RETINA-INSPIRED SPATIO-TEMPORAL FILTERING FOR DYNAMIC VIDEO CODING

Effrosyni Doutsi[*†], *Marc Antonini*[*] *member IEEE, and Panagiotis Tsakalides*[†]

[*]Université Côte d'Azur, I3S, CNRS
am@i3s.unice.fr

[†]Institute of Computer Science, FORTH
edoutsi@isc.forth.gr

2022 IEEE International Conference on Image Processing (ICIP) | 978-1-6654-9620-9/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICIP46576.2022.9897355

ABSTRACT

The goal of this work is to propose a simple yet efficient way to dynamically transform a sequence of images according to the functional properties of the visual system. To achieve this goal, we extend to video sequences the Retina-Inspired Filter (RIF), which we have recently proposed for still images. Under the assumption that the input signal remains constant for a given time, the RIF decomposition was proven to be invertible, meaning that the image could be perfectly recovered. In this paper, we relax this assumption into a piece-wise constant input and we prove that RIF can be applied to a Group Of Pictures (GOP). Under the condition that a GOP consists of frames without strong pixel motion, we mathematically prove and experimentally show that when RIF is applied to GOP, whatever the size of the GOP is, we are still able to perfectly recover the video frames and at the same time simplify the complexity of the whole process. In addition, we show that while the GOP size increases, the memory cost required to store this amount of frames is sufficiently reduced.

Index Terms— Center-surround filter, dynamic transform, video compression, retina, neuro-inspired filtering, perception-based processing.

I. INTRODUCTION

It is no secret that the internet traffic today is due mostly to videos. In fact, by 2022 online videos are expected to make up more than 82% of all consumer internet traffic. Thus, there are several open challenges that need to be addressed including adaptive video streaming, video pre-processing and storage, and video understanding. In this work, we are interested in video compression which is one of the most crucial steps in the entire pipeline of video streaming. The state-of-the-art in video compression is the Versatile Video Coding (VVC) or H.266 that was specifically designed for 4K and 8K streaming, achieving significant bit rate reduction in the neighborhood of 50% over its predecessor, the HEVC, and 75% over AVC, currently the most-widely used format. However, VVC is yet another computationally expensive compression scheme that treats a dynamic signal in a stationary manner. A number of intra and inter frame predictions are required, and the motion compensation process is based on a large number of comparisons among sequential frames, when the spatio-temporal redundancy is extremely high such as in Ultra High Definition (UHD) video.

Seeking for alternative and computationally more efficient solutions capable of processing a video stream in a dynamic manner, researchers focused on biological systems whose performance seems to be beyond the current state-of-the-art. The brain with its visual system has been considered a special biological "device" to mimic as it deals in real time with the UHD visual stimulus that is dynamically captured and transformed into *spike trains*, a very compact form of electrical impulses. There have been several image compression architectures motivated by neuroscience models including the Rank Order Coder [1][2], the bio-inspired Analog-to-Digital Converter [3], and the neuro-inspired image codec [4]. These methods have been applied to still images achieving performance comparable to the most widely-used image compression standards. However, none of the aforementioned architectures has ever been extended to videos to study their efficiency to dynamically filter and encode a group of sequential frames instead of a single frame.

In this work, we study the behavior of the Retina-Inspired Filter (RIF) [5], which is the first component of the neuro-inspired image codec we introduced in [4], when applied to videos. The RIF filter approximates the structure and functions of a group of cells that shape the Outer Plexiform Layer (OPL) of the retina, responsible for capturing the luminance of light that will be transformed into a sequence of electrical impulses whose density varies along time depending on the input value. If the input value is constant in time, the spike density will remain the same [6]. Under the assumption that the input signal is constant in time, the RIF filter was proven to be an invertible transform according to the frame theory as the RIF decomposition layers form a frame [7]. Consequently, the input signal can be perfectly recovered without any loss of information. Here, we employ the RIF filter to a Group of Pictures taking advantage

This work was funded by the Hellenic Foundation for Research and Innovation (HFRI) and the General Secretariat for Research and Technology (GSRT) under grant agreement No. 330 (BrainSIM) and by the Université Côte d'Azur and Campus France within the framework of the Make Our Planet Great Again (MOPGA) project with grant agreement No. 979447 (BRIEFING).

of its dynamic behavior. The advantage of considering a GOP instead of a single frame can be easily understood if one considers some special case-studies such as video surveillance, teleconferencing and telecommuting, where the spatiotemporal redundancy of GOP is extremely high as the background of the scene is most of the times almost static.

In this paper, we analytically prove that the RIF filter can be applied to a GOP if we consider the video stream as a piecewise constant signal in time. In addition, according to the frame theory if the GOP consists of frames that remain almost constant, it is possible to perfectly recover the input signal. Additionally, we experimentally show that the GOP size is highly important when we also consider some compression mechanism after the RIF filtering, because the entropy can be highly reduced while the reconstruction quality remains high. In the rest of the paper, we first provide an introduction to the RIF filter. Then, we present the mathematical framework regarding the RIF transform on videos and we study via experiments the effect of the GOP size. Finally, we draw some conclusions and we give a short discussion about future work.

II. BACKGROUND ON RETINA-INSPIRED FILTER

The RIF filter, $K(\boldsymbol{x},t)$, is connected to the dynamic behavior of the retina and increases the sharpness of the visual stimulus during filtering before its transmission to the brain. The visual stimulus is a 3D spatio-temporally varying signal $I(\boldsymbol{X},t)$ with $\boldsymbol{X} \in \mathbb{R}^3$ and $t \in \mathbb{R}$. This 3D visual stimulus is projected onto the retina via the lens hence it is finally simplified into a 2D luminance $f(\boldsymbol{x},t)$ where $\boldsymbol{x} \in \mathbb{R}^2$ and $t \in \mathbb{R}$. As a result, the impact of the retina can be described as the following spatio-temporal transform:

$$A(\boldsymbol{x},t) = K(\boldsymbol{x},t) \overset{\boldsymbol{x},t}{*} f(\boldsymbol{x},t), \tag{1}$$

where $\overset{\boldsymbol{x},t}{*}$ denotes the spatio-temporal convolution between the projection of the visual stimuli and the retina transform, defined as the difference of two spatio-temporal functions:

$$K(\boldsymbol{x},t) = C(\boldsymbol{x},t) - S(\boldsymbol{x},t), \tag{2}$$

$$C(\boldsymbol{x},t) = w_c G_{\sigma_c}(\boldsymbol{x}) V(t), \tag{3}$$

$$S(\boldsymbol{x},t) = w_s G_{\sigma_s}(\boldsymbol{x}) \left(V \overset{t}{*} E_{\tau_S} \right)(t), \tag{4}$$

where w_c and w_s are constant parameters, $G_{\sigma_c}(\boldsymbol{x})$ and $G_{\sigma_s}(\boldsymbol{x})$ are spatial Gaussian filters with standard deviation σ_c and $\sigma_s = 3\sigma_c$, standing for the center and surround areas in the OPL respectively, $V(t)$ is a temporal lowpass filter, and $E_{\tau_S}(t)$ is an exponential temporal filter whose exact description and properties can be found in [5].

It was proven in [5] that under the assumption that the input signal is an image visible for a given time T,

$$f(\boldsymbol{x},t) = f(\boldsymbol{x}) \mathbf{1}_{[0,T]}(t) \tag{5}$$

for all $\boldsymbol{x} \in \mathbb{R}^2$ and all $t \in \mathbb{R}$, it is possible to simplify (1) into a spatial convolution between the temporally constant input signal $f(\boldsymbol{x})$ and the RIF filter $\phi(\boldsymbol{x},t)$ that preserves all the properties of $K(\boldsymbol{x},t)$ in space and time:

$$A(\boldsymbol{x},t) = \phi(\boldsymbol{x},t) \overset{\boldsymbol{x}}{*} f(\boldsymbol{x}), \tag{6}$$

In fact, the RIF transform forms a group of spatio-temporal Weighted Difference of Gaussian (WDoG) filters when it is applied to an image visible for a given time T, defined as:

$$\phi(\boldsymbol{x},t) = a(t) G_{\sigma_c}(\boldsymbol{x}) - b(t) G_{\sigma_s}(\boldsymbol{x}) \tag{7}$$

for all $\boldsymbol{x} \in \mathbb{R}^2$ and for all $t \in \mathbb{R}^+$, where $a(t)$, $b(t)$ are the two time-varying weights that influence the shape of the Difference of Gaussians in time as follows:

$$a(t) = w_c R_c(t) = w_c \mathbf{1}_{[0,+\infty)}(t) \int_{\max\{0,t-T\}}^{t} V(u)du, \tag{8}$$

$$b(t) = w_s R_s(t) = w_s \mathbf{1}_{[0,+\infty)}(t) \int_{\max\{0,t-T\}}^{t} (V \overset{t}{*} E_{\tau_S})(u)du. \tag{9}$$

It is worth to note that $a(t)$ and $b(t)$ have almost the same shape but the latter starts evolving with a short delay due to the exponential term $E_{\tau_S}(t)$.

III. RIF APPLIED TO A GROUP OF FRAMES

In this work, we are interested in relaxing the initial assumption that the input signal is constant for a given time T considering instead that the input varies in time.

Proposition 1. *Assume* $f(\boldsymbol{x},t) = \sum_{i=1}^{M} f_i(\boldsymbol{x}) \mathbf{1}_{[t_i,t_{i+1}]}(t)$*, for all* $\boldsymbol{x} \in \mathbb{R}^2$ *and* $t \in \mathbb{R}$*, a video stream that consists of a* M *frames that is written as a piece-wise constant signal in time where* $f_i(\boldsymbol{x})$ *is the* $i-th$ *frame which appears on time* t_i*. and is only visible when* $t_i \leq t < t_{i+1}$*. Then, the spatio-temporal convolution in (1) turns into a spatial convolution of the RIF filter* $\phi(\boldsymbol{x},t)$ *with a group of* M *frames, where each frame* $f_i(\boldsymbol{x})$ *is filtered by a sub-group of the RIF layers* $\phi_i(\boldsymbol{x},t)$ *which appear during the time window* $T_c = t_{i+1} - t_i$*, when each frame remains visible:*

$$A(\boldsymbol{x},t) = \sum_{i=1}^{M} \phi_i(\boldsymbol{x},t) \overset{x}{*} f_i(\boldsymbol{x}), \tag{10}$$

where

$$\phi_i(\boldsymbol{x},t) = a_i(t) G_{\sigma_c}(\boldsymbol{x}) - b_i(t) G_{\sigma_s}(\boldsymbol{x}) \tag{11}$$

Proof. The proof is omitted due to the lack of space. However, it is straight-forward to be verified if one simply replaces the input signal in (6) with the piece-wise constant signal as defined in Proposition making use of the *distributive* property of convolution. □

In the special case where $T_c = T$ the time during which the frame $f_i(\boldsymbol{x})$ is visible equals the time that the RIF

filter evolves, meaning that the RIF filter is applied to each different frame separately and it is only re-initialized at time t_{i+1} the next frame $f_{i+1}(\boldsymbol{x})$ appears. In this special case the size of the GOP is $M = 1$ which is the worst case scenario in terms of computational complexity and memory cost. However, in the case where $T_c \leq T$, during the time evolution of the RIF filter there are more than one frames that appear. For example, let's suppose there are $M = 3$ different frames, $f_1(\boldsymbol{x})$, $f_2(\boldsymbol{x})$ and $f_3(\boldsymbol{x})$, that appear during the time window T at time t_1, t_2 and t_3, respectively. Then, each frame will be filtered by a different sub-group of the RIF $\phi_1(\boldsymbol{x}, t)$, $\phi_2(\boldsymbol{x}, t)$ and $\phi_3(\boldsymbol{x}, t)$ where $\phi_1(\boldsymbol{x}, t)$ consists of $T/3$ layers associated to the low frequencies captured by RIF in the beginning, $\phi_2(\boldsymbol{x}, t)$ are the next $T/3$ band-pass frequency layers and $\phi_3(\boldsymbol{x}, t)$ is a group of last $T/3$ high frequency layers. The higher the number of M the less the number of the RIF layers T/M associated to each RIF sub-group (see Fig. 1).

In this work, we are interested in applying the RIF to a GOP of the size $M > 1$. As a first approximation of this attempt in this paper we only consider the scenario when there is absence of motion (globally or locally) within the GOP frames. In fact there are two different study-cases: (i) when $f_i(\boldsymbol{x})$ is the full size $i - th$ frame of a UHD video, where there is almost no pixel motion compared to the frames before and/or after, and (ii) when $f_i(\boldsymbol{x})$ is a spatial region within a video frame, where the visual content corresponds to the static background that remains constant in time.

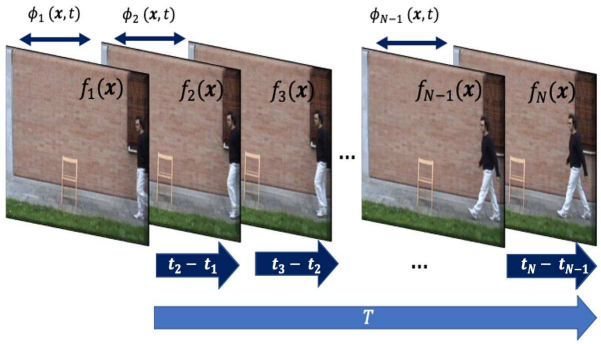

Fig. 1. This example illustrates the way the RIF filter is applied to a GOP of size $M > 1$ where each frame $f_i(\boldsymbol{x})$ appears for a given time $T_c = [t_{i+1} - t_i]$. As the RIF filter evolves for time T, each frame will be filter with T/M RIF layers.

Proposition 2. *Assume $f_1(\boldsymbol{x}), f_2(\boldsymbol{x}), \ldots, f_M(\boldsymbol{x})$ a GOP where $f_1(\boldsymbol{x}) = f_2(\boldsymbol{x}) = \ldots = f_M(\boldsymbol{x}) = f(\boldsymbol{x})$. Then, the RIF decomposition $A(x, t)$ is a frame as we have proven in [7] according to the frame theory because there exist a lower and an upper bound.*

Proof. The proof is exactly the same as in [7] with the only difference that the RIF decomposition was generated by M different frames with almost the same content. □

According to the above Proposition, we are able to perfectly recover the content of the visual scene by partially generating the RIF frame in time merging the information coming from a sequence of frames that appear in different times. The only condition to do so is that the sequence of frames should be either some frames with no pixel motion or sub-regions that belong to the constant background of the scene.

IV. EXPERIMENTAL RESULTS

This section aims at presenting the performance of the RIF filter when applied to a video stream. In our experiments we used surveillance videos from the VISOR database [8] and videos taken from the Xiph.org [9] using video where there was either a long sequence of constant frames (i.e. the "man with a dog" from VISOR or the "bowing" from Xiph.org) or the background of the visual scene was constant in several sub-regions (i.e. "deadline" "akiyo" or "vidyo4"). Our goal is twofold: (i) we show that we are able to provide high reconstruction quality when the GOP size is $M > 1$ and (ii) we show that the dynamic RIF is highly beneficial in terms of compression. To be able to provide results in terms of the rate-distortion curve we decided to use a spike-based quantizer that we released in [10]. This quantizer approximates the spike generation mechanism of neurons by an electrical resistor-capacitor circuit that enables to count the number of the emitted spikes within the observation window T_c. The number N of spikes that is defined below is the only information sent to the receiver:

$$N = N(I) = \begin{cases} 0, & \text{if} \quad RI \leq \lambda, \\ \left\lfloor \frac{T_c}{d(I)} \right\rfloor, & \text{if} \quad RI > \lambda. \end{cases} \tag{12}$$

$$d(I) = \begin{cases} +\infty, & \text{if} \quad RI < \lambda, \\ h(I) = -\tau \ln \left[1 - \frac{\theta}{RI} \right], & \text{if} \quad RI > \lambda. \end{cases} \tag{13}$$

where $I = A(\boldsymbol{x}, t)\ \forall \boldsymbol{x}$ and $\forall t$ is the RIF frame, $d(I)$ is the time each input value $A(\boldsymbol{x}, t)$ requires to excite the neuron in order to emit a spike, θ is a threshold value standing for the membrane potential of the neuron to be excited, R is the membrane resistance, $\tau = R * C$ and C is the membrane capacitance. We have proven in [10] that the spike-base quantizer might be uniform or non-uniform, depending on the parameter tuning.

Experiment 1: The experimental setup is related to the size of the GOP when the RIF filter is applied to a piece-wise constant input. We show in Fig. 2 that the reconstruction quality remains sufficiently high when the GOP increases. This Figure 2 illustrates the reconstruction quality of 5 different videos when the GOP size varies between $M = 1$

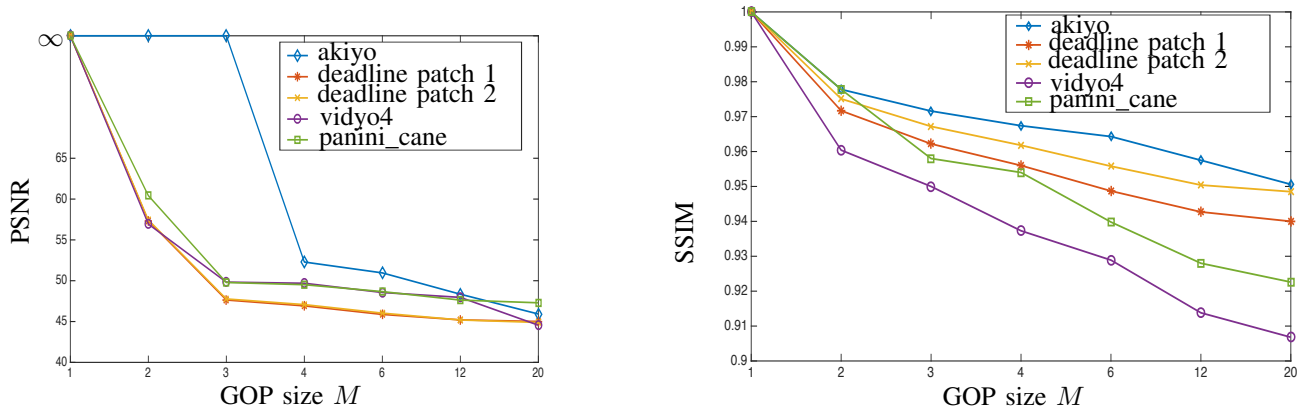

Fig. 2. Evaluation of the reconstruction results when the RIF filter is applied on a GOP that consist of constant frames or sub-regions that belong to the constant background. The reconstruction quality was measured by (left) the Peak Signal to Noise Ratio (PSNR) in the left-side sub-figure and (right) the Structure Similarity Index (SSIM) in the right-side sub-figure.

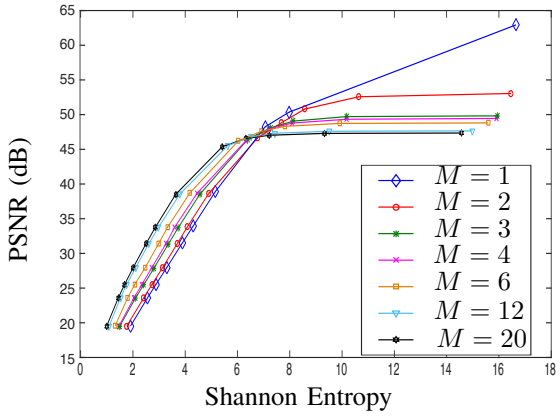

Fig. 3. This graph illustrates the rate-distortion performance of the a neuro-inspired compression architecture that consists of the RIF filter applied to a GOP and the spike-based quantization that is used to reduce the memory cost of the RIF frame.

and $M = 20$. We used two different image quality metrics, the Peak Signal to Noise Ratio (PSNR) [11] and the Structure Similarity Index (SSIM) [12]. According to the experimental results, we confirm the initial assumption that the RIF filter is able to recover a representative frame of a GOP with the minimum possible distortion, as evaluated by the two image quality metrics. It is worth to note that for $M = 12$ and $M = 20$, the PSNR is above 45dB while the SSIM metric remains above 0.9 degrees. Both these values correspond to an almost perfect recovery without any artifacts or any kind of distortion that is possible to be perceived by the human eye.

Experiment 2: The goal of this experiment is to show that the RIF filter and its dynamic properties are highly beneficial to video compression architectures as they are able to: (i) reduce the computational cost, which is mainly caused by the exhaustive comparisons between sequential frames that take place in every state-of-the-art video compression standard; (ii) reduce the memory cost; while (iii) retaining high reconstruction quality. To evaluate the reconstruction quality after the spike-based compression, we used the Shannon Entropy [13] computed on the quantized RIF frame and the PSNR metric. We used the first 92 frames of the "panini_cane" video from VISOR database where there is no pixel motion. We set different GOP sizes in order to compute the RIF frames and then we quantized these frames using the spike-based compression. The entropy was computed on the number of spikes N generated by the spike-based quantization of the RIF frame. Then, the inverse spike-based de-quantization and the RIF transform was applied in order to reconstruct the representative frame of the GOP. The results are depicted in Fig. 3 where we show that although the GOP size increases, the rate-distortion curve becomes more efficient as for the same number of bits we are able to achieve a higher reconstruction quality.

V. CONCLUSION

This paper has introduced an initial study of a dynamic filter that shares the properties of the early visual system when applied to a piece-wise constant signal as an extension of the temporally constant signal. It is experimentally shown that when a sequence of video frames is filtered with this retina-inspired transform, if the motion between the frames is smooth based on the dynamic decomposition, it is possible to perfectly reconstruct the input frames. There are several challenges to be addressed as an extension of this work including considering a piece-wise constant input with pixel motion for the RIF and providing some mathematical proof that it is possible to de-correlate the RIF frame to recover the input signal.

VI. REFERENCES

[1] R. VanRullen and S. J. Thorpe, "Rate coding versus temporal order coding: what the retinal ganglion cells tell the visual cortex," *Neural Computation*, vol. 13, no. 6, pp. 1255–1283, 2001.

[2] K Masmoudi, M. Antonini, and P. Kornprobst, "Exact Reconstruction of the Rank Order Coding using Frames Theory," 2011.

[3] K. Masmoudi, M. Antonini, and P. Kornprobst, "Streaming an image through the eye: The retina seen as a dithered scalable image coder," *Signal Processing: Image Communication*, vol. 28, no. 8, pp. 856–869, 2013.

[4] E. Doutsi, G. Tzagkarakis, and P. Tsakalides, "Neuro-inspired Compression of RGB Images," 2019.

[5] E. Doutsi, L. Fillatre, M. Antonini, and J. Gaulmin, "Retina-inspired Filter," *IEEE Transactions on Image Processing*, vol. 27, no. 7, pp. 3484–3499, 2018.

[6] W. Gerstner, *A framework for spiking neuron models: The spike response model*, vol. 4, North-Holland, 2001.

[7] E. Doutsi, L. Fillatre, M. Antonini, and J. Gaulmin, "Retina-inspired filtering for dynamic image coding," *IEEE International Conference in Image Processing (ICIP)*, pp. 3505–3509, 2015.

[8] R. Vezzani, "The Video Surveillance Online Repository VISOR," https://aimagelab.ing.unimore.it/visor/index.asp, 2007.

[9] "Xiph.org Video Test media," https://media.xiph.org/video/derf/, 2005.

[10] Effrosyni Doutsi, Lionel Fillatre, Marc Antonini, and Panagiotis Tsakalides, "Dynamic image quantization using leaky integrate-and-fire neurons," *IEEE Transactions on Image Processing*, vol. 30, pp. 4305–4315, 2021.

[11] A M Eskicioglu and P S Fisher, "Image quality measures and their performance," *IEEE Transactions on Communication*, vol. 43, no. 12, pp. 2959–2965, 1995.

[12] Z Wang, E. P. Simoncelli, and A C Bovik, "Multiscale structural similarity for image quality assessment," *37th Conference on ACSSC*, vol. 2, pp. 1398–1402, 2003.

[13] R. M. Gray and D. L. Neuhoff, "Quantization," *IEEE Transactions on Information Theory*, vol. 44, no. 6, pp. 2325–2384, 1998.