Super-resolution applied to Distributed Video Coding with Spatial Scalability

Bruno Macchiavello, Fernanda Brandi, Ricardo L. de Queiroz and Debargha Mukherjee

***Resumo*— A codificação distribuída de vídeo é um novo paradigma que permite obter codificadores de vídeo com complexidade reversa. É sabído que o desempenho de qualquer codificador de vídeo distribuído depende da qualidade da informação lateral gerada no decodificador através de estimação de movimento. Dando continuidade a trabalhos passados, neste artigo propomos o uso de um método de super-resolução para interpolar os quadros não-chave usando os quadros chave como referência, com o objetivo de melhorar a geração da informação lateral em um codificador de vídeo distribuído com escalabilidade espacial. O *framework*, proposto em trabalhos anteriores, permite a redução da complexidade de codificação a partir da redução da resolução espacial nos quadros que não são usados como referência, seguido de uma codificação Wyner-Ziv do resíduo Laplaciano. A geração de informação lateral é iterativa, na primeira iteração a alta-freqüência dos blocos caçados, mediante a procura da menor distorção, dos frames chave é adicionada aos respectivos blocos de baixa resolução nos quadros não-chave. São apresentados resultados usando o codificador estado-da-arte H.264/AVC.**

***Palavras-Chave*— super-resolução, Wyner-Ziv, escalabilidade espacial.**

***Abstract*— Distributed Video Coding is a new video coding paradigm that enables video codecs with reversed complexity. It is well-known that the performance of any distributed video coder is heavily dependent on the quality of the side information generated by motion-estimation at the decoder. As a continuation of previous works, in this paper we propose to use a super-resolution method to up-sample the non-key frames using the key frames as reference, in order to improve the side information generation in a distributed video coding scheme with spatial scalability. The framework, proposed in previous works, enables reduced encoding complexity by reduced spatial-resolution encoding of the non-reference frames, followed by Wyner-Ziv coding of the Laplacian residue. The side information generation is iterative, in the first iteration the high-frequency data of matching blocks from the key frames are added to the low-resolution blocks of the interpolated downsampled frames. Results are presented using the state-of-the-art video coding standard H.264/AVC.**

***Keywords*— Super-resolution, Wyner-Ziv, spatial scalabilty.**

I. Introduction

The paradigm of distributed source coding (DSC) is based on two information theory results, the theorems by Slepian and Wolf [1] and Wyner and Ziv [2] for coding correlated sources lossless and lossy respectively. It has recently become the focus of different kinds of video coding schemes [3]–[8]. A review on DSC applied to video coding, distributed video coding (DVC), can be found elsewhere [9]. DVC is a promising tool in creating reversed complexity codecs for power constrained devices. Currently, digital video standards are based on discrete cosine block transform and predictive interframe coding. Typically, the encoder has high complexity [10], mainly due to the need for mode search and motion estimation in finding the best predictor, whereas the decoder complexity is low. On the other hand, DVC allows for a reversed complexity codec, where the decoder is more complex than the decoder. This scheme fits the scenario where real-time encoding is required in a limited-power environment, such as mobile hand-held devices.

A mixed resolution framework that can be implemented as an optional coding mode in any existing video codec standard was proposed [11]–[13]. In that framework the encoding complexity is reduced by lower resolution encoding, while the residue is Wyner-Ziv (WZ) encoded by cosets. Also, the proposed framework considers the following realistic usage scenarios for video communication using mobile power-constrained devices. First, it is not necessary for the video encoder to always operate in a reversed complexity mode. Thus, this mode may be turned on only when available battery power drops. Second, while complexity reduction is important when battery power drops, it should not be achieved at a substantial cost in bandwidth. Hence, the amount of complexity reduction may be reduced in the interest of a better rate-distortion trade-off. Third, since the video communicated from one mobile device may be received and played back in real-time on another mobile device, the decoder in a mobile device must support a mode of operation where at least a low quality version of the reversed complexity bit-stream can be decoded and played back immediately with low complexity. Off-line processing may be carried for retrieving the higher quality version.

However, it is well-known that the performance of this or any other WZ codec is heavily dependent on the quality of the side information (SI) generated at the decoder. An iterative SI generation method was introduced to be applied on the mixed resolution DVC framework [11]. In this work, as a continuation of [11], we propose a new SI generation method. This method is based on a previous study in super-resolution using key frames [14]. The main idea is to restore the high-frequency information of an interpolated reconstructed block of the low resolution encoded frame. This is done through searching in the high resolution encoded key frames for a similar block, and by adding the high-frequency of the chosen

Bruno Macchiavello, Fernanda Brandi, Ricardo L. de Queiroz are with the Departamento de Engenharia Elétrica, Faculdade de Tecnologia, Universidade de Brasília, Brasília, Brasil, E-mails: {bruno,fernanda,queiroz}@image.unb.br. Debargha Mukherjee is with Hewlett Packard Labs, Palo Alto, California, USA, E-mail: debargha.mukherjee@hp.com. This work was supported by HP Brasil and by CNPq under grant 47.3696/2007-0.

block to the interpolated one. We present the results for this new SI generation implemented on an WZ coding mode for H.264/AVC.

II. Framework

The mixed resolution framework [11]–[13] can be implemented as an optional coding mode in any existing video codec standard (results using H.263+ can be also found [11],[12]). In this framework, the reference frames (key frames) are coded exactly as in a regular codec as I-, P- or reference B-frames, at full resolution. For the non-reference P- or B-frames, called non-reference WZ frames or non-key frames, the encoding complexity is reduced by low resolution (LR) encoding, as shown in Fig 1.

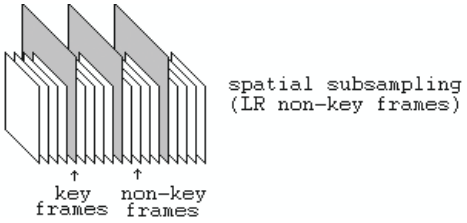

Fig. 1. Illustration of key frames in spatial scalable video.

The architecture of the WZ encoder is shown in Fig. 2, the non-reference frames are decimated and coded using decimated versions of the reconstructed reference frames in the frame store. Then, the Laplacian residual, obtained by taking the difference between the original frame and an interpolated version of the LR layer reconstruction, is Wyner-Ziv coded to form the enhancement layer. Since the reference frames are regularly coded, there are no drift errors. Ideally, the number of non-references frames and the decimation factor can be varied dynamically based on the complexity reduction target. At the decoder, Fig. 3, high quality versions of the non-reference frames are generated originally by a multi-frame motion-based mixed super-resolution mechanism [11]. The interpolated LR reconstruction is subtracted form this frame to obtain the side-information Laplacian residual frame. Thereafter, the WZ layer is channel decoded to obtain the final reconstruction. Note that for encoding and decoding the LR frame, all reference frames in the frame store and syntax elements are first scaled to fit the non-reference LR coded frame. The channel code used is based on memoryless cosets. A study for optimal coding parameter selection for coset creation can be found elsewhere [15], [16]. There, a mechanism to estimate the correlated statistics form the coded sources is described.

III. Super Resolution for Side Information Generation

As mentioned, at the decoder, the SI is generated iteratively. However, the first iteration is different form the other ones and represents the main contribution of this work. In the first iteration, similar to an example-based algorithm [17], we seek to restore the high-frequency information of an interpolated block through searching in previous decoded key frames for a

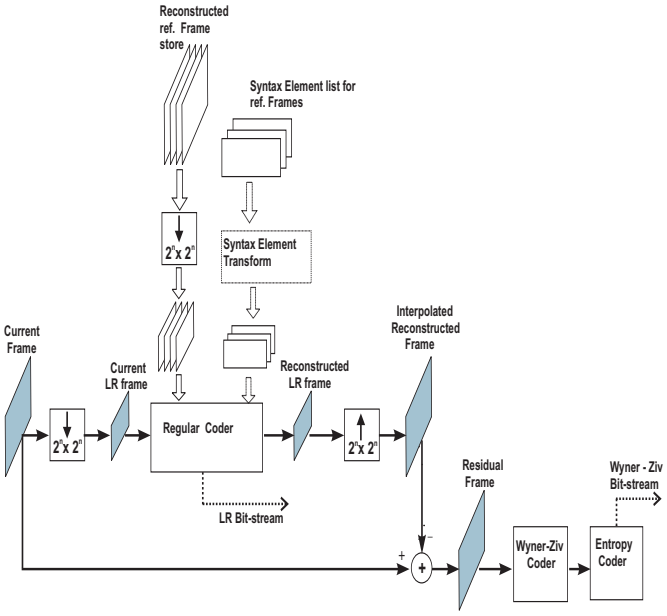

Fig. 2. Encoder of the WZ Mixed Resolution Framework

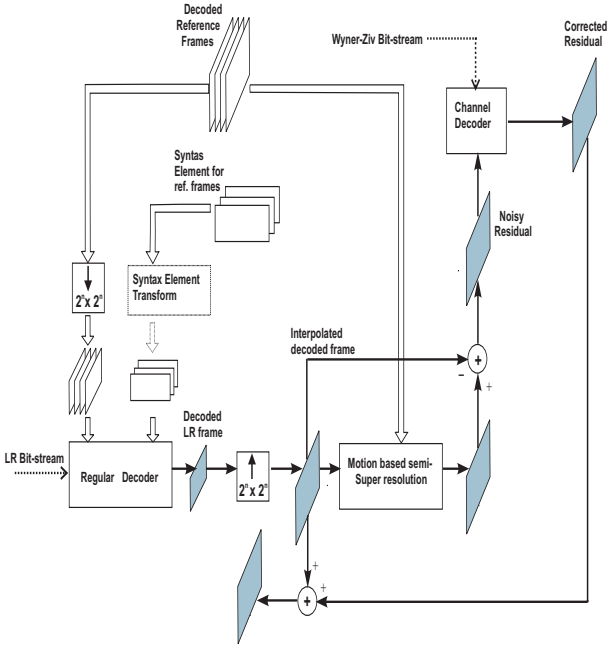

Fig. 3. Decoder of the WZ Mixed Resolution Framework

similar block, and by adding the high-frequency of the chosen block to the interpolated one.

The original sequence of frames at a high resolution have both key frames and non-key frames. The framework will encode the non-key frames at a lower resolution and the key frames at regular resolution. At the decoder, the video sequence is received with mixed resolution; the decoder will interpolate the non-key frames to obtain all the decoded frames at the desired resolution. However, note that the decoded non-key frames have lost high-frequency content, since they have been coded at lower resolution and then a simply interpolation

has been applied. Our algorithm will try to recover the lost high frequency content using temporal information from the key frames.

First, the past and future references frames in the frame store, of the current non-key frame, are low-pass filtered. The low-pass filter is implement through a down-sampling followed by an up-sampling process (using the same decimator and interpolator applied to the non-key frames). At this point we have both key and non-key frames interpolated form a LR version. Next, a block-matching algorithm is applied using the interpolated decoded frame. The block-matching algorithm works as follows. For every 8×8 block in the interpolated decoded frame, the best sub-pixel motion vectors in the past and future filtered frames are computed. If the corresponding best predictor blocks are denoted as B_p and B_f in the past and future filtered frames respectively, several predictor candidates are calculated as $\alpha B_p + (1-\alpha)B_f$, where α assumes values between 0 and 1, for our implementation we use $\alpha = \{0.0, 0.25, 0.5, 0.75, 1.0\}$. Then, if the sum of absolute differences (SAD) of the best predictor of a particular macroblock is lower than a threshold T, the corresponding high-frequency of the matched block of the key frame is added to the block to be super-resolved (see Fig. 4). The high-frequency is generated by subtracting the decoded key frames from their filtered version.

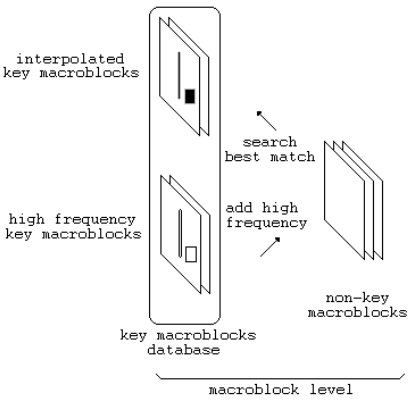

Fig. 4. After searching for a best match in the database, we add the corresponding high-frequency to the block to be super-resolved

Different from [14], the high frequency content is first scaled by a confidence factor before being added to the LR block. The confidence factor is calculated based on the SAD obtained from the block matching algorithm and the rate (R_n) spent by the coder in order to encode the current block. If the lower SAD calculated during the block matching algorithm has a high value it is unlikely that the high frequency of the key frame block matches exactly the lost high frequency of the non-key frame block. Then, it is intuitive to think that lower SAD gives us more confidence in our match. Also, if, at the encoder side, a large amount of bitrate is spent to code a particular block, it is because it wasn't found a good match on the reference frames. Thus, the more bitrate the less the confidence. The scaling factor that will multiply each pixel of the high frequency block, before adding it to the block to be super-resolved, is defined as:

$$c = 1 - ((SAD + \lambda R_n)/c_{max}), \quad (1)$$

where c_{max} represents the threshold T, and λ is a Lagrange multiplier. Note, that if $SAD = R_n = 0$ then $c = 1$, that means that all the high-frequency content will be added. In the other hand, if $(SAD + \lambda R_n) = c_{max}$ then $c = 0$, so no high frequency will be added. The values of c_{max} and λ were empirically found using different test sequences.

After the first iteration, parameters may change. From iteration to iteration the strength of the low-pass filter should be reduced (in our implementation the low-pass filter is eliminated after one iteration). The grid for block matching is offset from iteration to iteration to smooth out the blockiness and to add spatial coherence. For example, the shifts used in four passes can be (0, 0), (4, 0), (0, 4) and (4, 4) (see Fig. 5). It is important to note that after the first iteration we already have a frame with high frequency content. Hence, after the first iteration the SI generation is similar to the work presented at [11], where we replace the entire block for the unfiltered matched block on the key frames, instead of just adding high-frequency. Then, after the first iteration the threshold T (or c_{max} in (1)) is drastically reduced, and will continue to be reduced gradually so that fewer blocks are changed in later iterations.

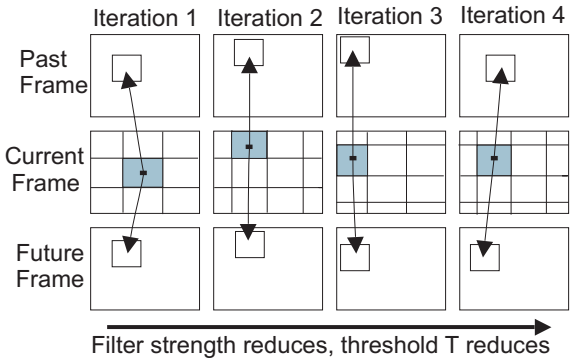

Fig. 5. SI generation for non-reference WZ frames. Threshold reduces, and the grid is shifted from iteration to iteration.

IV. RESULTS

The proposed SI generation using super-resolved frames as in [14] was implemented in an optional WZ coding mode on the KTA software implementation of H.264/AVC [18]. In our entire tests we use fast motion estimation, along with CAVLC entropy coder and no rate-distortion optimization. The parameters used to obtain the confidence value were set to $\lambda = 0.1$ and $c_{max,i} = \{500, 80, 60\}$, where i represents the iteration number and $1 \leq i \leq 3$. The H.264 codec was set in $IBPBP...$ mode, where the B-frames were non-reference frames. For the WZ coding mode, the B-frames where downsampled by a factor of 2×2 (quarter resolution). In Fig. 6 we compare the SI generated after only one iteration of the proposed method against the previous one [11] for 299 frames of the test sequences. The PSNR curves include key and non-key frames. It can be seen that the new SI

generation significantly improves performance, mainly due to the confidence value that scales the high-resolution content. For all Figures the rate measures the number of bits for the entire 299 frames, it does not measure the number of bits per second.

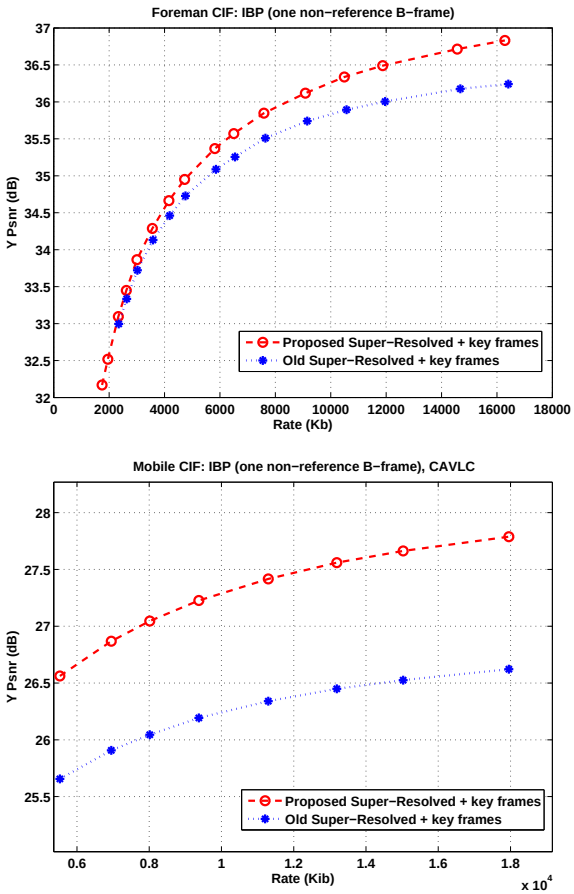

Fig. 6. Results for super-resolved frames along with key frames, using the new and old mehod for Foreman and Mobile CIF sequences.

In Fig. 7 we compare regular H.264/AVC, the WZ coding mode after three iterations and the method using one iteration of super-resolution. As expected, using the WZ layer improves the performance compared to just super-resolve frames. However, the performance of the WZ mode is completely dependent on the SI generation.

Finally, in Fig. 8 we can see the performance of the WZ coding mode after three iterations for two more test sequences. It can be seen that the WZ coding mode is competitive. Curiously, for the Silent CIF sequence, the WZ mode outperforms the regular H.264 at low rates.

V. Conclusions

In this work we have introduced a new SI generation method for a Wyner-Ziv coding mode with spatial scalability presented in previous works. The SI generation uses a confidence value to scale the amount of high-frequency content that will be added to the block to be super-resolved. The results show an improvement in comparison of the previous developed algorithm. The WZ coding mode is competitive and may outperform regular codec for low-movement sequences at low rates.

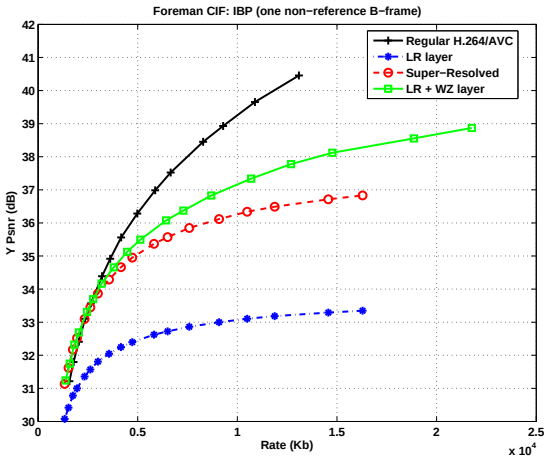

Fig. 7. Results for Foreman sequence, comparing regular H.264/AVC, WZ coding mode, and the super-resolved frames.

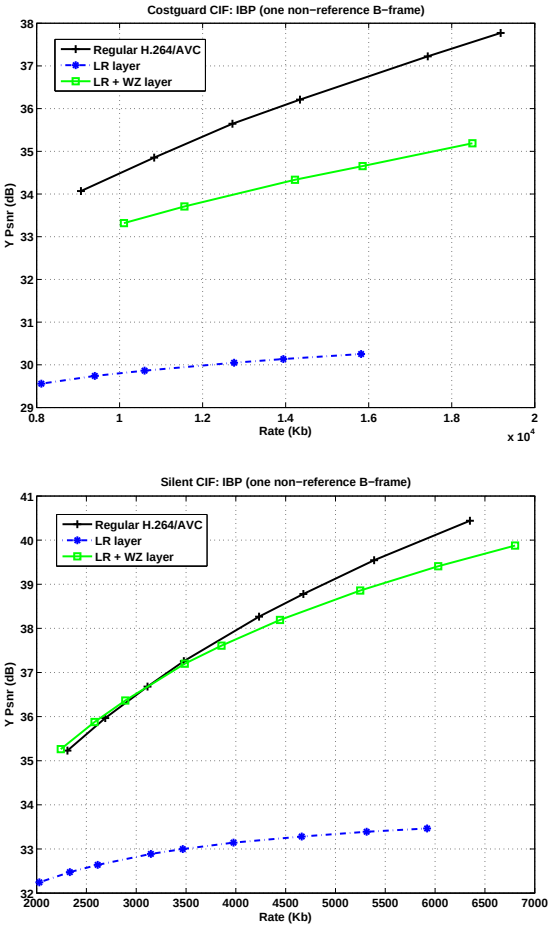

Fig. 8. Results for WZ coding mode for Coastguard and Silent CIF sequences.

REFERENCES

[1] J. Slepian and J. Wolf, "Noiseless coding of correlated information sources", *IEEE Trans on Inf. Theory*, Vol. 19, No. 4, pp. 471-480, Jul 1973.

[2] A. Wyner and J. Ziv, "The rate-distortion function for source coding with side information at the decoder", *IEEE Trans. on Inf. Theory*, Vol. 2, No. 1, pp. 1-10, Jan 1976.

[3] A. Aaron, R. Zhang, and B. Girod, "Transform-domain Wyner-Ziv codec for video," *Proc. SPIE Visual Communications and Image Processing*, vol. 5308, pp. 520-528, January 2004.

[4] R. Puri and K. Ramchandram, "Prism: A new robust video coding architecture based on distributed compression principles", *Allerton Conference on Communications, Control and Computing*, 2002.

[5] Q. Xu and Z. Xiong, "Layered WynerZiv video coding," *IEEE Trans on Image Processing*, vol. 15, no. 12, pp. 3791-3809, Dec 2006.

[6] H. Wang, N. M. Cheung, and A. Ortega, "A framework for adaptive scalable video coding using Wyner-Ziv techniques", *EURASIP Journal on Applied Signal Processing*, pp. 1-18, 2006.

[7] X. Wang and M. T. Orchard, "Desing of trellis codes for source coding with side information at the decoder", *Proceedings of IEEE Data Compression Conference*, pp. 361-370, 2001.

[8] A. Aaron and B. Girod, "Compression with side information using turbo codes", *Proceedings of IEEE Data Compression Conference*, pp. 252-261, 2002.

[9] B. Girod, A.M. Aaron, S. Rane, and D. Rebollo-Monedero, "Distributed video coding," *Proceedings of the IEEE*, vol. 93, no. 1, pp. 71-83, Jan 2005.

[10] T. Weigand, G. Sullivan, G. Bjontegaard, and A. Luthra, "Overview of the H.264/AVC video coding standard," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 13, no. 7, pp. 560-576, 2003.

[11] B. Macchiavello, R. de Queiroz, and D. Mukherjee, "Motion-based side-information generation for a scalable Wyner-Ziv video coding," *Proc. of the IEEE Int. Conf. on Img. Proc.*, pp. VI-413-VI-416, 2007.

[12] D. Mukherjee, B. Macchiavello, and R. L. de Queiroz, "A simple reversed-complexity Wyner-Ziv video coding mode based on a spatial reduction framework," *Proc. of SPIE Visual Communications and Image Processing*, vol. 6508, pp. 65 081Y1-65 081Y12, Jan. 2007.

[13] D. Mukherjee, "A robust reversed complexity Wyner-Ziv video codec introducing sign-modulated codes," *HP Labs Technical Report, HPL-2006-80*, May 2006.

[14] F. Brandi, R. L. de Queiroz and D. Mukherjee, "Super resolution of video using key frames," *Proc. of International Symposium on Circuits and Systems*, Seattle, WA, USA, May 2008.

[15] D. Mukherjee, "Optimal parameter choice for Wyner-Ziv coding of laplacian sources with decoder side-information," *HP Labs Technical Report*, HPL-2007-34.

[16] B. Macchiavello, D. Mukherjee, and R. de Queiroz, "A statistical model for a mixed resolution Wyner-Ziv framework," *Picture Coding Symposium*, Lisboa, Portugal, November 2007.

[17] W.T. Freeman, T.R. Jones, and E.C. Pasztor, "Example-Based Super-Resolution," *IEEE Computer Graphics and Applications*, Vol. 22, pp. 56-65, 2002.

[18] J. Jung and T.K. Tan, "KTA 1.2 software manual," *VCEG-AE08*, January, 2007.