Technical Report: MPEG-2 Prediction Residue Analysis

David Vázquez-Padín and Fernando Pérez-González

1 Introduction

This technical report complements the work in [1]. Based on the use of synthetic signals and autoregressive models to characterize temporal dependencies and inter predictions, here we build a semi-analytic model that explains the evolution of the prediction residue in the MPEG-2 double compression scenario assumed in [1]. As outlined in [1], this characterization provides valuable insights on the behavior of the Variation of Prediction Footprint (VPF), exploited by different methods (e.g., [2, 3]), and also shows that the performance of these VPF-based techniques for GOP size estimation and double compression detection depend on the deadzone width of the scalar quantizers used for encoding each type of MacroBlock (MB).

In the following, we first formulate and model the MPEG-2 video double quantization problem in Section 2 to keep this report self-contained. The distinct evolution of the variance of the inter-prediction residue is then analytically characterized in Section 3 through the use of the semi-analytic model, and finally, Section 4 concludes this report and hints at possible new research directions to be explored in the future.

2 Problem Formulation and Modeling

Let us consider a video double compression scenario where the two encodings are performed with the same MPEG-2 encoder. During the first compression, we assume that the input video sequence is compressed with a constant GOP of length G_1 and a fixed quantization parameter $\mathrm{Q}_1 \in \{2, \ldots, 31\}$. Similarly, the succeeding second compression is conducted with a GOP of length G_2 (different from any integer multiple of submultiple of G_1) and a fixed quantization parameter $\mathrm{Q}_2 \in \{2, \ldots, 31\}$. For the sake of simplicity, we assume that no temporal shift is introduced between both encodings and we discard the use of B-frames, leaving the analysis of bipredictive residues for a future work.

In MPEG-2, the MBs of an I-frame can only be encoded by means of a single intra-coding mode that does not perform any spatial/temporal prediction and is denoted by I-MB. In the case of P-frames, besides the use of I-MBs, two inter-coding modes are available to perform temporal (or motion-

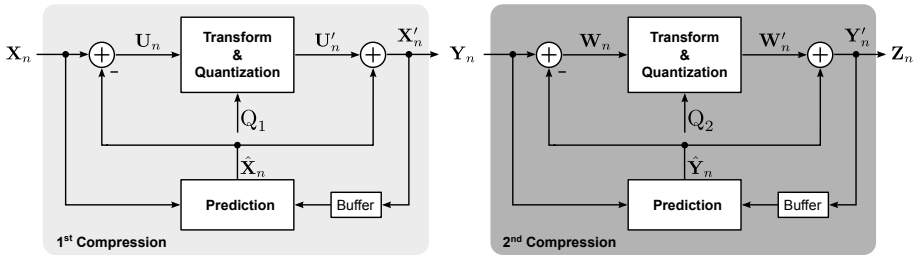


Figure 1: Double compression scheme: the left block diagram shows the first compression stage and, correspondingly, the right block depicts the structure of the second compression stage.

compensated) predictions from the last decoded frame: P-MB, which encodes the motion vector and the prediction residue, and S-MB, which efficiently signals those inter-predicted MBs that yield a zero-valued motion vector and null residual data. Accordingly, the set of available coding modes in this case is $\mathcal{C} \triangleq \{\text{I-MB, P-MB, S-MB}\}$. To identify the type of encoding a frame has undergone at a particular time index n during the first compression, we define the sets I_1 and P_1, which respectively contain the time indices of I- and P-frames.

Under this setting, the block diagram illustrated in Fig. 1 summarizes the main variables involved in the whole double encoding process. The left scheme in Fig. 1 models how a given MB at time index n, denoted by $\mathbf{X}_n$, is predicted based on a set of previously coded and reconstructed samples stored in a buffer.[1] Depending on the coding mode $\mathrm{c} \in \mathcal{C}$ selected by the encoder, the prediction $\hat{\mathbf{X}}_n$ is computed as

$$\hat{\mathbf{X}}_n = \begin{cases} 0, & \text{if c = I-MB} \\ \mathbf{X}'_{n-1}(\mathbf{m}), & \text{otherwise} \end{cases}, \tag{1}$$

where $\mathbf{X}'_{n-1}(\mathbf{m})$ denotes the MB extracted from the reference frame (previously decoded at time index $n-1$) with the relative displacement that the motion vector $\mathbf{m}$ points out. The first case in (1) reflects that no prediction is used for I-MBs, while the second case is valid for representing the motion-compensated prediction of P-MBs and also that of S-MBs, provided that $\mathbf{m} = (0, 0)$.

After the prediction, a residue is obtained as $\mathbf{U}_n = \mathbf{X}_n - \hat{\mathbf{X}}_n$, which is later transformed applying the Discrete Cosine Transform (DCT) on an 8×8 block-basis. In the DCT domain, each (i, j)-th coefficient with $i, j \in \{0, \ldots, 7\}$ is quantized with a distinct quantization step size and a configurable deadzone width. The quantization step size is modified by two mechanisms: a weighting matrix $\mathbf{S}$ to improve the perceptual quality of the encoded videos and a scale factor (controlled by the quantization parameter Q_1) that globally adapts the

[1] The position indices showing the location of the MB $\mathbf{X}_n$ within the frame are omitted for the sake of clarity.

size of each quantization step, having

$$\Delta_1(i,j) \triangleq \mathrm{Q}_1 \frac{S_{i,j}}{8}, \quad \forall i,j \in \{0,\ldots,7\}, \tag{2}$$

where $S_{i,j}$ represents the (i,j)-th element of the matrix $\mathbf{S}$. MPEG-2 supports the use of different quantization weighting matrices for intra- and inter-coding modes. As an example, the MPEG-2 encoder implementation from the FFmpeg library [4] uses by default the following weighting matrix $\mathbf{S}^{\mathrm{I}}$ for intra-coding modes

$$\mathbf{S}^{\mathrm{I}} = \begin{pmatrix} 8 & 16 & 19 & 22 & 26 & 27 & 29 & 34 \\ 16 & 16 & 22 & 24 & 27 & 29 & 34 & 37 \\ 19 & 22 & 26 & 27 & 29 & 34 & 34 & 38 \\ 22 & 22 & 26 & 27 & 29 & 34 & 37 & 40 \\ 22 & 26 & 27 & 29 & 32 & 35 & 40 & 48 \\ 26 & 27 & 29 & 32 & 35 & 40 & 48 & 58 \\ 26 & 27 & 29 & 34 & 38 & 46 & 56 & 69 \\ 27 & 29 & 35 & 38 & 46 & 56 & 69 & 83 \end{pmatrix},$$

whereas for inter-coding modes a weighting matrix $\mathbf{S}^{\mathrm{P}}$ is adopted with $S^{\mathrm{P}}_{i,j} = 16, \forall i,j \in \{0,\ldots,7\}$. Regarding the quantizer deadzone, its width is defined as a function of the applied quantization step, i.e.,

$$w_1(i,j) \triangleq \alpha \Delta_1(i,j), \tag{3}$$

where $\alpha \in [1,2]$ is the parameter that allows the control of the deadzone width. The use of a wider deadzone generally leads to a lower bitrate because more transform coefficients are quantized to zero, but in contrast a higher degree of distortion is introduced. Hence, in practice, the use of wider deadzones is recommended for small magnitude signals, such as the ones resulting from the use of inter-coding modes, whereas tighter deadzones are more convenient for intra-coding modes to retain more details in key reference frames. For instance, the MPEG-2 encoder in [4] uses by default different values of α for intra- and inter-coding modes, namely: $\alpha_{\mathrm{I}} = \frac{5}{4}$ and $\alpha_{\mathrm{P}} = 2$, respectively. Further in this work, the impact on the variance of the prediction residue in the second compression stage will be analyzed for $\alpha_{\mathrm{P}} = 2$ and different values of α_{I}, keeping them constant across both compressions.

Now, using (2) and (3), the quantization of a given AC coefficient u from the DCT of $\mathbf{U}_n$ can be written (omitting the position indices) as

$$u_q \triangleq \begin{cases} \mathrm{sgn}(u) \left\lfloor \frac{|u| + \Delta_1^{\mathrm{I}}\left(1 - \frac{\alpha_{\mathrm{I}}}{2}\right)}{\Delta_1^{\mathrm{I}}} \right\rfloor, & \text{if c = I-MB} \\ \mathrm{sgn}(u) \left\lfloor \frac{|u| + \Delta_1^{\mathrm{P}}\left(1 - \frac{\alpha_{\mathrm{P}}}{2}\right)}{\Delta_1^{\mathrm{P}}} \right\rfloor, & \text{otherwise} \end{cases}, \tag{4}$$

where $|\cdot|$ is the absolute value operator, $\lfloor \cdot \rfloor$ denotes the floor function, and $\mathrm{sgn}(\cdot)$ represents the sign function which returns -1 for a negative number, 0

for the number zero, and +1 for a positive number. The notation Δ_1^{I} and Δ_1^{P} has been used to remark that different quantization steps can be employed in each coding mode, depending on which quantization weighting matrix $\mathbf{S}^{\mathrm{I}}$ or $\mathbf{S}^{\mathrm{P}}$ is used, respectively.

According to the MPEG-2 standard, the de-quantized version u' of a quantized AC coefficient u_q is given by

$$u' = \begin{cases} \operatorname{sgn}(u_q)\left\lfloor \Delta_1^{\mathrm{I}} |u_q| \right\rfloor, & \text{if c=I-MB} \\ \operatorname{sgn}(u_q)\left\lfloor \Delta_1^{\mathrm{P}} |u_q| + \frac{\Delta_1^{\mathrm{P}}}{2} \right\rfloor, & \text{otherwise} \end{cases}, \tag{5}$$

where a specific process is followed depending on the applied coding mode, i.e., the reconstructed values for I-MBs are distributed on an equally spaced grid (determined by Δ_1^{I}), while for inter-coded MBs the first nonzero reconstructed value is shifted a distance of $\frac{3}{2}\Delta_1^{\mathrm{P}}$ from zero. The use of different rules for de-quantization contributes to improving the coding performance, and as we will further see in Section 3, it also affects the variance of the prediction residue in the second compression stage. As a last step in the reconstruction process, the samples in the pixel domain $\mathbf{X}'_n$ are recovered by adding back the de-quantized and inverse transformed samples $\mathbf{U}'_n$ to the prediction $\hat{\mathbf{X}}_n$, such that $\mathbf{X}'_n = \mathbf{U}'_n + \hat{\mathbf{X}}_n$.

The above description straightforwardly extends to the second compression block on the right of Fig. 1: the source and predicted samples are denoted by $\mathbf{Y}_n$ and $\hat{\mathbf{Y}}_n$, respectively, the residue signal by $\mathbf{W}_n$ and its reconstructed version by $\mathbf{W}'_n$; in this case, the quantization parameter is denoted by Q_2, the quantization steps by Δ_2^{I} and Δ_2^{P}, and the recovered samples are accordingly represented by $\mathbf{Y}'_n$.

To make this problem analytically tractable, the upcoming analysis will be theoretically supported by focusing on a single DCT coefficient (i.e., the one at position $(1,0)$) and assuming that the input samples from $\mathbf{X}_n$ in the DCT domain follow a Laplacian distribution with mean μ_X and variance σ_X^2 (the rationale behind the use of the Laplacian model is provided in [5]). Certainly, this model is too simplistic and lossy, but the theoretical conclusions derived from it will be applicable on a large set of real video sequences.

3 Prediction Residue Analysis

The use of de-synchronized GOPs in a double encoding scheme causes the VPF effect unveiled in [2], which leads to periodic changes in the distribution of certain MB types in double compressed videos, specifically, at P-frames that were originally encoded as I-frames. In view of the straight connection between the presence of the VPF and the MB type selection process implemented by the encoder, we focus on the nowadays most common strategy for MB coding-mode selection, which is based on Lagrangian optimization [6] and consists in solving

the following minimization problem

$$\text{MB type} = \arg\min_{\mathrm{c}\in\mathcal{C}} D(\mathbf{Y}_n, \mathbf{Y}'_n) + \lambda_{\mathrm{c}} R(\mathbf{Y}'_n), \tag{6}$$

where λ_{c} denotes the Lagrange multiplier of the coding mode $\mathrm{c}\in\mathcal{C}$, the distortion $D(\mathbf{Y}_n, \mathbf{Y}'_n)$ is the Sum of Squared Differences (SSD) between the reconstructed block $\mathbf{Y}'_n$ and its source $\mathbf{Y}_n$, and the rate $R(\mathbf{Y}'_n)$ measures the number of required bits to reconstruct $\mathbf{Y}'_n$. From (6) and since

$$D(\mathbf{Y}_n, \mathbf{Y}'_n) \triangleq \|\mathbf{Y}_n - \mathbf{Y}'_n\|_2^2 = \|\mathbf{W}_n + \hat{\mathbf{Y}}_n - (\mathbf{W}'_n + \hat{\mathbf{Y}}_n)\|_2^2 = \|\mathbf{W}_n - \mathbf{W}'_n\|_2^2,$$

we know that the selection of a particular MB type depends on the variance of the prediction residue. So, to predict the strength of the VPF on those P-frames originally encoded as I-frames, we need to analyze the evolution of the difference of the variance $\mathrm{Var}\,(\mathbf{W}_n)$ under $n\in\mathsf{I}_1$ and $n\in\mathsf{P}_1$, i.e.,

$$\mathrm{Var}\,(\mathbf{W}_n)\,|_{n\in\mathsf{I}_1} - \mathrm{Var}\,(\mathbf{W}_n)\,|_{n\in\mathsf{P}_1}, \tag{7}$$

where a larger difference value yields a stronger VPF. In the following, we carry out a comprehensive analysis of (7), particularizing the cases in which solely predictions of type intra (Sect. 3.1) or inter (Sect. 3.2) are used during the second compression.

Let us characterize the prediction residue $\mathbf{W}_n = \mathbf{Y}_n - \hat{\mathbf{Y}}_n$ by first describing the input signal $\mathbf{Y}_n$, which can be expressed as

$$\begin{aligned}\mathbf{Y}_n &= \mathbf{X}'_n \\ &= \mathbf{U}'_n + \hat{\mathbf{X}}_n \\ &= \mathbf{X}_n + (\mathbf{U}'_n - \mathbf{U}_n)\,,\end{aligned}$$

where the relation $\hat{\mathbf{X}}_n = \mathbf{X}_n - \mathbf{U}_n$ has been used in the last step. The above equation shows that the input signal at the second compression stage can be seen as the source signal $\mathbf{X}_n$ with an added quantization error $(\mathbf{U}'_n - \mathbf{U}_n)$ that depends on the selected type of frame (and the correspondingly applied coding modes) during the first compression.

When $n \in \mathsf{I}_1$, only I-MBs can be used to encode $\mathbf{X}_n$, which implies that $\mathbf{U}'_n - \mathbf{U}_n = \mathbf{X}'_n - \mathbf{X}_n$, since from (1) we have that $\mathbf{U}_n = \mathbf{X}_n$ and $\mathbf{U}'_n = \mathbf{X}'_n$. This particular quantization error is denoted by $\mathbf{E}_n^{\mathrm{I}_1} \triangleq \mathbf{X}'_n - \mathbf{X}_n$. On the other hand, when $n \in \mathsf{P}_1$, we have that $\mathbf{E}_n^{\mathrm{P}_1} \triangleq \mathbf{U}'_n - \mathbf{U}_n$. In consequence, the input signal can be rewritten as

$$\mathbf{Y}_n = \begin{cases} \mathbf{X}_n + \mathbf{E}_n^{\mathrm{I}_1}, & \text{if } n \in \mathsf{I}_1 \\ \mathbf{X}_n + \mathbf{E}_n^{\mathrm{P}_1}, & \text{if } n \in \mathsf{P}_1 \end{cases}. \tag{8}$$

The subsequent sections separately describe the prediction $\hat{\mathbf{Y}}_n$ as a function of the two coding modes, i.e., intra or inter, that can be applied under the second compression.

3.1 Intra-prediction residue analysis

The use of an intra-coding mode during the second compression yields $\hat{\mathbf{Y}}_n = 0$, such that $\mathbf{W}_n = \mathbf{Y}_n$. Hence, from (8), the variance of the resulting prediction residue can be expressed as

$$\mathrm{Var}\left(\mathbf{W}_n\right) = \begin{cases} \mathrm{Var}\left(\mathbf{X}_n\right) + \mathrm{Var}\left(\mathbf{E}_n^{\mathrm{I}_1}\right), & \text{if } n \in \mathsf{I}_1 \\ \mathrm{Var}\left(\mathbf{X}_n\right) + \mathrm{Var}\left(\mathbf{E}_n^{\mathrm{P}_1}\right), & \text{if } n \in \mathsf{P}_1 \end{cases}, \tag{9}$$

where we assume that the quantization errors $\mathbf{E}_n^{\mathrm{I}_1}$ and $\mathbf{E}_n^{\mathrm{P}_1}$ have negligible correlation with the source signal $\mathbf{X}_n$. This assumption typically holds whenever the probability density function (pdf) of the source signal is smooth and its variance is much larger than the employed quantization step sizes, which is generally the case in practice. By inserting the relationship (9) in (7), the strength of the VPF can be evaluated in this case by means of

$$\mathrm{Var}\left(\mathbf{E}_n^{\mathrm{I}_1}\right) - \mathrm{Var}\left(\mathbf{E}_n^{\mathrm{P}_1}\right).$$

The variance of both quantization errors can be analytically described conforming to the model discussed in Section 2, as follows:

1. $\mathrm{Var}\left(\mathbf{E}_n^{\mathrm{I}_1}\right)$: given the definition of $\mathbf{E}_n^{\mathrm{I}_1}$, its variance is proportional (except for a constant normalization factor) to the distortion between the reconstructed samples $\mathbf{X}'_n$ and their source $\mathbf{X}_n$ when the SSD measure is considered, thus having

   $$\mathrm{Var}\left(\mathbf{E}_n^{\mathrm{I}_1}\right) \propto D\left(\mathbf{X}_n, \mathbf{X}'_n\right) \triangleq \|\mathbf{X}_n - \mathbf{X}'_n\|_2^2.$$

   Since the DCT adopted in MPEG-2 has orthogonal basis (discarding rounding effects in the DCT calculation), the above distortion can be directly computed in the transformed domain as

   $$D\left(\mathbf{X}_n, \mathbf{X}'_n\right) = \sum_{i=0}^{7}\sum_{j=0}^{7} D_X(i,j),$$

   where $D_X(i,j)$ represents the distortion of the (i,j)-th DCT coefficient. In line with the quantization model introduced in Section 2, the values of $D_X(i,j)$ (except for the DC coefficient) are given by

   $$\begin{aligned} D_X(i,j) = &\int_{-\frac{\alpha_{\mathrm{I}}}{2}\Delta_1^{\mathrm{I}}}^{\frac{\alpha_{\mathrm{I}}}{2}\Delta_1^{\mathrm{I}}} x^2 f_X(x)\, dx \\ &+ 2\sum_{k=1}^{\infty}\int_{(k-1+\frac{\alpha_{\mathrm{I}}}{2})\Delta_1^{\mathrm{I}}}^{(k+\frac{\alpha_{\mathrm{I}}}{2})\Delta_1^{\mathrm{I}}} \left(x - \left\lfloor k\Delta_1^{\mathrm{I}} \right\rfloor\right)^2 f_X(x)\, dx, \end{aligned} \tag{10}$$

   where $f_X(x)$ denotes the pdf of the corresponding AC coefficient which, as discussed at the end of Section 2, can be modeled as a Laplacian distribution with mean μ_X and variance σ_X^2.

From the above equation we observe that for a fixed value of σ_X^2, the evolution of $\mathrm{Var}\left(\mathbf{E}_n^{\mathrm{I}_1}\right)$ mostly depends on the deadzone width (configured through the parameter α_{I}) and on the quantization step Δ_1^{I} (controlled by the quantization parameter Q_1 as in (2)). More specifically, we can state that larger values of α_{I} (or, equivalently, wider deadzones) yield larger values of $\mathrm{Var}\left(\mathbf{E}_n^{\mathrm{I}_1}\right)$, which also increases with Q_1, since larger values of Q_1 produce coarser quantization steps Δ_1^{I}.

2. $\mathrm{Var}\left(\mathbf{E}_n^{\mathrm{P}_1}\right)$: the variance of the quantization error $\mathbf{E}_n^{\mathrm{P}_1}$ satisfies the following relation

$$\begin{aligned}\mathrm{Var}\left(\mathbf{E}_n^{\mathrm{P}_1}\right) \propto\ & \mathrm{p(I\text{-}MB)}D(\mathbf{X}_n, \mathbf{X}'_n) + \mathrm{p(P\text{-}MB)}D(\mathbf{U}_n, \mathbf{U}'_n) \\ & + \mathrm{p(S\text{-}MB)}D(\mathbf{X}_n, \mathbf{X}'_{n-1}),\end{aligned}$$

where p(c) denotes the probability of using the coding mode $\mathrm{c} \in \mathcal{C}$ per frame. Regarding the distortion terms, both $D(\mathbf{X}_n, \mathbf{X}'_n)$ and $D(\mathbf{X}_n, \mathbf{X}'_{n-1})$ can be computed as in the previous case through (10),[2] whereas $D(\mathbf{U}_n, \mathbf{U}'_n)$ can be obtained by accumulating the distortion of each (i,j)-th DCT coefficient, i.e., $D(\mathbf{U}_n, \mathbf{U}'_n) = \sum_{i=0}^{7}\sum_{j=0}^{7} D_U(i,j)$, where $D_U(i,j)$ is given by

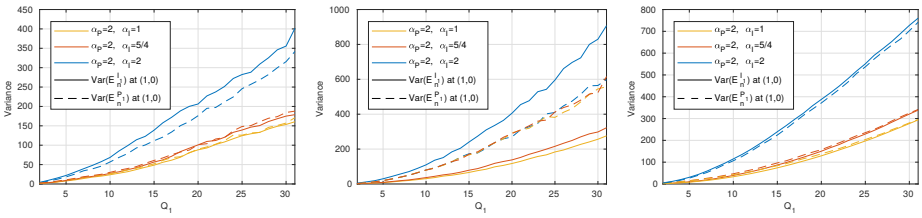
$$\begin{aligned}D_U(i,j) = & \int_{-\frac{\alpha_{\mathrm{P}}}{2}\Delta_1^{\mathrm{P}}}^{\frac{\alpha_{\mathrm{P}}}{2}\Delta_1^{\mathrm{P}}} u^2 f_U(u)\ du \\ & + 2\sum_{k=1}^{\infty}\int_{(k-1+\frac{\alpha_{\mathrm{P}}}{2})\Delta_1^{\mathrm{P}}}^{(k+\frac{\alpha_{\mathrm{P}}}{2})\Delta_1^{\mathrm{P}}} \left(u - \left\lfloor k\Delta_1^{\mathrm{P}} + \frac{\Delta_1^{\mathrm{P}}}{2}\right\rfloor\right)^2 f_U(u)\ du.\end{aligned}$$

In the above equation, $f_U(u)$ denotes the pdf of the (i,j)-th DCT coefficient which can be modeled as a Laplacian distribution with zero mean and variance σ_U^2 [7].

Similarly to the previous case, $\mathrm{Var}\left(\mathbf{E}_n^{\mathrm{P}_1}\right)$ also monotonically increases with Q_1, although now the effect of the deadzone width is a function of both parameters α_{I} and α_{P}. The influence of each deadzone width will depend on the type of scene to be encoded, since its content will finally rule the probability of using each type of MB. For instance, in static scenes, it is common to have $\mathrm{p(S\text{-}MB)} \gg \mathrm{p(P\text{-}MB)} + \mathrm{p(I\text{-}MB)}$, so $\mathrm{Var}\left(\mathbf{E}_n^{\mathrm{P}_1}\right)$ will increase with Q_1 and α_{I}, and will not be excessively affected by α_{P}, whereas in dynamic scenes, where almost all the coded MBs have a non-zero motion vector with $\mathbf{U}'_n \neq 0$, such that $\mathrm{p(P\text{-}MB)} \gg \mathrm{p(S\text{-}MB)} + \mathrm{p(I\text{-}MB)}$, the evolution of $\mathrm{Var}\left(\mathbf{E}_n^{\mathrm{P}_1}\right)$ will be mostly governed by α_{P} instead of α_{I}.

The above findings can be checked in Fig. 2, where the evolution of $\mathrm{Var}\left(\mathbf{E}_n^{\mathrm{I}_1}\right)$ and $\mathrm{Var}\left(\mathbf{E}_n^{\mathrm{P}_1}\right)$ for the $(1,0)$-th DCT coefficient is shown for two videos gathered from [8]: the static video *akiyo* in Fig. 2(a), and the dynamic video *mobile* in

[2] Note that for $D(\mathbf{X}_n, \mathbf{X}'_{n-1})$, only an approximation would be obtained through (10), but still valid in practice since $\mathbf{X}_n \approx \mathbf{X}_{n-1}$ for S-MBs.



(a) (b) (c)

Figure 2: Evolution of $\mathrm{Var}(\mathbf{E}_n^{\mathrm{I}_1})$ (*solid*) and $\mathrm{Var}(\mathbf{E}_n^{\mathrm{P}_1})$ (*dashed*) for $\alpha_{\mathrm{P}} = 2$ and varying α_{I} and Q_1: (a) static video (`akiyo`), (b) dynamic video (`mobile`), (c) synthetic model for static video ($\sigma_X^2 = 2500$).

Fig. 2(b). In the static case reported in Fig. 2(a), the evolution of $\mathrm{Var}\left(\mathbf{E}_n^{\mathrm{P}_1}\right)$ at $(1,0)$ mostly depends on α_{I} due to the role of the S-MBs that fundamentally make reference to MBs from the previously encoded I-frame, thus yielding a quantization error $\mathbf{E}_n^{\mathrm{P}_1}$ very similar to $\mathbf{E}_n^{\mathrm{I}_1}$. On the other hand, when dealing with dynamic scenes as in Fig. 2(b), $\mathrm{Var}\left(\mathbf{E}_n^{\mathrm{P}_1}\right)$ at $(1,0)$ does not depend on α_{I}, instead it is governed by α_{P} given that the curves for different values of α_{I} are close to each other.

From the above analysis, it follows that for static video sequences, the difference $\mathrm{Var}\left(\mathbf{E}_n^{\mathrm{I}_1}\right) - \mathrm{Var}\left(\mathbf{E}_n^{\mathrm{P}_1}\right)$ is small and so $\mathrm{Var}\left(\mathbf{W}_n\right)|_{n\in\mathsf{I}_1} \approx \mathrm{Var}\left(\mathbf{W}_n\right)|_{n\in\mathsf{P}_1}$, while for dynamic videos it is harder to define a similar relation. In fact, the varying nature of the prediction residue with dynamic videos complicates the modeling (i.e., at least a motion estimation of the scene would be needed), thus we leave its study for a future work. In constrast, the nearly constant behavior of $\mathrm{Var}\left(\mathbf{W}_n\right)$ under the intra prediction (independently of the type of frame used in the first compression), implies that for low-motion videos, the presence of the VPF is ultimately guided by the behavior of $\mathrm{Var}\left(\mathbf{W}_n\right)$ under the inter prediction, which we analyze in the following Section 3.2.

Prior to address the evolution of the residue under the inter-prediction setting, we introduce the proposed semi-analytic model to predict the behavior of $\mathrm{Var}\left(\mathbf{E}_n^{\mathrm{I}_1}\right)$ and $\mathrm{Var}\left(\mathbf{E}_n^{\mathrm{P}_1}\right)$ for static video sequences. To that end, we use a first-order autoregressive process x_n for modeling the temporal dependencies of the source signal $\mathbf{X}_n$. Troughout the paper we will consider three distinct time indices: $n-2$, $n-1$, and n, and we will correspondingly define three stochastic processes: x_{n-2}, x_{n-1}, and x_n. Thus, assuming the stochastic process x_{n-2} follows a Laplacian distribution with zero mean and variance $\sigma_X^2 = 2500$, we generate the remaining stochastic processes as $x_{n-1} = \rho x_{n-2} + r_{n-1}$ and $x_n = \rho x_{n-1} + r_n$, where we take $\rho = 0.99$ to simulate the high temporal correlation between adjacent frames in low-motion videos and we define the residue signal r_n as a zero-mean Gaussian process with variance $\sigma_R^2 = 10$. On the other hand, the inter-prediction process is also characterized through a first-order autoregressive process, such that $\hat{\mathbf{X}}_n = \mathbf{X}'_{n-1}(\mathbf{m})$ is modeled through the stochastic process $\hat{x}_n = \rho_{\mathrm{P}} x'_{n-1} + \nu_n$, where we take $\rho_{\mathrm{P}} = 0.88$ to simulate the effect of a non-

perfect motion estimation and we make use of a zero-mean Gaussian process ν_n with variance $\sigma_\nu = 1$ to mimic the effect of the motion compensation through $\mathbf{m}$. During the first compression, we assume that at time index $n-2$ there is always an I-frame, then at $n-1$ a P-frame, and finally at n, it will depend on the case under study, i.e., $n \in \mathsf{I}_1$ or $n \in \mathsf{P}_1$. For the second compression, an I-frame is also assumed at $n-2$, and a P-frame at $n-1$.

Now, from the definition of $\mathbf{E}_n^{\mathrm{I}_1}$, we generate the samples of the corresponding synthetic signal as $e_n^{\mathrm{I}_1} = x'_n - x_n$, where all the samples from x_n are first quantized as in (4) for I-MBs (using Δ_1^{I} and α_{I}) and then are accordingly reconstructed through (5) to obtain x'_n. On the other hand, the synthetic signal for $\mathbf{E}_n^{\mathrm{P}_1}$, must consider the different coding modes from $\mathcal{C}$ that can be used when a P-frame is encoded. Without loss of generality, we assume that the number of I-MBs in static videos is negligible, such that p(I-MB) $\to 0$, which is commonly the case in practice. Regarding the modeling of P-MBs, we assume that p(P-MB) decreases as Q_1 grows (which is the expected behavior in low-motion videos), so we use p(P-MB) $= p_1$ with $p_1 \triangleq 0.15 + 0.7e^{-9\frac{\mathrm{Q}_1}{\mathrm{Q}_{\max}}}$, where $\mathrm{Q}_{\max} = 31$ denotes the maximum allowed quantization parameter. Accordingly, we set p(S-MB) $= 1 - p_1$ and we finally compute the synthetic signal for $\mathbf{E}_n^{\mathrm{P}_1}$ as

$$e_n^{\mathrm{P}_1} = \begin{cases} u'_n - u_n, & \text{if c = P-MB} \\ x'_{n-1} - x_n, & \text{otherwise} \end{cases},$$

where $u_n = x_n - \hat{x}_n$ and u'_n is obtained by first quantizing u_n as in the second case of (4) and then reconstructing its samples through (5) (in both cases, using Δ_1^{P} and α_{P}). Finally, to obtain x'_{n-1}, the whole encoding process of a P-frame must be applied, so with probability $1 - p_1$, the samples of x'_{n-1} stem from the use of the first case in (4)-(5) over the samples of x_{n-1} (using Δ_1^{I} and α_{I}). On the other hand, with probability p_1, the corresponding signal $u_{n-1} = x_{n-1} - \hat{x}_{n-1}$ must be computed, where $\hat{x}_{n-1}$ is given through $\hat{x}_{n-1} = \rho_{\mathrm{P}} x'_{n-2} + \nu_{n-1}$ with x'_{n-2} being the result of reconstructing the whole signal x_{n-2} as the encoding of an I-frame. Once u_{n-1} is computed, then the second case in (4) and (5) must be applied to generate u'_{n-1}, which finally allows the calculation of $x'_{n-1} = u'_{n-1} + \hat{x}_{n-1}$.

The result of using this semi-analytic approach to predict the behavior of $\mathrm{Var}\left(\mathbf{E}_n^{\mathrm{I}_1}\right)$ and $\mathrm{Var}\left(\mathbf{E}_n^{\mathrm{P}_1}\right)$ in the DCT domain at $(1,0)$ for static video sequences is illustrated in Fig. 2(c), which closely resembles its empirical counterpart in Fig. 2(a).

3.2 Inter-prediction residue analysis

In this case, $\hat{\mathbf{Y}}_n$ is the result of an inter prediction, i.e., $\hat{\mathbf{Y}}_n = \mathbf{Y}'_{n-1}(\mathbf{m})$. Assuming that the estimated motion scene through $\mathbf{m}$ coincides in the two consecutive compressions (which is reasonable in practice, provided that the content of the scene remains unchanged across both compressions), $\hat{\mathbf{Y}}_n$ can be

expressed as

$$\begin{aligned}\hat{\mathbf{Y}}_n &= \mathbf{Y}_{n-1}(\mathbf{m}) + \mathbf{E}_{n-1}^{\mathsf{P}_2} \\ &= \mathbf{X}_{n-1}(\mathbf{m}) + \mathbf{E}_{n-1}^{\mathsf{P}_1} + \mathbf{E}_{n-1}^{\mathsf{P}_2},\end{aligned} \tag{11}$$

where $\mathbf{E}_{n-1}^{\mathsf{P}_1} \triangleq \mathbf{U}'_{n-1}(\mathbf{m}) - \mathbf{U}_{n-1}(\mathbf{m})$ and $\mathbf{E}_{n-1}^{\mathsf{P}_2} \triangleq \mathbf{W}'_{n-1}(\mathbf{m}) - \mathbf{W}_{n-1}(\mathbf{m})$ represent the quantization errors that result from the first and second compression, respectively. Using (8) and (11) in the definition of $\mathbf{W}_n$, we obtain

$$\mathbf{W}_n = \mathbf{R}_n + \mathbf{E}_n, \quad \text{with} \quad \mathbf{E}_n = \begin{cases} \mathbf{E}_n^{\mathsf{I}_1} - \mathbf{E}_{n-1}^{\mathsf{P}_1} - \mathbf{E}_{n-1}^{\mathsf{P}_2}, & \text{if } n \in \mathsf{I}_1 \\ \mathbf{E}_n^{\mathsf{P}_1} - \mathbf{E}_{n-1}^{\mathsf{P}_1} - \mathbf{E}_{n-1}^{\mathsf{P}_2}, & \text{if } n \in \mathsf{P}_1 \end{cases},$$

where $\mathbf{R}_n \triangleq \mathbf{X}_n - \mathbf{X}_{n-1}(\mathbf{m})$ represents the prediction residue without any quantization error and $\mathbf{E}_n$ comprises all the quantization errors that emerge during the two successive compressions. Now, assuming that these quantization errors have negligible correlation with $\mathbf{R}_n$, we can approximate the variance of $\mathbf{W}_n$ as

$$\mathrm{Var}\left(\mathbf{W}_n\right) = \mathrm{Var}\left(\mathbf{R}_n\right) + \mathrm{Var}\left(\mathbf{E}_n\right).$$

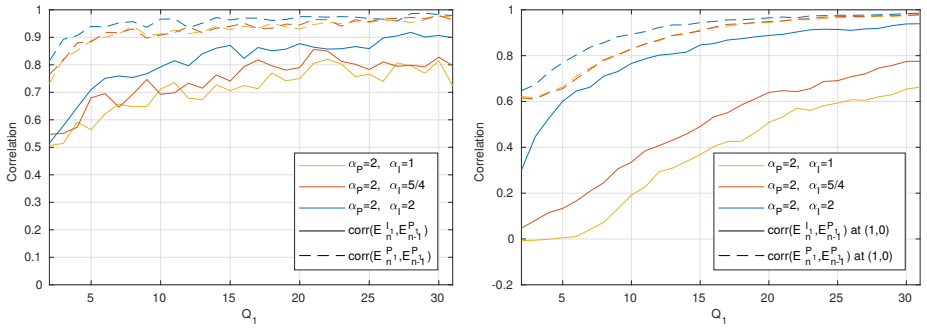
Therefore, in this case, (7) becomes $\mathrm{Var}\left(\mathbf{E}_n\right)|_{n\in\mathsf{I}_1} - \mathrm{Var}\left(\mathbf{E}_n\right)|_{n\in\mathsf{P}_1}$, which after deriving the expression for $\mathrm{Var}\left(\mathbf{E}_n\right)$, can be expressed as

$$\begin{aligned}&\mathrm{Var}\left(\mathbf{E}_n\right)|_{n\in\mathsf{I}_1} - \mathrm{Var}\left(\mathbf{E}_n\right)|_{n\in\mathsf{P}_1} \\ &\quad= \underbrace{\mathrm{Var}\left(\mathbf{E}_n^{\mathsf{I}_1}\right) - \mathrm{Var}\left(\mathbf{E}_n^{\mathsf{P}_1}\right) - 2\left(\mathrm{cov}\left(\mathbf{E}_n^{\mathsf{I}_1}, \mathbf{E}_{n-1}^{\mathsf{P}_1}\right) - \mathrm{cov}\left(\mathbf{E}_n^{\mathsf{P}_1}, \mathbf{E}_{n-1}^{\mathsf{P}_1}\right)\right)}_{\text{depends on } \mathrm{Q}_1} \\ &\quad\quad \underbrace{-2\left(\mathrm{cov}\left(\mathbf{E}_n^{\mathsf{I}_1}, \mathbf{E}_{n-1}^{\mathsf{P}_2}\right) - \mathrm{cov}\left(\mathbf{E}_n^{\mathsf{P}_1}, \mathbf{E}_{n-1}^{\mathsf{P}_2}\right)\right)}_{\text{depends on } \mathrm{Q}_1 \text{ and } \mathrm{Q}_2}.\end{aligned} \tag{12}$$

In the above equation, some of the terms uniquely depend on the quantization parameter used during the first compression Q_1, whereas the rest depends on the two quantization parameters Q_1 and Q_2 applied in the double compression scheme. We first center our attention on the terms that only depend on Q_1, except for $\mathrm{Var}\left(\mathbf{E}_n^{\mathsf{I}_1}\right)$ and $\mathrm{Var}\left(\mathbf{E}_n^{\mathsf{P}_1}\right)$, which have already been described in Section 3.1. Without loss of generality, we analyze the behavior of the two covariance functions through their corresponding correlations:[3]

1. $\mathrm{cov}\left(\mathbf{E}_n^{\mathsf{I}_1}, \mathbf{E}_{n-1}^{\mathsf{P}_1}\right)$: as hinted in Section 3.1, because of the large amount of S-MBs that show up in static videos, the correlation between $\mathbf{E}_n^{\mathsf{I}_1}$ and $\mathbf{E}_{n-1}^{\mathsf{P}_1}$ is expected to be significant. Moreover, the larger the value of Q_1, the higher becomes the number of S-MBs and, as a consequence, $\mathrm{corr}\left(\mathbf{E}_n^{\mathsf{I}_1}, \mathbf{E}_{n-1}^{\mathsf{P}_1}\right)$ grows with Q_1. This behavior can be observed in Fig. 3(a),

[3]Note that for two arbitrary random variables A and B, their covariance and correlation relate as follows: $\mathrm{corr}\left(A, B\right) = \mathrm{cov}\left(A, B\right)/(\mathrm{Var}(A)\mathrm{Var}(B))$.



(a) (b)

Figure 3: Evolution of corr $\left(\mathbf{E}_n^{\mathrm{I}_1}, \mathbf{E}_{n-1}^{\mathrm{P}_1}\right)$ (*solid lines*) and corr $\left(\mathbf{E}_n^{\mathrm{P}_1}, \mathbf{E}_{n-1}^{\mathrm{P}_1}\right)$ (*dashed lines*) as a function of α_{I}, α_{P}, and Q_1: (a) `akiyo`, (b) synthetic model.

where the empirical correlation measured from the low-motion video `akiyo` is depicted. The largest correlation value is achieved when $\alpha_{\mathrm{I}} = \alpha_{\mathrm{P}}$ and its value proportionally decreases as α_{I} moves away from α_{P}.

In this case, the semi-analytic model for $\mathbf{E}_n^{\mathrm{I}_1}$ and $\mathbf{E}_{n-1}^{\mathrm{P}_1}$ is derived in the same way as discussed in Section 3.1 with the peculiarity that now the latter quantization error is obtained at $n-1$ instead of n, but the same procedure is followed and the same values of p_1 are employed. By doing so, the correlation between the derived synthetic signals $e_n^{\mathrm{I}_1}$ and $e_{n-1}^{\mathrm{P}_1}$ is depicted in Fig. 3(b), where it can be observed that the proposed synthetic model does not follow very well the empirical cases when $\alpha_{\mathrm{I}} \in \{1, \frac{5}{4}\}$, so probably an adjustment of the model parameters would be needed in these cases.

2. cov $\left(\mathbf{E}_n^{\mathrm{P}_1}, \mathbf{E}_{n-1}^{\mathrm{P}_1}\right)$: the corresponding correlation is computed between the quantization errors that arise from two inter-predictive frames at different time indices, i.e., $\mathbf{E}_n^{\mathrm{P}_1}$ and $\mathbf{E}_{n-1}^{\mathrm{P}_1}$, during the first compression. Given that $\mathbf{U}_n$ is very close to $\mathbf{U}_{n-1}$ in static scenes, we expect a very high correlation in this case, which is empirically confirmed in Fig. 3(a).

   Fig. 3(b) collects the resulting correlation after generating the synthetic versions of the quantization errors $\mathbf{E}_n^{\mathrm{P}_1}$ and $\mathbf{E}_{n-1}^{\mathrm{P}_1}$. Differently from the previous point, now the synthetically computed correlations follow very well the trend of the empirical ones, except for small values of Q_1.

The analysis in Section 3.1 of the variance terms in (12) reveals that for low-motion videos the value of Var $\left(\mathbf{E}_n^{\mathrm{I}_1}\right)$ − Var $\left(\mathbf{E}_n^{\mathrm{P}_1}\right)$ is generally small. Hence, its effect is negligible in the evolution of (12). Additionally, since Var $\left(\mathbf{E}_n^{\mathrm{I}_1}\right) \approx$ Var $\left(\mathbf{E}_n^{\mathrm{P}_1}\right)$, this implies that cov $\left(\mathbf{E}_n^{\mathrm{I}_1}, \mathbf{E}_{n-1}^{\mathrm{P}_1}\right)$ − cov $\left(\mathbf{E}_n^{\mathrm{P}_1}, \mathbf{E}_{n-1}^{\mathrm{P}_1}\right)$ is proportional to the difference of the corresponding correlations. From the example shown

in Fig. 3(a), one can observe that such difference is nearly constant for distinct values of Q_1 independently of the considered relation between α_I and α_P, so we can ensure that the terms dependent on Q_1 do not cause prominent changes in (12) for low-motion videos. As a consequence, the appearance of the VPF is fundamentally determined by the two covariance terms that jointly depend on the quantization parameters Q_1 and Q_2, which we describe below:

1. $\mathrm{cov}\left(\mathbf{E}_n^{\mathrm{I}_1}, \mathbf{E}_{n-1}^{\mathrm{P}_2}\right)$: in this case, we need to consider the correlation between the quantization errors that arise during two distinct compression stages: $\mathbf{E}_n^{\mathrm{I}_1}$ during the first compression, whose synthetic modeling has already been detailed in Section 3.1), and $\mathbf{E}_{n-1}^{\mathrm{P}_2}$ during the second stage. The synthetic signal for $\mathbf{E}_{n-1}^{\mathrm{P}_2}$ can be derived following a similar procedure to that of $\mathbf{E}_n^{\mathrm{P}_1}$ (described in Section 3.1), but taking now into account that the input signal has been compressed once and that the recompression of a video sequence alters the probabilities of each MB type. In fact, the probability of having a P-MB during the second compression is now lowered to $p_2 \triangleq 0.15 + (p1 - 0.15)e^{-9\frac{\mathrm{Q}_2}{\mathrm{Q}_{\max}}}$ because in double compressed videos we expect to find prediction residues closer to zero as an effect of the first compression, and so we count on finding a smaller number of P-MBs. Hence, using $\mathrm{p(P\text{-}MB)} = p_2$ and $\mathrm{p(S\text{-}MB)} = 1 - p_2$, we compute the synthetic signal as

$$e_{n-1}^{\mathrm{P}_2} = \begin{cases} w'_{n-1} - w_{n-1}, & \text{if c = P-MB} \\ y'_{n-2} - y_{n-1}, & \text{otherwise} \end{cases},$$

where $w_{n-1} = x'_{n-1} - \hat{y}_{n-1}$, $\hat{y}_{n-1} = \rho_\mathrm{P} x'_{n-2} + \nu_{n-1}$, and w'_{n-1} is obtained by first quantizing w_{n-1} as in the second case of (4) and then reconstructing its samples through (5) (in both cases, using Δ_2^P and α_P). On the other hand, we have that $y_{n-1} = x'_{n-1}$ (see Sect. 3.1) and also that y'_{n-2} is the result of reconstructing the whole signal x'_{n-2} as the encoding of an I-frame with Q_2.

To check the validity of the model, we compare in Fig. 4 the obtained synthetic results of $\mathrm{corr}\left(\mathbf{E}_n^{\mathrm{I}_1}, \mathbf{E}_{n-1}^{\mathrm{P}_2}\right)$ fixing α_P and varying the values of α_I, Q_1, and Q_2 (*upper panels*) with the ones stemming from video sequence `akiyo` (*center panels*). The similarity among the depicted results supports the validity of the model. Although it is hard to predict the exact correlation between two quantization errors of different nature, we give some hints on why there are regions in the correlation maps shown in Fig. 4 that share the same correlation sign. Let us consider a sample located at the same position in $\mathbf{E}_n^{\mathrm{I}_1}$ and $\mathbf{E}_{n-1}^{\mathrm{P}_2}$. As long as the sample sign is retained in the succeeding compressions, this sample contributes positively in the resulting correlation. However, a change of sign across both compressions yields a negative contribution. Therefore, the regions in the correlation map with the same sign indicate that under these cases the two quantization errors share the same direction.

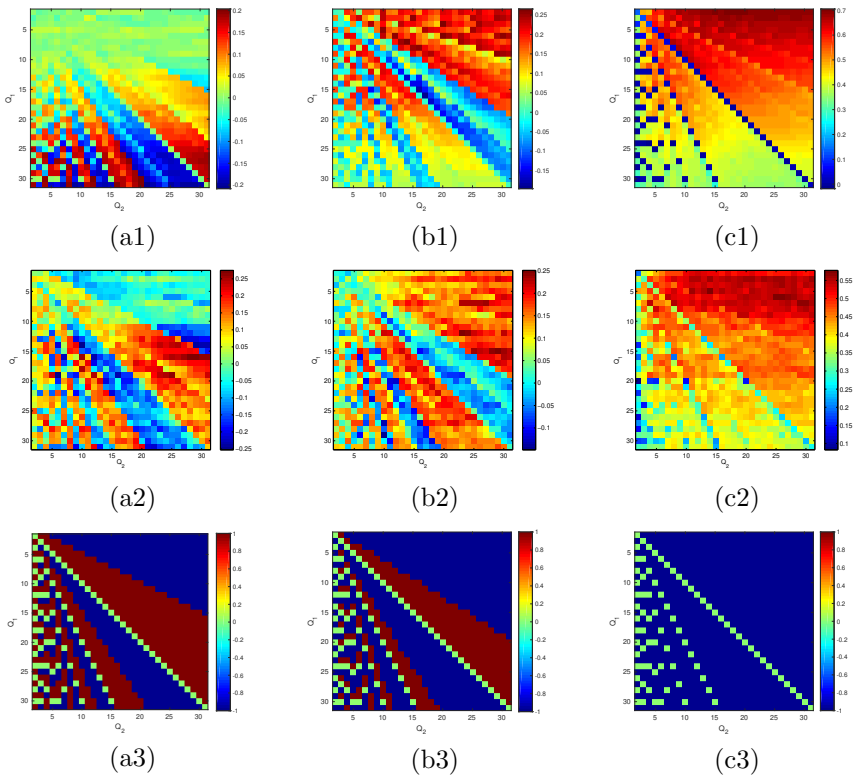


Figure 4: Evolution of corr $\left(\mathbf{E}_n^{\mathrm{I}_1}, \mathbf{E}_{n-1}^{\mathrm{P}_2}\right)$ for a fixed $\alpha_{\mathrm{P}} = 2$ and varying α_{I}, Q_1, and Q_2. The upper panels show the obtained results with synthetic signals, while the center panels correspond to the static video `akiyo`. In the lower panels, only the sign of the correlation is shown for those samples that have been quantized to the centroid Δ_1^{I} in the first compression stage. (a1-a3) $\alpha_{\mathrm{I}} = 1$, (b1-b3) $\alpha_{\mathrm{I}} = \frac{5}{4}$, (c1-c3) $\alpha_{\mathrm{I}} = 2$.

Now, what delimits the particular shape of such regions is the relation between the selected width for the quantizer deadzones and the different reconstruction procedures that result in $\mathbf{E}_n^{\mathrm{I}_1}$ and $\mathbf{E}_{n-1}^{\mathrm{P}_2}$. As an example, we show in the lower panels of Fig. 4 the evolution of $\mathrm{sgn}\,(d' - d)$, where we set $d = \Delta_1^{\mathrm{I}}$ and we obtain d' as its reconstructed version after quantizing d with Δ_2^{I} (the rule for I-MBs is followed in both cases). This basic example shows some of the particular shapes that arise in the mentioned correlation maps and reflects, for instance, that in Fig. 4(c3) no change of sign occurs for $\mathrm{Q}_2 > \mathrm{Q}_1$ and $\alpha_{\mathrm{I}} = \alpha_{\mathrm{P}} = 2$, while in Figs. 4(a3) and 4(b3) the flip of sign takes place at $\mathrm{Q}_2 > (2/\alpha_{\mathrm{I}})\mathrm{Q}_1$. This last relation will determine the limit beyond which the strength of the VPF vanishes for $\alpha_{\mathrm{I}} \in \{1, \frac{5}{4}\}$.

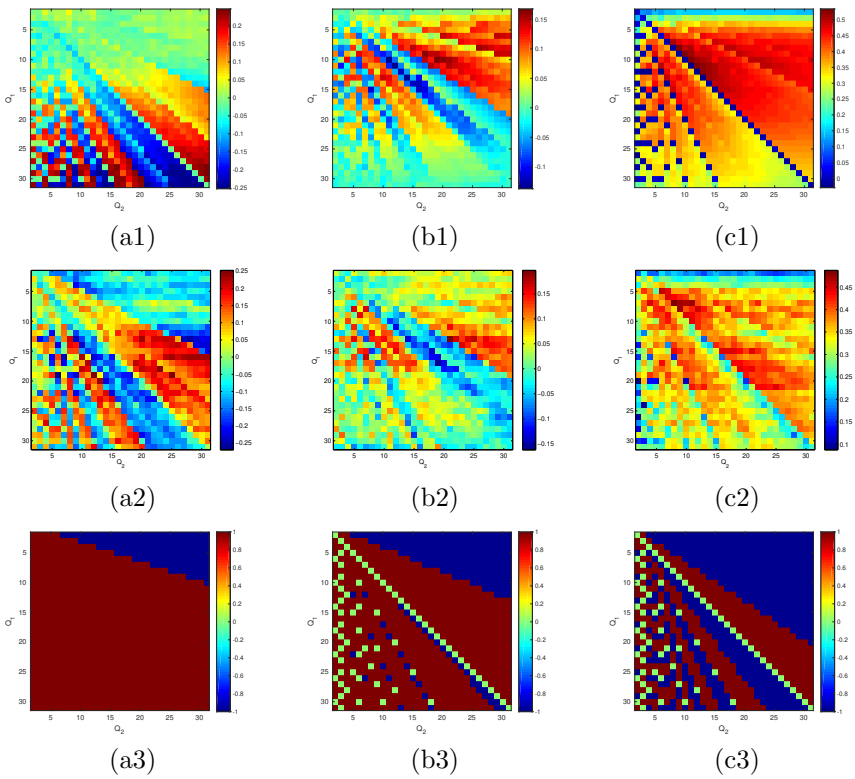


Figure 5: Evolution of corr $\left(\mathbf{E}_n^{\mathrm{P}_1}, \mathbf{E}_{n-1}^{\mathrm{P}_2}\right)$ for a fixed $\alpha_{\mathrm{P}} = 2$ and varying α_{I}, Q_1, and Q_2. The upper panels show the obtained results with synthetic signals, while the center panels show the corresponding results for the static video `akiyo`. In the lower panels, only the sign of the correlation is shown for those samples that have been quantized to the centroid $\Delta_1^{\mathrm{P}} + \frac{\Delta_1^{\mathrm{P}}}{2}$ in the first compression stage. (a1, a2) $\alpha_{\mathrm{I}} = 1$, (b1, b2) $\alpha_{\mathrm{I}} = \frac{5}{4}$, (c1, c2) $\alpha_{\mathrm{I}} = 2$.

2. cov $\left(\mathbf{E}_n^{\mathrm{P}_1}, \mathbf{E}_{n-1}^{\mathrm{P}_2}\right)$: here we analyze the correlation between the quantization error $\mathbf{E}_n^{\mathrm{P}_1}$ (whose synthetic version has been detailed in Section 3.1) and the one described in the previous point $\mathbf{E}_{n-1}^{\mathrm{P}_2}$. The upper panels of Fig. 5 report the obtained values of corr $\left(\mathbf{E}_n^{\mathrm{P}_1}, \mathbf{E}_{n-1}^{\mathrm{P}_2}\right)$ using the described model for a fixed α_{P} and varying the parameters α_{I}, Q_1, and Q_2. The center panels in Fig. 5 show their empirical counterparts, which have been extracted from the low-motion video `akiyo`. Again, the model remarkably follows the shape of the empirical correlations. In this case, by considering a similar example as in the previous point for $\mathrm{sgn}(d' - d)$, but taking now $d = \Delta_1^{\mathrm{P}} + \frac{\Delta_1^{\mathrm{P}}}{2}$ and reconstructing d' with Δ_2^{P} following the rule for P-MBs, one can observe in Fig. 5(c3) that the evolution of $\mathrm{sgn}(d' - d)$ shows a

change of sign at $Q_2 > (3/2)Q_1$, that will determine the limit beyond which the VPF vanishes for $\alpha_I = \alpha_P$.

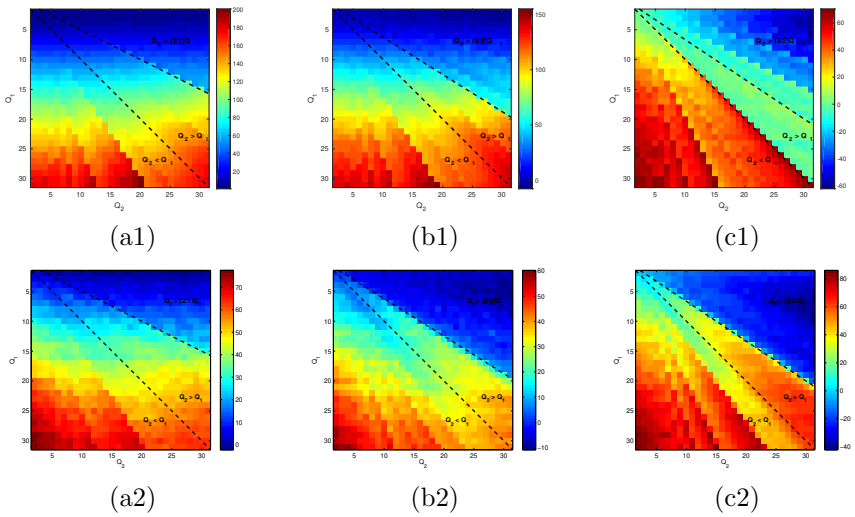
Once modeled and discussed all the terms in (12), we finally compare in Fig. 6 the resulting synthetic versions of the difference $\mathrm{Var}(\mathbf{W}_n)|_{n\in \mathsf{I}_1} - \mathrm{Var}(\mathbf{W}_n)|_{n\in \mathsf{P}_1}$ for the considered values of α_I (*upper panels*) with the ones empirically obtained after processing 14 videos from [8][4] (*lower panels*). The synthetic models show a very high degree of similarity with respect to their empirical counterparts, except for the case $\alpha_I = 2$, where the model possibly needs some adjustment. Still, this allows us to predict the strength of the VPF under each particular configuration of α_I and α_P, and also to infer from which relation between Q_1 and Q_2 the VPF is more likely to show up or not. In this sense, the limits in the correlation maps previously discussed in the two above points, show the boundary beyond which the difference $\mathrm{Var}(\mathbf{W}_n)|_{n\in \mathsf{I}_1} - \mathrm{Var}(\mathbf{W}_n)|_{n\in \mathsf{P}_1}$ drops and, as a consequence, the VPF vanishes. In particular, we have seen that for $\alpha_I \in \{1, \frac{5}{4}\}$, this limit is achieved at $Q_2 > (2/\alpha_I)Q_1$ (see Figs. 6(a2)-(b2)), while it moves downward to $Q_2 > (3/2)Q_1$ for $\alpha_I = \alpha_P$, as can be observed in Fig. 6(c2). Given that in all cases, the limit is above $Q_2 > Q_1$, this explains why the VPF-based approaches are able to satisfactorily work in the challenging video forensic scenarios where the second compression applied is stronger than the first one. The reader is referred to [1] for checking how this theoretical analysis serves to predict the performance of the VPF-based methods.

4 Conclusions

In this report we have delved into the analysis of the prediction residue computed in the second stage of an MPEG-2 double compression scheme. The characterization of the quantization process and the different parameters that control the deadzone width of the quantizers have made possible the derivation of a semi-analytic model that through the use of synthetic signals allows us to explain why the VPF shows up and how this footprint behaves depending on the quantization strength applied in each compression stage. One of the most valuable outcomes from the above analysis is the justification of why the approaches that exploit the VPF are able to successfully work on challenging scenarios where the second compression applied is stronger than the first one, while other available techniques typically fail. Nevertheless, the obtained synthetic results are not always consistent with their empirical counterparts, so a review of the semi-analytic model detailed in this report is still necessary, so as to understand what can be causing such inconsistencies.

Furthermore, as pointed out throughout the report, there is room for improving the above analysis, for instance, the model should be extended to encompass other video coding standards (e.g., MPEG-4 and H.264), to address

[4] In particular, we considered the following 14 CIF resolution videos: `akiyo`, `bridge-close`, `bridge-far`, `container`, `foreman`, `hall`, `highway`, `mother-daughter`, `mobile`, `news`, `silent`, `paris`, and `waterfall`.



(a1) (b1) (c1)

(a2) (b2) (c2)

Figure 6: Evolution of $\mathrm{Var}\left(\mathbf{W}_n\right)|_{n\in \mathsf{I}_1} - \mathrm{Var}\left(\mathbf{W}_n\right)|_{n\in \mathsf{P}_1}$ for a fixed $\alpha_{\mathrm{P}} = 2$ and varying α_{I}, Q_1, and Q_2. The upper panels show the obtained results with synthetic signals, while the lower panels show the corresponding average difference from 14 real videos in [8]. (a1, a2) $\alpha_{\mathrm{I}} = 1$, (b1, b2) $\alpha_{\mathrm{I}} = \frac{5}{4}$, (c1, c2) $\alpha_{\mathrm{I}} = 2$.

more complex coding settings (e.g., including B-frames, adaptive bitrate controls, etc.), and also to cover more complex type of scenes, such as the dynamic ones. Finally, since the proposed semi-analytic model has proved to be valid, the corresponding closed-form expressions should also be derived in a future work.

References

[1] D. Vázquez-Padín and F. Pérez-González, "Prediction residue analysis in mpeg-2 double compressed video sequences," in *Proceedings European Signal Processing Conference (EUSIPCO)*, 2019, (submitted).

[2] D. Vázquez-Padín, M. Fontani, T. Bianchi, P. Comesaña, A. Piva, and M. Barni, "Detection of video double encoding with GOP size estimation," in *Proc. of the 4th IEEE International Workshop on Information Forensics and Security (WIFS)*, December 2012, pp. 151–156.

[3] D. Vázquez-Padín, M. Fontani, F. Pérez-González, D. Shullani, A. Piva, and M. Barni, "Video integrity verification and GOP size estimation via generalized variation of prediction footprint," *IEEE TIFS*, 2019 (under review).

[4] [Online]. Available: http://ffmpeg.org/

[5] E. Y. Lam and J. W. Goodman, "A mathematical analysis of the DCT coefficient distributions for images," *IEEE Transactions on Image Processing*, vol. 9, no. 10, pp. 1661–1666, Oct 2000.

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

[7] F. Bellifemine, A. Capellino, A. Chimienti, R. Picco, and R. Ponti, "Statistical analysis of the 2D-DCT coefficients of the differential signal for images," *Signal Processing: Image Communication*, vol. 4, no. 6, pp. 477 – 488, 1992.

[8] [Online]. Available: https://media.xiph.org/video/derf/