The remainder of this paper is organized as follows. In Section II we review the texture dependent spatial correlation model and the marginal rate distortion bounds of a single video frame, as proposed in our previous work. In Section III we study the temporal correlation between pixels located in nearby frames of a video sequence. We reveal the approximate independence of the spatial and temporal correlation between these pixels and propose a model to quantify their overall correlation coefficients. In Section IV we calculate the conditional rate distortion bounds of video sequences based on the new correlation coefficient model and compare them to the *inter-frame* coding in AVC/H.264. We conclude this paper in Section V.

II. Previous work: correlation model in the spatial domain

In our previous work [3], [4] 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.

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 [30], but with a generalized block size and 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. Also the local textures need not to be restricted to those defined 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. 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 $-I \leq \Delta i \leq I$, $-J \leq \Delta j \leq J$. 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.

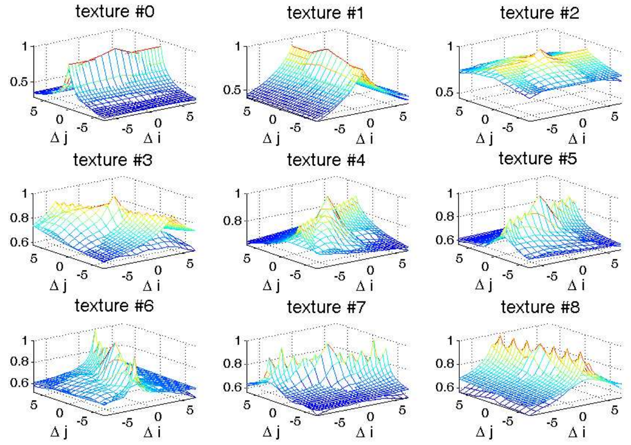


Fig. 1. The loose surfaces (the mesh surfaces with less data points) are $\hat{\rho}_s(\Delta i, \Delta j|y)$, the approximate correlation coefficients of two pixel values in the first frame from paris.cif, averaged among the blocks that have the same local texture; the dense surfaces are $\rho_s(\Delta i, \Delta j|y)$, the correlation coefficients calculated using the proposed conditional spatial correlation model, along with the optimal set of parameters

The following is the formal definition of the new spatial correlation coefficient model 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

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

where

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

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 $a(y) + b(y) \leq 1$.

For each local texture, we choose the combination of the five parameters a, b, α, β and γ 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 their corresponding MAEs are presented in Table I. (The local textures are calculated for each one of the 4 by 4 blocks; the available local textures are chosen to be those implemented in AVC/H.264; 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. In Fig. 1 we plot $\rho_s(\Delta i, \Delta j | y)$ of all the local textures for the same image 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 between two pixels on the local texture and fits the average approximate correlation coefficients $\hat{\rho}_s(\Delta i, \Delta j | y)$ very well.

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. 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. 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.

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. 2 for paris.cif. This plot shows 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. In [4], we calculate the conditional rate distortion function when universal side information on local texture is available at both the encoder and the decoder. This side information, when available, can save as much as 1 bit per pixel for selected videos at low distortions. This rate distortion bound is compared to the operational rate distortion functions generated in *intra-frame* coding using the AVC/H.264 video coding standard.

TABLE I
THE OPTIMAL PARAMETERS FOR ONE FRAME IN PARIS.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

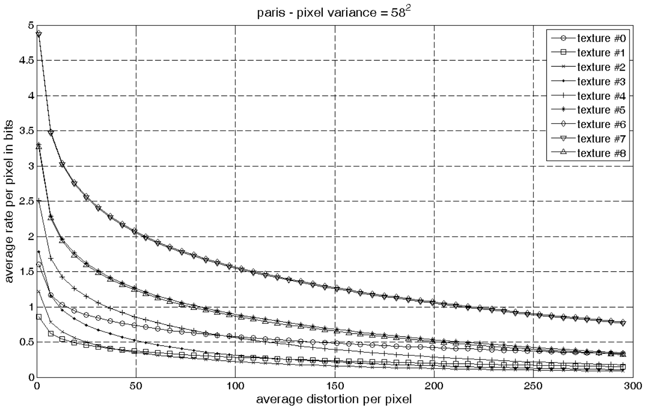


Fig. 2. Marginal rate distortion functions for different local textures, $R_{\underline{X},\underline{S}|Y=y}(D_y)$, for a frame in paris.cif

III. 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 $\hat{\rho}(\Delta i, \Delta j, k_1, k_2|y)$. In Fig. 3 we plot $\hat{\rho}(\Delta i, \Delta j, k_1 = 1, k_2 = 16|y)$, 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 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 we proposed in our previous work.

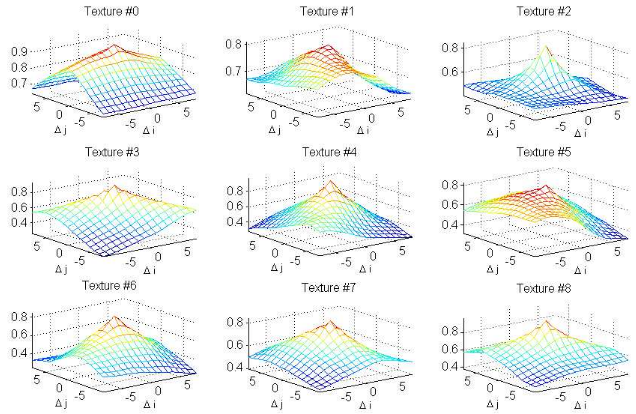


Fig. 3. $\hat{\rho}(\Delta i, \Delta j, k_1 = 1, k_2 = 16|y)$, the overall approximate correlation coefficients of two blocks, each in the 1^{st} and 16^{th} frames of silent.cif, respectively, averaged among the blocks that have the same local texture

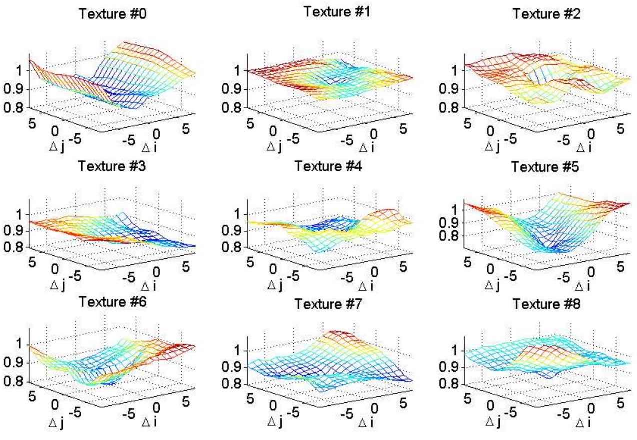


Fig. 4. $\frac{\hat{\rho}(\Delta i, \Delta j, k_1=1, k_2=16|y)}{\hat{\rho}(\Delta i, \Delta j, k_1=k_2=1|y)}$, the element by element fraction of the overall approximate correlation coefficient over the spatial approximate correlation coefficient of the first frame, of the video paris.cif

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., $\hat{\rho}(\Delta i, \Delta j, k_1 = k_2 = 1|y)$. The results for paris.cif are plotted in Fig. 4. As shown in this figure (note that the scales in this figure are different than those in Figs. 1 and 3), although the fractions are not exactly constant across all the values of Δi and Δj, their variations are much smaller than the variations of the overall approximate correlation coefficients and the spatial approximate correlation coefficients. As a result, we calculate the temporal approximate correlation coefficients, denoted by $\hat{\rho}_t(k_1, k_2|y)$, as the fractions of $\hat{\rho}(\Delta i, \Delta j, k_1, k_2|y)$ over $\hat{\rho}(\Delta i, \Delta j, k_1 = k_2|y)$, then averaged over all values of Δi and Δj.

Now let us take a closer look at the temporal approximate correlation coefficients $\hat{\rho}_t(k_1, k_2|y)$ for all frames and local textures of interest. For example, if we investigate the correlation among 16 frames of a video and there are 9 different local textures, we need to calculate and store a $16 \times 16 \times 9$ matrix in order to specify the temporal correlation among all pixels within these 16 video frames. One attempt to reduce the dimension of this matrix is to take the averages of $\hat{\rho}_t(k_1, k_2|y)$ over all local textures y, the result of which is plotted in Fig. 5 for paris.cif. Looking at this plot, we notice that when $k_2 > k_1$, $\hat{\rho}_t(k_1, k_2)$ is almost a constant for all values of k_1 and k_2 with the same shift $\Delta k := k_2 - k_1$. We therefore further take average of $\hat{\rho}_t(k_1, k_2)$ over all values of k_1 and k_2 with the same temporal shift Δk which results in the curve plotted in Fig. 7. As seen from this plot, $\hat{\rho}_t(\Delta k)$ stably descends as Δk increases from $\Delta k \geq 0$ and it is not quite symmetric with respect to $\Delta k = 0$. Another attempt to reduce the dimension of $\hat{\rho}_t(k_1, k_2|y)$ is to take its average over all values of k_1 and k_2 with the same shift Δk first for each local texture y. Averages taken this way are plotted in Fig. 6. $\hat{\rho}(\Delta k|y)$ shown in this figure appears to be different for different local textures. This behavior is interesting and is currently under investigation. For simplicity, 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 all local texture y's, to specify approximately the temporal correlation coefficient between two video frames with index difference Δk. In the next section, we will show that for paris.cif, the rate distortion bounds when either $\hat{\rho}(\Delta k|y)$ or $\hat{\rho}(\Delta k)$ is used are very similar in values.

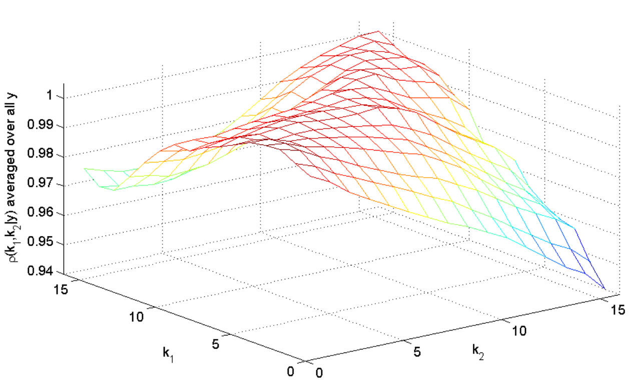


Fig. 5. $\hat{\rho}_t(k_1, k_2)$, the average of $\hat{\rho}_t(k_1, k_2|y)$ over all texture y's

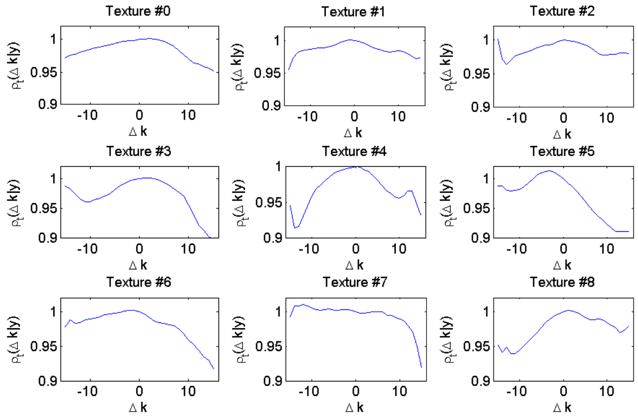


Fig. 6. $\hat{\rho}_t(\Delta k|y)$, the average of $\hat{\rho}_t(k_1, k_2|y)$ over all k_1 and k_2 with the same shift $\Delta k = k_2 - k_1$

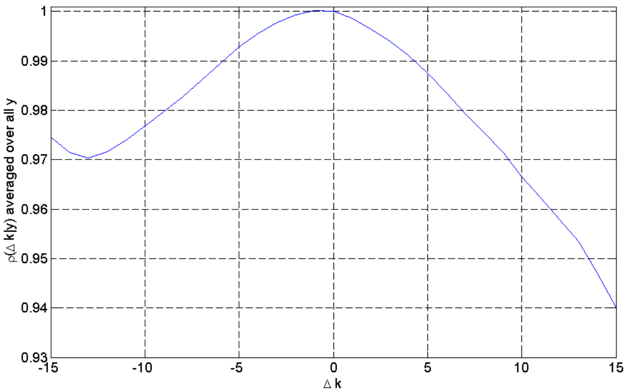


Fig. 7. $\hat{\rho}_t(\Delta k)$, the average of $\hat{\rho}_t(k_1, k_2|y)$ over all k_1 and k_2 with the same shift $\Delta k = k_2 - k_1$ and all local texture y's, for paris.cif. This average is used to specify approximately the temporal correlation coefficient between two video frames with index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

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

Definition 3.1: The correlation coefficient of two pixel values within a digitized video, with spatial offsets Δi and Δj, and temporal offset Δk, is defined as

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

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 as the approximate temporal correlation coefficients $\hat{\rho}_t(\Delta_k|y)$, averaged over all local texture y's.

IV. 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 3.1 and compare these bounds to the *inter-frame* coding of AVC/H.264.

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 MN 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). If we investigate the rate distortion bounds of a few frames k_1, k_2, $\ldots$, k_l, the video source across all these frames is defined as a long vector $\underline{V}$, where

$$\underline{V} = [\underline{X}_{k_1}^T, \underline{S}_{k_1}^T, \underline{X}_{k_2}^T, \underline{S}_{k_2}^T, \ldots, \underline{X}_{k_l}^T, \underline{S}_{k_l}^T]^T. \tag{IV.5}$$

We use Y to denote the information of local textures formulated from a collection of natural images and Y is considered as universal side information available to both the encoder and the decoder. Again, 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 coefficients between each two entries of $\underline{V}$ can be calculated using Definition 3.1.

The conditional rate distortion function of $\underline{V}$ with side information Y is

$$R_{\underline{V}|Y}(D) = \min_{p(\underline{\hat{v}}|\underline{v},y): D(\underline{V},\underline{\hat{V}}|Y)\le D} I(\underline{V};\underline{\hat{V}}|Y), \tag{IV.6}$$

where

$$\begin{aligned}&D(\underline{V},\underline{\hat{V}}|Y) = \textstyle\sum_{\underline{v},\underline{\hat{v}},y} p(\underline{v},\underline{\hat{v}},y)D(\underline{v},\underline{\hat{v}}|y)\\ &\text{and}\\ &I(\underline{V};\underline{\hat{V}}|Y) = \textstyle\sum_{\underline{v},\underline{\hat{v}},y} p(\underline{v},\underline{\hat{v}},y) log\dfrac{p(\underline{v},\underline{\hat{v}}|y)}{p(\underline{v}|y)p(\underline{\hat{v}}|y)}.\end{aligned} \tag{IV.7}$$

It can be proved [31] that the conditional rate distortion function in Eq. (IV.6) can also be expressed as

$$R_{\underline{V}|Y}(D) = \min_{D_y' s: D(\underline{V},\underline{\hat{X}}|Y)=\sum_y D_y p(y)\le D} \sum_y R_{\underline{V}|y}(D_y)p(y), \tag{IV.8}$$

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.

We calculate three types of theoretical rate distortion bounds in this section.

1) *With texture, one ρ_t for all textures:* this rate distortion bound is defined in the above Eq. (IV.6) and correlation coefficients are exactly those defined in Definition 3.1.
2) *With texture, one ρ_t for each texture:* this rate distortion bound is also what is defined in Eq. (IV.6), but when using Definition 3.1 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, i.e., for paris.cif, we use the values plotted in Fig. 6 rather than those in plotted Fig. 7.
3) *Without texture:* This rate distortion bound does not take into account the local texture as side information. For this rate distortion bound, we first take the average of the texture dependent correlation coefficients as defined in Definition 3.1 over all local textures,

then calculate $R_{\underline{V}}(D)$ which is a straightforward rate distortion problem of a source with memory that has been studied extensively.

For all the above three cases, we first decorrelate the entries of the video source $\underline{V}$ by taking eigen value decomposition of their respective correlation matrices. The reverse water-filling theorem [32] is then utilized to calculate the rate distortion bound of $\underline{V}$, whose entries are independent Gaussian random variables after decorrelation.

In Fig. 8 we plot these three rate distortion bounds for paris.cif and the operational rate distortion curves for paris.cif, inter-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. 8, 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 at all distortion levels. 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. Also 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 possible coding efficiency improvement in video codec design by involving texture information even for inter-coded frames.

V. Conclusions

We revisit the classic problem of developing a correlation model for natural videos and studying their rate distortion bounds. In our previous work [3], [4] we addressed the difficult task of modeling the correlation in video sources by proposing a new spatial correlation model for two close pixels in one frame of digitized natural video sequences that is conditional on the local texture. This new spatial correlation model is dependent upon five parameters whose optimal values can be calculated for a specific image or video. 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 5%. With the new block-based local-texture-dependent spatial correlation model, we first studied the marginal rate distortion functions of the different local textures. These marginal rate distortion functions were shown to be quite distinct from each other. Classical results in information theory were 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 demonstrated 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, was shown indeed to be a valid lower bound with respect to the operational rate distortion curves of *intra-frame* coding in AVC/H.264.

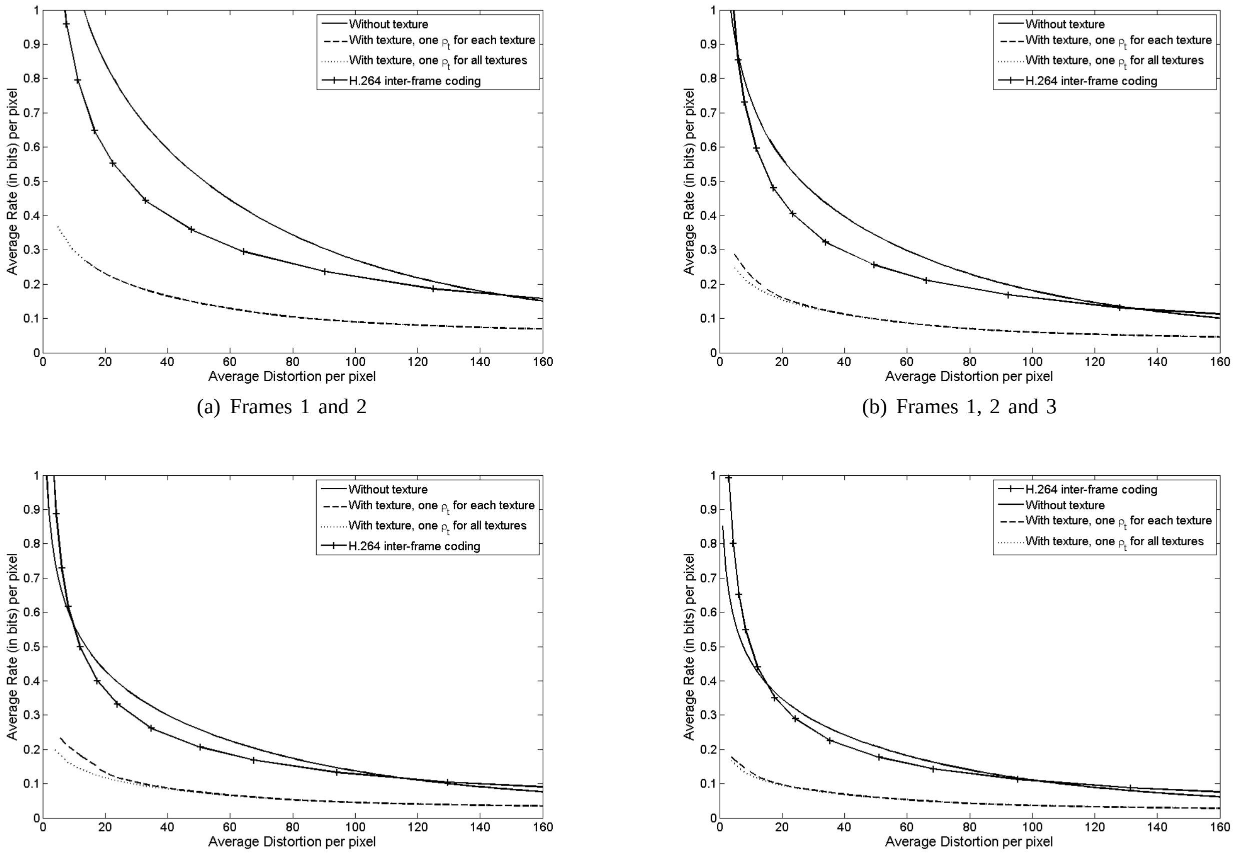
In this paper we extend the correlation coefficient modeling and rate distortion analysis from pixels within one video frame to pixels that are located in nearby video frames. The correlation coefficient of two pixels in two nearby video frames, denoted by ρ, is modeled as the product of ρ_s, the texture 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. We show that for two pixels located in nearby video frames, their spatial correlation and their temporal correlation are approximately independent. Therefore ρ_t does not depend on the textures of the blocks the two pixels are located in and is a function of the indices of the two frames. With ρ_t calculated for the nearby frames of a video, we again derive the conditional rate distortion function when the side information of local textures is available at both the encoder and decoder. We demonstrate that by involving this side information, 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. 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 *inter-frame* coding 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] J. Hu and J. D. Gibson, “New block-based local-texture-dependent correlation model of digitized natural video,” *Proceedings of the Fortieth Asilomar Conference on Signals, Systems, and Computers,* Oct. 2006.



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

[4] ——, "New rate distortion bounds for natural videos based on a texture dependent correlation model," *IEEE International Symposium on Information Theory,* Jun. 2007.

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

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

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

[21] 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.

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[26] 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.

[27] 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.

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

[29] 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.

[30] 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.

[31] 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.

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