Prediction Residue Analysis in MPEG-2 Double Compressed Video Sequences

David Vázquez-Padín, Fernando Pérez-González
Signal Theory and Communications Department, *University of Vigo*, Vigo, Spain, {dvazquez, fperez}@gts.uvigo.es

***Abstract*—In video forensics, the study of the prediction residue across successive frames is key to verify the integrity of digital videos. Focusing on an MPEG-2 double compression scheme, we analyze how the variance of the prediction residue evolves during the second compression depending on the type of frame (either I or P) employed in the first encoding and exploring different compression strengths and deadzone widths for quantization. This analysis reveals that the width of the quantizer deadzones actually affects the performance of existing methods based on the Variation of Prediction Footprint (VPF) for double compression detection and Group Of Pictures (GOP) size estimation. The predicted behavior from the theoretical characterization of the prediction residue is confirmed through experimental results with real video sequences.**

***Index Terms*—Prediction residue analysis, double compression detection, GOP size estimation, video forensics, MPEG-2.**

I. Introduction

Video edition and composition is becoming increasingly accessible due to the availability of many video editing software tools not only for desktop or laptop computers, but also for smartphones which nowadays represent one of the main acquisition devices to share videos online. In addition, the recent use of deep learning techniques to automatically create realistic forgeries rises an important concern about the trust of video contents, thus demanding more research in the emerging field of video forensics. Fortunately, existing video editing tools do not have the capability to directly work on the compressed domain (except for some basic cut and paste of clips within a video stream), so when creating realistic forgeries a recompression of the whole sequence of frames is needed, which will consequently alter the normal evolution of the prediction residue across successive frames.

In this context, Wang and Farid were the first to study the periodic artifacts induced in the prediction residue after the insertion/removal of frames from a digital video in [1]. This scheme has been further improved in [2], expanding this idea and proposing an automatic way to detect such a periodic artifacts. Still, not only the addition or deletion of frames alters the prediction residue, in [3] it has been shown that a desynchronization in the GOPs used between the first and second compression also modifies the evolution of the prediction residue, yielding the so-called VPF, which can be captured through the calculation of the relative frequency of each type of MacroBlock (MB) across time. The exploitation of this feature made possible the detection of double compressed videos and the estimation of the GOP size employed during the first compression. Later on, resorting to a rate-distortion analysis, the link between the presence of the VPF and the variance of the prediction residue has been established in [4], leading to an improved VPF acquisition process that outperforms [3]. The work in [4] has nevertheless left out of such analysis the influence on the prediction residue of the deadzone width used for quantization. As we will later confirm, the use of certain deadzone widths can either positively or negatively affect the performance of VPF-based approaches. Hence, with the aim of completing the analysis initiated in [4], here we study the variance of the prediction residue as a function of the deadzone widths that an encoder can use for quantization.

In particular, assuming an MPEG-2 double compression scheme, in this paper we characterize the variance of the prediction residue obtained in the second compression stage as a function of the type of frame employed in the first encoding (either I or P) and for different quantization deadzone widths. This analysis provides valuable insights on the behavior of the VPF exploited in [3], [4] and also indicates how the different deadzone widths can favor or impair double compression detection and GOP size estimation.

The paper is structured as follows: first, we formulate the video double compression problem in Sect. II, then the evolution of the variance of the prediction residue is analytically characterized in Sect. III, and finally the derived theoretical insights are experimentally validated for GOP size estimation and double compression detection in Sect. IV. Lastly, Sect. V concludes the paper and hints at possible new research paths.

II. Problem Formulation and Modeling

Let us consider the video double compression scenario shown in Fig. 1, where the same MPEG-2 encoder is used for both encodings. For the first compression, we assume a constant GOP of length G_1 and a fixed quantizer scale factor $\mathrm{Q}_1 \in \{2, \ldots, 31\}$. Similarly, the second compression is conducted with a GOP of length G_2 (different from any integer multiple or submultiple of G_1) and $\mathrm{Q}_2 \in \{2, \ldots, 31\}$. 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 their analysis for a future work.

In MPEG-2, the MBs of an I-frame can only be encoded by a single intra-coding mode that does not perform spatial prediction and is denoted by I-MB. In the case of P-

This work is partially funded by the Agencia Estatal de Investigación (Spain) and the European Regional Development Fund (ERDF) under the project WINTER (TEC2016-76409-C2-2-R), and by the Xunta de Galicia and the European Union (ERDF) under projects Agrupación Estratéxica Consolidada de Galicia accreditation 2016-2019 and Grupo de Referencia ED431C2017/53.

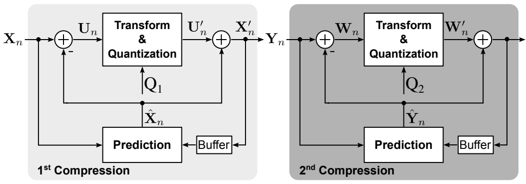


Fig. 1. Block diagram of the assumed video double compression scheme.

frames, besides the use of I-MBs, two inter-coding modes are available to perform temporal (or motion-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 left scheme in Fig. 1 models how a given MB at time index n, denoted by $\mathbf{X}_n$, is first 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 previously decoded frame at time index $n-1$ with the relative displacement pointed by motion vector $\mathbf{m}$. 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 $\mathbf{m}$ is null. Then, the prediction 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 controlled by the quantizer scale factor Q_1 as follows

$$\Delta_1(i,j) \triangleq (1/8)\mathrm{Q}_1 S_{i,j}, \quad \forall i,j \in \{0,\ldots,7\}, \tag{2}$$

where $S_{i,j}$ represents the (i,j)-th element of a weighting matrix $\mathbf{S}$ that improves the perceptual quality of the encoded videos, supporting the use of a different matrix for intra (i.e., $\mathbf{S}^{\mathrm{I}}$) and inter (i.e., $\mathbf{S}^{\mathrm{P}}$) coding modes. Regarding the quantizer deadzone, its width is defined as $w_1(i,j) \triangleq \alpha\Delta_1(i,j)$, where $\alpha \in [1,2]$ is the parameter that allows the control of the deadzone width. In practice, tighter deadzones are used for intra-coding modes (e.g., $\alpha_{\mathrm{I}} = \frac{5}{4}$ in [5]) to retain more details, whereas wider deadzones are considered for inter-coding modes which lead to small magnitude signals (e.g., $\alpha_{\mathrm{P}} = 2$ in [5]). Now, using (2) and the definition of the deadzone width, the quantization of an AC coefficient u from the DCT of $\mathbf{U}_n$ can be written (omitting position indices) as

$$u_q \triangleq \begin{cases} \mathrm{sgn}(u)\lfloor(|u|+\Delta_1^{\mathrm{I}}(1-\alpha_{\mathrm{I}}/2))/\Delta_1^{\mathrm{I}}\rfloor, & \text{if c = I-MB} \\ \mathrm{sgn}(u)\lfloor(|u|+\Delta_1^{\mathrm{P}}(1-\alpha_{\mathrm{P}}/2))/\Delta_1^{\mathrm{P}}\rfloor, & \text{otherwise} \end{cases}, \tag{3}$$

where $|\cdot|$ is the absolute value operator, $\lfloor\cdot\rfloor$ denotes the floor function, and $\mathrm{sgn}(\cdot)$ represents the sign function. 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 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

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

where a different reconstruction offset is used depending on the applied coding mode to improve coding efficiency. 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 $\mathbf{W}_n$, its reconstructed version by $\mathbf{W}'_n$, and the recovered samples are accordingly represented by $\mathbf{Y}'_n$.

III. Prediction Residue Analysis

The use of de-synchronized GOPs in a double encoding scheme causes the VPF effect unveiled in [3], 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

$$\text{MB type} = \arg\min_{\mathrm{c}\in\mathcal{C}} D(\mathbf{Y}_n, \mathbf{Y}'_n) + \lambda_{\mathrm{c}} R(\mathbf{Y}'_n), \tag{5}$$

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 (5) 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.,

$$\mathrm{Var}\,(\mathbf{W}_n)\,|_{n\in\mathsf{I}_1} - \mathrm{Var}\,(\mathbf{W}_n)\,|_{n\in\mathsf{P}_1}, \tag{6}$$

where a larger difference value yields a stronger VPF. With the aim of characterizing the prediction residue $\mathbf{W}_n = \mathbf{Y}_n - \hat{\mathbf{Y}}_n$,

[1] For the sake of clarity, position indices of $\mathbf{X}_n$ within the frame are omitted.

let us first describe the input signal $\mathbf{Y}_n$, which depending on the selected type of frame and the applied MB coding modes during the first compression, can be expressed as

$$\mathbf{Y}_n = \mathbf{X}_n + (\mathbf{U}'_n - \mathbf{U}_n) = \begin{cases} \mathbf{X}_n + \mathbf{E}_n^{\mathrm{I}_1}, & \text{if } n \in \mathsf{I}_1 \\ \mathbf{X}_n + \mathbf{E}_n^{\mathrm{P}_1}, & \text{if } n \in \mathsf{P}_1 \end{cases}, \tag{7}$$

where $\mathbf{E}_n^{\mathrm{I}_1} \triangleq \mathbf{X}'_n - \mathbf{X}_n$, since when $n \in \mathsf{I}_1$ we have from (1) that $\mathbf{U}_n = \mathbf{X}_n$ and $\mathbf{U}'_n = \mathbf{X}'_n$, while when $n \in \mathsf{P}_1$ we have that $\mathbf{E}_n^{\mathrm{P}_1} \triangleq \mathbf{U}'_n - \mathbf{U}_n$. The subsequent sections separately describe the prediction $\hat{\mathbf{Y}}_n$ as a function of the two prediction types (intra or inter) that can be applied in the second compression.

A. Intra-prediction residue analysis

The use of the intra-coding mode I-MB during the second compression yields $\hat{\mathbf{Y}}_n = 0$, such that $\mathbf{W}_n = \mathbf{Y}_n$. Hence, from (7), the variance of the prediction residue results in

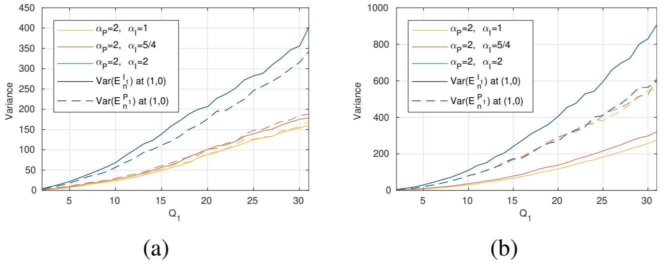
$$\mathrm{Var}\left(\mathbf{W}_n\right) = \begin{cases} \mathrm{Var}\left(\mathbf{X}_n\right) + \mathrm{Var}\left(\mathbf{E}_n^{\mathrm{I}_1}\right), & \text{if } n \in \mathsf{I}_1 \\ \mathrm{Var}\left(\mathbf{X}_n\right) + \mathrm{Var}\left(\mathbf{E}_n^{\mathrm{P}_1}\right), & \text{if } n \in \mathsf{P}_1 \end{cases}, \tag{8}$$

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 in practice since the probability density function (pdf) of the source signal is smooth and its variance is much larger than the employed quantization step sizes. By inserting (8) in (6), the strength of the VPF can be evaluated by means of the difference $\mathrm{Var}\left(\mathbf{E}_n^{\mathrm{I}_1}\right) - \mathrm{Var}\left(\mathbf{E}_n^{\mathrm{P}_1}\right)$.

The variance of both quantization errors is proportional to the distortion introduced during the DCT quantization process. Since the DCT adopted in MPEG-2 has orthogonal basis (if rounding effects are disregarded), such distortion can be computed in the DCT domain. Hence, for an arbitrary MB $\mathbf{Z}_n$, the SSD distortion caused by a quantizer with deadzone parameter α, reconstruction offset ϕ, and step size Δ, can be obtained as $D(\mathbf{Z}_n, \mathbf{Z}'_n) = \sum_{i=0}^{7}\sum_{j=0}^{7} D_Z(i,j)$, where $D_Z(i,j)$ stands for the distortion of the (i,j)-th DCT coefficient, given by

$$D_Z(i,j) = \int_{-\frac{\alpha}{2}\Delta}^{\frac{\alpha}{2}\Delta} z^2 f_Z(z)dz + 2\sum_{k=1}^{\infty}\int_{(k-1+\frac{\alpha}{2})\Delta}^{(k+\frac{\alpha}{2})\Delta} (z - z')^2 f_Z(z)dz, \tag{9}$$

where $z' \triangleq \lfloor (k+\phi)\Delta \rfloor$ and $f_Z(z)$ is the pdf of the (i,j)-th DCT coefficient. Now, given the definition of $\mathbf{E}_n^{\mathrm{I}_1}$, we have that $\mathrm{Var}\left(\mathbf{E}_n^{\mathrm{I}_1}\right) \propto D(\mathbf{X}_n, \mathbf{X}'_n)$, which can be computed through (9) using $\alpha = \alpha_{\mathrm{I}}$, $\Delta = \Delta_1^{\mathrm{I}}$, and $\phi = 0$ (see (3)-(4)), and a Laplacian pdf as $f_Z(z)$ [7]. Accordingly, for a given pdf, we can infer that the value of $\mathrm{Var}\left(\mathbf{E}_n^{\mathrm{I}_1}\right)$ increases with Q_1 (i.e., coarser quantization steps Δ_1^{I} are obtained through (2)) and also with α_{I} (i.e., due to wider quantizer deadzones). On the other hand, the variance of the quantization error $\mathbf{E}_n^{\mathrm{P}_1}$ satisfies the relation: $\mathrm{Var}\left(\mathbf{E}_n^{\mathrm{P}_1}\right) \propto \mathrm{p(I\text{-}MB)}D(\mathbf{X}_n, \mathbf{X}'_n) + \mathrm{p(P\text{-}MB)}D(\mathbf{U}_n, \mathbf{U}'_n) + \mathrm{p(S\text{-}MB)}D(\mathbf{X}_n, \mathbf{X}'_{n-1})$, 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,[2] whereas $D(\mathbf{U}_n, \mathbf{U}'_n)$ can be obtained through (9) setting $\alpha = \alpha_{\mathrm{P}}$, $\Delta = \Delta_1^{\mathrm{P}}$, and $\phi = 1/2$ (see (3)-(4)), and a Laplacian pdf as $f_Z(z)$ [8]. Therefore, the evolution of $\mathrm{Var}\left(\mathbf{E}_n^{\mathrm{P}_1}\right)$ does not only depend on the quantization parameters and the pdf of each DCT coefficient as for $\mathrm{Var}\left(\mathbf{E}_n^{\mathrm{I}_1}\right)$, but also on the type of scene to encode which rules the probability of each MB type. 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}. This 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 [9]: the static video *akiyo* in Fig. 2(a), and the dynamic video *mobile* in Fig. 2(b).



(a)

(b)

Fig. 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*).

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 usually 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 below.

B. 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 no forgery is introduced between both compressions), $\hat{\mathbf{Y}}_n$ can be written as

$$\hat{\mathbf{Y}}_n = \mathbf{Y}_{n-1}(\mathbf{m}) + \mathbf{E}_{n-1}^{\mathrm{P}_2} = \mathbf{X}_{n-1}(\mathbf{m}) + \mathbf{E}_{n-1}^{\mathrm{P}_1} + \mathbf{E}_{n-1}^{\mathrm{P}_2}, \tag{10}$$

where $\mathbf{E}_{n-1}^{\mathrm{P}_1} \triangleq \mathbf{U}'_{n-1}(\mathbf{m}) - \mathbf{U}_{n-1}(\mathbf{m})$ and $\mathbf{E}_{n-1}^{\mathrm{P}_2} \triangleq \mathbf{W}'_{n-1}(\mathbf{m}) - \mathbf{W}_{n-1}(\mathbf{m})$ represent the quantization errors that result from the first and second compression, respectively. Using (7) and (10) in the definition of $\mathbf{W}_n$, we obtain

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

[2] Note that for $D(\mathbf{X}_n, \mathbf{X}'_{n-1})$, only an approximation would be obtained through (9), but still valid in practice since $\mathbf{X}_n \approx \mathbf{X}_{n-1}$ for S-MBs.

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. Assuming that $\mathbf{E}_n$ has negligible correlation with $\mathbf{R}_n$, we can approximate the variance of $\mathbf{W}_n$ as $\text{Var}\left(\mathbf{W}_n\right) = \text{Var}\left(\mathbf{R}_n\right) + \text{Var}\left(\mathbf{E}_n\right)$, and so the difference in (6) becomes

$$\begin{aligned}&\text{Var}\left(\mathbf{E}_n\right)|_{n\in \mathrm{I}_1} - \text{Var}\left(\mathbf{E}_n\right)|_{n\in \mathrm{P}_1}\\ &= \underbrace{\text{Var}(\mathbf{E}_n^{\mathrm{I}_1}) - \text{Var}(\mathbf{E}_n^{\mathrm{P}_1}) - 2(\text{cov}(\mathbf{E}_n^{\mathrm{I}_1}, \mathbf{E}_{n-1}^{\mathrm{P}_1}) - \text{cov}(\mathbf{E}_n^{\mathrm{P}_1}, \mathbf{E}_{n-1}^{\mathrm{P}_1}))}_{\text{depends on } \mathrm{Q}_1}\\ &\quad \underbrace{-2(\text{cov}(\mathbf{E}_n^{\mathrm{I}_1}, \mathbf{E}_{n-1}^{\mathrm{P}_2}) - \text{cov}(\mathbf{E}_n^{\mathrm{P}_1}, \mathbf{E}_{n-1}^{\mathrm{P}_2}))}_{\text{depends on } \mathrm{Q}_1 \text{ and } \mathrm{Q}_2}. \qquad (11)\end{aligned}$$

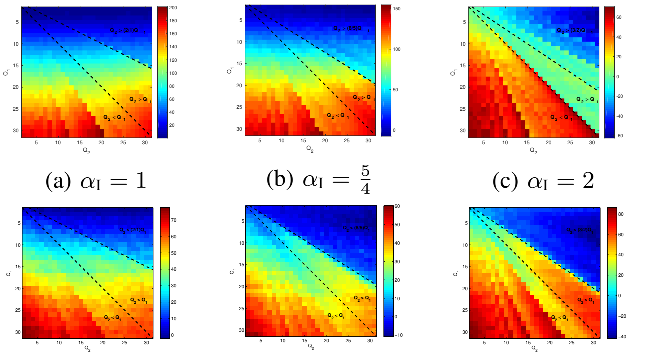
The expected evolution of the above terms $\text{Var}\left(\mathbf{E}_n^{\mathrm{I}_1}\right)$ and $\text{Var}\left(\mathbf{E}_n^{\mathrm{P}_1}\right)$ has already been analytically described for low-motion videos in Sect. III-A, showing that their difference has a negligible effect in the evolution of (11). Now, the derivation of analytical expressions for the remaining covariance terms is a complex task, so as a first step towards their modeling, we opted for a semi-analytic approach detailed in a tech report [10], based on the use of synthetic signals and autoregressive models for characterizing inter predictions and temporal dependencies, leaving the complete modeling for a future work. In brief, the analysis in [10] reveals that for static videos the difference between the above covariance terms that only depend on Q_1 is nearly constant for distinct values of Q_1, α_{I}, and α_{P}, thus ensuring that these terms do not cause prominent changes in (11). Hence, the presence of the VPF (and also its strength) is fundamentally determined by the last two covariance terms that jointly depend on Q_1, Q_2, and the relation among α_{I} and α_{P}. The reader is referred to [10] for a thorough description of each of these covariance terms.

To check the validity of the semi-analytic model derived in [10], we compare in Fig. 3 the resulting synthetic versions of the difference $\text{Var}\left(\mathbf{W}_n\right)|_{n\in \mathrm{I}_1} - \text{Var}\left(\mathbf{W}_n\right)|_{n\in \mathrm{P}_1}$ for $\alpha_{\mathrm{P}} = 2$, $\alpha_{\mathrm{I}} \in \{1, \frac{5}{4}, 2\}$, and varying Q_1 and Q_2 (*upper panels*), with the ones obtained empirically after processing the 14 videos to be described in Sect. IV (*lower panels*). The synthetic models show a very high degree of similarity with respect to their empirical counterparts, except for the case $\alpha_{\mathrm{I}} = 2$, where the model possibly needs some adjustment. Yet, from this characterization we can determine (through the relation between the deadzone widths) the marked boundaries for $\mathrm{Q}_2 > \mathrm{Q}_1$ beyond which $\text{Var}\left(\mathbf{W}_n\right)|_{n\in \mathrm{I}_1} - \text{Var}\left(\mathbf{W}_n\right)|_{n\in \mathrm{P}_1}$ drops and, as a consequence, the VPF vanishes. In particular, as discussed in [10], for $\alpha_{\mathrm{I}} \in \{1, \frac{5}{4}\}$ and $\alpha_{\mathrm{P}} = 2$ the limit is achieved at $\mathrm{Q}_2 > (2/\alpha_{\mathrm{I}})\mathrm{Q}_1$ (see Figs. 3(d)-(e)), while for $\alpha_{\mathrm{I}} = \alpha_{\mathrm{P}} = 2$ it moves downward to $\mathrm{Q}_2 > (3/2)\mathrm{Q}_1$, as can be observed in Fig. 3(f). Since in all cases, the boundary is located above $\mathrm{Q}_2 > \mathrm{Q}_1$, this explains why the VPF-based approaches are able to satisfactorily work in the challenging scenarios where the second compression is stronger than the first one.

IV. Experimental Results

The above findings on the dependence of the VPF with the quantization parameters: α_{I}, α_{P}, Q_1, and Q_2, should serve to predict the performance of the methods that essentially exploit the presence of the VPF to estimate the size of the GOP employed during the first compression and also to expose double compressed video sequences. Here, we consider the baseline method proposed in [3], i.e., "B-VPF", and its generalized version in [4], i.e., "G-VPF", that further improves the VPF acquisition by introducing in the peak extraction process information about the number of P-MBs with null motion vector, as described in Algorithm 1. Both methods share the same periodicity analysis proposed in [3].

The set of video sequences over which we conduct the experiments is composed of 14 uncompressed videos with CIF resolution [9], which are either static: with low-motion scenes captured by a fixed camera; or dynamic: with a large amount of motion caused by a focal length change or a moving camera position.[3] The experimental validations carried out in [3] and [4] prove the practical applicability of both methods, so here we mainly focus on validating the aforementioned findings in the constrained double compression framework illustrated in Fig. 1, where the same MPEG-2 encoder is used in both compressions under a fixed quantization procedure controlled first by Q_1 and then by Q_2. Still, we replicate the same set



(a) $\alpha_{\mathrm{I}} = 1$ (b) $\alpha_{\mathrm{I}} = \frac{5}{4}$ (c) $\alpha_{\mathrm{I}} = 2$

(d) $\alpha_{\mathrm{I}} = 1$ (e) $\alpha_{\mathrm{I}} = \frac{5}{4}$ (f) $\alpha_{\mathrm{I}} = 2$

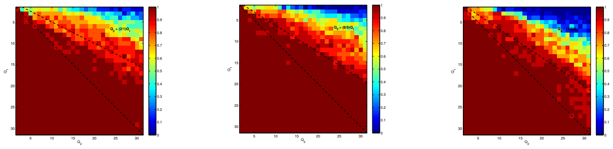
Fig. 3. Evolution of $\text{Var}\left(\mathbf{W}_n\right)|_{n\in \mathsf{I}_1} - \text{Var}\left(\mathbf{W}_n\right)|_{n\in \mathsf{P}_1}$ for a fixed $\alpha_{\mathrm{P}} = 2$ and varying α_{I}, Q_1, and Q_2. The upper panels show the obtained results with synthetic signals, while the lower panels show the average of 14 real videos.

Algorithm 1 Peak extraction pseudocode for G-VPF [4]

```
Inputs: $i_n,\ s_n,\ p_n$ // # of I-MBs, S-MBs, and P-MBs with $\mathbf{m} = (0,0)$
Output: $v_n$
  for $n = 1$ to $N-1$ do
    if ($\mathrm{E}\,(i_n, 1) = 1$ and $\mathrm{E}\,(-s_n, 1) = 1$ and $\mathrm{E}\,(p_n, 1) = 1$) then
      $v_n \leftarrow 0$
    else
      $v_n \leftarrow \sum_{k\in\{-1,1\}} \mathrm{E}\,(i_n, k)\,\mathrm{E}\,(-s_n, k)\,\mathrm{E}\,(p_n, k)$
    end if
  end for

  function $\mathrm{E}\,(a_n, k)$ // Extract up (down) peaks from $a_n$ with $k = 1$ ($k = -1$)
  if $a_n > \max(a_{n-1}, a_{n+1})$ then
    return $|a_n - a_{n-k}|$
  end if
  return 1
  end function
```

[3] Static: *akiyo, bridge-close, bridge-far, container, hall, mother-daughter, news, silent,* and *paris*. Dynamic: *foreman, highway, mobile,* and *waterfall*.



(a) $\alpha_{\mathrm{I}}=1$ (b) $\alpha_{\mathrm{I}}=\frac{5}{4}$ (c) $\alpha_{\mathrm{I}}=2$

Fig. 4. EMR for G-VPF under $\alpha_{\mathrm{P}}=2$ and varying α_{I} (MPEG-2 encoder).

of experiments using other two encoders, i.e., MPEG-4 from [5] and H.264 from [11], for checking to which extent the theoretical modeling for MPEG-2 also holds under different video coding standards. Since MPEG-2 and MPEG-4 share the same set of quantization parameters, we test each pair $(\mathrm{Q}_1, \mathrm{Q}_2)$ in the set $\mathcal{Q} = \mathcal{Q}_1 \times \mathcal{Q}_2$, where $\mathcal{Q}_i = \{2, \dots, 31\}$. In the case of H.264, we test the pairs $(\mathrm{Q}_1, \mathrm{Q}_2)$ that result from the Cartesian product between 30 integer values in the range $[4, 51]$. As common settings we settle the Lagrangian-based strategy for coding-mode decision, we fix the first and second GOP lengths to $\mathrm{G}_1 = 10$ and $\mathrm{G}_2 = 33$, respectively, and we use the same deadzone widths for the two consecutive encodings (fixing $\alpha_{\mathrm{P}} = 2$ and testing $\alpha_{\mathrm{I}} \in \{1, \frac{5}{4}, 2\}$). Finally, for each video we limit the analysis to the first 250 frames.

Fig. 4 shows the obtained Exact Match Rate (EMR) between the true GOP size G_1 and its estimated value $\hat{\mathrm{G}}_1$ by the G-VPF method under the MPEG-2 scenario. These results confirm the theoretical findings from Sect. III, given that the evolution of the EMR values is consistent with the representation in Fig 3, i.e., the larger the difference $\mathrm{Var}\left(\mathbf{W}_n\right)|_{n\in\mathrm{I}_1} - \mathrm{Var}\left(\mathbf{W}_n\right)|_{n\in\mathrm{P}_1}$, the higher the EMR. In fact, it can be appreciated how an almost perfect estimation is achieved for $\mathrm{Q}_2 > \mathrm{Q}_1$ upon reaching the marked boundaries (for each distinct value of α_{I}) and the increasing lost in performance beyond them.

In order to save space when comparing G-VPF against B-VPF in the tested scenarios, we define a digest measure denoted by $\overline{\mathrm{EMR}}$, that synthesizes the quality of the estimation under each setting by computing the average EMR across all the tested quantization parameters in the set $\mathcal{Q}$ as $\overline{\mathrm{EMR}} \triangleq \frac{1}{|\mathcal{Q}|}\sum_{(\mathrm{Q}_1,\mathrm{Q}_2)\in\mathcal{Q}} \mathrm{EMR}(\mathrm{Q}_1,\mathrm{Q}_2)$, where $\mathrm{EMR}(\mathrm{Q}_1,\mathrm{Q}_2)$ represents the obtained EMR for a particular pair $(\mathrm{Q}_1,\mathrm{Q}_2)$. The upper part of Table I reports the $\overline{\mathrm{EMR}}$ values obtained by the two methods in each scenario, from which it becomes clear that G-VPF always outperforms B-VPF, thus indicating that the use of P-MBs with $\mathbf{m} = (0,0)$ is key to better capturing the VPF. Focusing on the G-VPF results, the $\overline{\mathrm{EMR}}$ values for MPEG-2 decrease as α_{I} increases (consistently with Fig. 4), while this is not the case for MPEG-4 and H.264. The subtle differences between MPEG-2 and MPEG-4 suggest that the above modeling could be easily adapted to cover the particular structure of MPEG-4, while for H.264 a closer look to the novel coding elements of the standard with respect to MPEG-2, should be considered to understand their effect on the VPF. Interestingly, we also report in Table I the obtained $\overline{\mathrm{EMR}}$ results for each type of videos under MPEG-2, which essentially confirm that the assumed model for static videos does not match the characteristics of the dynamic ones.

To evaluate the detection performance of the two methods, we build for each encoder a dataset of 420 single compressed videos (which results from compressing the 14 videos with the 30 values of Q_1) and we randomly select other 420 videos from the total of 12,600 double compressed video sequences. The lower part of Table I collects the obtained values of Area Under the Curve (AUC) corresponding to the receiver operating characteristic of the G-VPF and B-VPF methods. In this case, the evolution of the AUC results follow the same course of their relative $\overline{\mathrm{EMR}}$ values, which reaffirms the conclusions drawn from the GOP size estimation analysis.

TABLE I

$\overline{\mathrm{EMR}}$ AND AUC VALUES FOR G-VPF VS. B-VPF UNDER DIFFERENT SCENARIOS. BOLD NUMBERS INDICATE THE BEST RESULTS PER ENCODER.

$\overline{\mathrm{EMR}}$								
	MPEG-2		**MPEG-4**		**H.264**		**MPEG-2** (G-VPF)	
α_{I}	G-VPF	B-VPF	G-VPF	B-VPF	G-VPF	B-VPF	Static	Dynamic
1	**0.9111**	0.6835	0.9065	0.6557	0.7262	0.6967	**0.9632**	0.8173
$\frac{5}{4}$	0.9063	0.7150	**0.9079**	0.6942	0.7274	0.6984	0.9444	0.8378
2	0.8630	0.7699	0.8826	0.7978	**0.7399**	0.6874	0.8651	0.8592
AUC								
	MPEG-2		**MPEG-4**		**H.264**		**MPEG-2** (G-VPF)	
α_{I}	G-VPF	B-VPF	G-VPF	B-VPF	G-VPF	B-VPF	Static	Dynamic
1	**0.9632**	0.8406	0.9509	0.8073	0.8556	0.8589	**0.9872**	0.9259
$\frac{5}{4}$	0.9588	0.8540	**0.9572**	0.8306	0.8589	0.8603	0.9733	0.9254
2	0.9243	0.8930	0.9382	0.8952	**0.8743**	0.8443	0.9111	0.9129

V. CONCLUSIONS

In this paper we have delved into the analysis of the prediction residue during the second stage of an MPEG-2 double compression scheme. The full characterization of the quantization process revealed how distinct deadzone widths affect the performance of VPF-based approaches. However, as noted along the paper, there is room for improving the above analysis by extending it to other video coding standards and by addressing more complex types of scenes and coding settings.

REFERENCES

[1] W. Wang and H. Farid, "Exposing digital forgeries in video by detecting double MPEG compression," in *MM&Sec*, 2006, pp. 37–47.

[2] M. C. Stamm, W. S. Lin, and K. R. Liu, "Temporal forensics and anti-forensics for motion compensated video," *IEEE TIFS*, vol. 7, no. 4, pp. 1315–1329, 2012.

[3] D. Vázquez-Padín, M. Fontani, T. Bianchi, P. Comesaña, A. Piva, and M. Barni, "Detection of video double encoding with GOP size estimation," in *IEEE WIFS*, 2012, pp. 151–156.

[4] 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).

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

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

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

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

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

[10] D. Vázquez-Padín and F. Pérez-González, "MPEG-2 prediction residue analysis," `arXiv:1906.07003 [cs.IT]`, June 2019.

[11] [Online]. Available: https://www.videolan.org/developers/x264.html