TRANSFORM-DOMAIN SUPER-RESOLUTION FOR MULTIVIEW IMAGES USING DEPTH INFORMATION

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

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

ABSTRACT

Mixed resolution formats have been employed in video encoding complexity reduction as well as data compression of stereoscopic video. High resolution frames within such formats may also be used as a means of enhancing lower resolution images. In this paper we present a super-resolution method for use in a mixed resolution, multiview video plus depth setup. High resolution views are initially projected onto the view point of low resolution images with the aid of available depth maps. The method introduces the use of transform-domain techniques for up-sampling the low resolution images and for appending high frequency content from projected views. The DCT is used for the necessary frequency decompositions. Results show gains over previous work for several test sequences and affirm the aptitude of transform-domain approaches for super-resolution within mixed resolution formats.

1. INTRODUCTION

Super-resolution (SR) methods attempt to achieve high resolution enlargements of an image. By aggregating information from multiple correlated images, these methods can overcome the inherent limitations of single frame up-sampling and interpolation. Typically, SR exploits subpixel precision shifts among low resolution images to form a high resolution image. However, in other applications, SR methods may resort to available high resolution images in order to estimate missing high resolution detail [1, 2]. For instance, Example-based SR [1] uses a training set of high resolution images to restore the high frequencies missing from patches in zoomed images. Likewise, SR has been used in the context of mixed resolution video to recover resolution of down-sampled frames by borrowing high frequency content from neighboring high resolution (key) frames [2]. The SR method presented herein is similar to those of the latter examples in that it is based on the use of available high resolution images to enhance low resolution ones.

Besides their usage in video encoding complexity reduction [2, 3], mixed resolution formats have a history in data size reduction for stereoscopic video [4, 5]. In this case, instead of temporally interspersing low and high resolution frames, a low resolution view is reserved for the left eye, for example, while high resolution is presented to the right eye. Although SR approaches have been used in temporally mixed resolution video to recover down-sampled frames, alleviating flickering during playback, mixed resolution (or asymmetric) stereoscopic video is generally displayed as is. This is justified by psychovisual studies [6, 7] which indicate that the overall sharpness and depth perception of the stereoscopic image is determined by the high resolution channel. Nevertheless, in more general multiview setups, mixed resolution alone may not be directly applicable and indeed few proposals consider so. Quality differences among views could be detrimental to some of the applications envisioned for mutliview systems. For example, due to its monoscopic nature, navigation between views in free-viewpoint video could suffer from perceivable quality differences between the mixed resolution views.

To overcome these limitations a SR method for use with mixed resolution, multiview video plus depth formats has been proposed [8]. The format is illustrated in Figure 1 and consists of multiple video sequences from different view points at alternating resolutions and corresponding depth maps. The depth maps [9] have been maintained at full resolution as they can be efficiently coded and represent a small portion of overall data size. Knowledge of depth maps is used to establish correspondences among views.

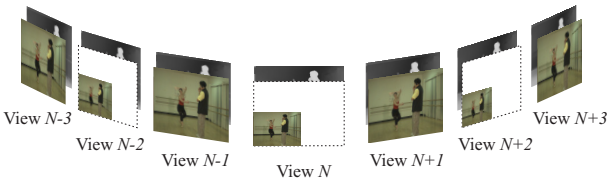


Figure 1: A mixed resolution, multiview video plus depth format.

The SR method of [8] applies linear interpolation filters in two operations: (i) to up-sample the low resolution image and (ii) to isolate high frequency content in neighboring high resolution views. Both operations were implemented in the spatial-domain, however, both also present competitive counterparts in the transform-domain. Previous research [10] has reported objective quality gains of up-sampling operations in the transform-domain over simpler, fixed-parameter interpolation methods such as bilinear. Visual quality has been further improved by combining a low-frequency preserving DCT-based up-sampling technique with a Wiener-based estimate of missing high frequency coefficients [11]. Besides promising results in up-sampling, the transform-domain is a natural medium for frequency decomposition.

This work has been partly supported by grants from FINATEC, FINEP and CNPq of the Brazilian Government.

In this paper we present a SR method for the mixed resolution, multiview video plus depth format. High frequency content from neighboring full resolution views is used to enhance low resolution images. We introduce transform-domain operations in up-sampling and isolating high frequency content. The proposed method preserves the low frequency DCT coefficients of the low resolution image and complements these with the high frequency DCT coefficients from projected high resolution views. View projection is accomplished with knowledge of depth information.

2. PROPOSED METHOD

An overview of the proposed transform-domain SR method is presented in the block diagram of Figure 2. High resolution images available within the mixed resolution, multiview video plus depth format are initially projected onto the view point corresponding to the low resolution image as indicated by the View Projection block. The Up-sampling block is responsible for enlarging the low resolution image to dimensions compatible with those of high resolution. Up-sampling techniques are investigated in both the spatial- and the transform-domains. Finally, the DCT-based SR block determines high frequency content from the projected view and super-resolves the low resolution image. The following subsections describe these steps in more detail.

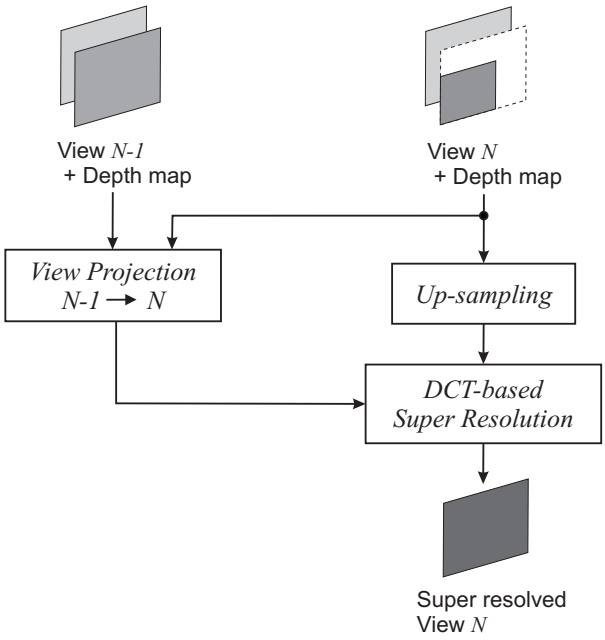


Figure 2: Block diagram of the proposed transform-domain SR method.

2.1 View Projection

Accurate view projection is an essential component of the SR proposal. Projection is a depth-based rendering technique which takes as inputs a full resolution image V_{N-1}, depth maps D_{N-1} and D_N and synthesizes an image corresponding to the Nth view $\widehat{V_N}$. Knowledge of intrinsic camera parameters $\mathbf{A}$, rotation matrix $\mathbf{R}$, translation vector $\mathbf{t}$ are first used to project pixel location $(\widehat{u},\widehat{v})$ of camera N onto 3D world coordinates (x,y,z) [12]:

$$(x,y,z)^T = \mathbf{R}_N {\mathbf{A}_N}^{-1}(\widehat{u},\widehat{v},1)^T D_N(\widehat{u},\widehat{v}) + \mathbf{t}_N. \quad (1)$$

Next, world coordinates are re-projected onto camera $N-1$ yielding (u,v):

$$(u*w, v*w, w)^T = \mathbf{A}_{N-1}{\mathbf{R}_{N-1}}^{-1}[(x,y,z)^T - \mathbf{t}_{N-1}]. \quad (2)$$

Generally, not all pixel correspondences between views are possible or available. A consistency check is employed to identify correspondence errors. Coordinates (u,v) are rounded to their nearest integers and projected back to camera N. If the Euclidean distance between the position resulting from this last projection and the original $(\widehat{u},\widehat{v})$ coordinates is below a specified threshold (typcally 1.0) the correspondence is accepted, otherwise it is rejected. View projection is completed by filling $(\widehat{u},\widehat{v})$ with the bilinearly interpolated sample from accepted correspondence position (u,v) in V_{N-1}. Positions with rejected correspondences are left as holes as exemplified in Figure 3. Note that for the purpose of DCT-based SR, projection holes may be filled with values from the up-sampled low resolution image. These holes cannot be super-resolved.

(a) (b)

Figure 3: (a) The *Ballet* sequence (view 0, frame 0) and (b) its projection onto view 1 (holes shown in white).

2.2 DCT-based Super-Resolution

The proposed transform-domain SR method is based on a DCT decomposition which determines the high frequency coefficients to be added to the low resolution image. Next we review the two basic transform-domain operations used in SR: DCT-based down-sampling and DCT-based up-sampling.

Let $\mathbf{b}$ be an $(m \times m)$ pixel block from an image. Equation (3) expresses the DCT coefficients of $\mathbf{b}$ as a partitioned matrix. $\mathbf{B}_{00}$ is an $(n \times n)$ sub-matrix containing the low frequency coefficients. $\mathbf{B}_{01}$, $\mathbf{B}_{10}$ and $\mathbf{B}_{11}$ are sub-matrices of sizes $(m-n \times n)$, $(n \times m-n)$ and $(m-n \times m-n)$, respectively, containing the high frequency coefficients

$$DCT\{\mathbf{b}\} = \left[\begin{array}{c:c} \mathbf{B}_{00} & \mathbf{B}_{01} \\ \hdashline \mathbf{B}_{10} & \mathbf{B}_{11} \end{array}\right]. \quad (3)$$

Down-sampling of the image block $\mathbf{b}$ is achieved by calculating the inverse DCT of the low frequency sub-matrix $\mathbf{B}_{00}$ while discarding the higher frequency components [10]. Due to the differences in direct and inverse DCT sizes, the sub-matrix $\mathbf{B}_{00}$ must be multiplied by a downsizing factor $s_{dsz} = n/m$ prior to computing the down-sampled image block of dimensions $(n \times n)$:

$$\mathbf{b}_{dsp} = IDCT\left\{s_{dsz}\left[\mathbf{B}_{00}\right]\right\}. \quad (4)$$

Up-sampling of an image block can be obtained by appending zeros to the missing high frequency coefficient slots and computing the inverse DCT [10]. For example, the up-sampling of $\mathbf{b}_{dsp}$ is obtained by assuming $\mathbf{B}_{01}$, $\mathbf{B}_{10}$ and $\mathbf{B}_{11}$ as zero sub-matrices and forming the $(m \times m)$ image block as

$$\mathbf{b}_{usp} = IDCT\left\{\left[\begin{array}{c:c} \mathbf{B}_{00} & \mathbf{0} \\ \hdashline \mathbf{0} & \mathbf{0} \end{array}\right]\right\}. \quad (5)$$

DCT-based up-sampling of an entire image is achieved by dividing the image into $(n \times n)$ blocks, determining the DCT coefficients of each block and appending zeros to the inverse DCT of each block as described in Equation (5).

DCT-based SR aims to outperform up-sampling methods by using estimated high frequency coefficients from neighboring views instead of assuming these as zero sub-matrices. As described in subsection 2.1, a high resolution view has been consistently projected onto the view point of the low resolution image. The projected view is the source of high frequency coefficients used to replace missing coefficients in the up-sampled low resolution image. For each block $\widehat{\mathbf{b}}$ within the projected view, DCT coefficients are formed as

$$DCT\left\{\widehat{\mathbf{b}}\right\} = \left[\begin{array}{c:c} \widehat{\mathbf{B}}_{00} & \widehat{\mathbf{B}}_{01} \\ \hdashline \widehat{\mathbf{B}}_{10} & \widehat{\mathbf{B}}_{11} \end{array}\right]. \quad (6)$$

High frequency sub-matrices $\widehat{\mathbf{B}}_{01}$, $\widehat{\mathbf{B}}_{10}$ and $\widehat{\mathbf{B}}_{\mathbf{11}}$ are used to complete the missing high frequency slots of the co-located block $\mathbf{b}_{usp}$ of the up-sampled low resolution image. Thus, the super-resolved image block $\mathbf{b}_{SR}$ is given by

$$\mathbf{b}_{SR} = IDCT\left\{\left[\begin{array}{c:c} \mathbf{B}_{00} & \widehat{\mathbf{B}}_{01} \\ \hdashline \widehat{\mathbf{B}}_{10} & \widehat{\mathbf{B}}_{11} \end{array}\right]\right\} \quad (7)$$

where $\mathbf{B}_{00}$ are the low frequency DCT components reminiscent of the up-sampled low resolution image block as given by Equation (5). By super-resolving each up-sampled low resolution block, an SR image of full dimensions and containg high frequency content from neighboring high resolution views is formed.

3. EXPERIMENTAL RESULTS

Performance of the proposed transform-domain SR method was evaluated on publicly available synthetic [13] and real [9] data sets. Due to the lack of high frequency content in some of the original sequences, the real data sets *Ballet* and *Breakdancers* were resized to 512×384 pixels and 256×192 pixels, respectively, prior to the evaluations. The multiview images, provided with depth or disparity maps, were down-sampled accordingly to form a mixed resolution format as illustrated in Figure 1. Each low resolution image is super-resolved with the immediately adjoining high resolution view. DCT-based operations employ the type-II DCT transforms in all cases with block sizes of 8×8 and down-sampling sizes of 4×4 ($m = 8$ and $n = 4$) resulting in a down-sampling ratio of 2.

The first tests compare up-sampling results between a high-performance linear interpolation filter (Lanczos kernel) used in [8] and DCT-based up-sampling as discussed in subsection 2.2. Up-sampling is the step prior to SR as indicated in Figure 2. Table 1 shows moderate underperformance, on average -0.32 dB, of the DCT approach when compared to Lanczos up-sampling. Note, however, that this DCT-based up-sampling scheme simply appends zeros as an estimate of its missing high frequency components.

Sequence	linear filter up-sampling	DCT-based up-sampling	PSNR gain
Ballet	34.01 dB	33.71 dB	-0.30 dB
Breakdancers	35.47 dB	34.95 dB	-0.52 dB
Barn1	27.76 dB	27.49 dB	-0.27 dB
Barn2	31.06 dB	30.74 dB	-0.32 dB
Bull	32.46 dB	32.23 dB	-0.23 dB
Map	28.00 dB	27.56 dB	-0.44 dB
Poster	26.46 dB	26.06 dB	-0.40 dB
Sawtooth	28.32 dB	27.93 dB	-0.39 dB
Venus	28.63 dB	28.40 dB	-0.23 dB

Table 1: PSNR comparison between linear interpolation filter (Lanczos kernel) and DCT-based up-sampling.

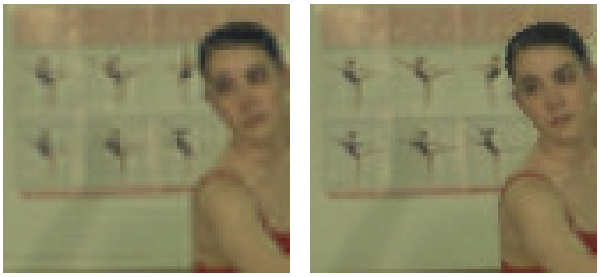
The second set of tests compares the proposed transform-domain SR method using DCT to a spatial-domain SR method similar to that of [8]. Spatial-domain SR employs linear interpolation filters (Lanczos kernel) both for up-sampling and for high frequency extraction as described in [8]. However, for the purpose of comparison, view projection of spatial-domain SR is made identical to that of the proposed SR method. Table 2 shows that the proposed SR method outperforms the spatial-domain SR with linear filtering for all but one of the tests sequences. Average PSNR gains over all sequences is 0.16 dB and as high as 0.5 dB for the *Bull* sequence. Observe that in spite of using a less sophisticated up-sampling technique whose performance is worse than up-sampling with linear filtering (see Table 1), DCT-based SR is still capable of outperforming the spatial-domain SR method.

Sequence	linear filter SR [8]	proposed DCT-based SR	PSNR gain
Ballet	36.18 dB	**36.31 dB**	0.15 dB
Breakdancers	38.69 dB	**38.84 dB**	0.15 dB
Barn1	35.83 dB	**36.22 dB**	0.39 dB
Barn2	38.40 dB	**38.50 dB**	0.10 dB
Bull	37.96 dB	**38.46 dB**	0.50 dB
Map	31.20 dB	**31.24 dB**	0.04 dB
Poster	33.93 dB	**34.09 dB**	0.16 dB
Sawtooth	**33.72 dB**	33.32 dB	-0.40 dB
Venus	35.61 dB	**35.99 dB**	0.38 dB

Table 2: PSNR comparison between spatial-domain SR with linear filtering [8] and proposed transform-domain SR with DCT.

A subjective evaluation of the proposed SR method is possible with the images of Figure 4. The DCT-based

SR is compared to DCT-based up-sampling for the *Ballet* sequence. High frequency details have been inserted by the SR method, sharpening contours on the ballerina's face as well as some of the background texture in the SR images. These enhancements reflect the gains achieved in terms of PSNR (2.60 dB). PSNR gains achieved by the proposed SR method over up-sampling alone can be computed by contrasting the second columns of Tables 1 and 2. For proper visual comparisons results are best viewed on a screen.



(a) (b)

Figure 4: Detail crops of the *Ballet* sequence, view 2: (a) DCT-based up-sampled image (33.71 dB) and (b) DCT-based SR image (36.31 dB).

For the synthetic sequence *Barn1* the objective PSNR difference between the proposed SR method and DCT-based up-sampling is 8.73 dB. Figure 5 permits a subjective comparison between results. Observe that the insertion of high-frequency components by the SR method results in sharper and more detailed image content.

(a) (b)

Figure 5: *Barn1* sequence, view 1: (a) DCT-based up-sampled image (27.49 dB) and (b) DCT-based SR image (36.22 dB).

4. CONCLUSION

This paper presents a novel transform-domain SR method for use in a mixed resolution, multiview video plus depth setup. A DCT-based technique is introduced for up-sampling of the low resolution image. The proposed SR method proceeds by projecting a high resolution view onto the view point of the low resolution image. High frequency DCT coefficients from the projected view are then used to complete the missing high frequency components of the low resolution view. The super-resolved image achieves significant PSNR and subjective quality gains over interpolated images.

Future work involves exploring the full potential of transform-domain methods for SR within mixed resolution formats. Among the envisioned efforts are the investigation of noise removal from projected views in the transform-domain, artifact reduction (e.g., DCT blocking) and image sharpening of the super-resolved image through manipulation of the high frequency DCT components.

REFERENCES

[1] 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.

[2] F. Brandi, R. de Queiroz, D. Mukherjee, "Super-resolution of video using key frames and motion estimation,", *Proc. IEEE Intl. Conf. on Image Processing*, San Diego, USA, Oct. 2007.

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

[4] H. Brust, A. Smolic, K. Mueller, G. Tech and T. Wiegand, "Mixed resolution coding of stereoscopic video for mobile devices,", *Proc. 3DTV Conference*, Potsdam, Germany, May 2009.

[5] M. G. Perkins, "Data compression of stereopairs," *IEEE Trans. on Communications*, Vol. 40, pp. 686-696, 1992.

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

[7] L. Stelmach, W. J. Tam, D. Meegan and A. Vincent, "Stereo Image quality: effects of mixed spatio-tamporal resolution," *IEEE Trans. Circuits and Systems for Video Tech.*, vol. 10, no. 2, pp. 188-193, Mar. 2000.

[8] 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.

[9] C. Zitnick, S. Kang, M. Uyttendaele, S. Winder and R. Szeliski, "High-quality video view interpolation using a layered representation," *ACM SIGGRAPH*, Los Angeles, USA, Aug. 2004.

[10] R. Dugad and N. Ahuja, "A fast scheme for image size change in the compressed domain," *IEEE Trans. Circuits and Systems for Video Tech.*, vol. 11, no. 4, pp. 461-474, Apr. 2001.

[11] 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.

[12] P. Kauff, N. Atzpadin, C. Fehn, M. Muller, O. Schreer, A. Smolic and R. Tanger, "Depth map creation and image-based rendering for advanced 3DTV services providing interoperability and scalability," *Image Communication*, pp. 217-234,2007.

[13] http://vision.middlebury.edu/stereo/data/