Dynamic Texture Synthesis using Linear Phase Shift Interpolation

Uday Singh Thakur, Karam Naser, Mathias Wien

Dynamic Texture Synthesis using Linear Phase Shift Interpolation

Uday Singh Thakur[1], Karam Naser[2] and Mathias Wien[1]
[1]Institute for Nachrichtentechnik, RWTH Aachen University, Aachen, Germany
[2]IRCCyN UMR CNRS 6597, University of Nantes, Nantes Cedex 3, France
{thakur, wien}@ient.rwth-aachen.de
karam.naser@univ.nantes.fr

***Abstract*—Dynamic texture motions like flowing water, motion of leaves etc. have a complex random character. A sequence containing such a content is challenging to encode even when using state of the art High Efficiency Video Coding (HEVC) especially, if the available bandwidth is limited. It is observed that often when predicting dynamic textures, codec switches to intra prediction. At lower rates, dynamic texture content shows visually annoying blurring and blocking artifacts.**

For dynamic textures, both spatial and temporal details are perceptually of less importance. This property of the Human Visual System (HVS) can be exploited when coding dynamic texture content, as suggested in this paper. At the encoder, preprocessing is done by skipping even numbered B pictures. At the decoder side, skipped pictures are synthesized using linear phase shift interpolation of the complex wavelet coefficients, from the adjacent already decoded pictures. Subjective evaluation of proposed approach is done by using a pair wise comparison test between the proposed results and the conventional HEVC decoded bitstream at similar bitrates. The evaluation results show that viewers prefer the proposed result over conventional HEVC.

I. Introduction

Textures are categorized into two categories based on motion i.e. static or dynamic [1]. The former refers to an object in a scene which exhibits spatial homogeneity such as a fabric, surface of a stone and does not have any local motion over time (excluding potential camera motion). The latter are sequences of images of moving scenes that exhibit certain stationarity properties in time; these include flowing water, leaves in motion, fire, etc. We focus primarily on dynamic textures, as such a type of motion is extremely challenging to compress, when using state of the art High Efficiency Video Coding (HEVC) [2], [3].

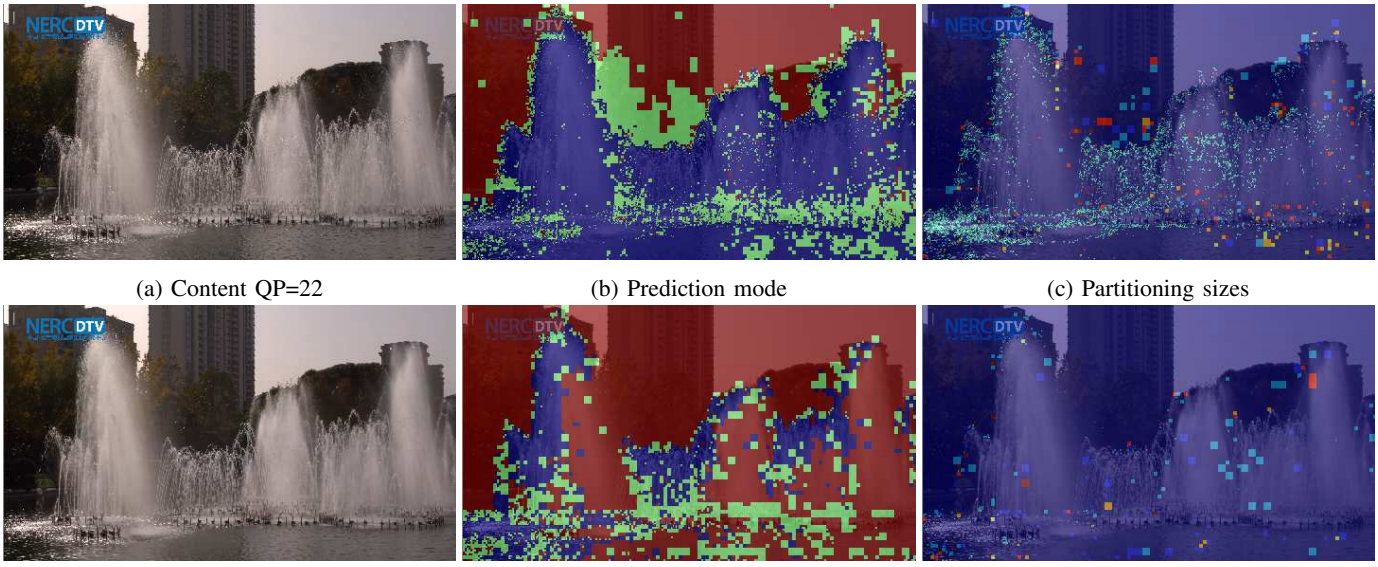
It is mostly observed that when encoding dynamic textures, motion compensation performance of HEVC is not efficient, as a consequence the residual has high energy. One of the reason is rapid change in the signal over time. Often, the codec switches to Intra mode over dynamic textured regions instead of using Inter prediction when predicting B pictures as shown in Fig. 1b. It is also observed that Coding Tree Units (CTU's) are partitioned in to smaller Coding Units (CU's) over such regions as shown in Fig. 1c, this further adds to the bitrate.

Accordingly, such a motion is difficult to predict using standard motion compensation algorithms, due to its complex random character. To solve this problem, several texture coding techniques have been proposed in the past. This can be generally classified into two categories: Top-down and bottom-up approaches. Top-down approaches start from the idea that fine details inside textures are perceptually irrelevant and therefore, they can be replaced with an equivalent content given a set of statistical constraints. On the other hand, the bottom-up approaches takes into the account the perceptual properties of the textured signals and distribute the bitrate adaptively, or allocate the distortions, according to their perceptual sensitivity. A well known example of top-down approaches is texture synthesis. It has been utilized many times in video coding after the inspiring work by Ndjiki-Nya et.al. [4]. In this approach, the texture is removed at the encoder side and synthesis parameters are sent to the decoder. Similarly, texture synthesis was used for interpolating textures between pictures [5], or replacing the texture with a simplified version [6], [7], [8] and [9].

For bottom-up approaches, there has been a lot of focus put into static textures, while neglecting the dynamic ones. An example of this [10], [11], is to consider the sensitivity of each region of the scene in distributing the bitrate. A recent investigation on dynamic textures has shown that there is a large amount of perceptual redundancies, that can be exploited to provide significant bitrate saving in HEVC [12]. Previously, motion of dynamic texture was predicted using texture synthesis model based on Auto Regressive Moving Average (ARMA) prediction H.264/AVC [1]. ARMA is a linear model and its performance is limited to simple motions which are now well handled by HEVC. More complicated motion still needs more advanced modeling.

Recently, phase based methods have shown promising results in video motion processing and interpolation [14], [15]. In this paper, we have used phase of complex wavelet coefficients for dynamic texture synthesis coding, for every 3 consecutive B pictures, 1 B picture in the middle is skipped at the encoder and later synthesized at the decoder using wavelet transform. Skipping B pictures for such a content will have minor effects on the motion compensation performance of HEVC, due to Intra prediction being mostly used when predicting B pictures, comprising of dynamic texture content. It is a matter of fact that, intra coded signal takes more bits compared to inter and therefore, by skipping 50% of the B pictures provides $27 - 45\%$ saving in bitrate for dynamic texture content as shown in Fig. 2.

Rest of the paper is organized as follows: Section II reviews



(a) Content QP=22 (b) Prediction mode (c) Partitioning sizes

(d) Content QP=37 (e) Prediction mode (f) Partitioning sizes

Fig. 1: HEVC's behavior over dynamic texture content for both small and large QP values. (a) shows the B picture from the sequence, Fountain [13] at QP=22 random access configuration. (b) shows the prediction mode selected by HEVC, Inter predicted regions are shown with color green, Intra in purple (mostly observed over dynamic texture content) and maroon shows region where skip mode is chosen. (c) shows the block partitioning sizes, smaller partitions are observed over dynamic texture content. Similarily, we have also shown the behavior of HEVC at large QP (QP=37) in (d), (e) and (f). It is clearly evident that HEVC's motion compensation performance when predicting dynamic texture content is not effective.

the phase video interpolation followed by its application to dynamic texture synthesis is discussed in Section III. Section IV discusses the experimental results. Conclusions are drawn in Section V.

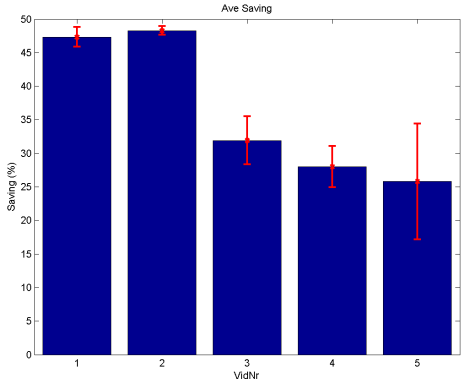


Fig. 2: Average saving in bitrate for homogeneous dynamic texture sequences, when 1 picture is skipped between 3 consecutive B pictures. The comparison is averaged at QP=22, 27, 32 and 37

II. Phase shift video interpolation

The presented approach in this Section is based on the work [15]. Phase shift between two pictures corresponds to the motion in a video. Interpolating such a phase shift can synthesize another intermediate picture. We have exploited this work for benefiting dynamic texture coding. In this section we will give a brief overview of the algorithm given in [15] as it is important to understand the further work.

A. Steerable Pyramid

In general for two-dimensional functions, one can separate the sinusoids into bands not only according to the frequency ω, but also according to spatial orientation θ, using the complex-valued steerable pyramid [16]. The steerable pyramid filters resemble Gabor wavelets and, when applied to the discrete Fourier transform of an image, they decompose the input image into a number of oriented frequency bands $R_{\omega,\theta}$. The remaining frequency content which has not been captured in the pyramid levels is summarized in (real valued) high and lowpass residuals. It is to be noted that for this work these pyramids have been generalized to arbitrarily scaling factor λ due to better cross-scale phase correlation. Amplitude $A_{\omega,\theta}$ and phase $\phi_{\omega,\theta}$ of above computed complex coefficients are subsequently calculated.

B. Phase shift interpolation with confidence based shifting

Assuming that motion is encoded in the phase shift, interpolating it requires computation of phase difference ϕ_{diff} between the phases ϕ_1 and ϕ_2 of two adjacent pictures. Due to periodicity of the phase value, resulting angular values are between $[-\pi, \pi]$ which correspond to smaller angular differences between the two input phases. Displacement corresponding to a phase difference of more than π leads to a phase ambiguity. To overcome this ambiguity, the method for

shift correction is proposed in [15], based on the assumption that the phase difference between two pyramid levels does not differ arbitrarily, i.e. phase differences between levels can be used as a confidence measure that quantifies whether the computed phase shift is reliable. The shift correction helps to interpolate the motion of high frequency content more robustly. The corrected phase shift $\widetilde{\phi}_{\text{diff}}$ is linearly interpolated to synthesize any motion in between phases ϕ_1 and ϕ_2.

III. Extension of Phase Interpolation to Dynamic Texture Content

Motion of dynamic textures have a complex random character, as a result they are challenging to encode. Conventional motion compensation algorithms used in state of the art HEVC completely fail when performing inter prediction for B pictures as a result, residual has high energy. Based on the Rate Distorion decision mechanism the codec often switches to Intra, shown in Fig. 1b. This behavior of codec opens up door for novel ways of modeling such a content.

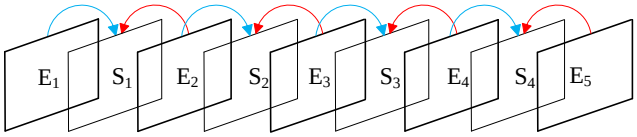


Fig. 3: Dynamic texture synthesis: Every S picture is reconstructed at the decoder from the closest neighbors i.e. decoded E pictures using complex wavelets

A. Skipping B pictures

The goal of the algorithm is to first skip and later synthesize the skipped pictures. This is a pre-processing step, in which we skip even numbered pictures during encoding. In general, if the original picture count is even, the remaining pictures left after preprocessing is given by $\big((\text{count}/2)+1\big)$ else, if the original picture count was odd the remaining picture count is given by $\big((\text{count}/2)-0.5\big)$.

B. Synthesis of B pictures

At the decoder side, the bitstream is decoded followed by skipped B picture synthesis as shown in Fig. 3, skipped picture e.g S_1 is synthesized from the two adjacent already decoded neighbor pictures E_1 and E_2 using the algorithm mentioned in Section II. The decoded E pictures are decomposed using multiscale oriented linear filters [16]. The resulting coefficients are complex-valued: The real and imaginary parts correspond to even and odd-symmetric filter impulse responses. Phase difference between between corresponding oriented frequency bands of the two consecutive decoded E pictures is calculated using four quadrant inverse tangent, followed by linear interpolation of the phase shift.

This method does not require any side information as it is simple linear interpolation of phases between bands of similar frequency and orientation. The output of interpolation is the phase of picture to be synthesized but, to complete the reconstruction of B picture, we will also need its magnitude,

(a) Reference (b) Linear blending (c) Proposed

Fig. 4: Subjective comparison of results for magnitude interpolation. (a) shows the reference, (b) linear blending of coefficient magnitudes as given in [15], (c) is the proposed result. It is clearly evident that proposed method reconstructs sharper interpolated picture as compared to the linear blending.

(a) Reference (b) HEVC (c) Proposed

Fig. 5: Subjective comparison of HEVC and Synthesis at similar bitrates. (b) shows the conventional HEVC decoded picture, with blockiness and blurriness. (c) shows the synthesized B picture at similar bitrate, with more spatial details compared to the HEVC

lowpass and high frequency information. The magnitude of the interpolated B picture cannot be just linearly blended for dynamic textures as given [15]. Doing such blending reduces the luminance of picture and introduces blurriness as shown in Fig. 4b. The reason is that motion in case of dynamic textures is very rapid, with highly varying contrasts over time. As a result, the magnitude widely varies between consecutive frames and averaging such a magnitude makes the reconstruction look blurry. In the proposed solution for the intermediate picture's magnitude, the magnitude is not linearly blended where is it changing beyond a certain threshold ρ.

Algorithm 1 Algorithm for synthesis of magnitude

```
if |A1 − A2| > ρ then
    if A1 > A2 then
        Asyn = A1
    else if A1 < A2 then
        Asyn = A2
    end if
else
    Asyn = (A1 + A2)/2
end if
```

In algorithm 1 A_1 and A_2 are magnitudes of decoded picture E_1 and E_2. We use the greater of two magnitudes for the intermediate picture S_1 as perceptually it looked better and

sharper as shown in Fig. 4c. If the difference in magnitudes is less that ρ we simply blend the magnitudes. Lowband is linearly blended in this case and high frequency of the later picture E_2 is chosen. Final reconstruction is obtained by using inverse transform. The interpolated S picture between two E pictures is perceptually of high quality.

IV. Results and Discussion

For evaluation purposes a 256×256 sequence resolution consisting of dynamic texture is studied. These sequences were cropped from original $4k$ video sequences [17] such that the selected areas only include dynamic texture content. Temporally, the sequences have an extent of 500 ms. The purpose of selecting short term sequences is to have as much homogeneous contents as possible. Both spatial and temporal resolution have been selected subject to perceptual constraints, spatially the sequences are within the foveal vision, whereas temporally they are beyond the minimum fixation time (100-200 ms).

A. Bitrate Saving

In order to show the amount of possible saving in bitrate, we have compared the bitrate between HEVC's random access configuration and the proposed coding scheme at the same QPs. The results are averaged for four QP values $(22, 27, 32$ and $37)$ as shown in Fig. 2, the error bar in the figure corresponds to standard deviation of the four QPs.

The bar graph in Fig. 2 clearly shows that $27 - 45\%$ of the bitrate is allocated to the skipped B pictures. For our proposed scheme, we have used 120 Hz sequences, after skipping B pictures during encoding it changes down to 60 Hz. The skipped B pictures are then synthesized at the decoder using the algorithm mention in Section II and III. The resulting sequences have better spatial quality as shown in Fig. 5 at same bitrate, as it retains more spatial details compared to HEVC. Temporally, the motion is coherent with less blockiness compared to HEVC. Perceptual details inside dynamic texture are of less importance to the human observer and therefore, the approach is considered to be appropriate for such type of contents. Subjective evaluations are performed for quality assessment, to assess the quality of the proposed method in comparison to the HEVC at similar bitrates.

B. Objective Evaluation

Evaluating the proposed model objectively is not appropriate as the target of the proposed scheme is not to approximate the reference. This is because we are highly deviating from pixel fidelity when synthesis is used. However, for odd pictures which are encoded with HEVC, the objective score is always better, as we have saving in bitrate. However, synthesis causes reduction in PSNR for even pictures. This is quite expected as PSNR is not a proper measure of the perceived distortion for synthesis. In contrast, the visual quality of the synthesized picture is better than the reconstructed HEVC image at the same bitrate shown in Fig. 5.

TABLE I: Subjective test result.

	Subject	1	2	3	4	5	**Ave**
q1	Calming Water	-1	0	1	-1	1	**0**
	Drops on Water	-1	0	0	-1	0	**-0.4**
	Lamp Leaves-1	0	-1	1	1	1	**0.4**
	Lamp Leaves-2	0	1	1	1	0	**0.6**
	Treewills	1	1	1	1	-1	**0.6**
q2	Calming Water	0	-1	1	1	1	**0.4**
	Drops on Water	-1	0	1	1	0	**0.2**
	Lamp Leaves-1	0	0	0	1	1	**0.4**
	Lamp Leaves-2	1	1	1	0	-1	**0.4**
	Treewills	0	0	1	-1	1	**0.2**
	Calming Water	1	1	1	1	-1	**0.6**
q3	Drops on Water	-1	0	1	1	0	**0.2**
	Lamp Leaves-1	0	0	0	0	-1	**-0.2**
	Lamp Leaves-2	-1	-1	1	-1	0	**-0.4**
	Treewills	0	-1	0	0	1	**0**
q4	Calming Water	1	0	1	1	0	**0.6**
	Drops on Water	-1	-1	1	1	-1	**-0.2**
	Lamp Leaves-1	0	1	1	1	1	**0.8**
	Lamp Leaves-2	-1	0	0	0	0	**-0.2**
	Treewills	1	-1	1	-1	1	**0.2**
	Ave	**-0.1**	**-0.05**	**0.75**	**0.3**	**0.15**	**0.21**

C. Subjective Evaluation

For evaluation of the proposed model, we designed a specific subjective test to verify its usefulness. The test is a pair wise comparison in which the observers saw two sequences side by side as shown in Fig. 7, and were asked two compare and select the one that they preferred. The observers were also allowed to have no preference option, to reduce the amount of random selections. The subjective testing material consisted of five source sequences as shown in Fig. 6. The observers were allowed to repeat and play the sequences at their choice. The sequences were encoded at four compression levels (q1; q2; q3 and q4), QPs ranging in value between 22 to 37. The proposed scheme is compared to conventional HEVC random access configuration at same bitrate. Five subjects participated in the test, their respective responses are given in Table I. In this table, 1 refers to when the proposed model is preferred, -1 when default HEVC is preferred and 0 refers to no preference. Overall, we can see that the average is a positive number, which indicates that the subjects generally preferred the synthesis over the default HEVC.

V. Conclusion

The paper presents a novel scheme for B picture synthesis in the context of dynamic texture coding. HEVC switches to intra prediction in B pictures with smaller partitionings eventually, leading to high bitrate. The proposed scheme skips every even B picture during encoding process. The skipped pictures are synthesized using phase shift interpolation from already decoded neighboring pictures. Skipping half of B pictures saves $27 - 45\%$ bitrate. The paper has presented a novel idea, about new ways for exploration and modeling dynamic texture content from the perspective of benefitting video coding. In a realistic scenario structure and texture will be segmented, textured blocks in B pictures can be skipped and later synthesized at the decoder. Objective quality

(a) Calming Water (b) Drops on Water (c) Lamp Leaves-1 (d) Lamp Leaves-2 (e) Treewillis

Fig. 6: Subjective testing material

Fig. 7: Subjective evaluation view

assessment of such a content is challenging. Investigating decision mechanism based on perceptual metrics will be a stepping stone for this research.

Our future work will focus on skipping more than 1 B picture, and then synthesizing. This is currently challenging as dynamic textures have rapid motion and interpolating phase shift in between distant pictures can be ambiguous and produce annoying artifacts in the synthesized pictures.

ACKNOWLEDGEMENT

This work was carried out within the Marie Sklodowska Curie Training Network PROVISION (PeRceptually Optimized VIdeo CompresSION) and received funding from the European Union's Seventh Framework Program for research, technological developments and demonstration under grant agreement no 608231.

REFERENCES

[1] J. Ballé, A. Stojanovic, and J. R. Ohm, "Models for static and dynamic texture synthesis in image and video compression," *Selected Topics in Signal Processing, IEEE Journal of*, vol. 5, no. 7, pp. 1353–1365, 2011.

[2] M. Wien, *High Efficiency Video Coding – Coding Tools and Specification*, Springer, Berlin, Heidelberg, Sept. 2014.

[3] G. J. Sullivan, J. R. Ohm, W. J. Han, and T. Wiegand, "Overview of the high efficiency video coding (HEVC) standard," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 22, no. 12, pp. 1649–1668, 2012.

[4] P. Ndjiki-Nya, B. Makai, G. Blattermann, A. Smolic, H. Schwarz, and T. Wiegand, "Improved H.264/AVC coding using texture analysis and synthesis," in *Image Processing, 2003. ICIP 2003. Proceedings. 2003 International Conference on*. IEEE, 2003, vol. 3, pp. III–849.

[5] F. Zhang and D. R. Bull, "A parametric framework for video compression using region-based texture models," *Selected Topics in Signal Processing, IEEE Journal of*, vol. 5, no. 7, pp. 1378–1392, 2011.

[6] A. Dumitras and B. G. Haskell, "A texture replacement method at the encoder for bit-rate reduction of compressed video," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 13, no. 2, pp. 163–175, 2003.

[7] K. Naser, V. Ricordel, and P. Le Callet, "Local texture synthesis: A static texture coding algorithm fully compatible with HEVC," in *Systems, Signals and Image Processing (IWSSIP), 2015 International Conference on*. IEEE, 2015, pp. 37–40.

[8] U. S. Thakur and O. Chubach, "Texture analysis and synthesis using steerable pyramid decomposition for video coding," in *2015 International Conference on Systems, Signals and Image Processing (IWSSIP)*, Sept 2015, pp. 204–207.

[9] U. S. Thakur and B. Ray, "Image coding using parametric texture synthesis," in *2016 IEEE Workshop on Multimedia Signal Processing (MMSP)[accepted]*, Sept 2016.

[10] C. Sun, H-J Wang, H. Li, and T-h Kim, "Perceptually adaptive lagrange multiplier for rate-distortion optimization in H.264," in *Future Generation Communication and Networking (FGCN 2007)*. IEEE, 2007, vol. 1, pp. 459–463.

[11] M. Liu and L. Lu, "An improved rate control algorithm of H.264/AVC based on Human Visual System," in *Computer, Informatics, Cybernetics and Applications*, pp. 1145–1151. Springer, 2012.

[12] K. Naser, V. Ricordel, and P. Le Callet, "Estimation of perceptual redundancies of HEVC encoded dynamic textures," in *2016 Eighth International Conference on Quality of Multimedia Experience (QoMEX)*. IEEE, 2016, pp. 1–5.

[13] L. Song, X. Tang, W. Zhang, X. Yang, and P. Xia, "The sjtu 4k video sequence dataset," in *Quality of Multimedia Experience (QoMEX), 2013 Fifth International Workshop on*, July 2013, pp. 34–35.

[14] N. Wadhwa, M. Rubinstein, F. Durand, and W. T. Freeman, "Phase-based video motion processing," *ACM Trans. Graph. (Proceedings SIGGRAPH 2013)*, vol. 32, no. 4, 2013.

[15] S. Meyer, O. Wang, H. Zimmer, M. Grosse, and A. Sorkine-Hornung, "Phase-based frame interpolation for video," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 1410–1418.

[16] E P Simoncelli and W T Freeman, "The steerable pyramid: A flexible architecture for multi-scale derivative computation," in *Proc 2nd IEEE Int'l Conf on Image Proc*, Washington, DC, Oct 23-26 1995, vol. III, pp. 444–447, IEEE Sig Proc Society.

[17] M. A. Papadopoulos, F. Zhang, D. Agrafiotis, and D. Bull, "A video texture database for perceptual compression and quality assessment," in *Image Processing (ICIP), 2015 IEEE International Conference on*, 2015, pp. 2781–2785.