A Spatial-Focal Error Concealment Scheme for Corrupted Focal Stack Video

Kejun Wu[1], Yi Wang[1], Wenyang Liu[1], Kim-Hui Yap[1] and Lap-Pui Chau[2]

[1]School of Electrical and Electronic Engineering, Nanyang Technological University, 50 Nanyang Ave, Singapore

[2]Department of Electronic and Information Engineering, The Hong Kong Polytechnic University, Hung Hom, Hong Kong

{kejun.wu,ekhyap}@ntu.edu.sg,{wang1241,wenyang001}@e.ntu.edu.sg,lap-pui.chau@polyu.edu.hk

2023 Data Compression Conference (DCC) | 979-8-3503-4795-1/23/$31.00 ©2023 IEEE | DOI: 10.1109/DCC55655.2023.00017

Abstract

Focal stack image sequences can be regarded as successive frames of videos, which are densely captured by focusing on a stack of focal planes. This type of data is able to provide focus cues for display technologies. Before the displays on the user side, focal stack video is possibly corrupted during compression, storage and transmission chains, generating error frames on the decoder side. The error regions are difficult to be recovered due to the focal changes among frames. Conventional error concealment methods result in sharpness inconsistency between recovered regions and their spatial adjacent regions. Motivated by this, in this paper, we propose a spatial-focal error concealment scheme specialized for focal stack videos. The spatial adjacent regions around an error region are employed to reveal the prediction relations between error frame and focal adjacent frames. Gaussian blur filtering and Lucy-Richardson deblur filtering are applied to simulate the video focal changes. In this way, the error regions can be well recovered by exploiting the spatial-focal information. Experiment results show that the proposed scheme can achieve the highest objective quality in terms of PSNR and SSIM. It can also obtain the best subjective quality with sharpness consistency in recovered regions and without block effect.

1 Introduction

Focal stack image sequences are focused on successive depth of scene, which are captured by varifocal lens or focusing on multiple focal planes; thus, the sequences can be regarded as special videos in focal dimension. Focal stack videos and related varifocal technologies have become increasingly crucial from the recent advances of display technologies and consumer electronics, such as near-eye see-through displays [1], plenoptic imaging [2, 3] and head-mounted virtual reality (VR) devices [4]. For example, Chang et al. [5] uses focus-tunable lens to enable varifocal display by generating 1600 focal planes per second. The VR displays suffer from the long-standing vergence-accommodation conflict (VAC), resulting in visual discomforts of VR users. Focal stack videos and related varifocal technologies extend depth of field (DoF) for comprehensive observations [6], and adjust depth of field for solving VAC and mitigating visual fatigue [7].

However, focal stack videos are highly redundant due to dense sampling in depth dimension of scenes, which brings the immediate requirements of compression, storage and transmission. It is a common problem that the packets of video data get lost during transmission due to unreliable channels, and the contexts of video bitstream get

979-8-3503-4795-1/23/$31.00 ©2023 IEEE
DOI 10.1109/DCC55655.2023.00017

corrupt during compression and storage [8]. Therefore, focal stack videos have these data loss and corruption during compression, storage and transmission chains, resulting in error frames on the decoder side. Error concealment aims to recover the error decoded regions of images/videos by taking advantage of the adjacent information.

The error concealment for focal stack videos is a novel area that has not been sufficiently studied before. Different from common videos, focal stack videos are a special type of image sequences, the distinguishing feature of which is focal changes instead of motion among neighboring frames. Therefore, developing specialized error concealment method for focal stack videos is a novel and challenging task. In our previous work, we propose a Gaussian guided inter prediction method for focal stack compression, which adopts 2D Gaussian function to approximate the sharpness changes among frames [9, 10]. The method reflects that Gaussian distribution can well fit the image changes of focal stack videos. Inspired by this, we can take advantage of Gaussian distribution to provide assistance for error concealment of focal stack videos.

In this paper, we propose a spatial-focal error concealment scheme for focal stack videos. The raw videos are first encoded using closed group of picture (GOP) configurations. Error videos are then generated by simulating the block errors and slice errors, which are common error types on decoder side. Third, the previous normal GOP and next normal GOP are used to recover the middle error GOP. Specifically, to reveal the prediction relations between error region and its co-located regions from normal GOPs, we implicitly employ the spatial adjacent regions to represent the error region due to lacking of ground truth on decoder side. Spatial-focal adjacent prediction between GOPs is conducted by using image filtering. Gaussian filtering and Lucy-Richardson filtering are applied to the adjacent regions and their co-located regions to simulate the image changes in focal dimension. The derived prediction parameters of the adjacent regions are used for error region. By minimizing the prediction residuals, the error region can be well recovered with the aid of spatial adjacent regions and focal adjacent regions.

The remainder of this paper is organized as follows. Sec. 2 presents all details of the proposed focal stack video error concealment scheme. Experiments and discussions are given in Sec. 3. We summarize this paper in Sec. 4.

2 Methodology

2.1 Problem statement

Typical video error concealment can be divided into three fundamental categories, including spatial, temporal or hybrid spatial-temporal methods [11]. The spatial methods make use of the adjacent information of error regions inside a single frame, while temporal methods take advantage of the sequential adjacent information. As the fusion of these two types of methods, hybrid spatial-temporal methods exploit spatial and temporal information of videos simultaneously, which allows extracting more adjacent information; thus, this type of methods can achieve better performance but higher complexity.

However, focal stack videos are significantly different from typical videos. Typical

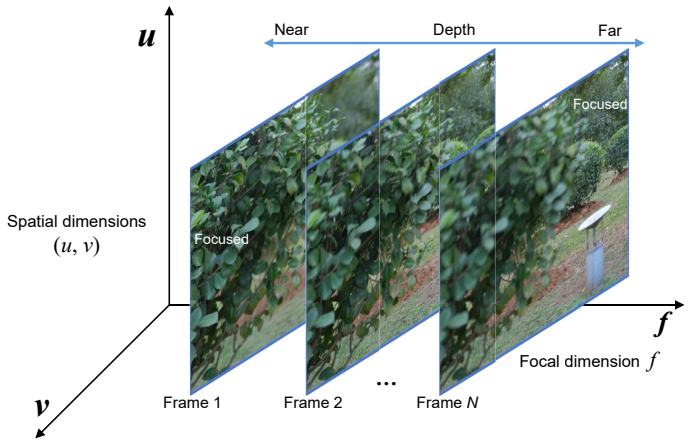


Figure 1: Focal stack video focusing on N planes with spatial dimensions (u, v) and focal dimension f.

videos have temporal changes, where motion of objects among sequential frames is the main characteristic. Focal stack videos, meanwhile, have focused and defocused changes among sequential frames shown in Fig. 1. Generally, typical videos are commonly represented as the 3D parameterization $\boldsymbol{I}(u, v, t)$. In the parameterization, (u, v) describes the pixel coordinates in spatial dimensions, and t signifies the display order in temporal dimension. Focal stack videos replace temporal information by focal information. Accordingly, focal stack videos can be modeled as a 3D parameterization $\boldsymbol{I}(u, v, f)$ shown in Fig. 1. The (u, v) indicates the pixel spatial position in a frame, and the f denotes the scene focal plane or camera imaging depth of a frame. Thus, the (u, v) and f dimensions are the spatial dimensions and focal dimension, respectively.

Inspired by hybrid spatial-temporal logic, we can exploit the spatial and focal information of focal stack videos to conceal the decoded video errors. In this case, the focal changes among frames of focal stack videos need to be well modeled. Gaussian blur can be used for simulating the focused-to-defocused changes in focal stack videos. On the contrary, defocused-to-focused changes can be simulated by deblur filter process, e.g. Lucy-Richardson deconvolution. In this way, the error regions can be recovered by applying filter for spatial adjacent and focal adjacent regions.

2.2 Focal stack video error concealment

We proposed a specialized spatial-focal error concealment scheme. The flowchart of proposed scheme is illustrated in Fig. 2. The scheme consists of error video generation, spatial-focal adjacent region prediction and error slice prediction, More details are presented in the following.

Firstly, in error video generation, focal stack video is compressed by h.264 with closed GOP. Due to closed GOP configuration, we can assume that the errors only occur in a certain GOP, so that the previous and next GOP can be correctly decoded to serve as the reference of error GOP. We simulate the video decoded errors by block errors and slice errors, which are common error types on decoder side.

Secondly, due to the fact that the decoder side lacks of the groundtruth of error

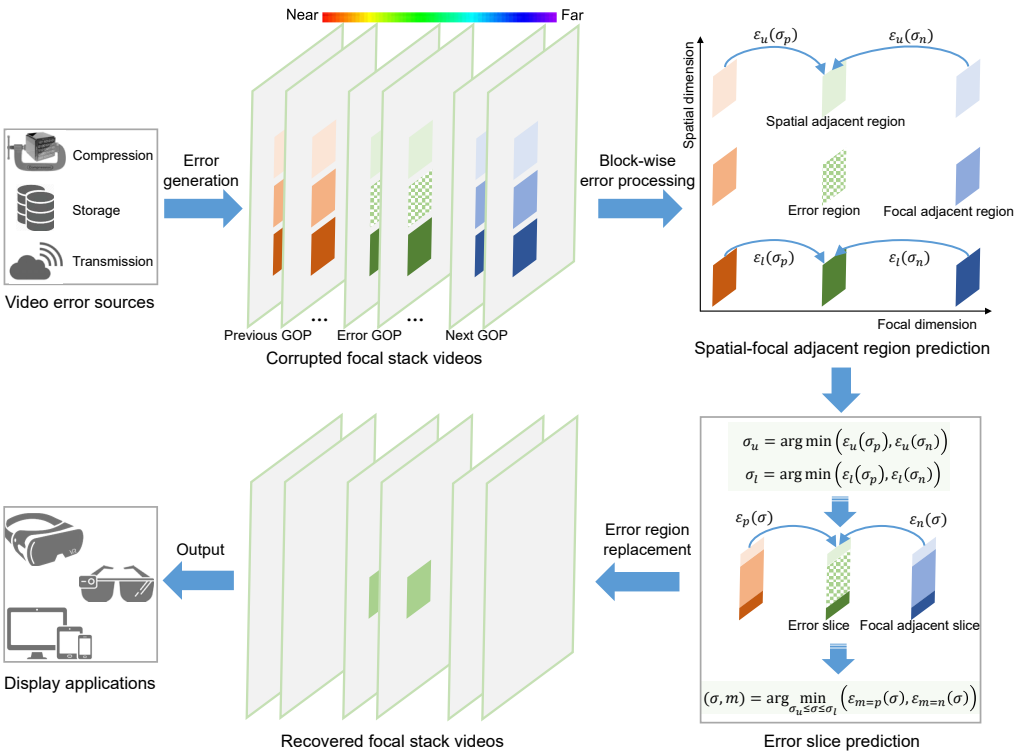


Figure 2: Overview of the proposed spatial-focal error concealment scheme for focal stack videos.

region, we conduct the spatial-focal adjacent prediction using adjacent regions. As shown in Fig. 2, these adjacent regions; for instance, the upper and lower adjacent regions of the error co-located regions are in lighter and darker colors, respectively. The spatial-focal adjacent prediction aims to implicitly reveal the prediction relations between error region and co-located regions from adjacent GOP. The co-located regions of focal stack video have different focusing degrees (sharper or more blurred). Gaussian blur filtering and Lucy-Richardson deblur filtering are applied to the adjacent co-located regions. Gaussian blur filtering is defined by follows:

$$R' = h \otimes R + n, \tag{1}$$

$$h = \frac{1}{2\pi\sigma^2} \exp(-\frac{x^2 + y^2}{2\sigma^2}), \tag{2}$$

where R represents a sharp region, R' denotes the Gaussian blur filtered results. The h stands for 2D Gaussian function. The σ is the standard deviation of 2D Gaussian function. The higher σ is, the more blurred a filtering result is. The n is the noise of images. For simplicity, we assume that there is no additive noise n in focal stack videos. On the other hand, Lucy-Richardson deblur filtering is as follows [12]:

$$R'_{t+1} = \left(\frac{R * k_\sigma}{R'_t \otimes k_\sigma} \right) R'_t, \tag{3}$$

where R represents a blurred region in this case, and R' is the deblurred result of region R. The t stands for the number of iteration. The k_σ signifies the PSF, which is Gaussian kernel with 5x5 size in this paper. Thus, σ is the only parameter for both Gaussian filtering and Lucy-Richardson filtering. The prediction residuals for a target region can be defined as follows:

$$\varepsilon = |R' - T|, \tag{4}$$

where ε stands for the residuals. The T denotes the target regions, R' is the filtered results of reference regions R. Target regions refer to the error region and the spatial adjacent regions, the reference regions indicate the co-located regions in adjacent frames from previous or next GOP.

The upper adjacent regions and lower adjacent regions from the previous and next GOPs are adopted in the filter processing, which is shown on the top right of Fig. 2. The $\varepsilon_u(\sigma_p)$ and $\varepsilon_u(\sigma_n)$ represent the prediction residuals for upper adjacent regions from the previous GOP and next GOP, respectively. Similarly, The $\varepsilon_l(\sigma_p)$ and $\varepsilon_l(\sigma_n)$ are the residuals for lower adjacent regions from the previous GOP and next GOP, respectively. We can obtain the best parameter σ_u for upper adjacent regions by minimizing the residuals of previous GOP $\varepsilon_u(\sigma_p)$ and next GOP $\varepsilon_u(\sigma_n)$. The best parameter σ_l for lower adjacent regions can be also obtained in this way as follows:

$$\sigma_u = \arg\min\left(\varepsilon_u(\sigma_p), \varepsilon_u(\sigma_n)\right), \tag{5}$$

$$\sigma_l = \arg\min\left(\varepsilon_l(\sigma_p), \varepsilon_l(\sigma_n)\right), \tag{6}$$

Thirdly, the error slice prediction is shown on the bottom-right of Fig. 2. Error region and its corresponding co-located regions are then spliced with part of adjacent regions to form multiple slices. The error slice is expected to be predicted by the co-located slices. The two optimized parameter σ_u and σ_l are used for determining the two bounds $\sigma_{min} = \min(\sigma_u, \sigma_l)$ and $\sigma_{max} = \max(\sigma_u, \sigma_l)$. Thus, the optimal parameter of error slice can be obtained by:

$$(\sigma, m) = \arg \min_{\sigma_{min} \le \sigma \le \sigma_{max}} (\varepsilon_{m=p}(\sigma), \varepsilon_{m=n}(\sigma)) \tag{7}$$

where the $\varepsilon_{m=p}$ and $\varepsilon_{m=n}$ denote the residuals predicted from previous GOP and next GOP, respectively. The optimal parameter σ and m can be computed by minimizing these residuals in the parameter searching range.

Finally, all the error slices are replaced by the predicted slices, the corrupted focal stack videos can be recovered. Various immersive display applications will benefit from the focal stack videos.

3 Experiments

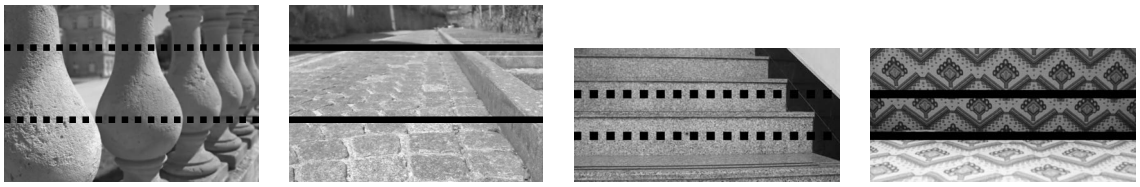
3.1 Experiment settings

There existing various error concealment methods for common image/video. The hybrid spatial-temporal video error concealment methods are simple and intuitive. The advanced error concealment methods make use of special prediction algorithm [13]

and estimation algorithm [14]. In the experiments, our proposed scheme is compared with image/video error concealment methods, including hybrid spatial-temporal (Hybrid) [15], sparse linear prediction (SLP) [13], kernel based minimum mean square error (KMMSE) [14].

Table 1: Details of test sequences and error distribution types

Scene	I01	I02	I03		I04	
Scene type	realistic		synthesized			
Frame number	15		15			
Resolution	624x432		1024x512			
Error types	block 16	slice 16	block 16	block 32	slice 16	slice 32
abbreviation	blk-16	slc-16	blk-16	blk-32	slc-16	slc-32



(a) I01-blk-16 (b) I02-slc-16 (c) I03-blk-32 (d) I04-slc-32

Figure 3: Thumbnails of Y-component of all test sequences and some examples of error distributions.

The experiments are conducted on 4 focal stack video test sequences, including 2 realistic scenes [16] and 2 synthesized scenes [9]. The realistic scenes are captured by Canon EOS60D camera by setting different focal planes, while the synthesized scenes are generated by refocusing from light field dataset [17]. These test sequences are with 15 frames. All frames are placed from near to far to form video sequences in YUV420 format. The H.264/AVC codec is commonly used for consumer electronics; for instance, smartphone, portable VR and other mobile devices. The codec packs video stream by slices, the minimum packet losing results in error of at least one 16x16 macroblock, or serious corruption of a slice [13]. In our experiment, the test sequences are encoded and decoded by H.264 codec with closed GOP configuration. The synthesized sequences and realistic sequences are compressed by lossy coding and lossless coding, respectively. For all the test sequences with 15 frames, the GOP size is specified as 5. The second GOP is corrupted, while the first GOP (previous GOP) and third GOP (next GOP) are uncorrupted shown in Fig. 2. The error distributions include block errors and slice errors, which are common error types on decoder side. The sizes of error regions include 16×16 and 32×32. For simplicity, only the Y-component of the YUV420 test sequences are corrupted. The detailed information of all test sequences is listed in Table 1. Thumbnails and error distributions are shown in Fig. 3, note that only partial error types are shown in this figure due to limited space.

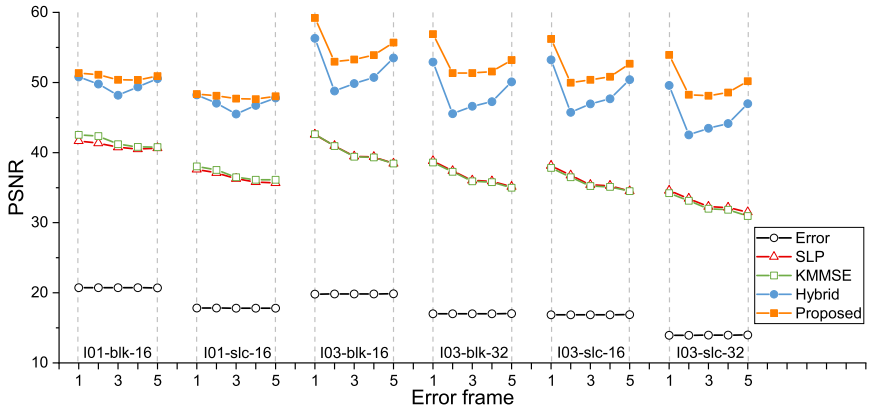
Table 2: Objective quality comparisons among all the sequences, all error types and all error concealment methods. The average PSNR and SSIM of all five error frames are calculated and presented.

Scene	Type	Error		Hybrid		SLP		KMMSE		Proposed	
		PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
I01	blk-16	20.73	0.9306	49.73	0.9978	40.98	0.9919	41.54	0.9923	50.83	0.9982
	slc-16	17.82	0.8980	47.05	0.9957	36.49	0.9802	36.86	0.9804	47.97	0.9964
I02	blk-16	18.24	0.9270	48.26	0.9963	41.58	0.9892	41.90	0.9897	49.42	0.9968
	slc-16	15.32	0.8942	45.30	0.9928	32.90	0.9672	32.60	0.9656	46.43	0.9936
I03	blk16	19.84	0.9445	51.83	0.9987	40.16	0.9860	40.15	0.9860	**55.02**	**0.9994**
	slc-16	16.87	0.9152	48.80	0.9975	36.01	0.9678	35.83	0.9672	52.01	0.9989
	blk-32	17.01	0.9150	48.49	0.9973	36.65	0.9682	36.51	0.9681	52.88	0.9991
	slc-32	13.96	0.8531	45.35	0.9947	32.79	0.9319	32.43	0.9310	49.81	0.9982
I04	blk-16	22.36	0.9493	49.28	0.9987	39.20	0.9927	39.54	0.9931	53.30	0.9994
	slc-16	19.33	0.9187	46.30	0.9975	29.98	0.9630	29.98	0.9643	50.29	0.9988
	blk-32	18.49	0.9176	46.08	0.9976	32.54	0.9735	33.04	0.9765	51.18	0.9991
	slc-32	15.45	0.8549	42.91	0.9954	24.84	0.9178	24.92	0.9205	48.04	0.9981

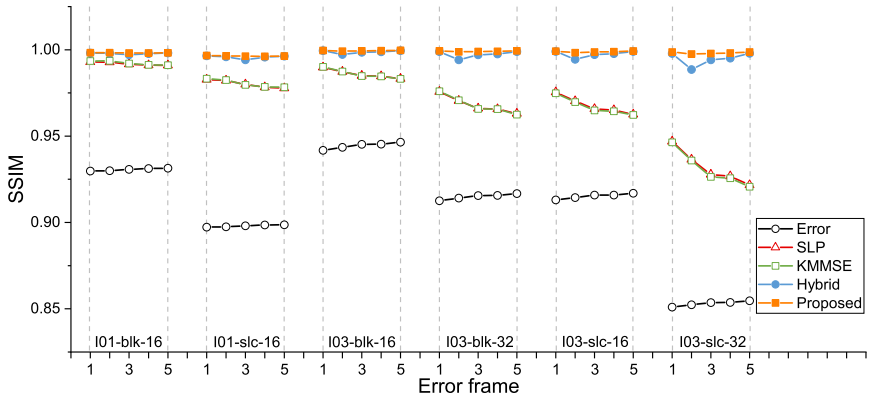
3.2 Objective performance

We analyze the objective performance of the proposed scheme and comparison methods in terms of PSNR and SSIM. The experimental results are shown in Table 2 and Fig. 4.

In Table 2, we compare all the test sequences, all error types and all error concealment methods. The "Error" refers to the corrupted videos, the remaining schemes refer to the error concealed videos using the proposed scheme and the comparison methods. We calculate the average PSNR and SSIM of all recovered error frames (five frames) with the original frames. We can find that all error concealed videos achieve PSNR and SSIM increments over the error videos. Our proposed scheme achieves the highest performance in terms of PSNR and SSIM in all test conditions. It can obtain as high as 55.02 dB PSNR and 0.9994 SSIM. In Fig. 4, the PSNR and SSIM of each recovered frames with the original frames are visualized and analyzed. Specifically, the "SLP" and "KMMSE" error concealment methods obtain relatively small increments since these methods only exploit spatial information in single error image. The "Hybrid" method realizes considerable performance improvement due to the use of spatial and temporal information. Our proposed scheme makes full use of the spatial and focal information according to the characteristic of focal stack videos. Gaussian blur and Lucy-Richardson deblur filtering are applied in our scheme, achieving the highest performance by simulating the frame changes of focal stack videos. For all error frames from 1 to 5, our error concealment scheme tends to have more increments for middle frames 2, 3 and 4 compared with the "Hybrid" method. This is because the middle frames yield long distance reference effect in the "Hybrid" method. The proposed error concealment scheme can mitigate the effect by using filtering.



(a) PSNR



(b) SSIM

Figure 4: Visualization of objective quality comparison among all methods. Test sequences I01 and I03 are selected to represent realistic scenes and synthesized scenes.

3.3 Subjective performance

Moreover, we assess the subjective quality for all the methods shown in Fig. 5. Due to limited space, we only illustrated the results of error types I01-blk-16 and I02-slc-32 for middle frame 3 in the figure. Close-ups are placed in the corners to magnify image details. The PSNR of error and concealed videos with the original video is also listed at the bottom.

It can be found that the concealed frames by all error concealment methods obtain considerable PSNR increments than the error frame. The proposed scheme gains the highest PSNR than all the comparison methods. As for visual experience, the proposed scheme achieves the highest subjective quality than all the comparison methods. Specifically, the recovered regions of "SLP", "KMMSE" and "Hybrid" yield noticeable block effect and texture artifacts, where the sharpness of these recovered regions is not consistent with the spatial adjacent regions in the error frame. By contrast, the recovered regions of the proposed scheme have proper sharpness consistency with adjacent regions, and have no block effect. This is because the proposed

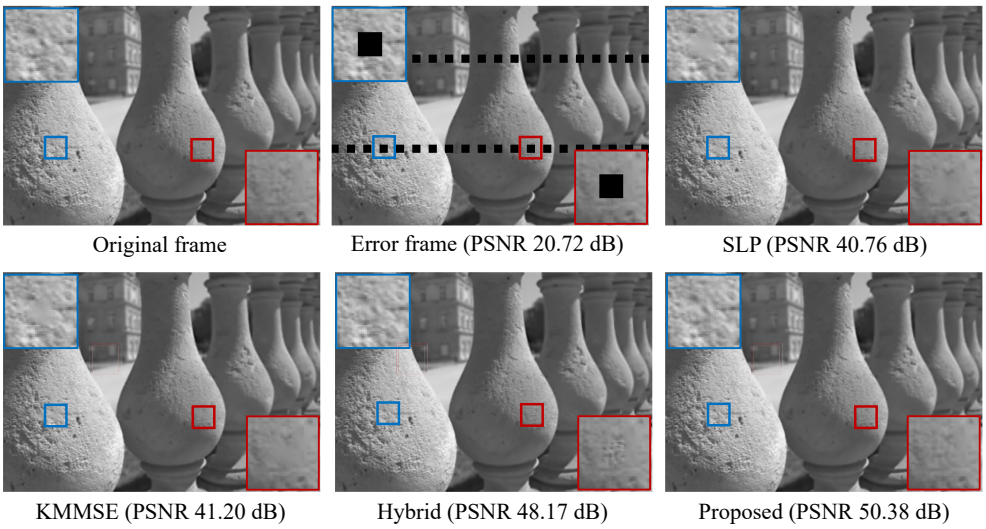


Figure 5: Subjective quality comparison of error frame 3 among the proposed scheme and comparison methods. Test sequences I01 is are selected for the comparison.

scheme splices error region with part of the spatial adjacent regions to form slices. The spliced slice contains adjacent information; thus, the recovered error regions are consistency with adjacent regions in sharpness and textures. Therefore, the proposed error concealment scheme achieves closer subjective quality to the original frames than other comparison methods.

4 Conclusions

Conventional error concealment methods are designed for common videos, which cannot efficiently exploit the distinctive focal changes of focal stack videos. In this paper, we propose an error concealment scheme for focal stack videos. The spatial adjacent and focal adjacent regions around an error region are employed to reveal the prediction relations in the video. Gaussian blur and Lucy-Richardson deblur filters are used to simulate the image sharpness changes. Error regions can be well recovered using the prediction relations of spatial-focal adjacent regions. Experiment results show that the proposed scheme can achieve the highest objective quality in terms of PSNR and SSIM. It can also obtain the best subjective quality with sharpness consistency and without block effect, achieving natural and smooth display experience.

Acknowledgment

This research / project is supported by the National Research Foundation, Singapore, and Cyber Security Agency of Singapore under its National Cybersecurity R&D Programme (NRF2018NCR-NCR009-0001). Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of National Research Foundation, Singapore and Cyber Security Agency of Singapore.

References

[1] Y. Zhou, J. Zhang, and F. Fang, "Design of the varifocal and multifocal optical near-eye see-through display," *Optik*, vol. 270, pp. 169942, 2022.

[2] M. Chen, M. Ye, Z. Wang, C. Hu, T. Liu, K. Liu, J. Shi, and X. Zhang, "Electrically addressed focal stack plenoptic camera based on a liquid-crystal microlens array for all-in-focus imaging," *Optics Express*, vol. 30, no. 19, pp. 34938–34955, 2022.

[3] K. Wu, Z. Liao, Q. Liu, Y. Yin, and Y. Yang, "A global co-saliency guided bit allocation for light field image compression," in *2019 Data Compression Conference (DCC)*, 2019, pp. 608–608.

[4] C. Ebner, S. Mori, P. Mohr, Y. Peng, D. Schmalstieg, G. Wetzstein, and D. Kalkofen, "Video see-through mixed reality with focus cues," *IEEE Transactions on Visualization and Computer Graphics*, vol. 28, no. 5, pp. 2256–2266, 2022.

[5] J. R. Chang, B. V. K. Vijaya Kumar, and Aswin C. Sankaranarayanan, "Towards multifocal displays with dense focal stacks," *ACM Transactions on Graphics*, vol. 37, no. 6, pp. 1–13, 2018.

[6] H. M. Kim, M. S. Kim, S. Chang, J. Jeong, H. G. Jeon, and Y. M. Song, "Vari-focal light field camera for extended depth of field," *Micromachines*, vol. 12, no. 12, pp. 1453, 2021.

[7] C. Chao, C. Liu, and H. Chen, "Learning coded apertures for time-division multiplexing light field display," *arXiv preprint arXiv:2107.06205*, 2021.

[8] V. Boussard, F. Golaghazadeh, S.Coulombe, F. X. Coudoux, and P. Corlay, "Robust h. 264 video decoding using crc-based single error correction and non-desynchronizing bits validation," in *2020 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2020, pp. 1098–1102.

[9] K. Wu, Q. Liu, Y. Yin, and Y. Yang, "Gaussian guided inter prediction for focal stack images compression," *2020 Data Compression Conference (DCC)*, pp. 63–72, 2020.

[10] K. Wu, Y. Yang, M. Yu, and Q. Liu, "Block-wise focal stack image representation for end-to-end applications," *Optics Express*, vol. 28, no. 26, pp. 40024–40043, Dec 2020.

[11] M. Kazemi, M. Ghanbari, and S. Shirmohammadi, "A review of temporal video error concealment techniques and their suitability for hevc and vvc," *Multimedia Tools and Applications*, vol. 80, no. 8, pp. 12685–12730, 2021.

[12] J. Jency Rubia and R. Babitha Lincy, "Digital image restoration using modified richardson-lucy deconvolution algorithm," in *International Conference on Image Processing and Capsule Networks*. Springer, 2020, pp. 100–112.

[13] J. Koloda, J. Østergaard, S. H. Jensen, V. Sánchez, and A. M. Peinado, "Sequential error concealment for video/images by sparse linear prediction," *IEEE Transactions on Multimedia*, vol. 15, no. 4, pp. 957–969, 2013.

[14] J. Koloda, A. M. Peinado, and V. Sánchez, "Kernel-based mmse multimedia signal reconstruction and its application to spatial error concealment," *IEEE Transactions on Multimedia*, vol. 16, no. 6, pp. 1729–1738, 2014.

[15] S. Ye, M. Ouaret, F. Dufaux, and T. Ebrahimi, "Hybrid spatial and temporal error concealment for distributed video coding," in *2008 IEEE International Conference on Multimedia and Expo*. IEEE, 2008, pp. 633–636.

[16] K. Wu, Y. Yang, Q. Liu, and X. P. Zhang, "Focal stack image compression based on basis-quadtree representation," *IEEE Transactions on Multimedia*, 2022.

[17] M. Rerábek and T. Ebrahimi, "New light field image dataset," in *Proc. 8th Int. Conf. Quality Multimedia Exper. (QoMEX)*, 2016.