A SPATIOTEMPORAL DECOMPOSITION OF A VIDEO STREAM BASED ON THE RETINA-INSPIRED FILTER

Effrosyni Doutsi[*], *Panagiotis Tsakalides*[†]

[*]Foundation for Research and Technology - Hellas, [†]University of Crete

ABSTRACT

The goal of this work is to propose a simple yet efficient way to dynamically transform a video stream 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 analytically prove that the RIF can be applied to a group of pictures (GOP). We experimentally show that the size of GOP is important when motion appears, as some artifacts are generated. However, in the absence of motion among the GOP frames we can still perfectly reconstruct the video frames reducing the computational cost of the whole process.

1. INTRODUCTION

It is no secret that the internet traffic today is due mostly to videos. In fact, by 2024 online videos are expected to make up more than 82% of all consumer internet traffic. Thus, several open challenges need to be addressed such as adaptive video streaming, video pre-processing, storage, and video understanding, among others. 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 a 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. Several 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 or 360-degree videos.

Seeking alternative and computationally more efficient solutions capable of processing a video stream dynamically, researchers focused on biological systems whose performance seems to be beyond the current state-of-the-art. The brain and the visual system have been considered as a very 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, also known as events, that can signal significant voltage changes at the membrane of neurons in time [1]. There have been several image compression architectures motivated by neuroscience models including the Rank Order Coder [2][3], the bio-inspired Analog-to-Digital Converter [4], and the neuro-inspired image codec [5]. 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 in dynamically filtering and encoding a group of sequential frames instead of a single frame.

In this work, we study the mathematical framework under which the Retina-Inspired Filter (RIF) [6] can be 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, the inner tissue of the eye, responsible for capturing and transforming the luminance of light into electrical impulses [1]. This filter was the first component of the neuro-inspired image codec we introduced in [5]. 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 (GOP) taking advantage 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 (i) the spatiotemporal redundancy of GOP is exceptionally high and/or (ii) the background of the visual scene is almost static during the whole period when the scene is captured.

This work was funded by the TITAN ERA Chair project (contract no. 101086741) within the Horizon Europe Framework Program of the European Commission and the Hellenic Foundation for Research and Innovation (HFRI) and the General Secretariat for Research and Technology (GSRT) under grant agreement No. 330 (BrainSIM).

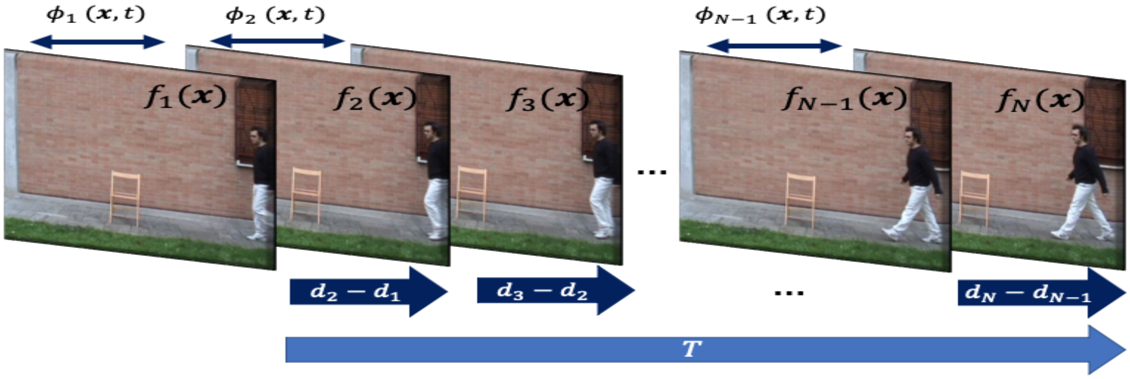


Fig. 1. This example illustrates the way the RIF filter is applied to a GOP of size N where each frame $f_i(\boldsymbol{x})$ appears for a given time $[d_{i+1} - d_i]$ while the RIF filter evolves for time T.

In this paper, we mathematically prove that under the assumption that the input video is a piecewise constant signal in time, the RIF filter can be efficiently applied to a GOP, instead of a single image as it was initially designed to. In addition, we experimentally show that according to the frame theory, if the GOP consists of frames that remain almost constant, it is possible to perfectly recover the input signal while the entropy is highly reduced even without considering any coding or quantization process. Additionally, we propose an experimental analysis regarding the size of the GOP considered by the RIF for different kinds of visual scenes; (i) when there is no motion presence among the frames of the GOP, (ii) when motion exists in time. As expected, in the second case some ghosting artifacts appear during the reconstruction of the visual scene which becomes stronger when the GOP size increases. However, these artifact issues can be addressed if one combines the RIF filter with spikes in order to dynamically capture the motion but this kind of experiment remains out of the spectrum of the current work.

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. Last but not least, we draw some conclusions and we give a short discussion about future work.

2. BACKGROUND ON RETINA-INSPIRED FILTER

The RIF filter is connected to the dynamic behavior of the retina, $K(\boldsymbol{x}, t)$, which enables the retina to increase the sharpness of the visual stimulus during filtering before its transmission to the brain. The visual stimulus is a 3D spatiotemporally 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 spatiotemporal 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 spatiotemporal convolution between the projection of the visual stimuli and the retina transform, defined as the difference between two spatiotemporal 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 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 [6].

It was proven in [6] that under the assumption that the input signal is an image visible for a given time $f(\boldsymbol{x}, t) = f(\boldsymbol{x})\mathbf{1}_{[0,T]}(t)$ 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:

$$A(\boldsymbol{x}, t) = \phi(\boldsymbol{x}, t) \overset{\boldsymbol{x}}{*} f(\boldsymbol{x}), \tag{5}$$

The RIF transform forms a group of spatiotemporal Weighted Difference of Gaussian (WDoG) filters when it is applied to an image visible for a given time T, defined as:

$$\phi(\boldsymbol{x}, t) \;=\; \begin{cases} a(t) G_{\sigma_c}(\boldsymbol{x}) - b(t) G_{\sigma_s}(\boldsymbol{x}) & \text{if } t \geq 0, \\ 0 & \text{otherwise,} \end{cases} \tag{6}$$

for all $\boldsymbol{x} \in \mathbb{R}^2$ and for all $t \in \mathbb{R}$, where G_{σ_c} and G_{σ_s} are two Gaussian functions with standard deviation σ_c and $\sigma_s = 3\sigma_c$, respectively, and $a(t)$, $b(t)$ are the two time-varying weights

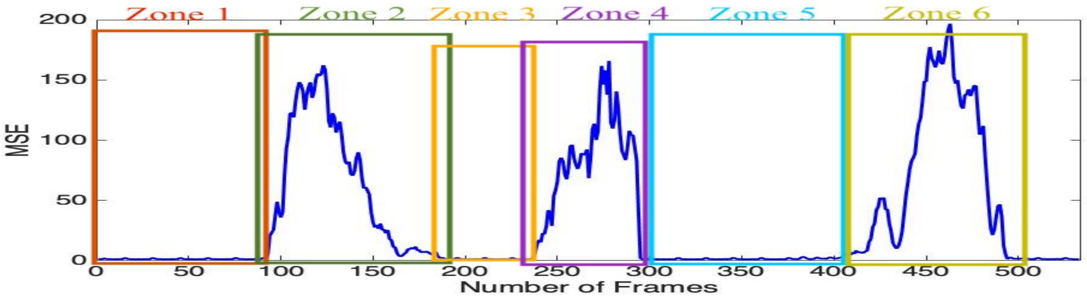
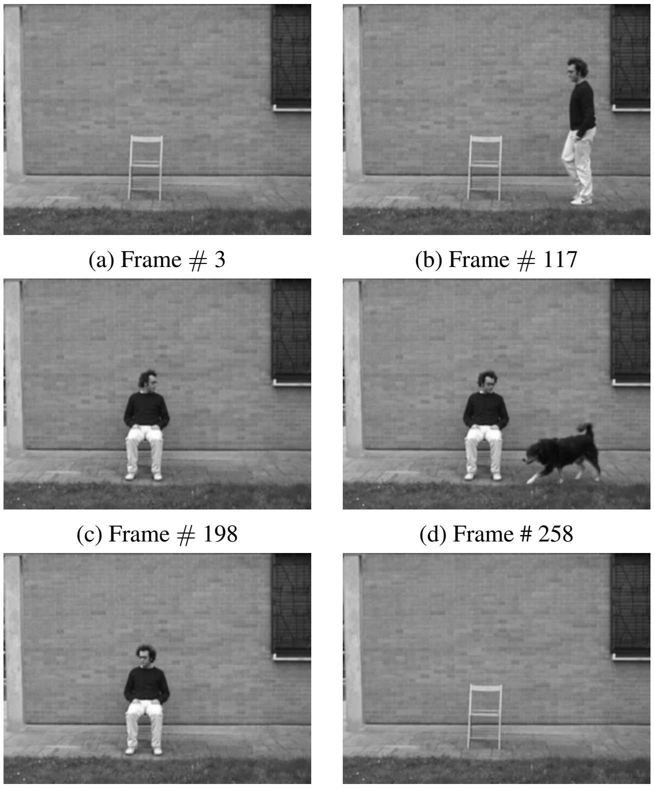


Fig. 2. Mean Square Error between every pair of sequential frames of the video ”man with a dog” taken by the VISOR database [8]. There are 6 different zones depending on the content of the visual scene.



(a) Frame # 3

(b) Frame # 117

(c) Frame # 198

(d) Frame # 258

(e) Frame # 320

(f) Frame # 500

Fig. 3. Representative frames of the 6 different time zones that occur according to the analysis of Fig. 2.

that influence the shape of the Difference of Gaussians in time as follows:

$$a(t) = w_c R_c(t) = w_c \mathbf{1}_{[0,+\infty)}(t) \int_{\max\{0,t-T\}}^{t} V(u)du, \quad (7)$$

$$b(t) = w_s R_s(t) = w_s \mathbf{1}_{[0,+\infty)}(t) \int_{\max\{0,t-T\}}^{t} (V \overset{t}{*} E_{\tau_S})(u)du. \quad (8)$$

It is worth noting 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.

3. RIF APPLIED TO A PIECE-WISE CONSTANT INPUT

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 $f(\boldsymbol{x},t) = \sum_{i=1}^{N} f_i(\boldsymbol{x}) \mathbf{1}_{[d_i,d_{i+1}]}(t)$, for all $\boldsymbol{x} \in \mathbb{R}^2$ and $t \in \mathbb{R}$, a piece-wise constant signal in time where $f_i(\boldsymbol{x})$ is the $i-th$ frame of a video stream that consists of N frames and $[d_i, d_{i+1}]$ is the time window where the $i-th$ frame is visible. Then, the spatio-temporal convolution in (1) turns into a spatial convolution of the RIF filter $\phi(\boldsymbol{x},t)$ with a group of N frames where each frame $f_i(\boldsymbol{x})$ is filtered by RIF for the time window $T_c = d_{i+1} - d_i$ it is visible:*

$$A(\boldsymbol{x},t) = \sum_{i=1}^{N} \phi_i(\boldsymbol{x},t) \overset{x}{*} f_i(\boldsymbol{x}). \quad (9)$$

Proof. The proof of the above proposition is given in the Appendix. □

Below we provide an intuitive analysis of the proposed mathematical proof. It was proven in [6] that the RIF transform has a dynamic spectrum that changes in time. Initially, the RIF filter was applied to a static image generating a bunch of spatiotemporal decomposition layers. In this work, we propose that the RIF decomposition is split into N groups. Then, each piece is applied to a different frame of the GOP. For instance, let's consider a video stream as a sequence of pictures

$$f_1(\boldsymbol{x}), f_2(\boldsymbol{x}), \ldots, f_N(\boldsymbol{x})$$

, each one of which appears at a given time $t_1, t_2 \ldots, t_N$, then each picture will be filtered by a different group of RIF

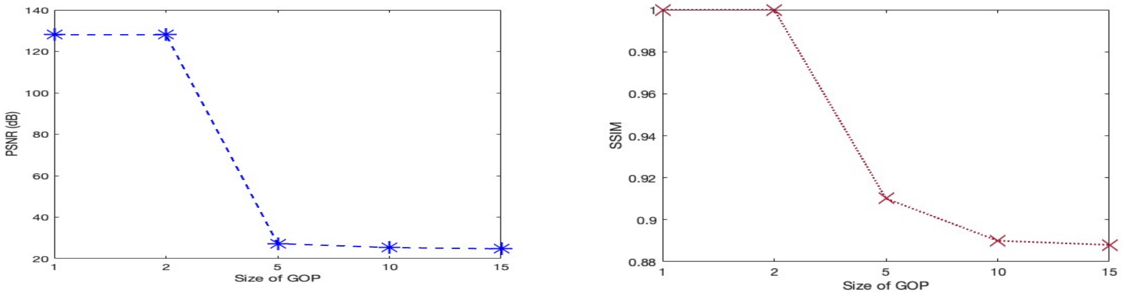


Fig. 4. This figure illustrates the reconstruction quality with respect to the size of GOP for two different evaluation metrics; PSNR (left) and SSIM (right). It is worth to be noticed that the reconstruction quality is perfect for GOP size $N = 1$ and $N = 2$. In addition, as expected, when the GOP size increases the reconstruction quality decreases. However, if we consider the SSIM metric, that is closer to the human visual perception, the quality still remains high.

layers

$$\phi_1(\boldsymbol{x}, t), \phi_2(\boldsymbol{x}, t), \ldots, \phi_N(\boldsymbol{x}, t),$$

respectively. At the end of the GOP, the RIF is reinitialized to be applied to the next GOP (see Fig. 1). An interesting special case arises when all the GOP frames are approximately equal to each other as a consequence of the absence of motion. In this case, RIF is applied to a GOP where $f_1(\boldsymbol{x}) = f_2(\boldsymbol{x}) = \cdots = f_N(\boldsymbol{x})$. In this scenario, the RIF decomposition forms a frame as per frame theory [7], allowing for perfect and highly efficient recovery of the input signal. This process generates a single image that represents the signal over T rather than N identical frames.

4. EXPERIMENTAL RESULTS

4.1. Dataset Description

This section aims to compute the performance of the RIF filter when applied to a video stream. In our experiments, we used surveillance videos from the VISOR database [8]. To illustrate the proposed method, we use the video titled 'man with a dog,' which contains a total of 543 frames. Figure 2 shows the Mean Square Error among every pair of sequential frames for the whole video stream. The shape of this curve is aligned to the content of the visual scene that could be mainly split into 6 different time zones; zone 1, zone 3, and zone 5 consist of frames that are almost constant since no motion appears in the visual scene. For the rest of the video frames, some moving objects show up. For computational reasons in our experiments, we used the first 360 frames of this video and we present some representative frames of the first 5 time zones (see Fig. 3).

4.2. GOP Size

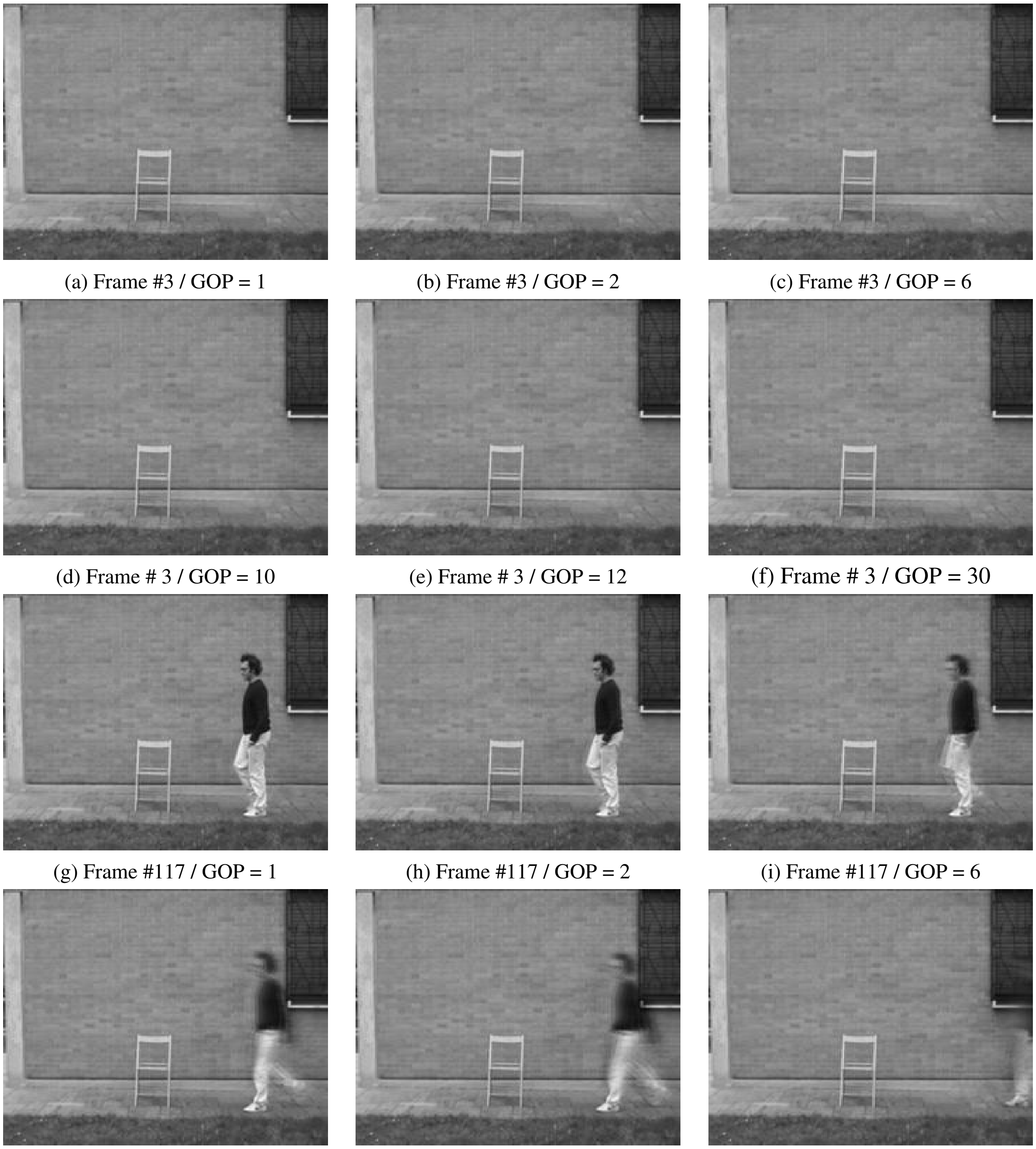
The experimental setup is related to the size of the GOP considered as the piece-wise constant input. The worst case scenario is when the size of GOP $N = 1$ which is equivalent to applying the RIF to each different frame of a video stream. Based on this initial assumption, Doutsi *et al* conducted a detailed analysis of the RIF filter [6] in both the spatial and frequency domains. While the reconstruction quality, in this case, is high due to frame theory [7], this process would undermine the retina's dynamic functionality in processing incoming light.

Let us now consider another scenario when the GOP size is $N > 1$. This approach enables the dynamic processing of a group of pictures in a manner akin to the biological retina. As expected when the GOP size increases some ghosting artifacts begin to appear as the RIF decomposition came from the spatio-temporal convolution of several frames. Unfortunately, the higher the GOP size is the more the artifacts. However, in the specific scenario where the GOP comprises frames with highly similar content and/or negligible motion, these artifacts vanish entirely. This is because the RIF adheres to frame theory, allowing it to accurately reconstruct a single frame that represents the signal over time T, defined by the GOP size.

Figure 5 illustrates some recovery results of the frames #3 and #117 when the GOP is of size $N \in \{1, 2, 5, 10, 15\}$ and Fig. 4 the reconstruction quality according to the Peak Signal to Noise Ratio (PSRN) and the Structure Similarity Index (SSIM) image quality metrics.

5. 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. However, if there is



(a) Frame #3 / GOP = 1

(b) Frame #3 / GOP = 2

(c) Frame #3 / GOP = 6

(d) Frame # 3 / GOP = 10

(e) Frame # 3 / GOP = 12

(f) Frame # 3 / GOP = 30

(g) Frame #117 / GOP = 1

(h) Frame #117 / GOP = 2

(i) Frame #117 / GOP = 6

(k) Frame #117 / GOP = 10

(l) Frame #117 / GOP = 12

(m) Frame #117 / GOP = 30

Fig. 5. This figure illustrates the recovery result of two frames (frame # 3 (a)-(f) and frame #117 (g)-(m)) when the size N of GOP belongs to the set $\{1,2,6,10,12\}$. Even though N increases in the first case where the moving scene is almost constant in time the recovery is perfect. On the other hand, if moving objects appear in the scene, the quality of the reconstruction reduces as N increases.

some motion among the sequential frames, the reconstruction depends on the size of the Group of Pictures.

There are several challenges to be addressed in order to improve the quality of the reconstruction via a more robust and efficient way to compute the deconvolution of a dynamic signal with a dynamic transform. In addition, it would be very interesting to study the entropy of the RIF transform as well as the behavior of this dynamic filter when it is followed by a dynamic quantization such as the Leaky Integrate-and-Fire model.

6. APPENDIX

We provide below the mathematical proof of Proposition 1

$$
\begin{aligned}
& A(\boldsymbol{x},t) = K(\boldsymbol{x}) \overset{x,t}{*} f(\boldsymbol{x},t) \\
= \quad & C(\boldsymbol{x},t) \overset{x,t}{*} \sum_{i=1}^{N} f_i(\boldsymbol{x}) \mathbf{1}_{[d_i,d_{i+1}]}(t) \\
& - S(\boldsymbol{x},t) \overset{x,y}{*} \sum_{i=1}^{N} f_i(\boldsymbol{x}) \mathbf{1}_{[d_i,d_{i+1}]}(t) \\
= \quad & w_c G_{\sigma_c}(\boldsymbol{x}) V(t) \overset{x,t}{*} \\
& \big(f_1(\boldsymbol{x}) \mathbf{1}_{[d_1,d_2]}(t) + \cdots + f_N(\boldsymbol{x}) \mathbf{1}_{[d_N,d_{N+1}]}(t)\big) \\
& - w_s G_{\sigma_s}(\boldsymbol{x}) \Big(V \overset{t}{*} E_{\tau_S})(t) \Big) \overset{x,y}{*} \\
& \big(f_1(\boldsymbol{x}) \mathbf{1}_{[d_1,d_2]}(t) + \cdots + f_N(\boldsymbol{x}) \mathbf{1}_{[d_N,d_{N+1}]}(t)\big) \\
= \quad & w_c G_{\sigma_c}(\boldsymbol{x}) V(t) \overset{x,t}{*} f_1(\boldsymbol{x}) \mathbf{1}_{[d_1,d_2]}(t) + \cdots + \\
& w_c G_{\sigma_c}(\boldsymbol{x}) V(t) \overset{x,t}{*} f_N(\boldsymbol{x}) \mathbf{1}_{[d_N,d_{N+1}]}(t) \\
& - w_s G_{\sigma_s}(\boldsymbol{x}) \Big(V \overset{t}{*} E_{\tau_S})(t) \Big) \overset{x,y}{*} f_1(\boldsymbol{x}) \mathbf{1}_{[d_1,d_2]}(t) + \cdots + \\
& w_s G_{\sigma_s}(\boldsymbol{x}) \Big(V \overset{t}{*} E_{\tau_S})(t) \Big) \overset{x,y}{*} f_N(\boldsymbol{x}) \mathbf{1}_{[d_N,d_{N+1}]}(t) \\
= \quad & \sum_{i=1}^{N} w_c G_{\sigma_c}(\boldsymbol{x}) V(t) \overset{x,t}{*} f_i(\boldsymbol{x}) \mathbf{1}_{[d_i,d_{i+1}]}(t) \\
& - \sum_{i=1}^{N} w_s G_{\sigma_s}(\boldsymbol{x}) \Big(V \overset{t}{*} E_{\tau_S})(t) \Big) \overset{x,y}{*} f_i(\boldsymbol{x}) \mathbf{1}_{[d_i,d_{i+1}]}(t) \\
= \quad & \sum_{i=1}^{N} w_c G_{\sigma_c}(\boldsymbol{x}) \overset{x}{*} f_i(\boldsymbol{x}) \; V(t) \overset{t}{*} \mathbf{1}_{[d_i,d_{i+1}]}(t) \\
& - \sum_{i=1}^{N} w_s G_{\sigma_s}(\boldsymbol{x}) \overset{x}{*} f_i(\boldsymbol{x}) (V \overset{t}{*} E_{\tau_S})(t)) \overset{t}{*} \mathbf{1}_{[d_i,d_{i+1}]}(t) \\
\overset{(10)(11)}{=} \quad & \sum_{i=1}^{N} w_c G_{\sigma_c}(\boldsymbol{x}) \overset{x}{*} f_i(\boldsymbol{x}) R_{c,i}(t) \\
& - \sum_{i=1}^{N} w_s G_{\sigma_s}(\boldsymbol{x}) \overset{x}{*} f_i(\boldsymbol{x}) R_{s,i}(t) \\
= \quad & \sum_{i=1}^{N} w_c R_{c,i}(t) G_{\sigma_c}(\boldsymbol{x}) \overset{x}{*} f_i(\boldsymbol{x}) \\
& - \sum_{i=1}^{N} w_s R_{s,i}(t) G_{\sigma_s}(\boldsymbol{x}) \overset{x}{*} f_i(\boldsymbol{x}) \\
= \quad & \sum_{i=1}^{N} a_i(t) G_{\sigma_c}(\boldsymbol{x}) \overset{x}{*} f_i(\boldsymbol{x}) \\
& - \sum_{i=1}^{N} b_i(t) G_{\sigma_s}(\boldsymbol{x}) \overset{x}{*} f_i(\boldsymbol{x}) \\
& \sum_{i=1}^{N} \left(a_i(t) G_{\sigma_c}(\boldsymbol{x}) - b_i(t) G_{\sigma_s}(\boldsymbol{x})\right) \overset{x}{*} f_i(\boldsymbol{x}) \\
= \quad & \sum_{i=1}^{N} \phi_i(\boldsymbol{x},t) \overset{x}{*} f_i(\boldsymbol{x})
\end{aligned}
$$

where $R_{c,i}(t)$ and $R_{s,i}(t)$ are given by:

$$R_{c,i}(t) = \mathbf{1}_{[0,+\infty)}(t) \int_{\max\{0,t-iT_c\}}^{t} V(u)du, \qquad (10)$$

$$R_{c,i}(t) = \mathbf{1}_{[0,+\infty)}(t) \int_{\max\{0,t-iT_c\}}^{t} (V \overset{t}{*} E_{\tau_S})(u)du. \qquad (11)$$

7. REFERENCES

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

[2] 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.

[3] K. Masmoudi, M. Antonini, and P. Kornprobst, “Frames for exact inversion of the rank order coder,” *IEEE Transactions on Neural Networks and Learning Systems*, vol. 23, no. 2, pp. 353–359, 2012.

[4] 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.

[5] E¿ Doutsi, G. Tzagkarakis, and P. Tsakalides, “Neuro-inspired compression of rgb images,” in *2019 27th European Signal Processing Conference (EUSIPCO)*, 2019, pp. 1–5.

[6] 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.

[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 and R. Cucchiara, “Video surveillance online repository (visor): an integrated framework,” *Multimedia Tools and Applications*, vol. 50, pp. 359–380, 2010.