SALIENCY-COGNIZANT ROBUST VIEW SYNTHESIS IN FREE VIEWPOINT VIDEO STREAMING

Bruno Macchiavello [*], *Camilo Dorea* [*], *Edson M. Hung* [*], *Gene Cheung* [#], *Wai-tian Tan* [o]

[*] Universidade de Brasilia, Brazil, [#] National Institute of Informatics, Japan,
[o] Hewlett-Packard Laboratories, USA.

ABSTRACT

In free viewpoint video, texture and depth maps from two camera-captured viewpoints are transmitted, so that at receiver, a novel virtual view chosen by the client can be synthesized via depth-image-based rendering (DIBR). When irrecoverable packet losses occur during transmission—typically affecting less important spatial regions in the video given unequal error protection (UEP) is deployed—appropriate error concealment strategies must be used at decoder to minimize resulting visual degradation in the synthesized view. Towards this goal, we propose a new optimization framework based on visual saliency to combine two different concealment techniques. First, given a pixel in the virtual view is typically constructed as a convex combination of corresponding pixels in the left and right captured views, weighted pixel blending (WPB) readjusts the weights in the linear sum to reflect the expected error in code blocks that contain the corresponding pixels. Second, exemplar-based patch matching (EPM) finds the most similar patches in the known spatial region to complete missing pixels in the unknown region. To choose between candidates constructed using the two techniques when filling a given pixel patch in the synthesized view, we first compute a weighted sum of expected error and visual saliency for each candidate patch. The candidate with the smaller sum (one with small expected error and visual saliency, so that even if errors do occur, they do not stand out visually) is selected for pixel completion. Experimental results show that our scheme can outperform the use of co-located blocks from a previous frame by up to 0.7dB in PSNR and improve subjective visual quality.

Index Terms— free viewpoint video, visual saliency, error concealment

1. INTRODUCTION

Free viewpoint video [1] can dramatically enhance a viewer's depth perception in the observed 3D scene by enabling *motion parallax* [2], where the viewpoint from which to render an image on a 2D display is continuously adjusted according to the current head position of the observer. The representation and coding of free viewpoint video have been studied extensively in recent years [3, 4]. In particular, the *depth-image-based rendering* (DIBR) approach [5], where a novel virtual view image is synthesized using texture and depth maps of two nearby captured views, has been widely adopted. In this paper, we study the complementary problem of error concealment for images synthesized using DIBR.

Error concealment in free viewpoint video is more challenging than single-view video, due to the complex relationship between synthesized view quality and information loss in texture and depth maps from two captured views. At the same time, it holds promise to yield good concealment results due to inherent data redundancy across views. Depending on factors such as whether the images are jointly compressed across views via inter-view prediction or independently compressed, loss patterns observed in the various texture and depth maps can be uncorrelated or highly correlated.

In the case where losses are uncorrelated across views, given a pixel in the virtual view is typically constructed as a convex combination of corresponding pixels in the left and right captured views [6], a concealment strategy called *weighted pixel blending* (WPB) first estimates the distortion on a per-block basis for the two received views, and then readjusts the weights in the linear sum to reflect the expected error in blocks that contain the corresponding pixels. It has been shown [7] experimentally that WPB performs well when packet losses across views are independent.

In the case when losses are highly correlated across views, e.g., due to inter-view prediction of video data [3], a reasonable concealment strategy is *exemplar-based patch matching* (EPM) [8], which finds the most similar patches in the known spatial region of the synthesized image to complete missing pixels in the unknown region. Given these two complementary concealment strategies, in this paper we propose a unifying framework based on visual saliency to combine them for the general scenario when both correlated and uncorrelated losses are possible. The use of visual saliency is motivated by the observation that an inappropriate concealment choice tends to be visually annoying and therefore of high visual saliency [9]. Further, since unequal error protection (UEP) is often employed in video streaming, when irrecoverable packet losses occur during transmission, they typically affect less important (less visually salient) spatial regions in the video. Thus, to choose between candidates constructed using the two techniques when filling a given pixel patch in the synthesized view, for each candidate patch we first compute a weighted sum of expected error and visual saliency. The candidate with the smaller sum is selected for pixel completion. Experimental results show that our scheme can outperform the use of co-located blocks from a previous frame by up to 0.7dB in PSNR.

The rest of this paper is organized as follows. We discuss related works in Section 2, followed by an overview of our system and problem formulation in Section 3 and 4, respectively. We then present experimental results in Section 5 followed by a conclusion.

2. RELATED WORK

As the compression technologies for multiview video [10, 11] and free viewpoint video [12, 4] mature, research in 3D video communication has been gradually shifting to the streaming and distribution aspects [13, 14]. For example, toward the goal of loss resiliency, [15, 16] proposed to exploit the flexibility in reference picture selection (RPS) [17] in H.264 video coding standard [18] to encode a visually important block in a current texture or depth frame us-

This work was partially supported by CNPq grant 310375/2011-8.

ing a reference frame further in the past as predictor, so that the probability of correct decoding can be improved. The error concealment problem—how to best recover lost information in streaming video given packet losses have already occurred [19]—has never been studied in the context of free viewpoint video, however. To the best of the authors' knowledge, we are the first to address this important problem in a formal manner.

Visual saliency for video—the likelihood that an observer will look at a particular spatial area during video playback—has been studied extensively in the visual science literature [20]. One obvious practical use of visual saliency for video streaming is unequal bit allocation, so that more visually salient spatial regions are encoded with more bits (higher quality) than other regions [21]. Recently visual saliency has also been used for error concealment of single-view video [9], so that even if an error-concealed block is wrong, the error does not stick out visually. Our current work can be considered as a non-trivial application of the low-saliency prior [9] to error concealment in free viewpoint video.

3. SYSTEM OVERVIEW

We assume a single-sender / single-receiver streaming architecture, where sender transmits texture and depth maps captured from left (view 0) and right view (view 1), so that the receiver can synthesize a novel image from a freely chosen intermediate virtual view v, $0 \leq v \leq 1$, via DIBR [5]. Because two reference images from relatively close viewpoints are used for view synthesis, disoccluded regions—spatial locations in virtual view that are not visible from either the left or the right view—tend to be small, and low-complexity algorithms such as [22] can perform hole-filling satisfactorily.

As done in our previous work [16], we assume also a low end-to-end delay application requirement (e.g., video conferencing). Thus, a retransmitted packet will inevitably be late for its playback deadline, and forward error correction (FEC) is a more suitable packet loss protection strategy than automatic repeat request (ARQ). Like the 2D video streaming system in [9], we assume sender performs unequal error protection (UEP), so that the more important (more visually salient) spatial regions of 3D video (both texture and depth maps) are protected with stronger FEC code. Hence, a typical irrecoverable packet loss event will affect only the low-saliency region. Important spatial regions can be automatically detected using saliency analysis such as [20]. We will assume that the important regions take up no more than 20 to 25% of each image.

4. PROBLEM FORMULATION

We first overview our optimization methodology. For each pixel patch in the virtual view, we can construct two possible candidates for robust view synthesis. The first candidate is constructed via *weighted pixel blending* (WPB), where corresponding pixels in the left and right captured views are weighted appropriately during pixel blending, taking into consideration of the expected errors in the transmitted texture and depth maps. The second candidate is constructed via *exemplar-based patch matching* (EPM), where the most similar patch in the already synthesized portion of the virtual view is identified and copied. Between these two candidates, the patch with the smaller weighted sum of expected error plus visual saliency value is selected as the winner for pixel completion. We next discuss the construction of the two candidates in order.

4.1. Weighted Pixel Blending

In a majority of the cases, a synthesized pixel (i, j) in the virtual view v has two corresponding pixels, (i, j^0) and(i, j^1), in the left (view 0) and right (view 1) captured views. Pixel (i, j) in virtual view v is assigned intensity $S_t^v(i, j)$ that is a convex combination of the two corresponding pixels [6]:

$$S_t^v(i,j) = (1-v)\, X_t^0(i,j^0) + v\, X_t^1(i,j^1) \tag{1}$$

where the weights $(1 - v)$ and v reflect the distance between the virtual view v to each of the two reference views.

In general, corresponding pixels $X_t^0(i, j^0)$ and $X_t^1(i, j^1)$ can be corrupted by packet losses. Let $e_t^0(i, j^0)$ and $\epsilon_t^0(i, j^0)$ be the estimated errors at pixel location (i, j^0) of the respective texture and depth maps of view 0. $e_t^0(i, j^0)$ and $\epsilon_t^0(i, j^0)$ can be computed at decoder recursively depending on observable packet loss events [16], as reviewed in Appendix. Given $e_t^0(i, j^0)$ and $\epsilon_t^0(i, j^0)$, we can estimate the distortion of left pixel $d_t^0(i, j^0)$ as follows:

$$d_t^0(i,j^0) = \max_{l=j^0-\epsilon_t^0,\ldots,j^0+\epsilon_t^0} e_t^0(i,l) + \left|X_t^0(i,l) - X_t^0(i,j^0)\right| \tag{2}$$

In words, (2) states that the estimated distortion $d_t^0(i, j^0)$ is texture error $e_t^0(i, l)$ *plus* the largest difference between texture pixel $X_t^0(i, j^0)$ and horizontally shifted pixel $X_t^0(i, l)$, where the range of l depends on depth error $\epsilon_t^0(i, j^0)$. This means that given e_t^0 and ϵ_t^0, distortion d_t^0 is small only if location (i, j^0) points to the interior of an object and the object's surface texture is smooth (slow varying), which agrees with intuition. Right estimated distortion $d_t^1(i, j^1)$ can be computed similarly.

Weights w_t^0 and w_t^1 are thus selected based on estimated distortion $d_t^0(i, j^0)$ and $d_t^1(i, j^1)$ of the corresponding left and right pixels. Specifically, $S_t^v(i, j)$ is now the adaptively blended pixel:

$$S_t^v(i,j) = w_t^0\, X_t^0(i,j^0) + w_t^1\, X_t^1(i,j^1) \tag{3}$$

where weights are computed based on *reliability* of the two corresponding pixels, r^0 and r^1:

$$\begin{aligned} w_t^0 &= \frac{r^0(1-v)}{r^0(1-v) + r^1 v} \\ w_t^1 &= \frac{r^1 v}{r^0(1-v) + r^1 v} \end{aligned} \tag{4}$$

Weights defined using (4) have the following properties: i) default to $(1 - v)$ and v in (1) when reliability $r^0 = r^1$; ii) converge to 1 and 0 when $r^0 \gg r^1$; and iii) sum to 1. Reliability r^v is computed using distortion d_t^v as follows:

$$r^v = \frac{1}{d_t^v + \bar{d}} \tag{5}$$

with parameter $\bar{d} > 0$, so that r^v is well defined even if $d_t^v = 0$.

4.2. Exemplar-Based Patch Matching

We follow similar notation as [8]. Denote the recovered region (source region) and the missing region (target region) of the image $\mathcal{I}$ by Φ and $\mathcal{I} - \phi = \Omega$, respectively. Further, denote the contour of the target region by $\delta\Omega$. For a square patch $\Psi_{\mathbf{p}}$ with center $\mathbf{p}$ on the contour $\delta\Omega$, we want to identify a similar patch $\Psi_{\mathbf{q}}$ with center $\mathbf{q}$ in Φ, so that the filled-in pixels in $\Psi_{\mathbf{p}}$ matches the pixels in $\Psi_{\mathbf{q}}$.

The order in which patches in the target region Ω is filled is very important. As done in [8], we define a priority $P(\mathbf{p})$ as follows:

$$P(\mathbf{p}) = C(\mathbf{p})D(\mathbf{p}) \tag{6}$$

where $C(\mathbf{p})$ and $D(\mathbf{p})$ are the confidence and data terms, respectively. $C(\mathbf{p})$ and $D(\mathbf{p})$ are computed as:

$$C(\mathbf{p}) = \frac{\sum_{\mathbf{q}\in\Psi_p\cap\Phi} C(\mathbf{q})}{|\Psi_\mathbf{p}|} \qquad D(\mathbf{p}) = \frac{|\nabla I_\mathbf{p}^\perp \cdot \mathbf{n}_\mathbf{p}|}{\alpha} \tag{7}$$

where $|\Psi_\mathbf{p}|$ is the area of $\Psi_\mathbf{p}$. $C(\mathbf{p})$ is initialized to be 1 if $\mathbf{p} \in \Phi$, and 0 otherwise. $\mathbf{n}_\mathbf{p}$ is the normal to contour $\delta\Omega$, and $\nabla I_\mathbf{p}^\perp$ is the isophote (direction and intensity) at point $\mathbf{p}$. $C(\mathbf{p})$ essentially computes the amount of confidence in pixels in patch $\Psi_\mathbf{p}$ that have already been recovered. Data term $D(\mathbf{p})$ is a "function of the strength of isophotes (linear structures) hitting the front $\delta\Omega$ at each iteration" [8], and is used to encourage propagation of linear structures. See [8] for details.

4.3. Low-Saliency Prior for Candidate Selection

Having described the two methods of constructing candidate pixels in the synthesized image, we now describe a procedure to order missing pixel patches to fill, and to select one of two candidates for each patch. We first perform regular DIBR-based pixel blending, as described in (1), for pixels with corresponding pixels having zero estimated distortions d_t^0 and d_t^1. Because of the UEP applied to different spatial regions as described in Section 3, it means that the regions with high saliency plus a subset of regions with low saliency will be synthesized.

The synthesized region and missing region will be the source region $\mathcal{I}$ and target region Ω as described in Section 4.2. We compute the priority term for each possible size 16×16 square patch $\Psi_\mathbf{p}$ with center $\mathbf{p}$ on the contour $\delta\Omega$, selecting one with the highest priority for filling first. We compute the two candidate solutions for filling patch $\Psi_\mathbf{p}$, $\Psi_\mathbf{p}^1$ and $\Psi_\mathbf{p}^2$ as described in Section 4.1 and Section 4.2. For each candidate, we compute the expected distortion $D()$ for each solution. For WPB, it is the average distortion of the synthesized pixels in the patch, where distortion for each synthesized pixel is the weighted sum of estimated distortions d_t^0 and d_t^1 of the two corresponding pixels in left and right views. For EPM, it is the average distortion of synthesized pixels in the copied patch $\Psi_\mathbf{q}$.

Having computed the expected distortions for the candidates, we select the candidate patch with the smaller weighted sum of distortion and visual saliency:

$$\min_{\mathbf{g}\in\{1,2\}} D(\Psi_\mathbf{p}^g) + \lambda Z(\Psi_\mathbf{p}^g) \tag{8}$$

where $Z(\Psi_\mathbf{p} s^g)$ is the computed saliency of candidate patch $\Psi_\mathbf{p}^g$, and $\lambda > 0$ is a pre-set parameter.

5. EXPERIMENTATION

The proposed framework is evaluated through synthesized image quality, in terms of both PSNR comparisons and visual inspection.

We assume that packet losses manifest themselves as losses of isolated macroblocks (MB) rather than contiguous regions due to application of Flexible Macroblock Ordering for error resilience. We further assume that losses occur only in the low-saliency parts of both the texture and depth maps of both the left and right views, due to application of unequal error protection. Low-saliency region of each frame consisted of the macroblocks with the lowest 75% of average pixel saliency. Models to estimate texture and depth errors due to error propagation for differentially coded video, such as from H.264 [18], are given in the Appendix.

A sample saliency image is shown in Fig. 1, as calculated by [9]. In the correlated loss experiment, MBs losses occurred simultaneously in the same corresponding locations in both left and right views. In the uncorrelated loss experiment, MB losses were randomly and independently inflicted upon low saliency regions at different time instants in each view. Simulations include losses of 5%, 10%, 20% and 30% of frame MBs for both texture and depth maps.

Fig. 1. Saliency values computed for 16×16 blocks for `Kendo` frame 11, view 1. Higher saliency values are shown as brighter.

In all, four error concealment strategies were considered. The baseline for comparison consists of copying co-located blocks from the previous frame when a MB loss is encountered. The second and third strategies use only EPM and only WPB, respectively, to fill missing patches of the synthesized image as described in Section 4.2 and 4.1. EPM uses a search window of size 64×64. Our proposed strategy selects between EPM and WPB candidate patches with the aid of a low saliency prior, using an empirically selected λ in (8).

Experiments were conducted for multiview sequences `Kendo` (1024×768 pixels) and `Akko & Kayo` (640×480 pixels) using the MPEG View Synthesis Reference Software (VSRS v3.5) [23]. For `Kendo`, views 1 and 3 were corrupted through MB losses and used for synthesis of the central view 2. For `Akko & Kayo`, views 47 and 49 were used to synthesize view 48. In both cases, the losses were introduced in 10 randomly selected frames among the first 20 frames, and the original central view was used as ground truth.

In Table 1 and 2 the average PSNR for each error concealment strategy for uncorrelated and correlated losses, respectively, are presented at various loss rates. The average PSNR is computed considering only the frames that are affected by losses. Note that in an error-free scenario, the average PSNR for is 35.83dB for `Kendo` and 29.02dB for `Akko & Kayo`. As observed in Table 1, WPB can effectively conceal uncorrelated errors, providing performance superior to co-located copying and EPM. This is due to the fact that when losses are uncorrelated, a projected pixel from one of the views will generally be un-corrupted and can be weighted heavily by WPB. When errors are correlated, WPB has performance similar to co-located block copying as seen in Table 2. In this case, EPM can outperform co-located and WPB, since it avoids the use of simultaneously loss-corrupted corresponding pixels. The proposed scheme successfully combines WPB and EPM. Note that since the errors are inflicted only in low-saliency regions due to UEP, the low-saliency

prior aids in selecting the appropriate candidate, outperforming both WPB and EPM, even when the MB loss pattern favors one of these strategies. Our scheme can outperform the use of co-located blocks by as much as 0.76 dB, the use of EPM by as much as 0.48 dB and WPB by 0.73 dB, in specific scenarios.

Table 1. Uncorrelated losses

	Kendo			
	5%	**10%**	**20%**	**30%**
Co-located	35.48 dB	35.33 dB	35.03 dB	34.72 dB
EPM	35.64 dB	35.55 dB	35.26 dB	35.09 dB
WPB	35.72 dB	35.62 dB	35.49 dB	35.35 dB
Proposed	35.74 dB	35.64 dB	35.68 dB	35.48 dB
	Akko and Kayo			
	5%	**10%**	**20%**	**30%**
Co-located	28.70 dB	28.54 dB	28.12 dB	27.64 dB
EPM	28.88 dB	28.64 dB	28.25 dB	27.74 dB
WPB	28.87 dB	28.75 dB	28.35 dB	28.00 dB
Proposed	28.88 dB	28.78 dB	28.46 dB	28.22 dB

Table 2. Correlated losses

	Kendo			
	5%	**10%**	**20%**	**30%**
Co-located	35.50 dB	35.30 dB	35.07 dB	34.66 dB
EPM	35.68 dB	35.62 dB	35.48 dB	35.29 dB
WPB	35.57 dB	35.37 dB	35.12 dB	34.69 dB
Proposed	35.69 dB	35.70 dB	35.62 dB	35.42 dB
	Akko and Kayo			
	5%	**10%**	**20%**	**30%**
Co-located	28.70 dB	28.36 dB	27.92 dB	27.61 dB
EPM	28.77 dB	28.56 dB	28.30 dB	27.91 dB
WPB	28.69 dB	28.33 dB	27.99 dB	27.59 dB
Proposed	28.81 dB	28.59 dB	28.41 dB	27.99 dB

For subjective comparisons, synthesized frames from our proposed scheme are compared to those from the co-located concealment strategy. For Kendo, shown in Fig. 2, uncorrelated losses at 20% rate were used. For Akko & Kayo, shown in Fig. 3, correlated losses at 30% rate were chosen. The figures present detail crops of areas with more significant quality differences. For both sequences, the proposed scheme presents noticeable superior visual quality. Artifacts have been eliminated from the background and swordsman in Kendo, while errors around the neck, arm, ball and others of Akko & Kayo have been significantly reduced.

6. CONCLUSION

Error concealment for free viewpoint video to minimize adverse effects to synthesized view quality due to irrecoverable packet losses is a challenging problem. In this paper, we first estimate the error for a texture or depth pixel in differentially coded view using a set of recursive equations. We use visual saliency to combine two previous error concealment techniques, weighted pixel blending (WPB) and exemplar-based patch matching (EPM), into one framework, so that the concealment candidate with the smaller weighted sum of expected error and visual saliency can be chosen to complete a loss-corrupted patch. Experiments show that using our framework, one can achieve better concealment quality than WPB or EPM alone, in terms of PSNR. Tests also show gains in subjective visual quality. Formal subjective evaluations are subject of future studies.

(a) (b)

Fig. 2. Frame 11 from Kendo (a) using co-located blocks, PSNR 34.70 dB and (b) using proposed scheme, PSNR 35.39 dB.

(a) (b)

Fig. 3. Frame 3 from Akko & Kayo (a) using co-located blocks, PSNR 26.87 dB and (b) using proposed scheme, PSNR 27.48 dB.

7. APPENDIX: COMPUTING RECURSIVE ERROR AT DECODER

We first estimate texture error $e_{t,m}$ for a MB m in frame t containing corresponding texture pixel (i, j^0) in view 0. Depending on whether packet containing MB m is correctly received or not, we write $e_{t,m}$ as:

$$e_{t,m} = \begin{cases} e^+_{t,m} & \text{if MB } m \text{ is correctly received} \\ e^-_{t,m} & \text{o.w.} \end{cases} \quad (9)$$

Suppose the packet containing MB m is correctly received. If MB m is intra-coded, then $e^+_{t,m} = 0$. Otherwise, given motion vector (MV) $v_{t,m}$ points to an off-grid reference block in a previous frame $\tau_{t,m}$, $e^+_{t,m}$ is computed as a weighted sum of errors of a neighborhood of on-grid blocks in reference frame $\tau_{t,m}$:

$$e^+_{t,m}(\tau_{t,m}, v_{t,m}) = \begin{cases} 0 & \text{if MB } m \text{ is intra} \\ \sum_{k \in v_{t,m}} \alpha_k e_{\tau_{t,m},k} & \text{o.w.} \end{cases} \quad (10)$$

If packet containing MB m is not correctly received, then the error $e^-_{t,m}$ is approximated as the error of block in the same location in previous frame $t-1$, *plus* an estimate of the frame-to-frame block difference δ:

$$e^-_{t,m} = e_{t-1,m} + \delta \quad (11)$$

See [16] for further details.

8. REFERENCES

[1] M. Tanimoto, M. P. Tehrani, T. Fujii, and T. Yendo, "Free-viewpoint TV," in *IEEE Signal Processing Magazine*, January 2011, vol. 28, no.1.

[2] C. Zhang, Z. Yin, and D. Florencio, "Improving depth perception with motion parallax and its application in teleconferencing," in *IEEE International Workshop on Multimedia Signal Processing*, Rio de Jeneiro, Brazil, October 2009.

[3] P. Merkle, C. Bartnik, K. Muller, D. Marpe, and T. Weigand, "3D video: Depth coding based on inter-component prediction of block partitions," in *2010 Picture Coding Symposium*, Krakow, Poland, May 2012.

[4] I. Daribo, D. Florencio, and G. Cheung, "Arbitrarily shaped sub-block motion prediction in texture map compression using depth information," in *2010 Picture Coding Symposium*, Krakow, Poland, May 2012.

[5] W. Mark, L. McMillan, and G. Bishop, "Post-rendering 3D warping," in *Symposium on Interactive 3D Graphics*, New York, NY, April 1997.

[6] D. Tian, P.-L. Lai, P. Lopez, and C. Gomila, "View synthesis techniques for 3D video," in *Applications of Digital Image Processing XXXII, Proceedings of the SPIE*, 2009, vol. 7443 (2009), pp. 74430T–74430T–11.

[7] B. Macchiavello, C. Dorea, M. Hung, G. Cheung, and W. t. Tan, "Loss-resilient texture & depth map coding in multiview video conferencing," in *e-print arXiv:1305.5464*, May 2013.

[8] A. Criminisi, P. Perez, and K. Toyama, "Region filling and object removal by exemplar-based image inpainting," in *IEEE Transactions on Image Processing*, September 2004, vol. 13, no.9, pp. 1–13.

[9] H. Hadizadeh, I. Bajic, and G. Cheung, "Saliency-cognizant error concealment in loss-corrupted streaming video," in *IEEE International Conference on Multimedia and Expo*, Mulbourne, Australia, July 2012.

[10] M. Flierl, A. Mavlankar, and B. Girod, "Motion and disparity compensated coding for multiview video," in *IEEE Transactions on Circuits and Systems for Video Technology*, November 2007, vol. 17, no.11, pp. 1474–1484.

[11] P. Merkle, A. Smolic, K. Muller, and T. Wiegand, "Efficient prediction structures for multiview video coding," in *IEEE Transactions on Circuits and Systems for Video Technology*, November 2007, vol. 17, no.11, pp. 1461–1473.

[12] M. Maitre, Y. Shinagawa, and M.N. Do, "Wavelet-based joint estimation and encoding of depth-image-based representations for free-viewpoint rendering," in *IEEE Transactions on Image Processing*, June 2008, vol. 17, no.6, pp. 946–957.

[13] E. Kurutepe, M. R. Civanlar, and A. M. Tekalp, "Client-driven selective streaming of multiview video for interactive 3DTV," in *IEEE Transactions on Circuits and Systems for Video Technology*, November 2007, vol. 17, no.11, pp. 1558–1565.

[14] A. M. Tekalp, E. Kurutepe, and M. R. Civanlar, "3DTV over IP: End-to-end streaming of multiview video," in *IEEE Signal Processing Magazine*, November 2007.

[15] B. Macchiavello, C. Dorea, M. Hung, G. Cheung, and W. t. Tan, "Reference frame selection for loss-resilient depth map coding in multiview video conferencing," in *IS&T/SPIE Visual Information Processing and Communication Conference*, Burlingame, CA, January 2012.

[16] B. Macchiavello, C. Dorea, M. Hung, G. Cheung, and W. t. Tan, "Reference frame selection for loss-resilient texture & depth map coding in multiview video conferencing," in *IEEE International Conference on Image Processing*, Orlando, FL, September 2012.

[17] G. Cheung, W.-T. Tan, and C. Chan, "Reference frame optimization for multiple-path video streaming with complexity scaling," in *IEEE Transactions on Circuits and Systems for Video Technology*, June 2007, vol. 17, no.6, pp. 649–662.

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

[19] Y. Chen, Y. Hu, O. Au, H. Li, and C. W. Chen, "Video error concealment using spatio-temporal boundary matching and partial differential equation," in *IEEE Transactions on Multimedia*, January 2008, vol. 10, no.1, pp. 2–15.

[20] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, November 1998, vol. 20, no.11, pp. 1254–1259.

[21] Z. Cheng and c. Guillemot, "Perceptually-friendly H.264/AVC video coding based on foveated just-noticeable-distortion model," in *IEEE Transactions on Circuits and Systems for Video Technology*, June 2009, vol. 29, no.6.

[22] I. Daribo and B. Pesquet-Popescu, "Depth-aided image inpainting for novel view synthesis," in *IEEE Multimedia Signal Processing Workshop*, Saint-Malo, France, October 2010.

[23] M. Tanimoto, T. Fujii, K. Suzuki, N. Fukushima, and Y. Mori, "Reference softwares for depth estimation and view synthesis," in *ISO/IEC JTC1/SC29/WG11 MPEG2008/M15377*, Archamps, April 2008.