Bipredictive Video Super-Resolution Using Key-frames

Karen F. de Oliveira, Fernanda Brandi, Edson M. Hung, Ricardo L. de Queiroz, Debargha Mukherjee

Universidade de Brasilia, Brazil

karen,fernanda,mintsu@image.unb.br, queiroz@ieee.org, debargha.mukherjee@hp.com

Abstract

Many scalable video coding systems use variable resolution frames to enable different decoding layers. Some of these systems also use frame down-sampling along with enhancement layers to reduce complexity. In order to do that, super-resolution methods associated with efficient interpolation processes may help to increase the quality of low-resolution frames. This work presents a super-resolution technique based on key frames. The goal is to restore the high-frequency information of down-sampled frames using high-resolution frames as references. The super-resolved frames can be used in scalable video coders, in variable quality coders, or in the side information generation for distributed coders. Results indicate substantial improvements over previous schemes.

I. Introduction

Super resolution is known as a process of obtaining a high resolution image from a set of low resolution observations [1]. A low resolution image is that in which there is a low density of pixels, that is, a low number of pixels per unit area, resulting in poor details. In the opposite, high resolution images have high pixel density, offering better detail information. By this way, a super resolution process receives a set of low resolution images that will be processed generating a high resolution image. In practice, this low resolution images could be a sequence of video frames or even different views of the same scene. A traditionally used super resolution method is the Bayesian one [2], in which all input images have the same resolution. A different approach was proposed in [3], in which a algorithm is used to extract examples of a database of images to increase the resolution of a given image.

In video compression, there is a trade off between quality and rate required to represent an image. In this way, the better the quality desired, the greater the amount of bits to be spent. Usually, this rate-distortion relation is controlled by a quantization parameter. Alternative video encoding systems have been proposed to take care of this relation. For example, a mixed resolution approach was proposed, in which just part of frames are downsampled allowing scalable video coding [4], or reducing complexity coding [5]. In these mixed resolution schemes [6],[7], two types of frames could be identified: frames that remain in high resolution format, called Key Frames (KFs); and the frames that have reduced resolution, the Non-Key Frames (NKFs). The KFs could be encoded like Intra frames (I), Predicted (P) or even Bipredicted (B), and will be used as reference on NKFs encoding process.

At the decoder side, high resolution frames are reconstructed, interleaved with low resolution ones, as shown in Figure 1. Therefore, itt's desired to use neighbor KFs to enhance the NKfs quality. This kind of solution could be applied in any encoding system in which a set of more degraded frames are interleaved with a less degraded ones like, for example, mixed quality codecs [8]. Also, the proposed method is relevant in the side-information generation process for certain distributed video coding architectures [9].

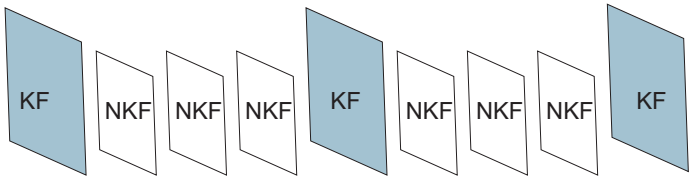

Fig. 1. Exemple of a video sequence in a mixed spacial resolution codec. A sequence of key frames (KF), in high-resolution, is interleaved with non-key frames (NKF), in low-resolution.

The solution proposed to enhance non-key frames quality by using high resolution ones is quite generic, covering different video encoding scenarios. The main goal of this work is presents a efficient technique to enhance low-quality decoded frames by using high-quality information extracted from available neighbor frames in full resolution.

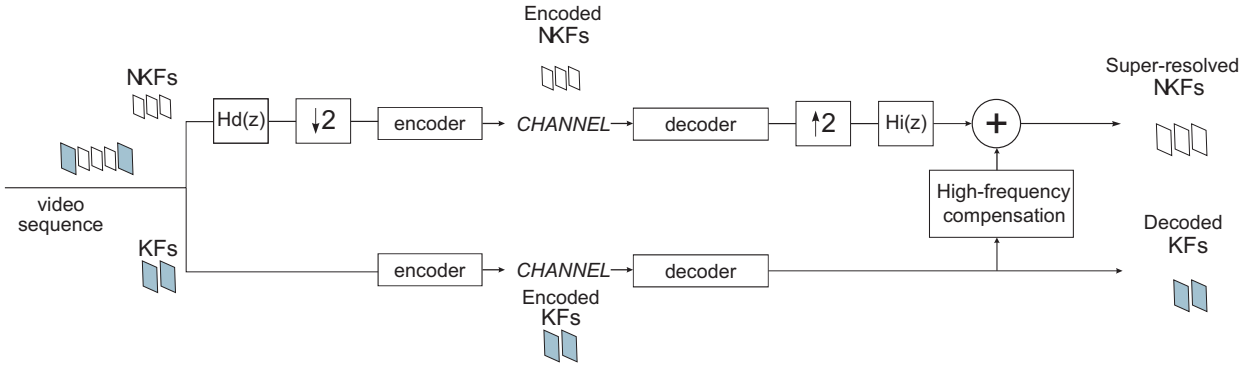

Fig. 2. General scheme for the super-resolution proposed method. At the encoder side, the key-frames (KFs) are independently encoded in full resolution and the non-key frames (NKFs) are filtered and down-sampled to remain at low-resolution. At the decoder side, the KFs and the NKFs are reconstructed and the NKFs are super-resolved.

II. Key-frame based super resolution

The proposed super resolution method is based on a mixed resolution approach [10], as depicted in Figure 2. At the encoder side, one can see the spliting between KFs and NKFs. The NKFs are filtered and down-sampled, what reduces their resolution, while the KFs are encoded in full resolution. Then, the whole mixed resolution codified sequence is transmitted by a channel. Already at the decodert's, after all frames have been decoded, the proposed super resolution will be applied to increase the NKFs quality. This method is basically composed of the following steps:

- NKFs interpolation, to have the same size of the KFs;
- KFs degradation, resulting from decimation and interpolation processes, followed by motion estimation between blocks of the NKFs and the blocks of the backward key-frame (KFb or Key-Frame backward), and the the blocks of the forward key-frame (KFf or Key-Frame forward);
- High-frequency information extraction from the NKFs blocks;
- Weighting of the high-frequency information of the KFb and KFf best matching blocks, according to the respective reliability measure, which is inversely proportional to the SSDs (Sum of Squared Differences);
- Summ of the wheighted high-frequency information to the NKFs interpolated blocks, in a process called high frequency compensation.

As shown in Figure 2, one can identify the presence of two filters, one for decimation, named Hd, and another for interpolation, named Hi, with the respective function to avoiding aliasing and imaging effects. For this, it was used a Lanczos3 filter, a windowed form of the sinc filter with three lobes preserved. This kind of filter is already implemented on MatLab ®software functions. In the next sections, each of the procedures performed after the decodification step to enhance the quality of the low resolution frames will be carefully described.

A. High-frequency compensation

After decodifying a mixed resolution video sequence, the NKFs should be interpolated. However, in this process, the interpolated frames loses contents of high-frequency, because they were encoded at a lower resolution. It's known that, in general, temporally adjacent or neighboring frames in a video sequence are stronger correlated. Based on it, a motion estimation process on the KFb and on KFf is proposed to take the high-frequency content of the two best matching blocks, which will be added to the actual NKF interpolated block. In a previous work [10], the motion estimation process used as reference in both KFb and Kff, but just the high-frequency information of the best match block of these two frames was added. In this work, we propose weighting the high-frequency information of the two best matchs, each of one neighbor KF, weighted by a reliability measure and then add the composed information to the interpolated block.

A central question of this work is how to do a efficient matching between the NKFs interpolated blocks and the KFs blocks, ensuring also a good high-frequency matching. This way, the KFs also will be subjected to the same process of degradation suffered by the NKFs, that is, it will be decimated and interpolated before the motion estimation process. Also, it was observed that to perform the search on the high-frequency information of the interpolated frames leads to better results than just doing on the interpolated frames, especially when it has a high density of KFs. Itt's important to note that by doing the procedure depicted in Figure 3 is the same as using the low frequency content to predict the high frequency one like in a subband decomposition. This is the same hierarchical concept explored by zerotrees image coders [11], [12].

To implement this prediction process, the degraded KFs and the interpolated NKFs will be subjected to a filtering operation, only for the motion estimation process, whose two-dimensional mask is:

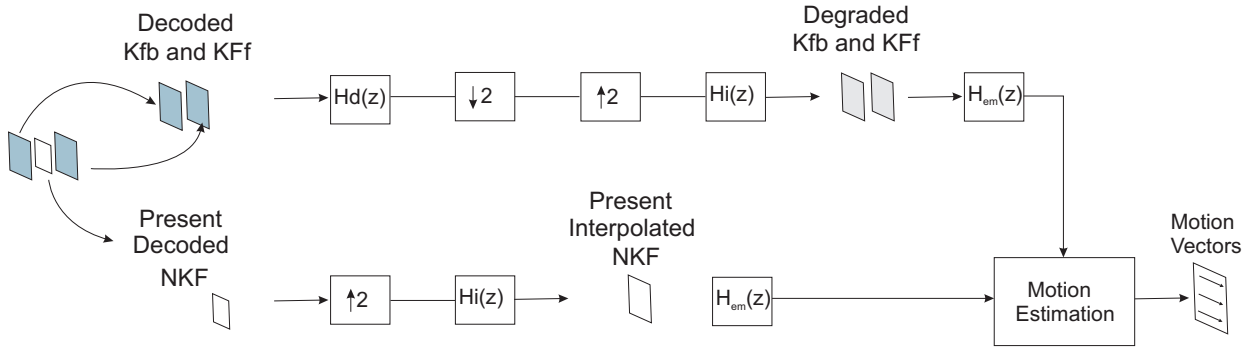

Fig. 3. Search of the NKFs blocks in KFb and in KFf. The NKFs are only interpolated, while the KFs are down-sampled and interpolated to suffer a degradation similar of the NKFs. Later, all frames are filtered by a high-pass filter (H_{em}), performing a motion estimation in the interpolated frames high-frequency information.

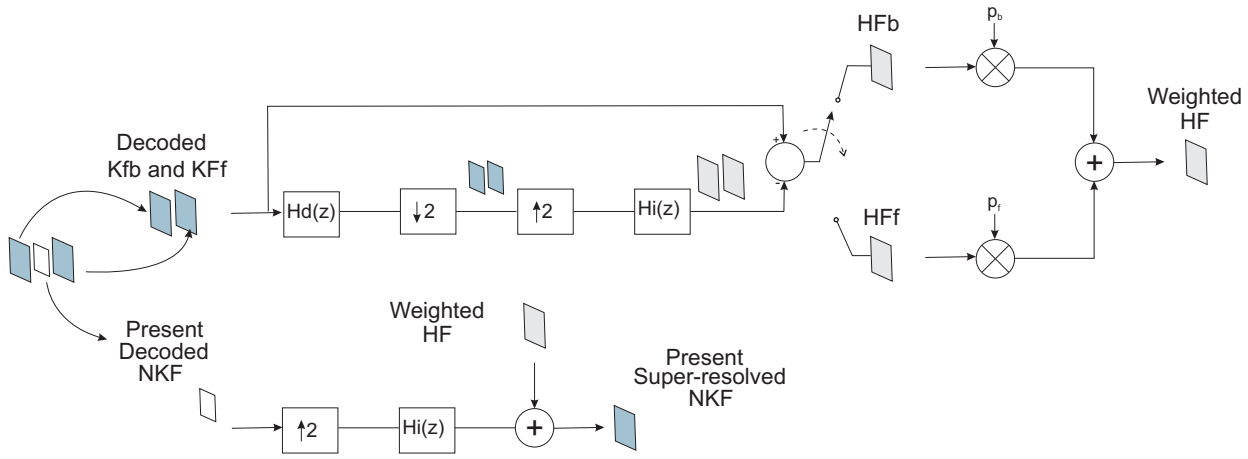

Fig. 4. High-frequency compensation of the NKFs. The KFs high-frequency content is obtained by the difference between the decoded KFs and its degraded versions. The KFb and KFf high-frequencies information are weighted and added to the interpolated NKF to enhance it.

$$H_{em} = \frac{1}{9}\begin{pmatrix} -1 & -1 & -1 \\ -1 & 8 & -1 \\ -1 & -1 & -1 \end{pmatrix}. \tag{1}$$

Further detailing the motion estimation process, the search algorithm was implemented with variable blocksize. Thus, the search is started with blocks of 16×16 pixels partitioned to the 8×8, if the sum of this sub blocks SSDs is greater than the macroblock SSD. Itt's important to note that itt's not necessary to implement a high-pass filter to extract the KFs high-frequency information. This content could be obtained just by doing the difference between the decoded KF and its degraded version. This high-frequency compensation procedure is shown in Figure 4

B. Determining the weights of the KFbs and KFfs high-frequency information

As mentioned, in the high-frquency compensation process of the NKFs, the high-frequency content of the best matching blocks of the KFb and the KFf should be weighted, in a manner that reflects the reliability that one have on each information. In order to do that, we propose to use weights inversely proportional to the SSDs obtained in the motion estimation process. However, these weights should also be normalized to not change the original range. Thus, the weights are given by the following expressions:

$$p_b = \frac{SSD_f}{SSD_b + SSD_f} \tag{2}$$

$$p_f = \frac{SSD_b}{SSD_b + SSD_f}, \tag{3}$$

in which SSD_b and SSD_f are, respectively, the SSDs obtained by the search in KFb and KFf. A similar weighting also itt's been used for the sequnces composed by frames with variable quality [8].

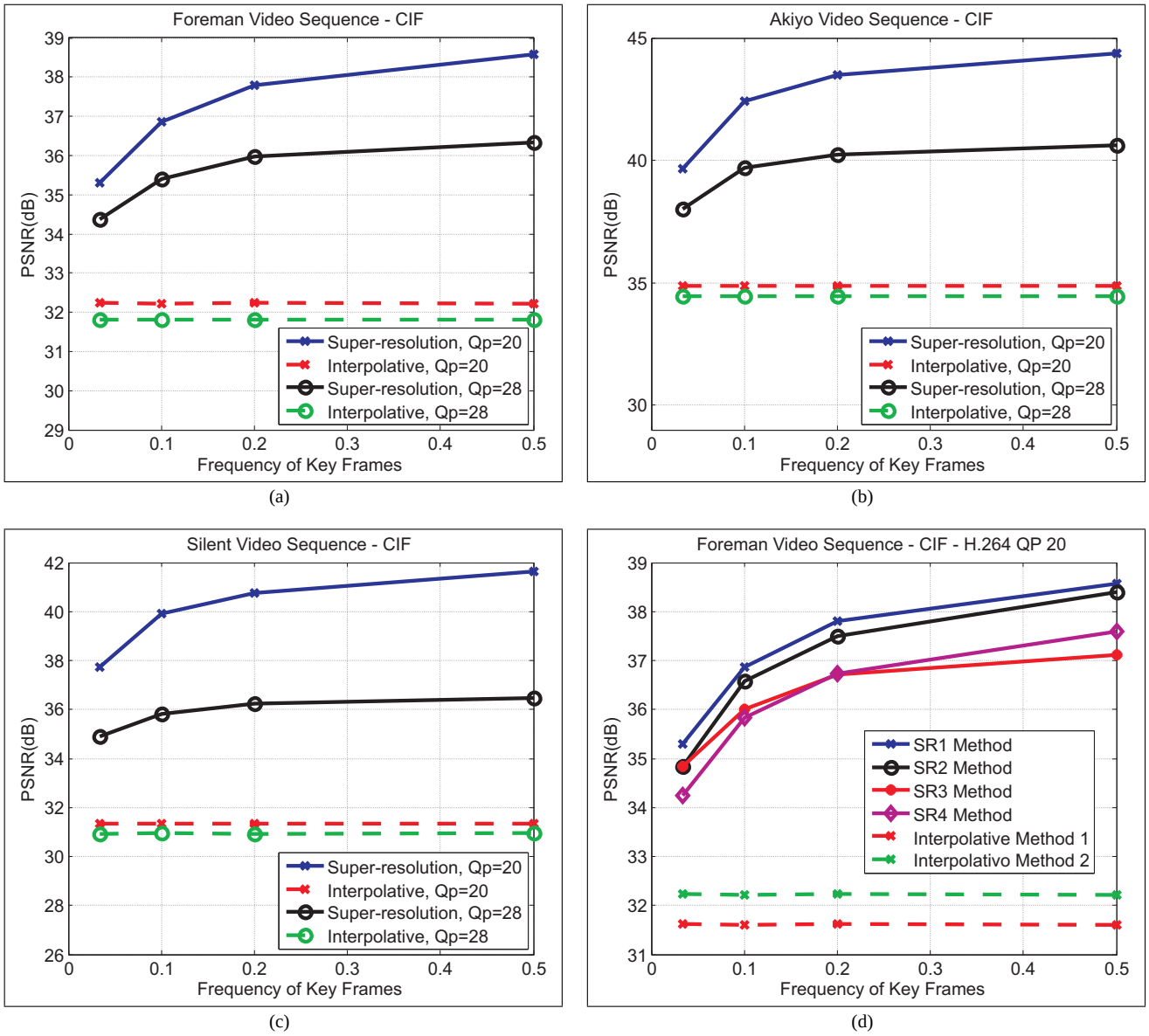

Fig. 5. Super-resolution results for the sequences: (a) Foreman, (b) Akiyo and (c) Silent. (d) Comparison between the improvements reached by the different contributions. The SR1 Method uses variable blocksize, biprediction and Lanczos3 filters for down-sampling and interpolation. The SR2 Method is also bipredictive and uses Lanczos3 filters, but just fixed blocksize. The SR3 Method uses fixed blocksize and Lanczos3 filters, but is only predictive. The SR4 Method uses fixed blocksize and is bipredictive, but downsamples with an average filter and interpolates with a bicubic kernel. The Interpolative Method 1 performs downsample with an average filter and bicubic interpolation, while the Interpolative Method 2 uses Lanczos3 filters.

III. Experiments

The proposed super-resolution method was applied to the following CIF video sequences: Foreman, Akyio and Silent. In order to do the tests, thirty frames of each sequence were encoded using H.264/AVC with Intra mode selected. Also, two values were setted for the quantization parameter (QP), 20 and 28, and the KFs quantity was uniformely varied in eac sequence ($1/30$, $3/30$, $6/30$ e $15/30$). In the motion estimation process, a bipredictive full-search method was used, considering a serch window of 32×32 *pixels* for 16×16 *pixels* blocks, and a search window of 16×16 *pixels* for 8×8 *pixels* blocks. As a criterion for the best matching, the least SSD was used, and the search was made in the high-frequency of the interpolated frames. The Figure 5(a), (b) and (c) shows the results obtained, comparing the proposed super-resolution method with intuitive method of simply interpolate the low-resolution frames (interolative method), by using the same Lanczos3 kernel.

Comparing the present results with those of previous works [10], one can note that were obtained gains of around 4dBs. Part of it should be given to improving the quality of interpolated frames. However, as more significant improvement factors could be pointed out the new techniques for bipredictive weighting, using weights inversely proportional to the SSDs; and the variable blocksize motion estimaton. To take the idea of gain increase promoted by each of these new contributions, the Figure 5(d) shows, for the Foreman sequence and $Qp = 20$, a comparison between our current best method (SR1 Method), that uses

a variable blocksize matching, with a 16×16 fixed blocksize method (SR2 Method). Itt's also shown another version with fixed blocksize and just using as reference one KF, this is, the one with the least SSD (SR3 Method). By the end, one also can see a drop in performance by changing Lanczos3 filters for downsampling with average followed by bicubic interpolation (SR4 Method).

IV. Work Underway

We continue to look for better modeling of the high-frequency compensation process. A more complex model was developed to take advantage of other available informations. This new model will be described below.

Let a frame F to be super-resolved using two key-frames, namely frames F_1 and F_2. Frames F_1 and F_2 are decimated and interpolated back, generating frames F_1' and F_2' and, while frame F is interpolated to generate frame F'. After block-match motion estimation search in F_1' and F_2', the best match to a given block L in F' are L_1 and L_2 in F_1' and F_2', respectively. The corresponding block in F_1 for L_1 is B_1. The same thing applies to F_2. The high-frequency blocks of the best matches are $H_1 = B_1 - L_1$ and $H_2 = B_2 - L_2$. Let D be a distance metric so that:

$$D_H = D(H_1, H_2)$$
$$D_1 = D(L, L_1)$$
$$D_2 = D(L, L_2)$$

So, we want to super-resolve block L into block B, by adding high-frequency content, i.e.

$$B = L + \beta H,$$

where H is high-frequency content, which is a function of H_1 and H_2, and β is a confidence factor.

As H is to be $H = \alpha H_1 + (1 - \alpha)H_2$, we suggest the use of a variable p to control how much privilege will be given to the best of the two matches, obtaining:

$$\alpha = \frac{D_2^p}{D_1^p + D_2^p}.$$

$p \to \infty$ means that we always pick the best match (smaller D_k wins), and $p = 0$ means $\alpha = 1/2$ no matter what. By doing $p = 1$, note that our current stage is taken. We suggest to make this variable p to be smoothly variable with a distance measure between the high-frequencies to be summed (DH), as:

$$p = e^{cD_H}$$

.

Also, Let $D_m = min\{D_1, D_2\}$. We suggest to control the amount of high-frequency information to be added by a reliability measure that we have on this content, this is

$$\beta = e^{-\sigma(D_m - \tau)^2}$$

if $D_m > \tau$, for a given threshold τ. If $D_m \leq \tau$, then $\beta = 1$.

The model is being tested and some results should be included in the final version of this paper.

References

[1] A. K. Katsaggelos, R. Molina e J. Mateos, "Super Resolution of Images and Video". *Synthesis Lectures on Image, Video, and Multimedia Processing*. ISBN 978-1598290844, Morgan and Claypool Publishers, 2007.

[2] C. A. Segall, A. K. Katsaggelos, R. Molina e J. Mateos, "Bayesian Resolution Enhancement of Compressed Video". *IEEE Transactions on image processing*, vol. 13, no. 7, 2004.

[3] W. T. Freeman, T. R. Jones e E. C. Pasztor, "Example-based super-resolution", *IEEE Computer Graphics and Applications*, vol. 22, pp. 56-65, 2002.

[4] H. Schwarz, D. Marpe e T. Wiegand, "Overview of the scalable video coding extension of the H.264/AVC standard", *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 17, no. 9, pp. 1103-1120, 2007.

[5] D. Mukherjee, B. Macchiavello e R. L. de Queiroz, "A simple reversed-complexity Wyner-Ziv video coding mode based on a spatial reduction framework", *Proc. IS&T/SPIE Symp. on Electronic Imaging, Visual Communications and Image Processing*, vol. 6508, pp. 65081Y1-65081Y12, 2007.

[6] B. Macchiavelo, D. Mukherjee e R. L. De Queiroz, "Iterative side-information generation in a mixed resolution Wyner-Ziv framework," preprint, IEEE Trans. Circuits and Systems for Video Technology, 2009.

[7] D. Mukherjee, "A robust reversed-complexity Wyner-Ziv video coding mode based on a spatial reduction framework", *HP Labs Techinical Report*, HPL-2006-80, 2006.

[8] E. M. Hung, R. L. de Queiroz e D. Mukherjee, "Codificação det'vídeo com complexidade reversa utilizando qualidade mista", *Simpósio Brasileiro de Telecomunicações (SBrT 2009)*.

[9] B. Macchiavello, F. Brandi, E. Peixoto, R. L. de Queiroz e D. Mukherjee, "Side-information generation for temporal and spatial scalable Wyner-Ziv codecs", *EURASIP Journal of Image and Video Processing*, vol. 2009, pp.1-11, 2009.

[10] F. Brandi, R. L. de Queiroz e D. Mukherjee, "Super resolution of video using key frames and motion estimation", *Proc. of International Conference on Image Processing*, 2008.

[11] J. M. Shapiro, "Embedded image coding using zerotrees of wavelets coefficients", *IEEE Transactions on Signal Processing*, SP-41:3445-3462, 1993.

[12] A. Said e W. A. Pearlman, "A new fast and efficient coder based on set partitioning in hierarchical trees", *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 243-250, 1996.