A Video Coding Framework with Spatial Scalability

Bruno Macchiavello, Eduardo Peixoto and Ricardo L. de Queiroz

***Resumo*— Um novo paradigma de codificação de vídeo, codificação distribuída de vídeo, tem sido o foco de vários estudos recentes. Neste trabalho, apresenta-se um *framework* simples de codificação de vídeo, baseado nos princípios da codificação distribuída, que pode ser aplicado a qualquer padrão de codificação de vídeo mediante pequenas modificações. O *framework* permite escalonabilidade espacial para os *frames* que não são usados como referência, e não requer um canal de retorno entre o codificador e decodificador frequentemente usado em codificação distribuída. No codificador, a complexidade é reduzida devido à codificação em baixa resolução dos *frames* que não são usados como referência. No decodificador, a informação lateral é gerada usando os *frames* de referência mediante estimação e compensação de movimento. O resultado da aplicação deste *framework* ao padrão H.263+ é mostrado.**

***Palavras-Chave*— escalonabilidade espacial, Wyner-Ziv, codificação de baixa complexidade**

***Abstract*— A new video coding paradigm, distributed video coding, has been the focus of many recent studies. In this paper we present a simple video coding framework, based on the principles of distributed coding, that can be applied to any video coding standards with minor modifications. The framework allows spatial scalability of the non-reference frames, and does not need any feedback channel between the encoder and decoder. The complexity in the encoder is reduced since the non-reference frames are coded at lower spatial resolution. At the decoder, side-information is generated using the reference frames through motion estimation and compensation. Results using the H.263+ standard are shown.**

***Keywords*— spatial scalability, Wyner-Ziv, low complexity coding**

I. INTRODUCTION

Today, most video compression methods are based on the discrete cosine transform (DCT) and on motion compensated prediction. The goal of these tools are the reduction of spatial and temporal redundancy, respectively. Typically, the encoder has a higher complexity than the decoder [1], mainly due to motion estimation. Recently, new applications have emerged in digital video streaming and broadcasting, like mobile wireless video communication. Along with them, new requirements have also emerged, such as bandwidth fluctuation and different Quality-of-Service (QoS). The need for scalable video coding has also increased. Scalable coding can adapt and optimize the quality of video for a range of bitrates rather than a fixed rate, and/or can also lower the complexity of the encoder [2]. In wireless applications, i. e. mobile camera phones, the computational complexity becomes a very important issue since it is essential to have low power consumption.

Distributed source coding (DSC) is a new coding paradigm that relies on the coding of two or more dependent random sequences in an independent way. Distributed coding exploits the source statistics at the decoder, enabling a lower complexity encoder and a more complex decoder. DSC is based on two important information theory results: the Slepian-Wolf theorem [3] and the Wyner-Ziv [4], [5] theorems. Distributed video coding (DVC) can fulfill the requirements of a low-power and low-complexity encoder, but with a high-complexity decoder. However, in wireless communications, it is important to have low power and low complexity in both the encoder and decoder, for those applications a transcoder becomes necessary. This transcoder can receive a sequence of DVC, transcode it to a particular standard, like MPEG-x or H.26x, and transmit it to a low complexity decoder terminal.

A review of DVC can be found in [6]. A pixel-domain encoding system was investigated in [7], [8], where it is assumed that certain regular spaced frames are known perfectly as side-information (SI) at the decoder, but not at the encoder. These frames are called key frames. At the encoder the key frames are encoded in intra-mode. At the decoder, the key frames are used as SI to decode the other frames using temporal information, as in "inter"-mode. The results have shown that it outperforms conventional intra-frame coding but it is significantly inferior than inter-frame coding. That work was extended to the transform-domain in [9], [10]. There, a blockwise DCT is also applied to the SI, and the encoding process can use motion compensation. The transform-domain Wyner-Ziv codec achieves better results that the pixel-domain codec. In both cases, a bank of turbo encoders and decoders were used to implement a Slepian-Wolf coder. Those works, as most of the DVC frameworks, use a feedback channel between the encoder and the decoder. The use of feedback channel requires that the decoder and the encoder should be working at the same time, which denies offline decoding of the sequence.

Our framework can be implemented as an optional coding mode in any existing video codec standard [1], [11], and works similar to DVC. We propose a framework with spatial scalability, that will generate SI using temporal information at the decoder. No error correction code will be applied, and no feedback channel is required. The encoding complexity is reduced due to lower resolution encoding. In related works of DVC [12], [13], spatial scalability is also exploited, and in [12], [14] no feedback channel is used. However, this work focus on a simple framework that can be easily implemented in any video coding standards, with minor modifications and reasonable results.

II. THE FRAMEWORK

The spatial scalable framework is based on complexity reduction applied only to non-reference frames. The reference frames (key frames) can be coded exactly as in a regular codec,

Bruno Macchiavello, Eduardo Peixoto and Ricardo L. de Queiroz¸ are with the Departamento de Engenharia Elétrica, Universidade de Brasília, Brasília-DF, Brasil, E-mails: bruno@image.unb.br, eduardo@image.unb.br, queiroz@ieee.org. This work was supported by a grant from HP Brasil

as $I-$, $P-$ or reference $B-$ frames. Then, there will not be any drifting error. The frame type of the key frames can be varied depending on the complexity reduction desired and the target rate-distortion. For instance, using only $I-$ reference frames, the encoder becomes less complex, because no motion estimation is applied to these frames. However, it increases the bitrate. Similarly, using $P-$ or reference $B-$ frames as key frames will yield better results in terms of rate-distortion but will increase the complexity.

The non-reference frames are first decimated. The decimation factor and number of non-reference frames between key frames can vary dynamically based on the complexity reduction required and the target quality. This frames can be coded as $I-$, $P-$ or non-reference $B-$ frames. Again, this can be set depending on the required complexity. However, in this case, even using motion estimation ($P-$ or non-reference $B-$ frames) the encoder will be less complex than a regular encoder, since the non-reference frames are at lower spatial resolution. Note that the reconstructed reference frames in the frame store also have to be decimated in order to use them as reference for the low resolution frames.

The coded key frames and low resolution non-reference frames form the base layer. The enhancement layer is formed by sending the difference between the DCT coefficients of the interpolated low resolution non-reference frame and those of the original non-reference frame. The computations required for decimation, interpolation and all the extra functions that are not present in the regular coder are not significant when compared with the computational effort of motion estimation in a full resolution frame. The encoder architecture is shown in Fig. 1

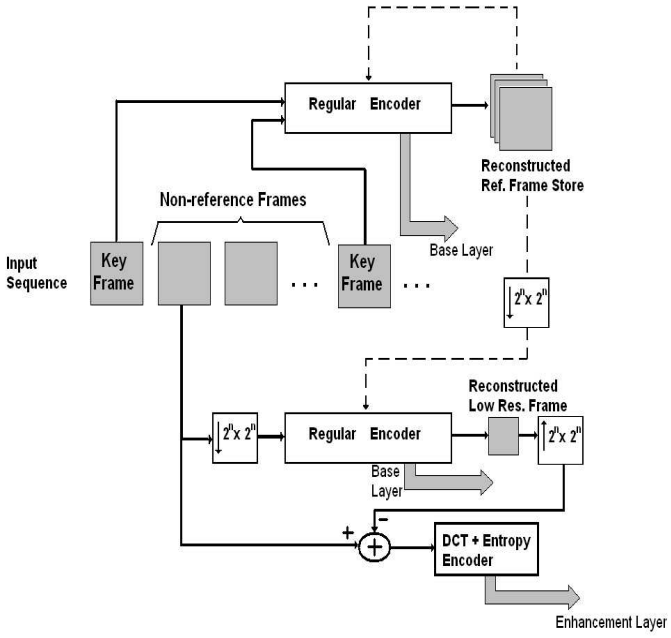

Fig. 1. Encoder of the Proposed Framework

At the decoder, if low decoding complexity is necessary we can use only the base layer, by interpolating the non-references frames coded at lower resolution. If the decoder does not have any complexity constraint the optional enhancement layer can be used. In order to correctly decode the enhancement layer, first we need to generate the SI using the decoded key frames. Then, the interpolated decoded frames of the non-references frames, along with the enhancement layer are used to perform the reconstruction of the frame. The decoder architecture is presented in Fig. 2

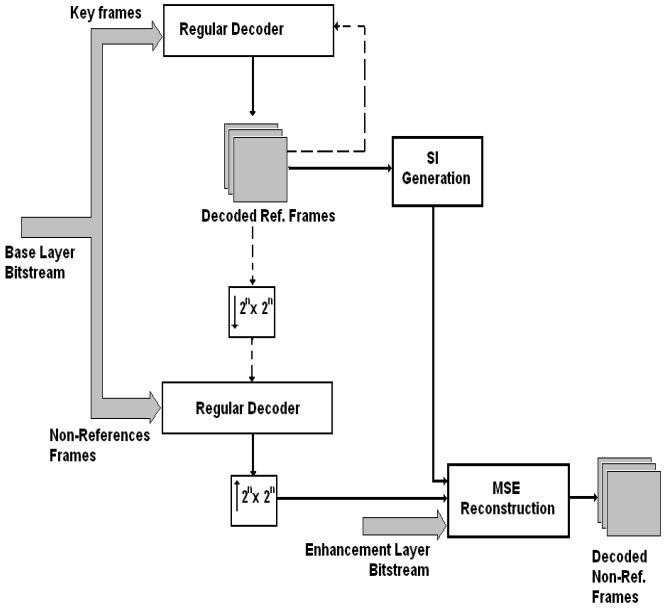

Fig. 2. Decoder of the Proposed Framework

III. Enhancement Layer

A. Coding

In the encoder, the enhancement layer is produced by first interpolating the reconstructed low resolution non-reference frame by the same factor used to decimate it, so the interpolated frame has the same resolution of the original frame. Then, a DCT is applied to the interpolated and original frames. Any transform size can be used. In this work, we use an 8×8 DCT. For each block, the difference between the 6 first DCT coefficients, in zig-zag scan order, of the original frame and the interpolated frame are computed. These coefficients are used to decode the enhancement layer. Our tests show that using more than the 6 first coefficients does not improve the final results, because the decimation of the non-reference frames degrades the high frequencies components of the DCT. The other DCT coefficients are taken from the SI generated frame, which is calculated using motion compensation interpolation on the full resolution frames. Thus, it is better, in terms of bitrate, to send only those 6 coefficients and set the others to zero. These residual coefficients are then quantized using the same quantization step of the regular coder. The distortion introduced to the residual frame will be the same one introduced to the regularly coded frames.

Note that DCT calculation and coefficient quantization are part of any video coding standard, and their functionality is already there at the regular codec.

The residual coefficients are sent to an entropy coder. No error correction code is applied and no feedback channel is required. The same entropy coder of the regular coder can be used to encode the residual coefficients.

B. Decoding

The base layer can be directly decoded using the regular decoder. For the low resolution non-reference frames, we only need to decimate frames in the frame store that are going to be used as reference, and to interpolate the result. This decoding process requires no motion estimation, or SI generation.

For decoding the optional enhancement layer, the first step is to generate the SI, as in any DVC framework. For SI generation, we use the previous and next key frames. Note that the key frames are at full resolution and the SI has better resolution that the decoded low resolution frames. Also, it is worth to mention that using more than one non-reference frame between the key frames makes the SI less accurate. This will reduce the final quality of the decoded frame. Nevertheless, as mentioned before, the more non-reference frames, the less complex the encoder. Hence, it is necessary to balance the complexity and quality requirements in order to select the number of non-reference frames.

After the SI generation, the residual coefficients are decoded and used to improve the interpolated decoded version of the frame. Since the residual coefficients represent the difference between the interpolated reconstructed frame and the original frame, they are added to the decoded interpolated version of the frame. A minimum mean squared error (MSE) reconstruction is applied using the improved interpolated frame and the SI.

C. Side Information generation

The SI generation is a crucial process in any DVC framework, as it is in our framework. An accurate SI generation allows us to obtain competitive results. In [15] there is a review on SI generation. Our method is based on the process proposed in [16], with some minor modifications.

For a current non-reference frame (F'_{2K}) between two key frames (F'_{2K-1}) and (F'_{2K+1}), the SI generation scheme uses the previous reconstructed key frame (F'_{2K-1}) as the reference and the next reconstructed key frame (F'_{2K+1}) as the source to calculate the forward motion vectors (MV_F). The, it uses the next reconstructed key frame (F'_{2K+1}) as the reference and the previous reconstructed key frame (F'_{2K-1}) as the source to calculate the backward motion vectors (MV_B). It then uses $\frac{MV_F}{2}$ on the previous reconstructed key frame (F'_{2K-1}) to generate frame P_F, and uses $\frac{MV_B}{2}$ on the next reconstructed key frame (F'_{2K+1}) to generate frame P_B. The final side estimation Y is considered as the weighted mean between P_F and P_B. A block size of 16×16 is used and the window search area is limited because longer motion vectors generate incorrect predictions when halved.

The final estimated frame is then calculated by a weighted arithmetic mean of the two compensated frames. If only one non-reference frame is used between two key frames, then the estimated frame will be the simple arithmetic mean. If more than one non-reference frame is used between two key frames, then weights will be applied to each set of motion vectors to calculate the compensated frames. The weights are inversely proportional to the distance between the current frame and the key frame.

D. MSE Reconstruction

The MSE Reconstruction is performed to each coefficient in the DCT domain. For further details on the reconstruction process, please refer to [17], [18]. Here, we will give a brief explanation, focusing on how the MSE reconstruction is performed in our work.

If Q denotes the quantized coefficients of the non-reference frame, Y represents the coefficients of the estimated SI frame, and X represents the coefficients of the original frame, then the final estimation $\hat{X}r$ is given as:

$$\hat{X}r_{YQ}(y,q) = E\{X|Y=y, Q=q\} \quad (1)$$

$$= \frac{\int_{xl(q)}^{xh((q)} x f_{X|Y}(x,y)dx}{\int_{xl(q)}^{xh(q)} f_{X|Y}(x,y)dx} \quad (2)$$

where $xl(q)$ and $xh(q)$ denote the high and low limits of the quantization bin represented by q. The variable Y is available at the decoder and, thus, its probability density function (PDF) can be computed. However, X is not available at the decoder. In order to model $f_{X|Y}(x,y)$, we start by defining a new random variable $Z = X - Y$. The random variable Z represents the noise between X and Y. As it can be seen in Fig. 3, the Laplacian residual model is well suited to represent the PDF of the variable Z for both the DC and AC bands.

$f_Z(z)$ is estimated as a Laplacian distribution, and since $f_Y(y)$ is available at the decoder we can obtain $f_{ZY}(z,y)$. Furthermore, the signal X can be described as $X = Y + Z$, so that we can deduce that $f_{XY}(x,y)$ can be simplify to:

$$f_{XY}(x,y) = f_{ZY}(x-y,y) \quad (3)$$

Then $f_{X|Y}(x,y)$ can be calculated as:

$$f_{X|Y}(x,y) = \frac{f_{XY}(x,y)}{f_Y(y)} \quad (4)$$

For each DCT coefficient, its reconstructed version is given by (1), where $f_{X|Y}(x,y)$ is given by (4) and (3).

IV. Results

The proposed framework was implemented using the H.263+ standard. We compared the results of the regular H.263+ codec working in $IIII...$ mode, and in $IPIPI...$ mode against our low complexity framework working in $IpIpIp...$ mode, where p represents the downsampled P frames. For testing, we use $p-$ frames at quarter resolution. We tested our framework using only the base layer, and using the enhancement layer along with the SI generation and MSE reconstruction. The test sequences were used in CIF format (352×288 pixels). Both codecs, regular and

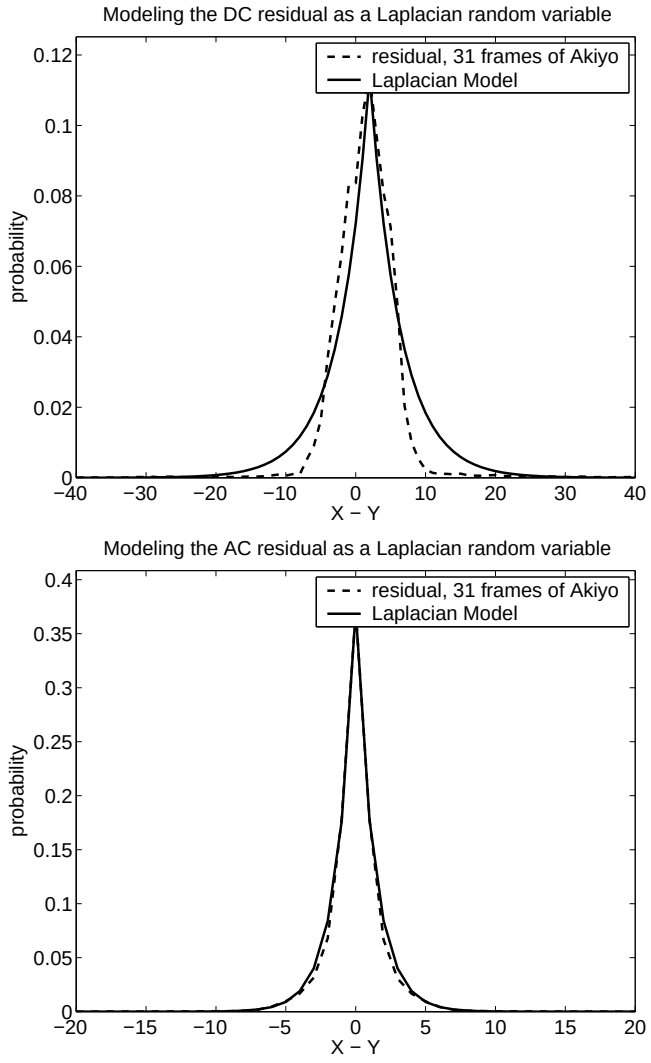

Fig. 3. DCT domain residual modeled as a Laplacian random variable. (a) DC Coefficients (b) AC Coefficients

proposed, have the same configurations which represent the base H.263+ standard without advanced or extended options, like advanced intracoding or the arithmetic entropy coder. In terms of complexity it is well known that the $P-$ frames are significantly more complex than the $I-$ frames due to motion estimation. If we ignore the additional complexity of the residual layer, which only involves the computation of the residual coefficients and the entropy coder, the encoding complexity of a low-resolution $p-$ frame will be roughly $1/4$ the encoding complexity of a full resolution $P-$ frame.

In Fig. 4, the results for the "Akiyo", and "Silent" sequences are shown. It can be seen that the proposed framework using the base and enhancement layers outperforms the regular codec in $III...$ form, and performs about $0.2 - 2.5$ dB lower than the regular codec at $IPIP...$ mode. This shows that our framework achieves low-complexity coding with high-complexity decoding yielding reasonable results. This is specially true for low rates (under 800 kbps) where the gap between the regular $IPIP$ codec and the low-complexity framework decreases. If low decoding complexity is also required, the base layer can be used. Note that using enhancement layer along with the MSE reconstruction and SI generation significantly improves the results when compared against using only the base layer.

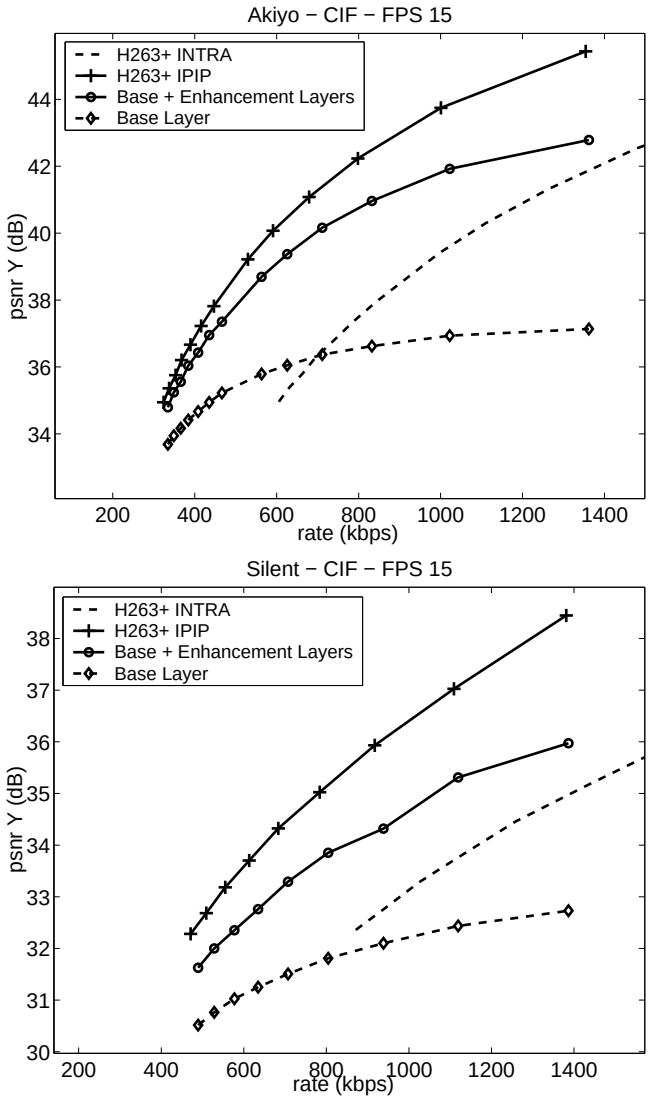

Fig. 4. Results on H.263+, comparing the regular Intra mode H.263+, the regular $IPIP$ mode, the base layer of the proposed framework in $IpIp$ mode and the base layer and enhancement layer with MSE reconstruction: (a) Akiyo CIF sequence (b) Silent CIF sequence

V. Conclusions

The basic framework of a simple distributed coding mode based on spatial scalability applied to H.263+ is presented. It can be incorporated into any other codec, notably H.264/AVC or MPEG-4. Future work would involve improving the side information generation process that can potentially yield better results. For instance a better side information generation process can be achieved using knowledge of the interpolated non-reference frames to perform the motion estimation and compensation.

REFERENCES

[1] 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*, v. 13, p. 560-576, 2003.

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

[3] J. Slepian and J. Wolf, "Noiseless coding of correlated information sources," *IEEE Transactions on Information Theory*, v. 19, p. 471-480, 1973.

[4] A. Wyner and J. Ziv, "The rate-distortion function for source coding with side information at the decoder," *IEEE Transactions on Information Theory*, v. 2, p. 1-10, 1976.

[5] A. Wyner and J. Ziv "Recent results in the Shannon theory ," *IEEE Transactions on Information Theory*, v. 1, p. 2-10, 1974.

[6] B. Girod, A.M. Aaron, S. Rane and D. Rebollo-Monedero "Distributed video coding," *Proceedings of the IEEE*, v. 1, p. 71-83, 2005.

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

[8] A.M. Aaron, S. Rane, R. Zhang and B. Girod "Wyner-Ziv coding for video: applications to compression and error resilience," *Proceedings of IEEE Data Compression Conference*, p. 93-102, 2003.

[9] D. Rebollo-Monedero, A. Aron and B. Girod "Transform for high-rate distributed source coding," *Asilomar Conference of Signals, Systems and Computers*, 2003.

[10] A. Aaron, R. Zhang and B. Girod "Trasnform-domain Wyner-Ziv codec for video," *Proc. SPIE Visual Communications and Image Processing*, 2004.

[11] G. Cote, B. Erol, M. Gallant and F. Kossentini "H.263+: video coding at low bit-rates," *IEEE Trans. Circuits Syst. Video Technology*, v. 8, p. 849-866, 1998.

[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*, v. 6508, p. 65081Y1-65081Y12, 2007.

[13] M. Wu, G. Hua and C. W. Chen 'Syndrome-based lightweight video coding for mobile wireless application," *Proc. Int. Conf. on Multimedia and Expo*, p. 2013-2016, 2006.

[14] L. W. Kang and C. S. Lu "Wyner-Ziv video coding with coding mode-aided motion compensation," *Proc. IEEE International Conf on Image Processing*, p. 237-240, 2006.

[15] Z. Li and L. Liu and E. J. Delp "Rate distortion analysis of motion side estimation in Wyner-Ziv video coding ," *IEEE Transactions on Image Processing*, v. 16, p. 98-113, 2007.

[16] Z. Li and E. J. Delp "Wyner-Ziv video side estimator: conventional motion search methods revisited ," *Proc. IEEE International Conf on Image Processing*, v. 1, p. 825-828, 2005.

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

[18] A. M. Aaron and R. Zhang and B. Girod "Wyner-Ziv coding of motion video," *Conference Record of the Thirty-Sixth Asilomar Conference on Signals, Systems and Computers*, v. 1, p. 240-244, 2002.