Retina inspired Video Codec

Effrosyni Doutsi, Lionel Fillatre, Marc Antonini, Julien Gaulmin

Retina-inspired Video Codec

Effrosyni Doutsi, Lionel Fillatre, Marc Antonini
Université Côte d'Azur, CNRS, I3S
Nice - Sophia Antipolis, France
doutsi@i3s.unice.fr

Julien Gaulmin
4G-TECHNOLOGY
Mouans Sartoux, France

***Abstract*—In this paper, we aim to propose a video codec based on the novel retina-inspired filter and retina-inspired quantizer which both perform according to the early visual system. The recently released non-separable spatiotemporal OPL retina-inspired filter enables to progressively extract different kind of information from the input signal which is the sequence of pictures of a video stream. This retina inspired transform has been proven to be a redundant frame which ensures a perfect reconstruction when no quantization appears. The reduction of this redundancy is achieved by a quantization which is inspired by the spike generation mechanism of ganglion cells. This mechanism has been approximated by the Rank Order Coder (ROC) and the Leaky-Integrate and Fire (LIF) models. The ROC model encodes the rank of the spikes and it has been proposed as a complete and very efficient codec for still-images. However, its limitations concerning the reconstruction method forced us to focus our attention on LIF which encodes the spike delays. We approximate the LIF by a scalar quantizer with a dead-zone. This is the first attempt to build a complete retina-inspired video codec which gives promising reconstruction results at low bitrate and high reconstruction quality.**

***Index Terms*—Retina-inspired filter, frame, Rank Order Coder (ROC), Leaky Integrate and Fire (LIF), video coding.**

I. Introduction

Tracking the development of video compression algorithms, one will come to the following two conclusions: the first one is that all the video compression algorithms (MJPEG, MJPEG2000, MPEG-1, MPEG-2, MPEG-4/H.264 (AVC) and MPEG-H/H.265 (HEVC)) have the same origin which is the JPEG algorithm proposed for image compression [1]. The designers used JPEG standard as a basis to encode and decode key-pictures of a video stream introducing at the same time other methods to reduce temporal and spatial redundancy. The second conclusion is related to the complexity of these algorithms which increases during the years [2]. To achieve more efficient compression algorithms and improve the bitrate, the designers proposed more complex solutions.

The increase of complexity is inefficient for a series of applications including video surveillance systems. As a result, being supported by high complexity algorithms they consume a lot of power. To our point of view, the basic drawback of the conventional architecture stems from the fact that videos are dynamic signals which are processed by methods proposed for static images (DCT, DWT, scalar quantization, etc.). As a result, we believe that a video should be processed dynamically.

This paper is the first attempt of proposing a novel video encoding architecture based on the dynamic properties of the retina processing which seems to be promising for compression. This novel codec is called retina-inspired video codec and it consists of the retina-inspired transform and the retina-inspired quantization. We propose to apply the retina-inspired codec to each picture of a video stream like MJPEG or MJPEG2000. Treating each picture of the video stream as a still-image which is flashed for a given time, this codec enables to progressively extract and encode different kind of details until the next picture (still-image) appears. Thus, the retina-inspired codec allows the tuning of the reconstruction quality along time. Of course, the retina-inspired codec does not reduce yet the temporal redundancy between pictures. This would be an interesting and definitely necessary extension in order to be comparable to the latest standards. We evaluate the efficiency of our video codec step-by-step comparing first of all the dynamic retina-inspired filter and then the retina-inspired quantizer to previous bio-inspired filtering and encoding methods proposed for images.

In this document, we show that the recently released dynamic filter, which is called the retina-inspired filter [3], is more efficient than a static transform like the filter bank which is used in the Rank Order Coder (ROC) and its extensions [4], [5], [6]. It is also proven in [7], [8], [9] that the retina-inspired filter is invertible according to the frame theory. Consequently, it enables a perfect reconstruction of the input signal if there is no quantization. Secondly, this paper introduces a Leaky Integrate and Fire (LIF) quantizer which reduces the spatiotemporal redundancy of the retina-inspired frame. We show that the LIF-quantizer is more efficient than other spike generation mechanisms like ROC encoder.

II. Background in Rank Order Coder (ROC)

We assume that a video $V(x,t)$ can be modeled as

$$V(x,t) = \sum_{i=1}^{N} f_i(x)\mathbf{1}_{[g_i,g_{i+1}]}(t), \tag{1}$$

where $x \in \mathbb{R}^2$, $t \in \mathbb{R}$, $f_i(x)$ stands for the i-th picture of the video, N is the total number of pictures which form the video stream and $\mathbf{1}_{[g_i,g_{i+1}]}(t)$ is the indicator function which is equal to 1 if $g_i \leq t \leq g_{i+1}$, and 0 otherwise. Let's call $T_i = g_{i+1} - g_i$ the duration for which a given picture $f_i(x)$ of the video stream appears. This time $T_i = T$ is the same for

every single picture of a video stream with a frame rate $1/T$. Thus, the total duration of the video is NT.

The ROC model proposed by Thorpe [4], [5] is a solution for coding each picture $f_i(x)$ independently as it is done in MJPEG or MJPEG2000. The ROC model is a bio-plausible generator and decoder of spikes. The spikes are generated by neurons. Thorpe assumed that the most informative spike which is emitted by a neuron is the first one. The first step of the ROC model is the convolution of a still-image, say $f(x)$ which stands for one of the pictures $f_i(x)$, with a filter bank $\{DoG^k(x)\}_{1\leq k\leq K}$:

$$A^k(x) = DoG^k(x) \overset{x}{*} f(x), \tag{2}$$

where $x \in \mathbb{R}^2$, $\overset{x}{*}$ denotes the spatial convolution and k is the layer index. Each filter is a Difference of Gaussian (DoG):

$$DoG^k(x) = G_{\sigma_c^k}(x) - G_{\sigma_s^k}(x), \tag{3}$$

where $G_{\sigma_c^k}(x)$, $G_{\sigma_s^k}(x)$ are two Gaussian filters with standard deviations σ_c^k and σ_s^k respectively. Thorpe assumed that all the layers are fed simultaneously to the neurons in order to spike.

Let $A(x) = (A^1(x), \ldots, A^L(x))$ be the input of the ROC model. Each contrast intensity $A^k(x_r)$, where x_r is a given spatial location, is converted in a spike train by a specific spiking neuron. For this purpose, the firing rate $\rho(A^k(x_r))$ of the spiking neuron is given by the Michaelis-Nenten function. Thorpe uses a Poisson process to produce the spike train which encodes $A^k(x_r)$. Each contrast intensity is associated to a specific spike train. The arrival time of the first spike within a spike train depends on the intensity of the coefficient $A^k(x_r)$. Then, the contrast intensity is linked to the arrival rank of the first spike: a stronger stimuli corresponds to a fast arrival of a spike (low rank) and it is assigned to a high weight during the reconstruction step. These weights were adjusted with a Look-Up-Table (LUT), which allows to look-up for the most likely intensity value with a given rank. This Look-Up-Table was experimentally defined after testing several grayscale images.

Contrast A	255	245	240
Rank A	1	2	3

Contrast B	20	19	10
Rank B	1	2	3

TABLE I: Two sets of contrast intensities with the same ranks.

It is important to note that the LUT is not accurate for a group of images with different statistical properties. Table I shows two examples of contrast intensity triplets which are assigned to the same rank. As a result, according to the LUT the reconstruction will be exactly the same for both of these examples, which is obviously wrong. This is the first limitation of the ROC model. The second drawback is related to its filter bank. If one considers the ROC model as a bio-inspired model, its filter bank is a very rough approximation of the dynamic OPL transform and, overall, the ROC filter bank is not invertible. Masmoudi in [6] was the first one who tackled these problems. He proposed a rectification function in order to obtain an invertible filter bank which leads to a perfect reconstruction. Masmoudi also created his own LUT which was learned for each input image.

III. Retina-inspired filter

The retina-inspired filter $\phi(x,t)$ proposed in [8], [3] is a non-separable spatiotemporal OPL retina-inspired filter. This filter behaves according to the dynamic transform which happens in the OPL of the retina tissue as it has been modeled by neuroscientists [10]. The retina-inspired filter is a DoG which varies with respect to time due to some temporal functions $a(t)$ and $b(t)$ as following:

$$\phi(x,t) = a(t)G_{\sigma_c}(x) - b(t)G_{\sigma_s}(x). \tag{4}$$

We propose to filter the video $V(x,t)$ with the retina-inspired filter. Let $A(x,t)$ be the filtered video. We have proven in [9] that $A(x,t) = 0$ if $t < g_1$ and, otherwise,

$$A(x,t) = \sum_{i=1}^{N} \phi(x, t-g_i) \overset{x}{*} f_i(x). \tag{5}$$

We also have proven in [9] that when $t \in [g_i, g_{i+1}]$,

$$A(x,t) \approx \phi(x, t-g_i) \overset{x}{*} f_i(x) = A_i(x,t), \tag{6}$$

provided that the filter $\phi(x,t)$ is vanishing sufficiently fast with respect to t. The convergence to 0 can be controlled by choosing adequately the parameters of the retina-inspired filter (see details in [9]). Under this assumption, we have proven in [9] that each picture $f_i(x)$ of the input video $V(x,t)$ can be perfectly reconstructed from its decomposition layers $A_i(x,t)$ according to the frame theory [11]. Hence, when all the neurons are able to emit a spike train which encodes each coefficient $A_i(x,t)$ with a high accuracy, our filter allows a perfect reconstruction. The decoder receives a stream of filtered subbands $A_i(x,t)$ coded under the form of spike trains and its goal is to reconstruct $f_i(x)$ from $A_i(x,t)$ over the time interval $t \in [g_i, g_{i+1}]$.

For numerical purpose, we need to discretize the retina-inspired filter in space and in time. Without any loss of generality, let us focus on the time interval $t \in [g_i, g_{i+1}]$. Let $x_1, \ldots, x_n \in \mathbb{R}^2$ be some sets of spatial sampling points and $t_1, \ldots, t_m \in [g_i, g_{i+1}]$ be temporal sampling points. The continuous spatial convolution $A_i(x,t)$ is then approximated by the discrete spatial circular convolution:

$$A_i(x_k, t_j) \quad = \quad \phi(x_k, t_j - g_i) \circledast f_i(x_k), \forall k, j. \tag{7}$$

Let us denote f_i the sampled version of the image $f_i(x)$ and A_i the sampled and vectorized version of the filtered image $A_i(x,t)$ for each retina-inspired decomposition layer $1 \leq j \leq m$:

$$f_i = (f_i(x_1), \ldots, f_i(x_n)),$$
$$A_i = (A_{i,1}(x_1), \ldots, A_{i,1}(x_n), \ldots, A_{i,m}(x_1) \ldots, A_{i,m}(x_n)),$$

where $M = n\,m$ is the total number of coefficients within a filtered image. The full set of the retina-inspired coefficients for a single picture i is given by $A_i = (A_{i,j}), \forall j$.

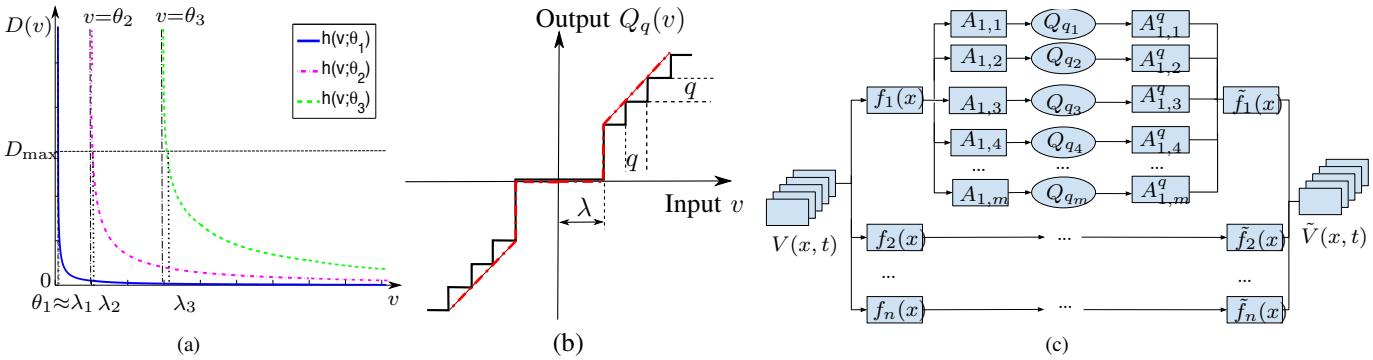

Fig. 1: (a) Delay $D(v)$ as a function of v: the quantization dead-zone $[0, \lambda]$ is imposed by $D_{\max}$ for different $\lambda \in \{\lambda_1, \lambda_2, \lambda_3\}$ ($C = 1$). (b) Black curve: LIF dead-zone uniform quantizer. Red curve: perfect LIF dead-zone quantizer. (c) Video Codec Schema.

IV. LIF Quantizer

The quantization, studied in this section, is necessary to compress the filtered image coefficients into single quantum values. As a result, we are interested in quantizing the frame coefficients $A_{i,j}$ in order to compress each picture of the input video stream. As it is mentioned in section II, the LIF model shares the same assumption with the ROC model concerning how informative the first spike of a neuron is. Instead of encoding the rank, the LIF model encodes the delay of the first spike. This is more efficient because it enables to overcome the limitations of the LUT: this encoding is a one-to-one mapping providing that the input intensity exceeds a given threshold.

The LIF neuroscientific model is detailed in [12]. For a frame coefficient value v (which stands for any coefficients $A_{i,j}$) the LIF computes the delay $D(v)$ before the first spike (and between the following spikes) in absence of noise. The delay $D(v)$ depends on the value of v as follows:

$$D(v) = \begin{cases} +\infty & \text{if} \quad v \leq \theta, \\ h(v;\theta) = -C \ln\left[1 - \dfrac{\theta}{v}\right] & \text{if} \quad v > \theta, \end{cases} \tag{8}$$

where C is a constant related to the capacity of the neurons. The threshold θ controls the redundancy reduction of the frame. If the input value does not exceed the threshold, there is no spike. Hence, the corresponding coefficient is not transmitted. When the frame coefficient intensity v exceeds the threshold, the delay $D(v)$ is given by the function $h(v;\theta)$ which is a continuous strictly decreasing function of v for $v > \theta$ (see Fig. 1 (a)). As a result if one knows exactly the value of the delay $D(v)$ of the first spike, he is able to perfectly reconstruct the value v using the function $h^{-1}(D(v), \theta)$, which is the inverse of $h(v, \theta)$.

In the case of a video, a picture $f_i(x)$ must be reconstructed before the following picture $f_{i+1}(x)$. Hence, the spikes coding the filtered picture A_i must be received before a maximum delay $D_{\max}$. In this paper, we choose $D_{\max}$ equal to the period between two consecutive pictures of the video stream, i.e. $D_{\max} = T$. According to the properties of $h(v;\theta)$, satisfying a maximum delay $D_{\max}$ is equivalent to encode only the values $A_{i,j}$ whose intensity is larger than $\lambda = h^{-1}(D_{\max};\theta)$. A short calculation shows that

$$\lambda = \lambda(D_{\max}) = \frac{\theta}{1 - e^{-\frac{D_{\max}}{C}}}, \tag{9}$$

which involves that $\lambda > \theta$ (see Fig. 1 (a)). In addition, the threshold θ can be viewed as a mean to control the percentage of neurons p which participate to the reconstruction of each picture during $D_{\max}$. Thus, θ and λ are inversely related to the percentage ($\lambda = 1/p$) which means that the more the neurons will spike, the smaller the width of the dead-zone of the LIF quantizer.

The LIF produces a spike train whose decoding allows us to estimate the input frame. Let $Q_q(v)$ be the decoded value from the LIF. We propose to approximate the full LIF model (spike train generation and decoding) by a scalar uniform quantizer with a dead-zone of width $2\,\lambda$ and step q (see Fig. 1 (b)). When the absolute value of the input intensity v is smaller than λ, there is no spike so the decoded value is 0. In this case the LIF quantizer is called perfect LIF since there is no quantization and the exact value of each intensity $v > \lambda$ is possible to be reconstructed (Fig. 1 (b)). However, in terms of compression this perfect reconstruction would cause a high entropy cost since the delay must be encoded perfectly. For this reason, we quantize the delay $D(v)$. The quantized delay $\widehat{D}(v)$ will cause an approximation $\widehat{v}$ of the intensity v. As a result, we assume that it is equivalent to quantize directly the intensity v. The smaller the quantization with a uniform quantizer of step q, the better the approximation $\widehat{v}$ (Fig. 1 (b)). The model of the dead-zone quantizer of step q is given as following:

$$Q_q(v) = \mathrm{sgn}(v) \max\left(0, \left\lfloor \frac{|v| - \lambda}{q} + 1 \right\rfloor\right), \tag{10}$$

where $\mathrm{sgn}(v)$ is the sign of the input value. For an input coefficient $A_{i,j}$, the output is the quantized value $A^q_{i,j} = Q_q(A_{i,j})$. Fig. 1 (c) illustrates the retina-inspired coding principle of a

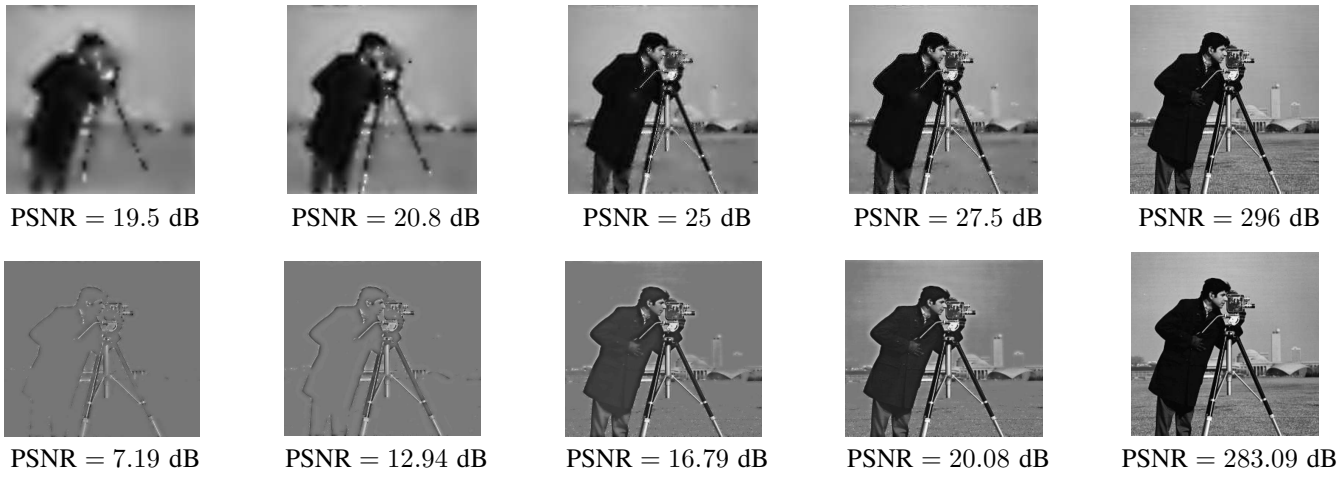

Fig. 2: These are some reconstruction results using different number of p which is the number of neurons which are allowed to spike (from left to right θ equals 0.5%, 1%, 5%, 10% and 100%). One the top line we illustrate the ROC encoder model as it was developed in [6]. The bottom line shows results of the LIF-quantizer using the retina-inspired filter.

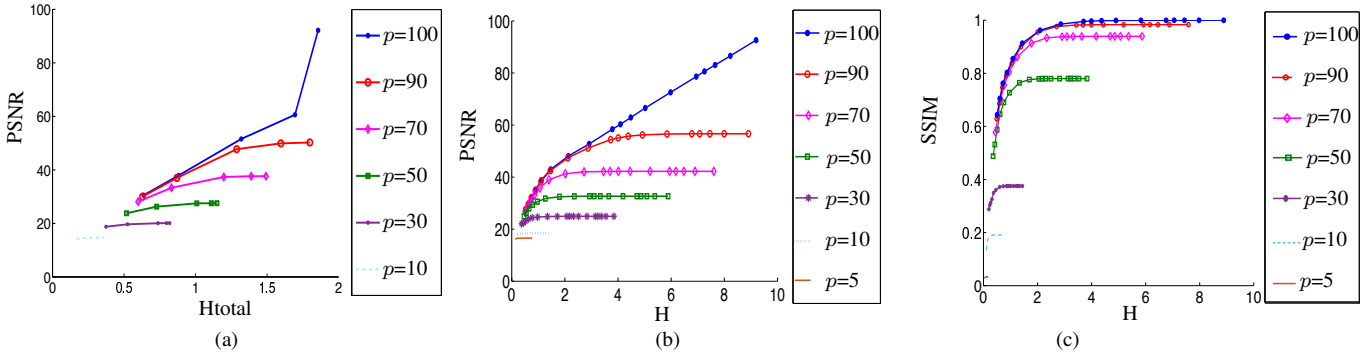

Fig. 3: (a) PSNR(dB) vs Htotal(Mbps) using the foreman video of total number of pictures 100 of the size 256x256 pixels (Pre-processing with the "Total Video Audio Converter") ($q \in \{1400, 800, 200, 50, 10, 1\}$) (b) PSNR(dB) vs H(bpp). (c) SSIM vs H(bpp). The % is controlled by p value and $q \in \{1400, 1200, 1000, 800, 600, 400, 200, 100, 50, 40, 30, 20, 10, 5, 4, 3, 2, 1\}$.

video stream. Each picture of the video stream is firstly retina-inspired filtered and then each retina-inspired decomposition layer is quantized by the LIF-quantizer. The de-quantized layers are used to reconstruct each picture of the video stream.

V. Reconstruction

The reconstruction is based on a classical image reconstruction approach which minimizes the MSE between the input images and the reconstructed ones. The reconstruction algorithm is detailed in [9] for non-quantized values. We exploit the same algorithm by replacing the non-quantized values by the output of the dead-zone uniform quantizer. Let $\tilde{V}$ be the reconstructed video considered as the set of reconstructed still-images $\tilde{f}_i$ for $i = 1, \ldots, N$. The quality of the reconstruction depends on the dead-zone semi-width λ and the quantization step q. The perfect reconstruction is only possible when q and λ (equivalent to θ when $D_{\max}$ is fixed) are small.

Fig. 2 illustrates the reconstruction results for different p using the perfect LIF quantizer ($q \approx 0$). We use a still-image to be comparable to the results in [6]. The reconstruction results is evaluated with the PSNR$(V, \tilde{V})$ metric expressed in dB and the SSIM$(V, \tilde{V})$ for a given bitrate which is measured by the total Shannon Entropy:

$$H = \frac{1}{N}\sum_{i=1}^{N} H_i = -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{n} p_k^i \log_2 \frac{1}{p_k^i}, \qquad (11)$$

where p_k^i is the probability the symbol k of the i^{th} picture to be occurred, H_i is the Shannon Entropy measured in bpp which also corresponds to the i^{th} picture. We consider that each input image is of a size 512x512 pixels and it is coded in grayscale levels from 0 to 255. Fig. 3 (b) and (c) illustrate respectively the evolution of PSNR and SSIM values for different q and p values using the cameraman as an input image. As it is expected, the smaller the q, the better the reconstruction quality which corresponds to high H. In addition, for a given q,

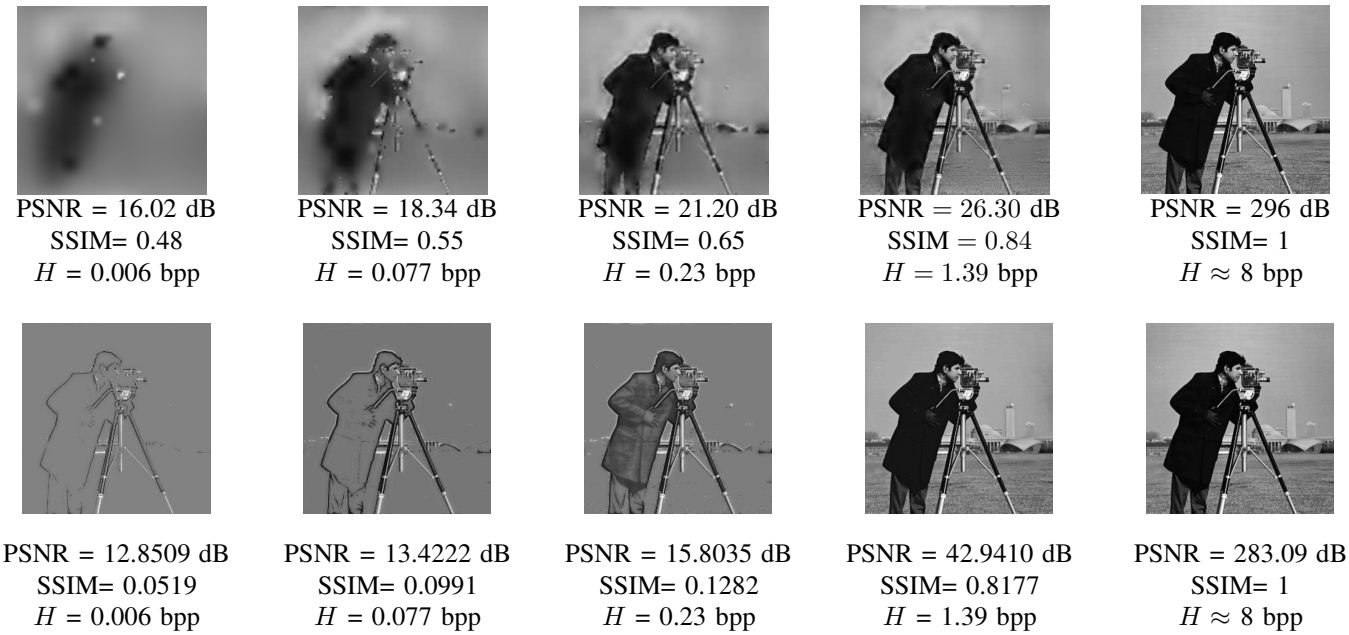

Fig. 4: Progressive Reconstruction. Top line: filter bank and the quantizer proposed in [13]. Bottom line: the retina-inspired filter with the LIF dead-zone quantizer. From left to right the coefficients which are used correspond to the first 20ms, 30ms, 40ms, 50ms and 150ms ($p = 100\%$).

the reconstruction quality increases while the dead-zone semi-width λ decreases because more neurons are able to spike. The higher the number of neurons, the better the reconstruction $\tilde{f}_i$ for each picture of the video stream. We also provide results for a well known video in Fig. 3 (a) for different kind of parameters. Interestingly, for low Htotal values, the PSNR is above 30dB.

The retina-inspired codec is motivated by the dynamic behavior of the visual system. This statement is defended by providing results for progressive reconstruction when only some of the decomposition layers are used for the reconstruction (Fig. 4). We managed to tune the parameters in order to provide the reconstruction results of the same entropy like the one proposed in [13]. One should notice both in Fig. 4 and Fig. 2 that the retina-inspired decomposition gives better visual results concerning the objects and the background of the image which are better structured even for very low values of p. Although the PSNR and SSIM values are lower using the retina-inspired frame with perfect LIF quantizer, our methods is more efficient in terms of contrast since all the object in the scene are better structured.

VI. Conclusion

This paper has introduced a novel retina-inspired codec which filters and encodes each picture of a video stream according to the way the visual system works. The reconstruction results are promising comparing to other retina-inspired coding methods while the visual quality is also higher.

References

[1] T. Sikora, "Mpeg digital video coding standards," *IEEE Signal Processing Magazine*, vol. 14, no. 5, pp. 82–100, 1997.

[2] D. Grois, D. Marpe, A.Mulayoff, and O.Hadar, "Performance comparison of h.265/mpeg-hevc, vp9, and h.264/mpeg-avc encoders," *30th Picture Coding Symposium 2013 (PCS 2013)*, December 2013.

[3] E. Doutsi, L. Fillatre, M. Antonini, and J. Gaulmin, "Retina-inspired filtering," *IEEE Transactions on Image Processing*, (submitted) 2016.

[4] S. J. Thorpe and J. Gautrais, "Rank Order Coding: A new coding scheme for rapid processing in neural network," *Computational Neuroscience: Trends in Research*, pp. 113–118, 1998.

[5] R. Van Rullen and S. J. Thorpe, "Rate coding versus temporal order coding: What the retinal ganglion cells tell the visual cortex," *Natural Neuroscience*, vol. 13, pp. 1255–1283, 2001.

[6] K. Masmoudi, M. Antonini, and P. Kornprobst, "Frames for exact inversion of the rank order coder," *IEEE Transaction on Neural Networks*, vol. 23, no. 2, pp. 353–359, 2012.

[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] E. Doutsi, L. Fillatre, M. Antonini, and J. Gaulmin, "Event-based coding of images using a bio-inspired frame," *International Conference on Event-Based Control, Communication and Signal Processing (EBCCSP)*, 2015.

[9] E. Doutsi, L. Fillatre, M. Antonini, and J. Gaulmin, "Video analysis and synthesis based on a retinal-inspired frame," *European Signal Processing Conference (EUSIPCO)*, 2015.

[10] A. Wohrer and P. Kornprobst, "Virtual retina: A biological retina model and simulator, with constrast gain control," *Journal of Computational Neuroscience*, vol. 26, no. 2, pp. 219–249, 2009.

[11] J. Kovacevic and A. Chebina, "An introduction to frames," *Signal Processing*, vol. 2, no. 1, pp. 1–94, 2008.

[12] Wulfram Gerstner and Werner Kistler, *Spiking Neuron Models: An Introduction*, Cambridge University Press, New York, NY, USA, 2002.

[13] 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.