New Block-Based Local-Texture-Dependent Correlation Model of Digitized Natural Video

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 spatial correlation model for natural images and video by proposing a conditional correlation model for relatively nearby pixels that is dependent upon four parameters. The conditioning is on local texture and the conditional correlation model is presented for each of the nine 4 by 4 intra-modes used in the AVC/H.264 video coding standard. We use this conditional correlation model to calculate the conditional rate distortion function when universal side information is available at both the encoder and the decoder. We demonstrate that this side information, when available, can save as much as 2 bits per pixel for selected videos at low distortions.**

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. 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. To study the new correlation model concretely we pick the blocksize as 4×4 and categorize the local texture using the 9 intra-frame prediction modes, which were introduced in the AVC/H.264 video coding standard [1]. The performance of the new correlation model is demonstrated through comparison with the approximated correlation coefficients of real video sequences (Section III).

We further study the marginal rate-distortion function of the different local textures, i.e., different intra-modes. These marginal rate-distortion functions are shown to be very distinct from each other. Classical results in information theory are then 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 "free" side information, the lowest rate that is theoretically achievable in video compression can be as much as 2 bits per pixel lower than that without the side information (Section IV). We conclude this paper and provide insights into future research in Section V.

II. Statistical models proposed in the past

A. In the spatial domain for images

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 one is

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

with Δi and Δj denoting offsets in horizontal and vertical coordinates. The parameters α and β control the magnitude of correlation in the horizontal and vertical directions, respectively, and their values can be chosen for different images [2]. 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 Kahrunen-Loeve transform (KLT).

The second correlation model studied previously is an isotropic function

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

This model implies that the correlation between two pixels within an image depends only on the Euclidean distance between them [3]. 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. The subscript s in both (II.1) and (II.2) emphasizes that these two correlation functions are in the spatial domain.

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 are ever advancing and the difference between the performance of various schemes is much more subtle than several decades ago.

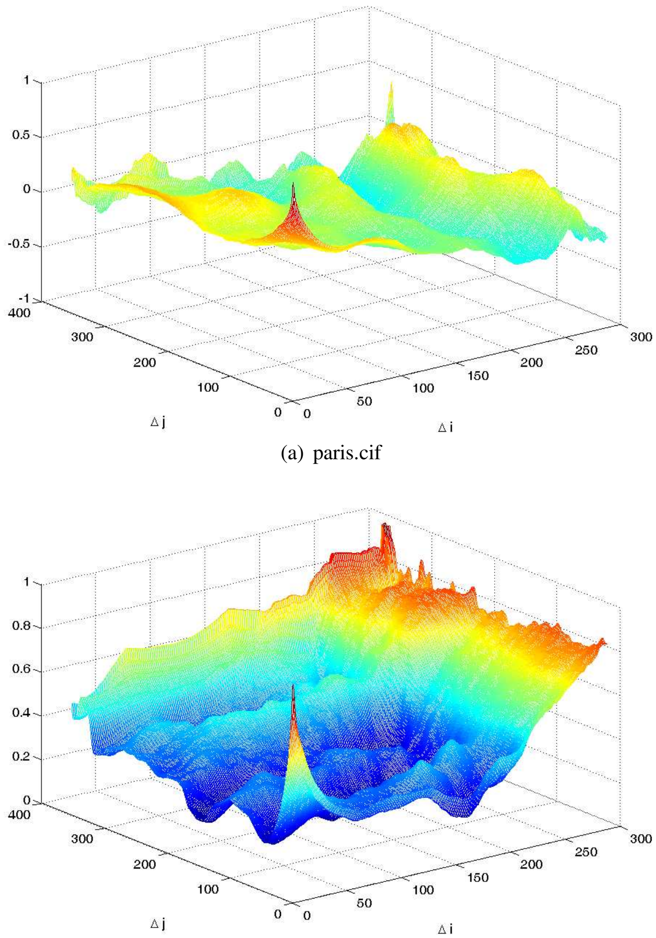
Figure 1 plots the approximated correlation coefficients $\hat{\rho}_s(\Delta i, \Delta j)$ of two digitized natural images, both of which are from two digitized natural video sequences, paris.cif and football.cif, respectively. For a digitized image let $X(i,j)$ denote its pixel value at the i^{th} row and the j^{th} column, and M and N denote the numbers of rows and columns, respectively, in the image. The approximated correlation coefficient $\hat{\rho}_s(\Delta i, \Delta j)$ of this image can be expressed as

$$\hat{\rho}_s(\Delta i, \Delta j) = \frac{1}{(M-\Delta i)(N-\Delta j)} \frac{\sum[X(i,j)X(i+\Delta i, j+\Delta j)]}{\sqrt{\sum[X^2(i,j)]\sum[X^2(i+\Delta i, j+\Delta j)]}}, \tag{II.3}$$

for $0 \le \Delta i \le M-1$, $0 \le \Delta j \le N-1$. The summations in (II.3) are taken over all pixels whose coordinates satisfy $0 \le i \le M-1-\Delta i$, $0 \le j \le N-1-\Delta j$. Calculated this way, $\hat{\rho}_s(\Delta i, \Delta j)$ is expected to approximate $\rho_s(\Delta i, \Delta j)$.

This research has been supported by the California Micro Program, Applied Signal Technology, Dolby Labs, Inc., Mindspeed, and Qualcomm, Inc., by NSF Grant Nos. CCF-0429884 and CNS-0435527, and by the UC Discovery Grant Program and Nokia, Inc..

However in Fig. 1 we can see that when Δi and Δj are larger than 50, which is much smaller than the image size we encounter in present applications, for example 352×288 in this figure, the approximated correlation coefficients $\hat{\rho}_s(\Delta i, \Delta j)$ are rather arbitrary 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 is far from being accurate.



(a) paris.cif

(b) football.cif

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

B. In the transformed domain for videos

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 [4]. If we use the absolute magnitude distortion measure $d(x, \hat{x}) = |x - \hat{x}|$, then there is a closed form rate distortion function which can be expanded into Taylor series and approximated by $R(D) \sim= aQ^{-1} + bQ^{-2}$. In this formula, the distortion is measured by the average quantization scale used in the frame.

This quadratic rate distortion function lays the foundation for the rate control schemes [5]–[7] that are adopted by the international video coding standards, such as ISO MPEG-4 and ITU-T H.263. 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 Laplacian model for DCT coefficients becomes less appropriate since emerging video compression schemes such as AVC/H.264 which offer significantly higher coding efficiency and better resilience to packet losses include a number of new schemes, such as 9 intra-frame prediction modes and different block sizes. Furthermore, the motion estimation and compensation scheme has evolved from full-pixel prediction based on only one previous frame, to quarter-pixel prediction with 6-tap FIR filter, based on multiple frames in both directions, with flexible weights among these prediction frames [1].

These new schemes and refinements stretch the Laplacian model of the DCT coefficients for two reasons. Firstly 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, 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. Secondly 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 has been to conduct optimization with a fixed QP. However, from the perspective of the rate control, the QP 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. A couple of recently proposed schemes following the same trend [8], [9] try to tackle this dilemma by either engaging a "two pass scheme" or defining a "basic unit". This is an ongoing research area and for more recent activities please refer to [10]. However a new and promising direction of solving this problem is to set up a statistical model of the video source in the spatial-temporal domain and the rate control and rate distortion optimization become a unified problem. This is the direction taken in our research and in this paper we focus on the correlation model and corresponding rate-distortion analysis in the spatial domain of video.

III. Definition of Block-Based Conditional Correlation Model

In this section we propose a new correlation model in the spatial domain of 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 2-D 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-A, 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 subtracted from the current block prior to encoding. For the luminance samples, P may be formed for each 4×4 sub-block or for a 16×16 MB. There are a total of 9 optional prediction modes for each 4×4 luminance block as shown in Fig. 2 and 4 optional prediction modes (mode 0 to 3 in Fig. 2) for a 16×16 luminance MB.

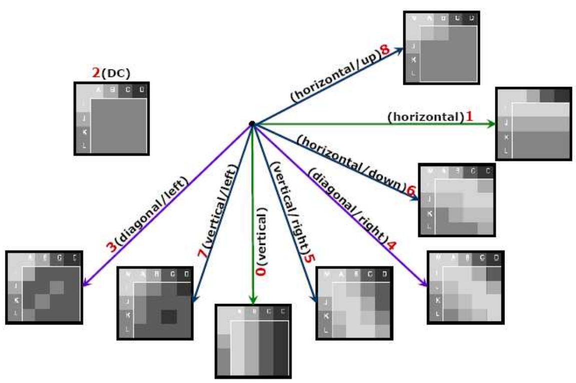


Fig. 2. The intra prediction modes for 4×4 blocks in AVC/H.264

Apart from its advantage in saving source bits, the intra-frame prediction reveals which one out of the 9 different local textures that are identified by their orientations (intra-modes) is the most similar to the texture of the current 4×4 block or 16×16 MB. It is reasonable to conjecture that the difference in local texture also affects the correlation between two close pixels. Let us look at Fig. 3 and focus on the loose surface (the mesh surface with less data points) in each subplot for the time being. In this figure we plot the approximate correlation coefficients of two pixel values $\hat{\rho}_s(\Delta i,\Delta j)$ in the image paris.cif according to the formula (II.3), averaged among the blocks that have the same intra-mode. We choose Δi and Δj to be very small to concentrate on the dependence of the statistics on local texture in an image.

Figure 3 shows that the average approximate correlation coefficients $\hat{\rho}_s(\Delta i,\Delta j)$ is very different among the blocks with different intra-modes. And for the blocks with one certain intra-mode, $\hat{\rho}_s(\Delta i,\Delta j)$ demonstrates a certain shape which agrees with the direction of the local texture. For example the intra-mode 0 in Fig. 2 has a vertical texture, correspondingly in Fig. 3 for the intra-mode 0, the approximate correlation coefficients are close to 1 for the same value of Δj, i.e., the pixels on the same column have similar values. If we average $\hat{\rho}_s(\Delta i,\Delta j)$ across all the blocks in the image, we will get what is shown in Fig. 1, but the important information about the local texture will be lost.

In the following we propose a new block-based correlation coefficient model which takes into account its dependence on the local texture. It is very important to point out that it is not necessary to use the intra-modes to identify the local texture, and there could be more than 9 categories for the local texture. We choose the intra-mode scheme that is adopted in the coding standard AVC/H.264 for its broad recognition.

Definition 3.1: The correlation coefficient of two pixel values with spatial offsets Δi and Δj is defined as

$$\begin{aligned}&\rho_s(\Delta i,\Delta j|Y_1=y_1,Y_2=y_2)\\ &=\tfrac{1}{2}(\rho_*(\Delta i,\Delta j|y_1)+\rho_*(\Delta i,\Delta j|y_2)),\end{aligned} \tag{III.4}$$

where

$$\rho_*(\Delta i,\Delta j|y)=a(y)+(1-a(y))e^{-|\alpha(y)\Delta i+\beta(y)\Delta j|^{\gamma(y)}}. \tag{III.5}$$

Y_1 and Y_2 are the intra-modes of the 4×4 blocks the two pixels are located in, respectively, and therefore they are integers between 0 and 8. The parameters a, α, β and γ are functions of the intra-mode Y.

It can be shown easily that this definition satisfies the 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 $\rho_s(0,0|Y_1=y_1,Y_2=y_2)=1$;
- $\rho_s(\Delta i,\Delta j|Y_1=y_1,Y_2=y_2)=\rho_s(-\Delta i,-\Delta j|Y_1=y_1,Y_2=y_2)$.

This correlation function discriminates all the 9 intra modes. Therefore the correlation between two pixel values in one image is dependent on the local texture where the two pixels are located. 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 four parameters a, α, β and γ.

The change of the four parameters a, α, β and γ from one block to another in the same frame, from one frame to another in the same scene and from one scene to another is currently under investigation. In this paper we choose the combination of the four parameters that jointly minimizes the mean absolute error (MAE) between the correlation coefficients approximated for a whole video frame and those calculated through the new model. These parameters for one frame in paris.cif and their corresponding MAE are presented in Table I. We can see from this table that the parameters associated with the new model are very distinct for different intra-modes while the MAE is always kept very small. In Fig. 3 we plot $\rho_*(\Delta i,\Delta j|y)$ of all the intra-modes for the same image paris.cif using these optimal parameters. 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 approximate correlation coefficients very well.

TABLE I

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

	a	γ	α	β	MAE
Mode0	0.3	0.5	0.0	1.2	0.027
Mode1	0.5	0.5	1.2	0.0	0.055
Mode2	0.7	0.2	0.0	-1.2	0.044
Mode3	0.7	0.5	-1.9	-1.0	0.057
Mode4	0.7	0.5	-1.0	1.9	0.047
Mode5	0.6	0.4	0.6	-1.8	0.037
Mode6	0.6	0.4	-1.9	0.6	0.033
Mode7	0.6	0.4	0.7	1.8	0.037
Mode8	0.7	0.6	-1.7	-0.4	0.057

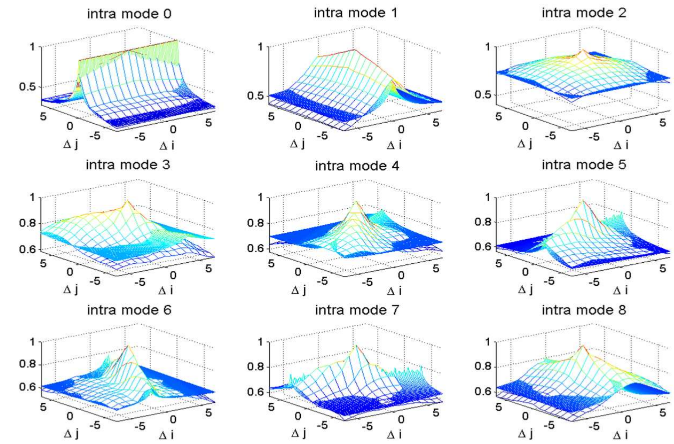


Fig. 3. The loose surfaces (the mesh surfaces with less data points) are the approximate correlation coefficients of two pixel values in the image paris.cif, averaged among the blocks that have the same intra-mode. The dense surfaces are the correlation coefficients calculated using the proposed conditional correlation model, along with the optimal set of parameters

IV. Rate-distortion functions and curves for one block and its surrounding pixels

In this section, we study the rate distortion curves and bounds for one 4×4 block and its surrounding pixels using the new correlation model in the spatial domain. The basic set up can be summarized in the block diagram in Fig. 4. $\underline{X}$ denotes the 4×4 block currently being processed. To simplify the derivation, $\underline{X}$ is considered as a vector source of length 16 instead of a two dimensional source of size 4 by 4. As discussed in Section III, the surrounding 13 pixels (9 on the top and 4 on the left, denoted by $\underline{S}$ of length 13) are used to form a prediction block for each one of the 9 intra modes, as

$$\underline{Z} = \underline{X} - P_d^{(A)} \underline{S}, \tag{IV.6}$$

where $P_d^{(a)}$ is a 16 by 13 matrix, different for each intra mode. A is the intra mode chosen for the current block which yields the smallest prediction error. $\underline{Z}$ and A are further coded and transmitted to the decoder, where inverse intra-prediction is performed, as

$$\hat{\underline{X}} = \hat{\underline{Z}} + P_d^{(\hat{A})} \hat{\underline{S}}. \tag{IV.7}$$

Y in Fig. 4 denotes the information of intra modes formulated from a collection of natural images and is considered as universal side information available to both the encoder and the decoder.

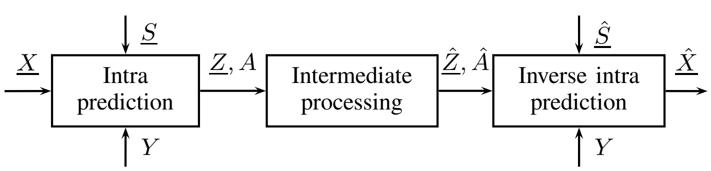


Fig. 4. Coding of one 4×4 block $\underline{X}$ based on the surrounding pixels $\underline{S}$

To study the minimum rate that can be achieved theoretically in video compression, we consider only the case where $\underline{S}$ and $\underline{X}$ are jointly coded with the average distortion constraint:

$$\begin{aligned} &\frac{1}{|\underline{S}| + |\underline{X}|} \left\{ E\left[\left\| \begin{pmatrix} \underline{X} \\ \underline{S} \end{pmatrix} - \begin{pmatrix} \hat{\underline{X}} \\ \hat{\underline{S}} \end{pmatrix} \right\|^2 \right] \right\} \\ &= \frac{1}{|\underline{S}| + |\underline{X}|} \left\{ E[\|\underline{S} - \hat{\underline{S}}\|^2] + E[\|\underline{X} - \hat{\underline{X}}\|^2] \right\} \leq D. \end{aligned} \tag{IV.8}$$

We study this case in two conditions: with or without the universal side information Y. Jointly coding $\underline{S}$ and $\underline{X}$ without the universal side information Y is the case normally studied in information theory. Therefore, we emphasize the latter condition where the universal side information Y is taken into account.

A. Not taking into account side information Y

This forms a straightforward rate-distortion problem of a source with memory which has been studied extensively. It can be expressed as

$$R_{withoutY}(D) = \min_{p(\hat{\underline{x}}, \hat{\underline{s}} | \underline{x}, \underline{s}):(IV.8)} \frac{1}{|\underline{S}| + |\underline{X}|} I(\underline{X}, \underline{S}; \hat{\underline{X}}, \hat{\underline{S}}). \tag{IV.9}$$

The correlation matrix is

$$K = E\left[\begin{pmatrix} \underline{X} \\ \underline{S} \end{pmatrix} \begin{pmatrix} \underline{X}^T & \underline{S}^T \end{pmatrix} \right] = \sum_{y=0}^{8} \sigma^2 \rho_s \left(\begin{pmatrix} \underline{X} \\ \underline{S} \end{pmatrix} | Y = y \right) P[Y = y], \tag{IV.10}$$

where the conditional correlation coefficients are exactly what the new model defines. The correlation matrix calculated using the above equation should be similar to those calculated from the old spatial correlation models (II.1) and (II.2).

The first order statistics of the universal side information Y, $P[Y = y]$, denotes the frequency of occurrence of each intra-mode, i.e., each texture in the natural images and videos. This information can be considered as available both at the encoder and decoder. In this section we plot all the rate-distortion functions and bounds for one luminance frame from each of the two video sequences paris.cif and football.cif as examples. Each pixel is coded using 8 bits. The variance of all the pixels in each image is calculated and employed. The new conditional correlation coefficient model (III.5) is used with the parameters shown in Table I. $P[Y = y]$ is calculated as the average over frames from ten natural video sequences commonly used as examples in video coding studies.

B. Taking into account side information Y

This forms a *conditional* rate-distortion problem of a source with memory:

$$\begin{aligned} R_{withY}(D) &= \min_{p(\hat{\underline{x}}, \hat{\underline{s}} | \underline{x}, \underline{s}, y):(IV.8)} \tfrac{1}{|\underline{S}| + |\underline{X}|} I(\underline{X}, \underline{S}; \hat{\underline{X}}, \hat{\underline{S}} | Y) \\ &= \min_{D_y : \sum_y D_y P[Y=y] \leq D} \textstyle\sum_y R_{\underline{X}, \underline{S} | Y=y}(D_y) P[Y = y]. \end{aligned} \tag{IV.11}$$

Because the proposed correlation model discriminates all the intra-prediction modes, we can calculate the marginal rate-distortion functions for all the intra-modes, $R_{\underline{X}, \underline{S} | Y=y}(D_y)$, as plotted in Fig. 5 for one frame in paris.cif. These plots show that the rate-distortion behavior for the blocks with different local texture, therefore different intra-modes, are very different. Without the conditional correlation coefficient model proposed in this paper, this difference cannot be calculated explicitly. The relative order of the 9 intra modes in terms of the average rate per pixel depends on the texture and the parameters associated with the correlation coefficient model for each intra-mode. For example, mode 2, which is DC prediction, consumes very little rate compared to other intra-modes.

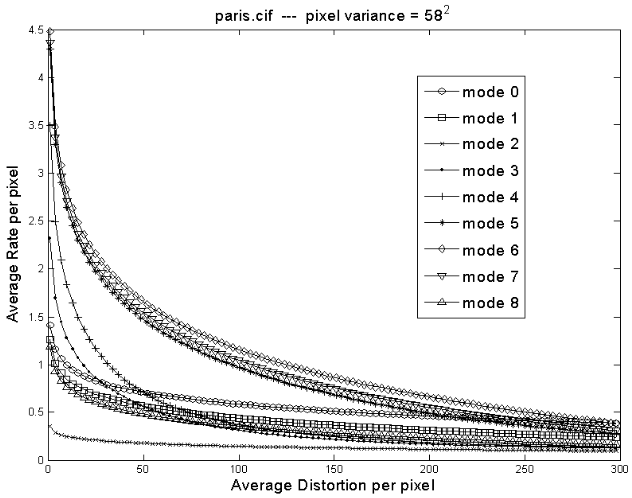
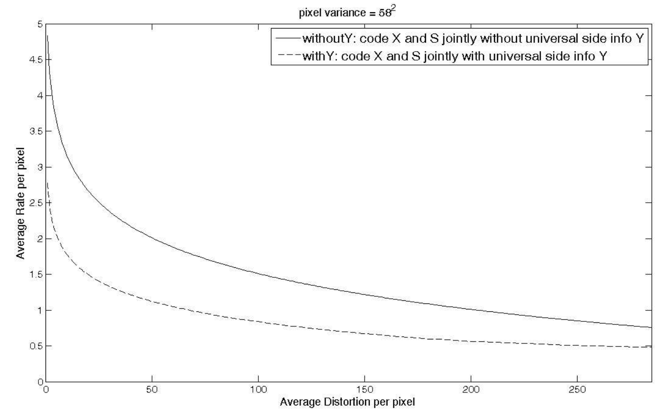


Fig. 5. Marginal rate-distortion functions of all the intra-modes, $R_{\underline{X},\underline{S}|Y=y}(D_y)$, for one frame in paris.cif

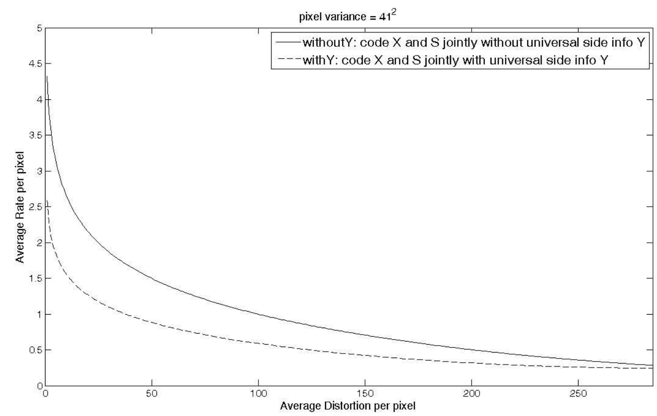
Utilizing the classical results for conditional rate distortion functions [11, Theorem 5], it can be proved that the minimum in the above equation (IV.11) is achieved at $D_y's$ where the slopes $\frac{\partial R_{\underline{X},\underline{S}|Y=y}(D_y)}{\partial D_y}$ are equal for all y and $\sum_y D_y P[Y = y] = D$. In Fig. 6 we plot this minimum $R_{withoutY}(D)$ as well as $R_{withY}(D)$ as solid lines and dashed lines, respectively. Comparing these two curves for both paris.cif and football.cif shows that engaging the first-order statistics of the universal side information Y saves as much as 2 bits per pixel at low distortion levels, which corresponds to about a reduction of 200 Kbits per frame for the CIF videos and 3 Mbps if the videos are played at a medium 15 frames per second. This difference fades quickly for football.cif but remains about quarter a bit per pixel for paris.cif at relatively higher distortion levels, about 350 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.

V. Conclusions

We propose a conditional correlation model for two close pixels in one frame of digitized natural video sequences, with the condition being the texture of the blocks where the two pixels are located. The blocksize is chosen as 4×4 and the local texture is categorized by the 9 AVC/H.264 intra-frame prediction modes in this paper. We further study the marginal rate-distortion function of the different local textures, i.e., different intra-modes. These marginal rate-distortion functions are shown to be very 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 "free" side information the lowest rate that is theoretically achievable in video compression can be as much as 2 bits per pixel lower than that without the side information. Future work includes the rate-distortion analysis of intra-frame prediction in the spatial domain and setting up a correlation coefficient model for two pixels that are not located in one frame.



(a) paris.cif



(b) football.cif

Fig. 6. Rate-distortion functions of the two conditions

References

[1] 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.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] Z. G. Li, L. Xiao, C. Zhu, and P. Feng, "A novel rate control scheme for video over the internet," *ICASSP'02, Florida, USA*, May 2002.

[10] 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.

[11] R. M. Gray, "Conditional rate-distortion theory," *Stanford Electron. Lab., Stanford, Calif.,*, vol. Tech. Rep. 6502-2, Oct. 1972.