TRANSFORM DOMAIN SEMI-SUPER RESOLUTION

Edson M. Hung[1], *Diogo C. Garcia*[2] *and Ricardo L. de Queiroz*[3]

Universidade de Brasilia
[1] Faculdade do Gama - Engenharia Eletrônica, [1,2] Departamento de Engenharia Elétrica,
[3] Departamento de Ciência da Computação. E-mail: {mintsu, diogo}@image.unb.br, queiroz@ieee.org

ABSTRACT

This paper presents a transform-based approach to semi-super resolution. The idea of semi-super resolution is to enhance low-resolution frames from a video sequence encoded with different resolutions among frames. The proposed framework uses a DCT-based down-sampling method at the encoder process. At the decoder, DCT-based up-sampling plus high-frequency information from adjacent frames in full resolution are used to enhance interpolated low-resolution frames. The results show an increase of the overall quality without using any extra information from the encoder. The proposed technique can also outperfom previous semi-super resolution methods operating in pixel domain.

Index Terms— Mixed resolution, semi-super resolution, example-based super resolution.

1. INTRODUCTION

The spatial scalability framework for distributed video coding was introduced in [1], in which the spatial resolution of the frames varies over time (as shown in Figure 1). In this sense, some of the frames are encoded at normal resolution (the key frames), and others, at a smaller resolution (the non-key frames), thus reducing the overall complexity at the encoder. The residuals among the spatially reduced frames and their original versions are also Wyner-Ziv coded [1] [2], generating an enhancement layer to be used at the decoder. However, this layer is only useful if a better estimate of the non-key frames can be generated at the decoder.

This introduced the problem of obtaining a higher resolution of certain frames in the sequence, based on other frames at full resolution. This process was called *semi super-resolution* (SSR) [1], to differentiate from the well-known process of *super-resolution* (SR), where a higher-resolution frame is obtained only from low-resolution frames [3]. Here we may refer to the SSR process simply as SR.

The proposed SSR solutions were mainly inspired by example-based SR approaches, where a database of full-resolution frames are used to restore the high-frequency of blocks in interpolated frames [4]. Since the key frames in the video sequence are at full resolution, they constitute the database of training images.

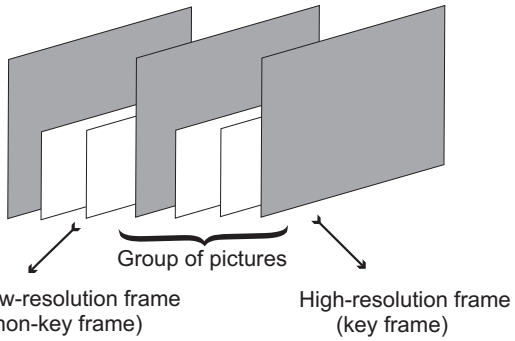


Fig. 1. Example of a mixed resolution video sequence.

Mukherjee [1] substituted blocks in the non-key frames with blocks from key-frames, based on block-based motion estimation between the interpolated non-key frames and the low-pass versions of the key frames. The sum of absolute differences (SAD) among these low-pass blocks was used to determine if the block substitution was reliable, when compared to a threshold. Brandi *et.al.* compared the high-frequency energy of blocks in non-key frames with certain thresholds, to evaluate whether these blocks needed to be super-resolved, and also added high-frequency content from key-frames in a block basis, instead of the replacing the whole block [5]. Oliveira *et.al.* introduced weights to ponder the high-pass contributions from blocks in different key frames, based on the sum of squared differences (SSD) between the low-pass versions of those blocks in key frames, and the interpolated block in the non-key frame [6].

Coding sequences at different resolutions allow for other possible applications, such as mixed-resolution stereoscopic and multiview coding. Studies on the subjective perception of depth in stereoscopic images [7] show that when one of the views is low-pass filtered, the binocular image's quality is not affected [8]. However, users tend to get eye fatigue after long periods of exposure to these mixed-resolution stereoscopic images. Furthermore, in the case of multiview sequences, the system may be used in free-viewpoint television, which is monoscopic by nature, and the quality reduction from full to low resolution views is very perceptible. Garcia *et.al.* used the SSR framework to enhance the low-resolution frames, for

color-plus-depth multiview imagery [9]. Rather than using block-based motion estimation (or in this case, disparity estimation), they relied on the depth maps to draw correspondences among views.

In this paper, we improve the previous SSR frameworks by altering the down/up-sampling methods, as well as the high-frequency motion compensation process. While the previous frameworks used bilinear and Lanczos interpolation methods, we use a block-based DCT interpolation method to better preserve the low-frequencies of each block, and further explore this fact to improve the way the high-frequency is added to the low-resolution blocks.

2. PROPOSED METHOD

In this section, we present the interpolation methods used for down-/up-sampling, as well as the proposed SSR framework. For down-sampling, we divide the frame into 8×8 blocks, and apply the 8×8 type-II DCT transform for each block. Next, the coefficients of the highest frequencies in the block are cut, so as to fit the remaining low-frequency coefficients in a 4×4 matrix. The 4×4 inverse type-II DCT is applied to this smaller block, and all pixels in the output are multiplied by a scaling factor $S_d = 0.5$, due to the different sizes between the forward and inverse DCT transforms. Figure 2 illustrates the method, which is used at the semi-super resolution process. It can also be used to generate the mixed resolution frames at the encoding process.

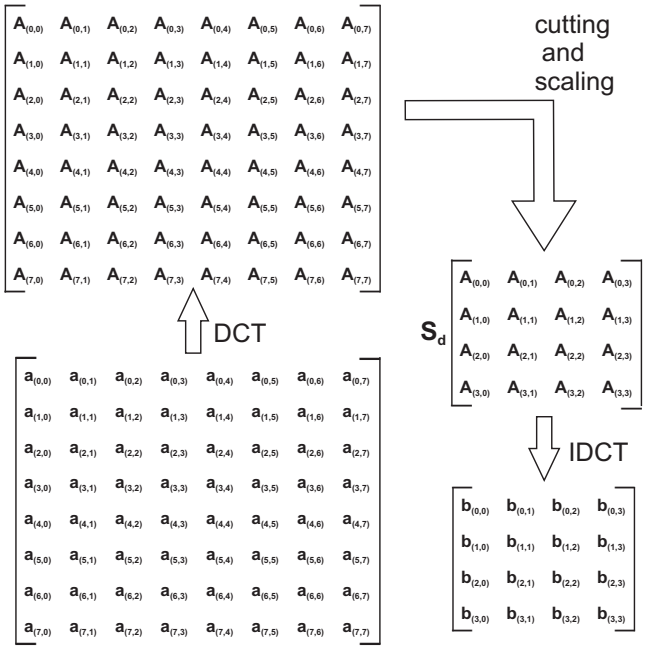


Fig. 2. Example of a 8×8 to 4×4 DCT sub-sampling.

Figure 3 shows the DCT up-sampling method, which is applied to each 4×4 block in the input frame. First, we apply the 4×4 type-II DCT, and multiply the coefficients by a scaling factor $S_u = 2$. Then, we consider the coefficients of this 4×4 block to be the low-frequency coefficients $C_{(i,j)}$ of an 8×8 block in the DCT domain. The high-frequency coefficients, $E_{(i,j)}$, are considered to be null in our up-sampling method. Then, the inverse 8×8 type-II DCT is applied to this block, generating an up-sampled block from the input 4×4 block.

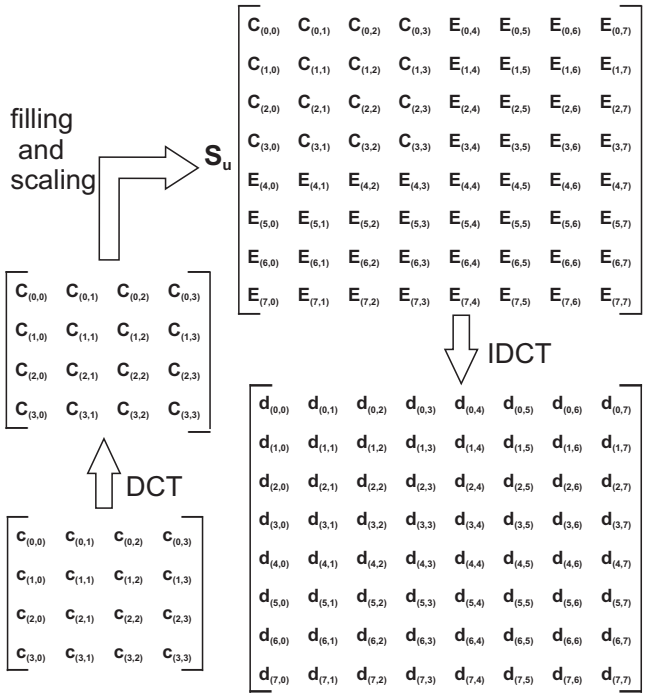


Fig. 3. Example of a 4×4 to 8×8 DCT up-sampling.

Despite of a higher PSNR objective quality compared to spatial domain interpolation methods, such as bilinear and bicubic interpolations, the subjective quality of this DCT up-sampling method is still not satisfactory, because zero appending in high frequency components usually leads to blocking artifacts [10]. Assuming we are working in a mixed-resolution framework, we propose to substitute the high-frequency sub-bands ($E_{(i,j)}$) with high-frequency coefficients from blocks in adjacent key frames, which have not been down-sampled. Motion estimation between key and non-key frames is performed to find the best matches. In this sense, we make the assumption that if the low-frequency coefficients of the blocks are similar, their high-frequency coefficients may also match.

Figure 4 shows a diagram of the proposed method, which works as follows: for every non-key frame $\widetilde{F}_n$, we are given one or more key frames, F_i, which are usually the closest key-frames to $\widetilde{F}_n$. We generate an up-sampled version of $\widetilde{F}_n$, defined as $\widetilde{F}_n^L$, and a low-pass version of each key frame F_i, defined as F_i^L. The latter frame is obtained through down-sampling followed by up-sampling of F_i, by the same factor employed on $\widetilde{F}_n$.

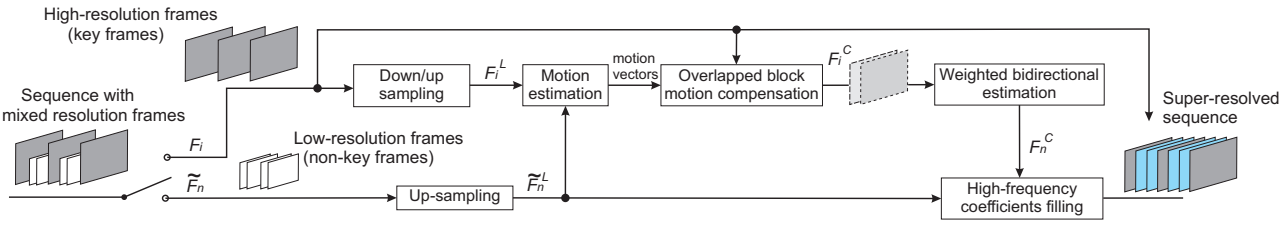


Fig. 4. The proposed architecture for the proposed transform domain semi-super-resolution.

Block-based motion estimation is then performed between $\widetilde{F}_n^L$ and F_i^L, with variable block sizes [6], generating a series of motion vectors. These are used for overlapped block motion compensation [11] from all key frames F_i to the non-key frame $\widetilde{F}_n$, generating frames F_i^C. This reduces blocking artifacts in the SSR process [11].

Having created the motion compensated frames, they are all fused together into one new frame, F_n^C, using weighted estimation. Each block from F_n^C is composed of a weighted sum of the blocks from each motion compensated frame F_i^C. In order to do so, we generate a low-pass version of each F_i^C, defined as F_i^{CL}. Then, we calculate on a block basis the SSDs between $\widetilde{F}_n^L$ and F_i^{CL}, defined as D_i. For every position (u, v) inside the current block, we have:

$$F_n^C(u,v) = \frac{\sum F_i^C(u,v)/D_i}{\sum 1/D_i}, \qquad (1)$$

where the summation is defined over all key frames F_i being used for the process.

F_n^C represents a version of $\widetilde{F}_n^L$ generated from nearby key frames F_i. As such, we assume that F_n^C has richer high-frequency information than $\widetilde{F}_n^L$. Assuming that $\widetilde{F}_n$ was generated using the previously described DCT-based down-sampling method, we apply the 8×8 DCT over all blocks in $\widetilde{F}_n^L$ and F_n^C, and substitute the null high-frequency coefficients $E_{(i,j)}$ from $\widetilde{F}_n^L$ by the same coefficients $E_{(i,j)}$ from F_n^C. Mantaining the low-resolution coefficients intact, except for the encoding quantization process.

3. EXPERIMENTAL RESULTS

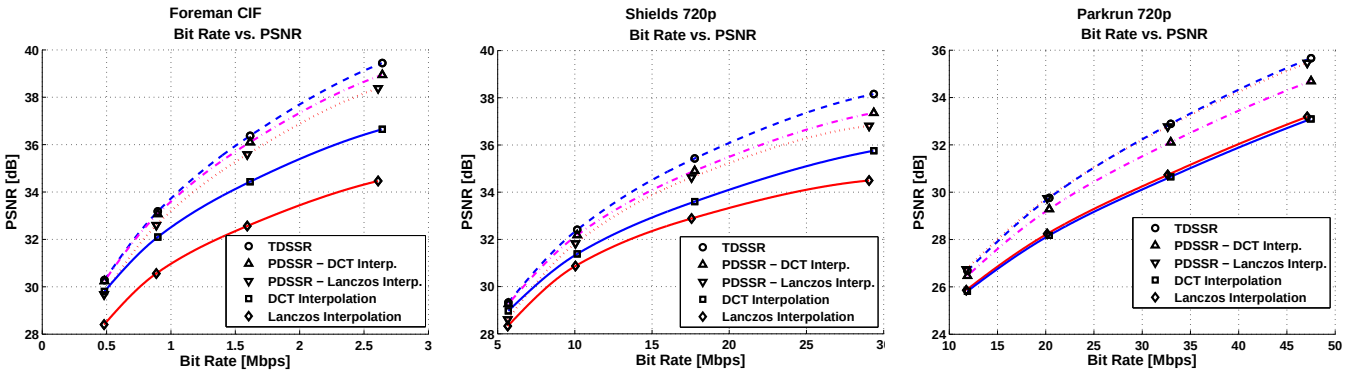
In order to evaluate the performance of the proposed technique, we processed video sequences with 300 frames, at CIF resolution (352×288 pixels) and high definition resolution (1280×720 pixels) encoded with H.264 (JM 15.1) and GOP length of 2. At the semi-super resolution we use a motion estimation window of 32×32 pixels and two key-frame references (the closest forward and backward key frames). The set of quantization parameters (QP) was $\{22, 27, 32, 37\}$ for PSNR curves comparisons [12]. The tests were performed to compare the transform-domain semi-super resolution (TDSSR) and pixel-domain semi-super resolution (PDSSR). Both techniques use the DCT interpolation in the encoding and decoding processes. Other simulations were made in order to compare the proposed technique to another in the literature [6]. In this case, the PDSSR was applied using the Lanczos interpolation as basis.

The curves in Figure 5 show good performance of the proposed method, specially for high-quality video compression. In Table 1 we compute the PSNR difference of the TDSSR and PDSSR-DCT. For example, for sequence Foreman the PSNR difference is 0.1943 dB. At the high-resolution sequences, the differences are greater than 0.35 dB. However, when we compare to the PDSSR using the Lanczos interpolation, the objective results get much better in most cases.

Table 1. PSNR comparison [12] of the proposed and previous techniques.

Sequence	Method vs. Lanczos Interpolation	PSNR difference
Foreman (CIF)	TDSSR-DCT	3.1168 dB
	PDSSR-DCT	2.9225 dB
	PDSSR-Lanczos	2.4748 dB
	DCT Interpolation	1.6530 dB
Shields (720p)	TDSSR-DCT	2.0111 dB
	PDSSR-DCT	1.6419 dB
	PDSSR-Lanczos	1.3002 dB
	DCT Interpolation	0.6432 dB
Parkrun (720p)	TDSSR-DCT	1.6502 dB
	PDSSR-DCT	1.0741 dB
	PDSSR-Lanczos	1.6551 dB
	DCT Interpolation	-0.1209 dB

Subjective results of the sequence Foreman, encoded with H.264 using $QP = 22$, are shown in Figure 6. In 6(a) an encoded high resolution frame is shown as an ideal result, in 6(b) we show the result of a Lanczos interpolation from an encoded low resolution frame and in 6(c) the PDSSR of the previous case. Finally, we show in 6(d) a DCT interpolation from an encoded low resolution frame and 6(e) the proposed TDSSR. Here we can observe that, despite of the objective gains, the proposed TDSSR method can sujbectively outperform the previous SSR and also the interpolated case in regions with textures. However, ringing artifacts from the DCT interpolated method were not completely removed after the TDSSR method.



(a) (b) (c)

Fig. 5. Comparison among the proposed semi-super resolution method and previous ones applied at sequences: (a) Foreman, (b) Shields and (c) Parkrun.

(a) (b) (c) (d) (e)

Fig. 6. Details of the Foreman sequence:(a) high resolution frame encoded with H.264, (b) Lanczos interpolation from an H.264 encoded low resolution frame, (c) PDSSR of the previous case, (d) DCT interpolation from an H.264 encoded low resolution frame and (e) TDSSR of the previous case.

4. CONCLUSION

In this paper, we propose a transform-based method to obtain a super-resolution of certain frames in the sequence, based on other frames with higher resolution. Here, we introduce a new way to add high-frequency coefficients into low-resolution blocks. Different from the previous works, we maintain intact the low-resolution frame coefficients, except for the video encoding process. In general, the results show a better performance compared to the previous semi-super resolution techniques.

5. REFERENCES

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

[2] B. Girod, A. Aaron, S. Rane and D. Rebollo-Monedero, "Distributed video coding," *Proceedings of the IEEE*, Special Issue on Video Coding and Delivery, vol. 93, no. 1, pp. 71-83, January 2005.

[3] S. C. Park, M. K. Park, and M. G. Kang, "Super-resolution image reconstruction: a technical overview," *IEEE Signal Processing Magazine*, 20(3):2136, May 2003.

[4] 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.

[5] F. Brandi, R. de Queiroz, D. Mukherjee, "Super resolution of video using key frames,", *Proc. IEEE Intl. Symp. on Circuits and Systems*, Seattle, USA, May 2008.

[6] K. F. Oliveira, F. Brandi, E. M. Hung, R. L. de Queiroz and D. Mukherjee, "Bipredictive video super–resolution using key–frames," *Proc. IS&T/SPIE Symp. on Electronic Imaging, Visual Information Processing and Communication*, San Jose, CA, USA, SPIE Vol. 7543, Jan. 2010.

[7] B. Julesz, "Foundations of cyclopean perception," University of Chicago Press, 1971.

[8] W. Tam, "Image and depth quality of asymmetrically coded stereoscopic video for 3D-TV," JVT-W094, San Jose, CA, April 2007.

[9] D. C. Garcia, C. C. Dorea, and R. L. de Queiroz, "Super-resolution for multiview images using depth information," *Proc. IEEE Intl. Conf. on Image Processing*, ICIP, Hong Kong, China, Sep. 2010.

[10] Z. Wu, H. Yu and C. W. Chen, "A New Hybrid DCT-Wiener-Based Interpolation Scheme for Video Intra Frame Up-Sampling," *IEEE Signal Processing Letters*, vol. 17, issue 10, pp. 827-830, Oct. 2010.

[11] E. M. Hung and R. L. de Queiroz, "Blocking-effect reduction in a reversed-complexity video codec based on a mixed-quality framework," *Intl. Telec. Symp.*, ITS, Manaus, Brazil, Sep. 2010.

[12] G. Bjontegaard, "Calculation of Average PSNR Differences between RD curves", *ITU-T SC16/Q6, Doc. VCEG-M33*, 13th VCEG Meeting, Austin, Texas, USA, April 2001.