Streaming From a Moving Platform with Real-Time and Playback Distortion Constraints

Giuseppe Cocco
LTS4 Signal Processing Laboratory
Laboratory of Intelligent Systems
École polytechnique fédérale de Lausanne
Lausanne, Switzerland
Email: giuseppe.cocco@epfl.ch

Laura Toni
Department of Electronic and
Electrical Engineering
University College London
London, UK
Email: l.toni@ucl.ac.uk

***Abstract*—Video streaming from remotely controlled moving platforms such as drones have stringent constraints in terms of delay. In some applications such videos have to provide real-time visual feedback to the pilot with an acceptable distortion while satisfying high-quality requirements at playback. Furthermore the output rate of the source encoder required to achieve a target distortion depends on the speed of the platform. Motivated by this, we consider a novel source model which takes the source speed into account and derive its rate-distortion region. A transmission strategy based on successive joint encoding, which efficiently takes the source correlation into account, is then considered for transmission over a block fading channel. Our numerical results show that such scheme largely enhances over an independent coding scheme in terms of on-line distortion while approaching the playback distortion performance of an optimal encoder as the group of pictures size grows.**

I. Introduction

The development of smartphones and WiFi-equipped cameras has boosted the advancement of a new digital era of mobile video communications, where each user creates and streams original video content rather than passively consuming it. Streaming from cameras mounted on flying drones or on the helmet of skiers is an ever increasing trend [1]. Beyond entertainment, drones with mounted cameras are being used to explore inaccessible areas of land and buildings [2] or for agricultural purposes [3].

The video stream from a drone has a two-fold purpose. On the one hand it serves as a feedback for the pilot to properly steer the aircraft. On the other hand, the video is stored at the drone control station for playback at a later time. The delay and distortion constraints imposed on the video stream are different in the two cases. When the video is used as feedback for the pilot, the delay must satisfy stringent constraints[1] while a certain degree of distortion, which we refer to as *on-line distortion*, is still acceptable. The delay requirements for video playback are more relaxed, since real time is not required. However, video quality in playback is of paramount importance for many applications such as face recognition for surveillance purposes or post-production in the advertisement and in the film industries. We refer to the distortion in this last case as *playback distortion*. Moreover, remotely controlled platforms such as drones are characterized by varying levels of mobility that impact the correlation between captured video frames: a camera moving at low speed acquires frames that are highly similar, since the portion of the scene present in the field of view (FOV) of the camera in consecutive frames is relatively large. Conversely, when the source speed is higher the correlation between consecutive frames decreases. Today's video codecs typically compress all frames within a group of pictures (GOP) jointly, exploiting its dynamic correlation [6]. Although efficient from a compression perspective, this implies that data is not transmitted until the whole GOP is captured. Faster codecs based on predictive coding, as in the *baseline profile* of H.264 standard, can be considered for real-time applications, but they suffer from error propagation if part of the stream is lost on the channel [7]. This is due to the interdependency introduced by the source encoder and to the fact that typically channel coding does not take such interdependency into account. From an information theoretical perspective, unlike in a point-to-point communication setup [8] separate source compression and channel coding is not always optimal if lossy communication channels are considered [9]–[11]. However, a joint source-channel coding would require a drastic redesign of current streaming systems and would pose serious challenges to terminals with relatively limited computational capabilities and power such as drones.

Motivated by such problem, in this work we study the source and channel coding of a moving Gauss-Markov source streamed over a block-fading channel using both on-line and playback distortion as performance metrics. We consider a variant of the Gauss-Markov source model presented in [12] modified to include the effect of motion in the source correlation and we derive its rate-distortion region. In order for the transmitter to take into account the interdependency generated by the source encoder, we propose to use a progressive joint-encoding scheme which allows to retrieve all frames received up to the decoding instant. Note that, although not done jointly to keep complexity low, source coding and channel coding are not independent in our setup. We then derive the on-line and

Giuseppe Cocco is partly funded by the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie Individual Fellowship grant agreement No. 751062.

The authors would like to thank Pascal Frossard of LTS4 laboratory for the useful discussions.

[1]Acceptable delays to allow for smooth maneuvering and avoid sickness, especially in haptic systems, are below 50 ms [4], [5].

the playback distortion in the source reconstruction in case the proposed joint channel encoder and a memoryless channel encoder are used. Unlike in [13], where an erasure channel with feedback is assumed, we consider a block fading channel with no feedback. Furthermore, unlike the present paper, the source models considered in [13] and [14] do not explicitly take the movement into account. Finally, we show numerically that applying a progressive joint encoder to a differential predictive coded modulation (DPCM)-coded stream allows to approach the distortion of a playback-optimal encoder as the size of the GOP grows and at the same time achieves a lower on-line distortion with respect to a memoryless coding.

II. System Model

We consider a moving camera system acquiring, compressing and streaming multimedia content in real-time. We assume that the camera moves in a straight trajectory and the speed v of the source is constant and known at the encoder [2]. The camera acquires video frames with a rate of F_r frames per second and applies lossy compression, generating GOPs of M frames each. GOPs have an IPPPP frame structure with one reference (I) frame and a number of P frames, each depending on the previous one only. To ensure low delay, each frame, once acquired, is source-encoded and immediately transmitted, i.e., before all frames in the GOP have been acquired. Frames are sent over a wireless fading channel with a coherence time that changes with the velocity of the source. Due to delay constraints, no feedback is sent to the transmitter and frames are not retransmitted if lost.

A. Source Model

The source model is a modified version of a spatially memoryless, spatially stationary and temporally correlated Gaussian (SSTCG) process [12]. In an SSTCG source the intensity of a pixel generated by the source is correlated with the value of the same pixel in other time instants (frames) but independent of the values of other pixels in the same or in other time instants[3]. Let n be the number of pixels in the source image. A new frame is generated by the source every inter-frame period, i.e., every $T_f = 1/F_r$ seconds. The t-th generated frame $\mathbf{X}_t = (X_t(1), X_t(2), \ldots, X_t(n-1), X_t(n))$ is an n-dimensional vector, which can be seen as the vectorization of a bi-dimensional $\sqrt{n} \times \sqrt{n}$ matrix. The elements of $\mathbf{X}_t$ are independent and identically distributed (i.i.d.) zero-mean Gaussian variables having variance σ_t^2. Such model is not able to account for the motion of the source. Therefore in the following we introduce the modified SSTCG model.

Due to the camera motion, the portion of the scene that is being observed moves inside the field of view (FOV). Thus the intensity of the pixels in consecutive frames corresponding to a given point in the scene is modelled as a temporal Markov process[4], i.e., $\forall t, t > 1$ we have

$$X_{t-1}(i^{(t-1)}) - X_t(i^{(t)}) - X_{t+1}(i^{(t+1)}), \tag{1}$$

where $i^{(t-1)}$, $i^{(t)}$ and $i^{(t+1)}$ represent the indices of the pixels in three consecutive frames corresponding to the same physical point in the scene. Such correspondence depends on the camera speed but dependency has been omitted to simplify the notation. Some of the points that are in the FOV in a given frame may not be present in the next one. In general a portion $\alpha(v)$, $0 \leq \alpha(v) \leq 1$, $v \geq 0$, of the $t-1$-th frame $\mathbf{X}_{t-1}$ is still present in frame t, although shifted and with some variations due to movements within the scene and change in perspective caused by the camera motion. Such portion is modelled as in expression (1). The remaining $\overline{\alpha}(v)$ portion of the scene is novel, i.e., it was not present in any of the previous frames. The source model is exemplified in Fig. (1). Assuming a Gaussian source, the overall source model is the following:

$$X_t(i^{(t)}) = \rho_s \frac{\sigma_t}{\sigma_{t-1}} X_{t-1}(i_\alpha^{(t-1)}) + N_t(i_\alpha^{(t-1)}) + N_{\overline{\alpha}}(i_{\overline{\alpha}}^{(t)}), \tag{2}$$

$\forall t, i \in \mathbb{N}, \alpha \in [0,1]$, where $\mathbb{N}$ is the set of natural numbers, $i_\alpha^{(t-1)}$ is the pixel index at time $t-1$ of a point in the scene that at time t is to be found in position $i^{(t)}$, $i_{\overline{\alpha}}^{(t)}$ is the pixel index at time t of a point that was not in the FOV at time $t-1$, ρ_s is the correlation coefficient accounting for the correlation between the frames' portion of size α representing the same part of the scene in $t-1$ and t, $N_t(i_\alpha^{(t-1)}) \sim \mathcal{N}(0, (1-\rho_s^2)\sigma_t^2)$ represents the innovation within such part of the frame t with respect to the frame $t-1$, while $N_{\overline{\alpha}}(i_{\overline{\alpha}}^{(t)}) \sim \mathcal{N}(0, \sigma_t^2)$ represents the novel portion of the scene that enters in the FOV at time t. Note that $i^{(t-1)}$ plays a similar role as the motion vectors in commercial standards such as H.264. The dependency of α on the speed is not explicitly indicated in Eqn. (2) to keep notation simple. For consistency, we assume $X_{t-1}(i_{\overline{\alpha}}^{(t)}) = 0$ and $N_{\overline{\alpha}}(i_\alpha^{(t-1)}) = 0$. Note that if the camera is not moving

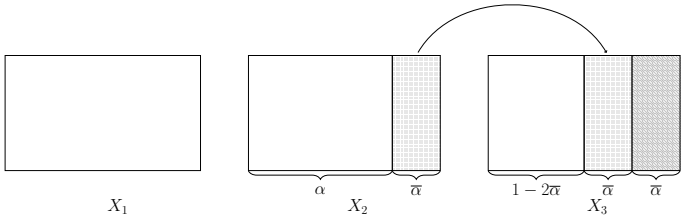

Figure 1. Source model. A portion α of X_2 is correlated with a portion of the same size in X_1. The remaining portion $\overline{\alpha} = 1 - \alpha$ of X_2 is independent of X_1. Due to the camera motion such portion is shifted in X_3 with respect to its position in X_2. The pixel values in such portion are correlated with those in the same portion in X_2. In the picture $\overline{\alpha} = 1/4$.

($v = 0$), in (2) $N_{\overline{\alpha}}(i_{\overline{\alpha}}^{(t)}) = 0$ and $i^{(t)} = i^{(t')}, \forall t, t'$, while in case of infinite speed $X_t(i^{(t)}) = N_{\overline{\alpha}}(i_{\overline{\alpha}}^{(t)})$, $\forall t, i$.

[2]These assumptions only affect the source correlation model as a function of the speed and the trajectory.

[3]We further discuss the impact of such assumption in Section (VI).

[4]A triplet of discrete random variables (rr.vv.) X, Y, Z forms a Markov chain in that order (denoted $X - Y - Z$) if their joint probability mass function satisfies $p(x,y,z) = p(x)p(y|x)p(z|y)$ [15]. The definition extends in a similar way to the case of continuous rr.vv..

B. Channel Model

The channel between the transmitter and the receiver is modelled as a frequency-flat block fading channel. The complex channel coefficient h is assumed to stay constant during a *coherence period* of T_c seconds and then takes a new value which is correlated with the previous one. A Markov autocorrelation model is assumed, so that the following holds:

$$h^{(r)} = \rho_c h^{(r-1)} + (1-\rho_c)w, \tag{3}$$

where $\rho_c \in [0,1]$ is the correlation coefficient between two consecutive channel coefficients, w is a zero-mean unit-variance complex Gaussian random variable (r.v.) which is independent of $h^{(r-1)}$, while the index r identifies the different coherence periods. According to the Clarke model [16], the faster the source, the shorter the coherence time. Specifically, we assume that $T_c = 1/f_D$ where f_D is the Doppler frequency defined as $f_D = f_0 v/c$, with f_0 being the central frequency of the communication channel while c is the speed of light. For the sake of clarity and without loss of generality, we consider in the following only the cases in which $T_f = kT_c$ and $T_f = T_c/k$, where k is a positive integer [5], corresponding to high-speed (shorter coherence time) and low-speed (longer coherence time) scenarios, respectively.

III. Transmitter Side

A. Source Encoder

Given a frame $\mathbf{X}_t$, the source encoder generates a compressed version that can be described with the least number of bits per symbol (bps) while satisfying a constraint on the error (*distortion*) between the corresponding reconstruction $\widehat{\mathbf{X}}_t$ and $\mathbf{X}_t$ [17]. We consider a per-frame Mean Squared Error (MSE) average distortion metric:

$$d_t^{(n)}\left(\mathbf{X}_t, \widehat{\mathbf{X}}_t\right) \triangleq \frac{1}{n}\sum_{i=1}^{n}\left(X_t(i) - \widehat{X}_t(i)\right)^2. \tag{4}$$

Let us consider a target distortion tuple $\mathbf{D} = (D_1, D_2, \ldots, D_M)$. It is required that, for large n, the average distortion for frame number t is lower than or equal to D_t, i.e.,

$$\lim_{n\to\infty} \mathbf{E}\left\{d_t^{(n)}\left(\mathbf{X}_t, \widehat{\mathbf{X}}_t\right)\right\} = \mathbf{E}\left\{d_t\left(\mathbf{X}_t, \widehat{\mathbf{X}}_t\right)\right\} \leq D_t, \tag{5}$$

where the average is taken with respect to the distribution of the source vectors. In the following we assume $D_t < \sigma_t^2$ $\forall t \in \{1,2,\ldots,M\}$. The source encoder we consider is a modified version of an idealized DPCM encoder. Such source encoder has been shown in [18] to be optimal, in the sense that it achieves the minimum sum-rate when an MSE distortion measure is adopted for all distortion values within the sum-rate-distortion region. Unlike in [18], the source encoding considered in the present paper depends on the speed v. The two coincide for $v = 0$. The details of the encoder are presented in the following for $v = 0$ and $v > 0$.

[5]Note that it can be $T_f/T_c < 1$, e.g., if $v \simeq 0$.

1) Source encoder with $v = 0$: When the first frame of a GOP $\mathbf{X}_1$ is captured and made available at the source encoder, it is quantized, using an encoding function $f_1^n(.)$, into the description $\mathbf{U}_1 = f_1^n(\mathbf{X}_1)$ which can be described using a source codebook of rate $R^{(1)}(D_1)$ bits per source symbol (pixel). The encoding of the first frame in each GOP is done independently of all previous frames. Once the encoding of the first frame is completed, the index of the descriptor $\mathbf{U}_1$ is passed on to the channel encoder. When, after T_f seconds, the second frame $\mathbf{X}_2$ is generated, the source encoder compresses it applying the encoding function $f_1^n(.)$, taking into account $\mathbf{X}_1$, $\mathbf{X}_2$ and $\mathbf{U}_1$ and outputs the descriptor $\mathbf{U}_2 = f_2^n(\mathbf{X}_1, \mathbf{X}_2, \mathbf{U}_1)$ having rate $R^{(2)}(D_2)$ bps. In general, the t-th frame in a GOP is compressed taking into account all available frames $\mathbf{X}^t = \mathbf{X}_1, \ldots, \mathbf{X}_t$ and all available encoder outputs $\mathbf{U}^{t-1} = \mathbf{U}_1, \mathbf{U}_2, \ldots, \mathbf{U}_{t-1}$. Frames are sequentially source-encoded in groups of M, where M is the product between the GOP duration expressed in seconds and the frame rate F_r expressed in Hz. This models an IPPPP video compressor in which a given frame within a GOP can be reconstructed only if all previous source-coded frames of the same GOP are available at the decoder. A similar source coding scheme was studied in [12], in which two correlated source vectors were successively generated and encoded. In [18] the rate-distortion region (RDR) for a generic number of frames with generic encoding and decoding delays is derived. In [13, Theorem 1] the distortion-rate region (DRR) for a generic number of Gauss-Markov sources is fully characterized. We recall such result in the following, stated in terms of RDR. Binary logarithms are considered throughout the paper.

Theorem 1. *[13] The rate-distortion region for M successive correlated Gauss−Markov sources and MSE distortion for a distortion tuple $\mathbf{D}$ is given by all rate tuples $\mathbf{R}$ that satisfy $R^{(t)} \geq R^{*(t)}$, where*

$$R^{*(t)}(D_t) = \frac{1}{2}\log^+\left(\frac{\sigma_{W_t}^2}{D_t}\right) \tag{6}$$

while

$$\sigma_{W_t}^2 = \begin{cases} \sigma_1^2, & \text{for } t = 1 \\ \rho_s^2 \frac{\sigma_t^2}{\sigma_{t-1}^2} D_{t-1} + (1-\rho_s^2)\sigma_t^2, & \text{for } t > 1. \end{cases} \tag{7}$$

and $\log^+(x) = \max(0, \log(x)), x > 0$.

From Theorem (1) the minimum distortion for the k-steps lookahead of successive correlated Gauss-Markov sources can be derived. In the following corollary an upper bound to such distortion is provided. A sketch of the proof is available in [19, Corollary 1].

Corollary 1. *Given t, $t > 0$, successive correlated Gauss-Markov sources of which the first $t-k$, $0 \leq k < t$, are source encoded within the RDR for a given distortion tuple $\mathbf{D} = (D_1, \ldots, D_{t-k})$ and the relative reconstructions $\widehat{X}_1, \ldots, \widehat{X}_{t-k}$ are available at the source decoder, the minimum distortion*

achievable for source X_t in case Gaussian descriptions are considered is:

$$\sigma^2_{W_{t,k}} = \rho_s^{2k} \frac{\sigma_t^2}{\sigma_{t-k}^2} D_{t-k} + \left(1 - \rho_s^{2k}\right) \sigma_t^2, \tag{8}$$

where $D_0 = 0$.

For the sake of clarity we assume in the following that all frames have to be recovered with a distortion at playback lower than or equal to D, i.e., $D_t = D \ \forall t$.

2) Source encoder with $v > 0$: In the generic case of $v \geq 0$, the rate-distortion region of the moving source model presented in Section (II-A) is given in the following theorem:

Theorem 2. *The rate-distortion region for the source model presented in Section (II-A) for a distortion D is given by all rate tuples $\mathbf{R}$ that satisfy $R^{(t)} \geq R'^{(t)}$, where*

$$R'^{(t)}(D) = \frac{1 - \tilde{t}\overline{\alpha}}{2} \log^+ \left(\frac{\sigma^2_{W_{\tilde{t}+1}}}{D} \right) + \frac{\overline{\alpha}}{2} \sum_{i=1}^{\tilde{t}} \log^+ \left(\frac{\sigma^2_{W_i}}{D} \right), \tag{9}$$

$\tilde{t} \triangleq \min\left\{t-1, \left\lceil \frac{1}{\overline{\alpha}} \right\rceil - 1\right\}$, while $\sigma^2_{W_{\tilde{t}}}$ is given in Eqn (7).

Proof. Let us consider a generic frame t. We distinguish two cases: (i) $t \leq \left\lceil \frac{1}{\overline{\alpha}} \right\rceil$ and (ii) $t > \left\lceil \frac{1}{\overline{\alpha}} \right\rceil$. In case (i) a frame can be divided into t parts, $t-1$ of them occupying a fraction $\overline{\alpha}$ of the frame each, and one occupying a fraction $1 - t\overline{\alpha}$. The latter is present since the first frame while each of the others entered the FOV in one of the successive frames. From Theorem 1 it follows that the part which is present in the FOV *since* t' frames must be encoded with a rate at least $R^{*(t')}(D)$. Thus, its contribution to the overall rate of frame t is

$$\begin{cases} \frac{\overline{\alpha}}{2} \log^+ \left(\frac{\sigma^2_{W_{t'}}}{D} \right), & \text{for } 2 \leq t' < t \\ \frac{1 - t\overline{\alpha}}{2} \log^+ \left(\frac{\sigma^2_{W_{t'}}}{D} \right) & \text{for } t' = t. \end{cases} \tag{10}$$

In case (ii) a frame is composed of $\left\lceil \frac{1}{\overline{\alpha}} \right\rceil$ parts, each of which entered the FOV at a different time. The contribution of the part present since t' frames to the overall rate for frame t is as in the first line of (10), for $t' \in \{1, \ldots, \left\lceil \frac{1}{\overline{\alpha}} \right\rceil - 1\}$, wile for $t' = \left\lceil \frac{1}{\overline{\alpha}} \right\rceil$ it is given by the second line. Defining $\tilde{t} \triangleq \min\left\{j-1, \left\lceil \frac{1}{\overline{\alpha}} \right\rceil - 1\right\}$ and adding up the contributions of the different parts leads to the expression (9). Achievability and converse follow directly from [13]. □

Note that in case $\alpha = 1$ the setup is the same as in Theorem 1 (Gauss-Markov sources) and Eqn. (9) takes the form of (6), while if $\alpha = 0$ the setup reduces to completely uncorrelated Gaussian frames and Eqn. (9) takes the form of the well-known rate-distortion function (RDF) for a Gaussian source.

As an example, let us consider the first two frames. The minimum rate required for the first frame ($t = 1$) to be reconstructed with distortion D is $R^{*(1)}(D) = 1/2 \log^+ \left(\sigma^2_{W_1}/D\right)$. In the successive frame ($t = 2$) a portion α is correlated with a portion of the same size in frame 1, while the remaining $\overline{\alpha}$ portion of frame 2 is completely novel. From (1) it follows that a rate $R^{*(2)}(D) = 1/2 \log^+ \left(\sigma^2_{W_2}/D\right)$ is sufficient to achieve a distortion D in the portion of width α, while the rest of the frame requires a rate $R^{*(1)}(D)$, since it was not present in frame 1. Note that such rate constraints are required in order to achieve a distortion D both in the part that was already present in frame 1 as well as in the novel part. The average rate in bps required for frame 2 is:

$$R^{(2)}(D) = \frac{\alpha}{2} \log^+ \left(\frac{\sigma^2_{W_2}}{D} \right) + \frac{\overline{\alpha}}{2} \log^+ \left(\frac{\sigma^2_{W_1}}{D} \right). \tag{11}$$

At this point, if $\overline{\alpha} < 1/2$, then the part of the frame present since $t = 1$ will be carried over to the third frame (case (i)), while it will not be present otherwise (case (ii)).

B. Channel Encoder

A new message is made available to the channel encoder by the source encoder every T_f seconds. The message is the result of the source compression applied to the last captured frame while taking into account all source frames belonging to the same GOP available so far, as well as the encoded frames. Due to the short-term delay constraint the message is transmitted right away, i.e., before the next message is available, and the transmission is completed within T_f seconds. We consider a joint encoder (JE) that every T_f seconds jointly encodes all available (compressed) frames in the current GOP. This is done by using a new codebook at each transmission. The set of codebooks is randomly generated according to an i.i.d. Gaussian distribution and revealed to the transmitter and to the receiver before transmissions start. At the beginning of time slot t a new message of length $nR^{(t)}(D)$ bits is made available from the source encoder. The channel encoder uses all source messages provided by the source encoder within the current GOP and uses them as index to select the codeword to be transmitted over the channel. Thus, the size of the codebook $\mathcal{C}^{(t)}$ used by the channel encoder during slot t is $\left|\mathcal{C}^{(t)}\right| = 2^{nR_\Sigma^{(t)}}$, where $R_\Sigma^{(t)} = \sum_{i=1}^{t} R^{(i)}(D)$. Note that we implicitly assumed that the number of symbols in the source vector to be compressed is the same as the number of symbols in a channel codeword. Each codeword $\mathbf{x}_t$ is transmitted with average power P. According to our channel model, the coherence time of the channel T_c depends on the speed. Since the length of the channel codeword is kept fixed, different fading levels may affect different parts of the codeword.

Note that the increase in complexity at the source encoder-channel encoder with respect to the approach used in commercial systems is limited. The only difference is the joint channel encoding of all available frames in the current GOP. Most of the complexity is moved to the channel decoder-source decoder side, which is appealing in setups in which the compressor-transmitter has limited computational capabilities.

IV. Receiver Side

A channel codeword is transmitted every T_f seconds. If $T_f > T_c$ (source moving at high speed) the codeword experiences different fading levels in different parts, while in case $T_f < T_c$ a constant fading level is observed. Under the assumptions of Section (II-B) a codeword will experience a number of distinct fading levels equal to $N_{\text{fade}} = \lceil T_f T_c \rceil$ each

affecting n/N_{fade} symbols. During the transmission of the t-th frame, the channel decoder observes the signal:

$$\mathbf{y}_t^{(r)} = h_t^{(r)}\mathbf{x}_t^{(r)} + \mathbf{w}_t^{(r)}, \tag{12}$$

for $r = 1, \ldots, N_{\text{fade}}$ where $\mathbf{x}_t^{(r)}$ is the portion of the codeword $\mathbf{x}_t$ affected by the r-th channel fading level $h_t^{(r)}$, which is modelled as in Eqn. (3), while $\mathbf{w}_t^{(r)}$ is a noise vector of i.i.d. complex Gaussian r.v. with zero mean and unit variance.

A. Channel Decoder

The channel decoder attempts to decode jointly all available messages relative to the current GOP each time the the transmission of a frame is completed. This is done by applying joint typicality decoding on the vectors $\mathbf{y}_1, \ldots, \mathbf{y}_t$, t being the index of the last received frame. The decoding is successful with high probability iff [20] [21]:

$$\begin{aligned} R^{(t)}(D) &\leq C_t \\ R^{(t)}(D) + R^{(t-1)}(D) &\leq C_t + C_{t-1} \\ &\vdots \\ R^{(t)}(D) + \ldots + R^{(1)}(D) &\leq C_t + \ldots + C_1, \end{aligned} \tag{13}$$

where $C_i = 1/N_{\text{fade}} \sum_{r=1}^{N_{\text{fade}}} \log_2 \left(1 + P \left|h_i^{(r)}\right|^2\right)$.

B. Source Decoder

The source decoder reconstructs the frames as data are made available from the channel decoder. A frame is reconstructed using the data retrieved from the channel together with the side information provided by previously decoded frames and is displayed in order to get a real-time (*on-line*) video steam. Once the transmission of a whole GOP has finished, the decoder reconstructs a *playback* version of the video taking into account all frames in the GOP that have been correctly decoded. Given the structure of the source and channel encoders, if conditions (13) are fulfilled for a given t the corresponding GOP can be reconstructed up to frame t at the desired distortion D. Formally, given a set of descriptors $\mathbf{U}^t$ made available from the channel decoder, the source decoder applies a decoding function $g_t^n(.)$ to obtain the reconstruction of frame t $\widehat{\mathbf{X}}_t = g_t^n(\mathbf{U}^t)$ which achieves a distortion D with high probability. If any of such conditions is not satisfied, frame t cannot be decoded. In this case the best reconstruction of the current frame based on the previously decoded ones, on the correlation structure and on the information about the speed v is created at the decoder[6]. Let t^- be the most recent frame that can be decoded at time $t, t \geq t^-$. The decoder outputs the best approximation of frame t: $\widehat{\mathbf{X}}_t = g_t^n(\mathbf{U}^{t^-})$ which achieves a distortion larger than or equal to D.

Note that, even if conditions (13) for frame t are not satisfied, they could be satisfied for some $t^+ > t$. In such case all frames in $\{1, 2, \ldots, t^+\}$ can be decoded. Although frames previous to t^+ do not contribute to decrease the on-line distortion, they decrease the playback distortion. On-line and playback distortion are formally defined in the following:

Definition 1. *The on-line distortion is the average across a GOP of the MSE of the different frames, each of which is reconstructed using the channel output relative to the frames up to the considered one only, i.e.:* $\widehat{\mathbf{X}}_t = g_t^n(\mathbf{U}^{t^-})$.

Definition 2. *The playback distortion is the average across a GOP of the MSE of the different frames, each of which is reconstructed using all channel outputs relative to the whole GOP, i.e.:* $\widehat{\mathbf{X}}_t = g_t^n(\mathbf{U}^{\min(t,t^*)})$*, where* t^* *is the largest* t *within an* M*-frames GOP for which conditions (13) hold.*

Using Theorem 2 we derived the average on-line and playback distortions across the frames in a GOP for the considered setup. They are given in the following Propositions where, for the sake of clarity, we set $\sigma_i^2 = \sigma^2$, $\forall i \in \{1, \ldots, M\}$.

Proposition 1. *The average on-line distortion* $\overline{D}_{ol}$ *across the frames of a GOP is:*

$$\overline{D}_{ol} = \frac{1}{M}\left(\sum_{i=1}^{M} \mathbb{1}_i D + \sum_{i=1}^{M} (1 - \mathbb{1}_i)\, D_{l,ol}^{(t)}\right) \tag{14}$$

where $\mathbb{1}_i$ *is the indicator function taking value* 1 *if the conditions (13) for* $t = i$ *are satisfied and* 0 *otherwise, while* $D_{l,ol}^{(t)}$ *is the distortion of frame* t *and is given by the expression in (6), where* $n_l(t) \in \{0, \ldots, t\}$ *is the number of consecutive lost frames up to number* t *(i.e., frames* $t - n_{l(t)} + 1, t - n_{l(t)} + 2, \ldots, t$ *are lost),* $\tilde{t} \triangleq \min\left\{t - 1, \lceil \frac{1}{\tilde{\alpha}} \rceil - 1\right\}$*, while* $\sigma^2_{W_{t,k}}$ *is given in Corollary* 1.

Proof. (Sketch) Given a frame t, if conditions (13) are fulfilled all messages up to t can be decoded. By Theorem 1 this allows to achieve a distortion D (first sum at the right hand side of Eqn. (14)). If conditions (13) are not fulfilled, the last message (output of the source encoder at time t) is not available at the source decoder. In this case the decoder locates the last available reconstruction achieving the desired distortion D and, using knowledge of the source correlation and α, creates an estimate of frame t. The distortion of such estimate is given in (6), which can be derived by using (8) and proceeding in a similar way as in the derivation of Theorem 2. □

Note that in (6), if $n_l(t) \leq \min\left\{t, \lceil \frac{1}{\tilde{\alpha}} \rceil\right\}$, the distortion is equal to the source variance σ^2. This takes into account the fact that, if the source is moving and many consecutive frames are lost, the current frame may be uncorrelated with respect to the last decoded one.

Proposition 2. *The average playback distortion* $\overline{D}_{pb}$ *across the frames of a GOP is:*

$$\overline{D}_{pb} = \frac{1}{M}\left(t^* D + \sum_{t=t^*+1}^{M} D_{l,ol}^{(t)}\right). \tag{15}$$

Proof. (Sketch) Since no delay constraints are in place, the decoder searches for the largest t for which conditions (13) are satisfied, indicated as t^*. All frames up to the t^*-th can be decoded and achieve the target distortion D. Successive

[6]This is done taking into account that the distortion in the reconstruction of a lost frame is equal to the variance of the innovation of the frame with respect to previous one [12].

$$D_{\text{l,ol}}^{(t)} = \begin{cases} \left(1-\tilde{t}\overline{\alpha}\right)\sigma^2_{W_{\tilde{t}+1,n_l(t)}} + \overline{\alpha}\sum_{i=\max\{\tilde{t}-n_l(t),n_l(t)+1\}}^{\tilde{t}} \sigma^2_{W_{i,n_l(t)}} + \min\{n_l(t),\tilde{t}\}\overline{\alpha}\sigma^2, & \text{if } n_l(t) \leq \min\left\{t, \lceil \frac{1}{\overline{\alpha}} \rceil\right\} \\ \sigma^2 & \text{otherwise,} \end{cases} \tag{6}$$

frames $\{t^*+1, t^*+2, \ldots, M\}$ are lost and their distortion can be calculated using Corollary (1) and considering a number of lost packets equal to $\{1, 2, \ldots, M-t^*\}$, respectively. □

Note that both $\overline{D}_{\text{ol}}$ and $\overline{D}_{\text{pb}}$ are functions of the channel coefficients, the target distortion D and the power P. The dependence on the speed of the source is implicit in the channel model as well as in the source model.

V. Benchmarks

1) Memoryless Transmission: We compare the performance of the JE system in terms of on-line distortion with that of a system that transmits each source-encoded frame independently. This approach, which we call memoryless transmission (MT) is similar to what is done in today's commercial systems, in which channel coding at the physical layer is done independently of the source coding [7]. With MT a message transmitted in slot t is successfully decoded with high probability if $C_t \geq R_t$. The on-line and the playback distortions for this scheme coincide and are given in the following proposition.

Proposition 3. *The average playback (and on-line) distortion $\overline{D}_{IC}$ across the frames of a GOP for the memoryless transmission scheme is:*

$$\overline{D}_{IC} = \frac{1}{M}\left(t^* D + \sum_{t=t^*+1}^{M} D_{l,ol}^{(t)}\right) \tag{16}$$

where t^* is the first frame that has been lost.

Proof. As in Proposition (15), since a frame can be reconstructed only if all previous ones are available, once a frame is lost all those that are received later on within the same GOP cannot be used an thus can be reconstructed with a distortion not lower than $D_{\text{l,ol}}^{(t)}$. □

2) Playback-optimal Coding: If the real-time requirement is removed, the encoder can wait to have a whole GOP available before transmitting it. The optimal way to do it if no Channel State Information (CSI) is available at the transmitter is to jointly encode the frames in a single codeword which is transmitted over the channel during the next MT_f seconds [22]. Such playback-optimal (PB-optimal) encoder is used as a benchmark for the playback distortion.

VI. Numerical Results

We compare the performance of the JE with the PB-optimal scheme in terms of the statistical mean of the average playback distortion $E\{\overline{D}_{\text{pb}}\}$ and with the MT scheme in terms of mean on-line distortion $E\{\overline{D}_{\text{ol}}\}$, $E\{.\}$ being the expectation operation. The expectation is calculated using the Monte Carlo method and averaging across the channel realizations. A total of 10^4 GOP transmissions have been simulated for each selected value of M. The corresponding number of channel blocks is calculated based on the speed and according to the channel model presented in Section (II-B). The relationship between α and the speed v is assumed to be $\alpha = [1-v]^+$, where $[x] = \max\{0, x\}$. The variance of each new portion entering the FOV is $\sigma_i^2 = \sigma^2 = 1\ \forall i$, the correlation coefficient ρ_s is set to 0.5 while the transmit power is $P = 2$ dB. The target playback distortion has been set to $D = D_{\text{pb}}^{\text{target}} = 0.2$.

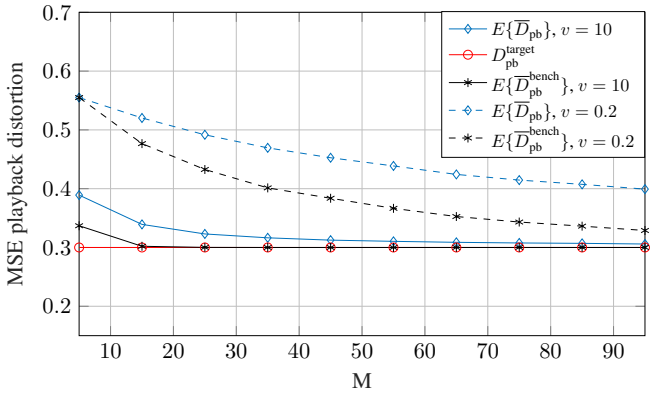

Figure 2. Playback distortion at high speed ($v = 10$) and low speed ($v = 0.2$). $\overline{D}_{\text{pb}}^{\text{bench}}$ is the distortion achieved by the PB-optimal scheme. The following parameters were used: $P = 2$ dB, $\rho_s = 0.5$, $\sigma^2 = 1$, $D = 0.2$.

In Fig. (2) the playback distortion is plotted versus the GOP size for two speed values, namely $v = 10$ and $v = 0.2$. In the high speed case the joint encoding scheme performs close to the benchmark and gets to within 2% of the target distortion for $M = 95$. In the low-speed case the convergence to the target distortion is slower. This is due to the lower time diversity experienced by each GOP during transmission. In Fig. (3) the

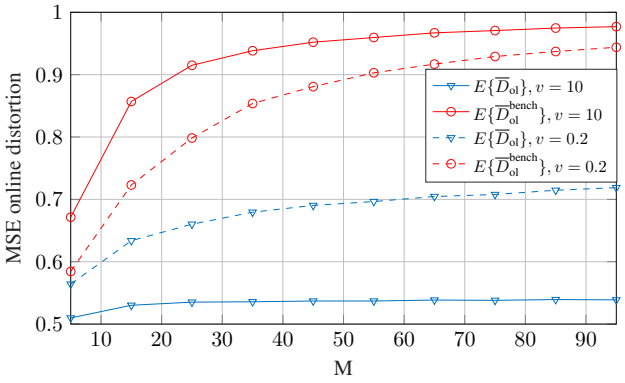

Figure 3. On-line distortion at high speed ($v = 10$) and low speed ($v = 0.2$). $\overline{D}_{\text{ol}}^{\text{bench}}$ is the distortion achieved by the MT scheme. The following parameters were used: $P = 2$ dB, $\rho_s = 0.5$, $\sigma^2 = 1$, $D = 0.2$.

on-line distortion for the joint encoding and for the benchmark

[7]There are multimedia streaming systems that apply packet level coding which takes source compression into account, but in such systems the physical layer is usually fixed.

MT scheme is plotted versus the GOP size for high and low speed. The JE scheme shows a significantly lower distortion with respect to the benchmark (up to around 44% in the high-speed case). This is thanks to the fact that the JE scheme naturally matches the DPCM source coding. As a matter of fact, the latter is very sensitive to frame losses, since all frames that follow a lost one in the same GOP cannot be used. The JE scheme can cope with this thanks to the fact that it allows to retrieve all frames from the first one (I frame) until the current one each time the decoding is successful. From the figure it can been also seen that the gain of the JE scheme with respect to the MT one is larger at higher speed despite the higher source decoder output rate. This is because, as the channel diversity increases due to the higher v, the JE scheme averages out the channel fluctuations more efficiently. From the two figures it can be seen that the playback distortion decreases monotonically with the GOP size M, while the on-line distortion increases as M increases, which indicates that there is a trade off between the two.

The results presented suggest that using a JE to transmit correlated sources over a block fading channel can help mitigate the impact of packet losses on the video quality and reveal a tradeoff between the online and the playback distortion. Care should be taken when comparing such results with the performance of real source encoders mounted on mobile platforms. In particular, the rate-distortion function of the considered source is likely to be an upper bound for that of practical codecs. This is due on the one side to the correlation among the pixels of real images, which decreases the rate required for a given distortion, and on the other side to the fact that we considered a Gaussian source, which is known to require a rate that is an upper bound for any other distribution for a given distortion [15]. Both points can be addressed by opportunely modifying the model, e.g., as done in [23] for the case of still images. Such modifications are not presented here for a matter of space and we leave them as a promising topic for future work, together with the impact of more general camera trajectories.

VII. Conclusion

We studied the source and channel coding of a moving Gauss-Markov source streamed over a block fading channel. We introduced a modified Gauss-Markov source model that takes the effect of motion into account and explicitly calculated its rate-distortion region. The interdependency generated by the DPCM is accounted for in the proposed scheme by applying a progressive joint-encoding scheme, which allows to retrieve all frames received up to the decoding instant. We derived the distortion in the source reconstruction obtained in case a JE and an MT scheme are used to transmit the frames. Our numerical results show that applying a progressive joint encoder to the DPCM-compressed stream coming from the considered source allows to approach the distortion of a playback-optimal compressor as the GOP size M grows while achieving an on-line distortion that is much smaller than that of the MT benchmark. Furthermore, we observed that, for the considered simulation setup, such gain increases with the speed v of the platform despite of the higher output rate of the source encoder, which is due to the fact that the JE averages out the channel fluctuations more efficiently when the time diversity increases. Finally, we observed that there is a trade-off between on-line and playback distortion, in that the former increases with M while the latter benefits from a larger GOP.

References

[1] DJI, https://www.dji.com/.

[2] FlyAbility, https://www.flyability.com/.

[3] Gamaya, http://gamaya.com.

[4] W. Amaya, H. Yu, and M. Moreno, "Ux impacts of haptic latency in automotive interfaces," Immersion Corporation, Tech. Rep., 2013.

[5] A. Cherpillod, S. Mintchev, and D. Floreano, "Embodied flight with a drone," *CoRR*, vol. abs/1707.01788, 2017. [Online]. Available: http://arxiv.org/abs/1707.01788

[6] *Recommendation ITU-T H.264; Infrastructure of audiovisual services Coding of moving video*, http://www.itu.int/rec/T-REC-H.264.

[7] R. Zhang, S. L. Regunathan, and K. Rose, "Video coding with optimal inter/intra-mode switching for packet loss resilience," *IEEE J. on Sel. Areas in Commun.*, vol. 18, no. 6, pp. 966–976, June 2000.

[8] C. E. Shannon, "A mathematical theory of communication," Bell Labs, Bell Sys. Tech. J., Tech. Rep., July 1948.

[9] M. Gastpar, B. Rimoldi, and M. Vetterli, "To code, or not to code: lossy source-channel communication revisited," *IEEE Trans. on Info. Theory*, vol. 49, no. 5, pp. 1147–1158, May 2003.

[10] S. Vembu, S. Verdú, and Y. Steinberg, "The source-channel separation theorem revisited," *IEEE Trans. on Info. Theory*, vol. 41, no. 1, pp. 44–54, Jan. 1995.

[11] D. Gündüz and E. Erkip, "Joint source-channel codes for MIMO block-fading channels," *IEEE Trans. on Info. Theory*, vol. 54, no. 1, pp. 116–134, Jan. 2008.

[12] H. Viswanathan and T. Berger, "Sequential coding of correlated sources," *IEEE Trans. Inform. Theory*, vol. 46, no. 1, pp. 236–246, Jan. 2000.

[13] A. Khina, V. Kostina, A. Khisti, and B. Hassibi, "Sequential coding of Gauss-Markov sources with packet erasures and feedback," in *IEEE Info. Theory Workshop (ITW)*, Kaohsiung, Taiwan, 6-10 Nov. 2017.

[14] E. Yang, L. Zheng, D. He, and Z. Zhang, "Rate distortion theory for causal video coding: Characterization, computation algorithm, and comparison," *IEEE Trans. on Info. Theory*, vol. 57, no. 8, pp. 5258–5280, Aug. 2011.

[15] T. Cover and J. Thomas, *Elements of Information Theory*. John Wiley and Sons, 2006.

[16] R. Clarke, "A statistical theory of mobile-radio reception," Bell Labs, Tech. Rep., 1968.

[17] T. Berger, *Rate distortion theory: a mathematical basis for data compression*, T. Kailath, Ed. Prentice-Hall, Inc. Englewood Cliffs, New Jersey, 1971.

[18] N. Ma and P. Ishwar, "On delayed sequential coding of correlated sources," *IEEE Trans. Info. Theory*, vol. 57, no. 6, pp. 3763–3781, June 2015.

[19] G. Cocco and L. Toni, "The sum-rate-distortion region of correlated Gauss-Markov sources," [Online] https://arxiv.org/abs/1804.03420, 2018.

[20] G. Cocco, D. Gündüz, and C. Ibars, "Streaming transmission over block fading channels with delay constraint," *IEEE Trans. on Wireless Commun.*, vol. 12, no. 9, pp. 4315–4327, Sep. 2013.

[21] V. V. Prelov, "Transmission over a multiple access channel with a special source hierarchy," *Problemy Peredachi Informatsii*, 1984.

[22] D. Tse and P. Viswanath, *Fundamentals of Wireless Communication*. Cambridge University Press, 2005.

[23] B. Girod, "Rate distortion theory," [Online] https://web.stanford.edu/class/ee368b/Handouts/04-RateDistortionTheory.pdf, 2000.