SYNTHESIS OF FINE DETAILS IN B PICTURE FOR DYNAMIC TEXTURES

Uday Singh Thakur[1]*, Madhukar Bhat*[1]*, Max Bläser*[1]*, Mathias Wien*[1]*, David Bull*[2] *and Jens-Rainer Ohm*[1]

[1]Institute für Nachrichtentechnik, RWTH Aachen University, Aachen, Germany
[2]Department of Electrical and Electronic Engineering, University of Bristol, BS8 1UB, UK
{thakur, blaeser, wien, ohm}@ient.rwth-aachen.de
madhukar.bhat@rwth-aachen.de
dave.bull@bristol.ac.uk

ABSTRACT

Dynamic textures are characterized with irregular motions that are often challenging for motion compensation as applied in the state of the art video codecs. Due to rapid and randomly evolving nature of such a signal, it is accompanied with very high energy in the residual. As a result, B-pictures as used in HEVC layer are relative expensive to code. This leads to an overall increase in the bitrate. Further, increasing $\mathrm{QP}_{\mathrm{offset}}$ worsens the quality by forcing lower rate to these B-pictures, leading to strong blurring and blocking artefacts. In this paper, we exploit Steerable Pyramid (SP) for coding pictures with $\mathrm{t}_{\mathrm{id}} > 2$ in a downsampled format. At the decoder side, details are synthesized for these low resolution pictures by adding back the high frequencies using motion compensation from the nearest key picture followed by an inverse SP transform. The paper synthesizes details for the dynamic textures that are expensive to code. Our investigation shows up to 31% saving in bitrate, while visual quality is kept acceptable.

Index Terms— Steerable Pyramids, Dynamic Textures

1. INTRODUCTION

Textures are broadly 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. The latter is a time-varying visual pattern that may or may not exhibit certain temporal stationarity. Perceptually, dynamic textures can be further categorized as discrete and continuous [2]. Discrete textures have discernible appearance in motion e.g. leaves fluttering in the wind. Continuous texture is characterized as deformable media which are practically indiscernible e.g. water waves, flames from a burning material, waving field of grass seen from considerable distance etc.

Dynamic textures are extremely challenging to encode even when using the state of the art, High Efficiency Video Coding (HEVC) [3, 4]. The encoding analysis performed on homogeneous dynamic texture patches in [5] points to the fact that the majority of bits are spent on residual coding. This implies that the bits used for coding additional information such as motion vectors and mode signalling have only minor impact to the final bitrate of the encoded sequence. Continuous dynamic textures are often coded by Intra mode selection, whereas discrete dynamic textures often cause small block partitioning [5, 6]. All these facts lead us to the conclusion that the conventional approach of motion compensation is not well suited for such signals.

In the past, textures have been coded using both top-down and bottom-up approaches. The top-down approach starts from the idea that fine details inside textures are perceptually of less importance and therefore, such details can be replaced with an equivalent content that satisfies a set of statistical constraints. On the other hand, bottom-up approach take into the account the perceptual properties of the signal and allocate the distortion according to perceptual sensitivity. Texture synthesis [7] is an example of top-down approach. Pioneering works of Dumitras et al. [8] and Ndjiki-Nya et al. [9] are examples that exploited texture synthesis as a tool in video coding for compression of textures. Recently, texture synthesis has been utilized in both image and video coding [1, 11, 10]. Commonly, the texture is removed at the encoder side and synthesis parameters are sent to the decoder. Similarly, dynamic texture as an auto regressive moving average process [12] was exploited for modelling complex motions in [13, 1]. Further, an improved 2D+t autoregressive model based coding scheme for dynamic textures was proposed in [14] that showed potential savings up to 49%. Dynamic texture modelling using linear phase-shift interpolation of Complex Wavelet coefficients for B-picture synthesis was proposed in [6]. Optical flow based motion characterization of dynamic textures is proposed in [15]. However, synthesizing temporally consistent motion still remains a challenge for all the above synthesis models. For bottom-up approaches, there has been a lot of focus on static textures, while neglecting dynamic ones. An example of this in [16, 17], considers the sensitivity of each region of the scene before distributing the bitrate. Further, an adaptive $\mathrm{QP}_{\mathrm{offset}}$ selection based on enoding statistics of static and dynamic textures is proposed in [18].

In this paper we employ the Steerable Pyramid (SP) decomposition [19] for coding B-pictures in a hierarchical structure with temporal identifier $(\mathrm{t}_{\mathrm{id}}) > 2$. For such pictures only their lowpass signal (downsampled by a factor of 2 both horizontally and vertically) is coded. At the decoder side, details (high frequencies) are reconstructed back for these low resolution pictures using motion compensation of the high signal bands, from their nearest key picture, which is followed by an inverse SP transform. The paper proposes to synthesize these fine details for the dynamic textures that are expensive to code.

The paper is organized as follows: Section 2 presents the framework and details of the proposed scheme. Experimental results are then presented in Section 3. Conclusions are drawn in Section 4.

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.

2. PROPOSED FRAMEWORK

In this section we explain in detail the entire architecture of the proposed framework. We start with a brief introduction to the complex valued SP decomposition. This is followed by prediction and encoding of SP lowpass into the bitstream. In the last step, we discuss about the necessary decoder side modification followed by the post processing step i.e. the synthesis of fine details.

2.1. Steerable Pyramid Decompostion

In general for two-dimensional functions, one can separate the sinusoidal component into bands not only according to the signal frequency but also according to spatial orientation, using decomposition such as the complex-valued SP [19]. The SP filters when applied to the discrete Fourier transform of an image, they decompose the input image into a number of oriented frequency bands. The remaining frequency content which has not been captured in the pyramid levels is summarized in (real valued) high and lowpass. The SP decomposition is an invertible process i.e. it can reconstruct the true signal back from its decomposition. We have exploited these pyramids for benefitting video compression. In this paper we use only single scale and orientation.

2.2. Prediction of lowpass signal

In typical video compression configuration of HEVC, the Random Access (RA) prediction structure employs hierarchical B-picture following the coding order as indicated in Fig. 1. The coding order specifies the order in which pictures are reconstructed at the decoder and thereby defines which pictures may be used as a reference. A set of pictures comprising the coding structure is termed as group of pictures (GOP). Each picture within a GOP has a temporal level indentifier ($\mathrm{t_{id}}$) associated to it. There is no dependency of pictures with lower $\mathrm{t_{id}}$ on picture with higher $\mathrm{t_{id}}$ and therefore, it gives us the freedom to manipulate the pictures with higher $\mathrm{t_{id}}$ without affecting the key pictures. In our experiment GOP size 16 is used.

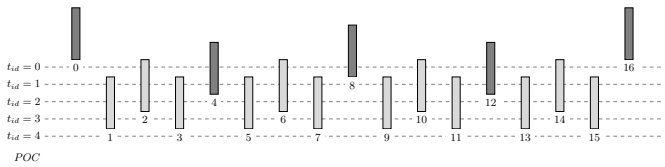

Fig. 1: Hierarchical-B coding structure with 5 temporal layers. Key pictures are indicated in dark grey shade. Light grey shade indicates pictures with $\mathrm{t_{id}} > 2$ which are coded in low resolution. $\mathrm{POC_{id}}$ below each picture shows the display order of each picture.

Statistics reveal that B-pictures are relatively expensive to code for dynamic textures due to random changing character of the signal [5, 6]. In our work, we exploit SP decompostion for B-pictures with $\mathrm{t_{id}} > 2$ as a pre-processing step before actual encoding begins. Both the decomposed high and bandpass signals are discarded from the original B-picture and only its lowpass is coded, instead of the true picture, see Fig. 2b. Our model is only proposed for the RA configuration and is implemented in two main stages. In the first stage, conventional encoding is performed for pictures with $\mathrm{t_{id}} < 3$. These are termed as key pictures which are shown in Fig. 1 with the dark grey shade. In the second stage only the lowpass signal is coded for the pictures with $\mathrm{t_{id}} > 2$. The lowpass is a downsampled signal and

(a) Original picture (b) Lowpass

Fig. 2: Lowpass of steerable pyramid to be coded for pictures with $\mathrm{t_{id}} > 2$

therefore, reference pictures are downsampled for the prediction of lowpass. No further changes are needed on the encoder side.

2.3. Reconstruction of the lowpass signal

To decode the above bitstream, modifications are needed at the decoder. Key pictures are conventionally decoded followed by SP decomposition of each reconstructed key picture to obtain its lowpass, which is stored in the picture buffer. It is required for reference in the reconstruction of pictures with $\mathrm{t_{id}} > 2$. After the reconstruction, the post-processing step is started.

2.4. Post-processing using Steerable Pyramid

The high frequency component for the pictures with $\mathrm{t_{id}} > 2$ that was discarded during the encoding is reconstructed back at this stage. It is important to reconstuct back these discarded frequencies, as otherwise, the picture would appear to be blurred. An example is shown in Fig. 3a, where the lowpass (512×512) is upsampled (without band and highpass) to the true sequence resolution i.e. (1024×1024). It is clearly visible in the figure that the upsampled picture lacks details and as a result looks blurred in comparison to the proposed method as shown in Fig. 3b. As already mentioned in Section 2.1, SP is an

(a) Upsampled (b) Proposed reconstruction

Fig. 3: Subjective comparison with and without proposed post-processing. (a) shows the upsampling of the lowpass signal without any band and highpass. In (b) fine details are synthesized after compensated band and highpass signal are added to the lowpass.

invertible transform i.e. it can perfectly reconstruct the true signal back from its SP decomposition. In order to exploit this property of SP, we first need to synthesize the highpass and bandpass signal for pictures with $\mathrm{t_{id}} > 2$.

Consecutive pictures should have a decent correlation in a video. Therefore, the probability of finding a close approximation of both

bandpass and highpass signals is maximum in the corresponding bandpass and highpass of the nearest key picture. In order to search the best matching high frequency signal in POC 1, it can be best predicted from POC 0 similarly, for POC 3 it can be best predicted from POC 4. For POC 2 we use the causal signal in the the display order i.e. POC 0 as it lies in center of two key pictures as shown in Fig. 1. The entire scheme is implemented in three main stages.

In the first stage, we decompose the key picture using SP decomposition. This separates the bandpass, highpass and lowpass of the key picture. The lowpass is further upsampled to the true sequence resolution. In the second stage, we perform block matching between the lowpass of the closest key picture $f(x, y)$ and non-key picture $g(x, y)$ e.g. lowpass of POC 0 and POC 1. This gives us integer pixel displacements $(\overline{\Delta x}, \overline{\Delta y})$. If $(\Delta x, \Delta y)$ is the true displacement, then $(\overline{\Delta x}, \overline{\Delta y})$ determined using block matching algorithm should be a good integer estimate of $(\Delta x, \Delta y)$. The block size i.e (6×6) is kept uniform throughout the entire process. We shift the image block by $\overline{\Delta x}$ pixels along x direction and $\overline{\Delta y}$ along y direction. This is followed by Taylor series approximation to refine the search, complete details are given in [20]. The method gives us sub-pixel motion vectors.

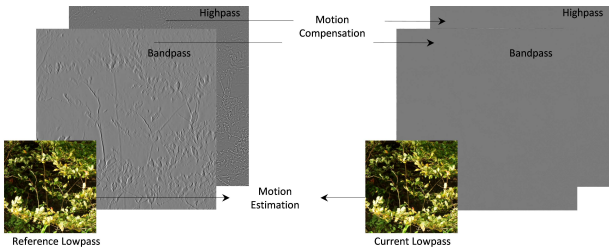

Fig. 4: Block matching is done between the lowpass pictures. Compensation is done for the corresponding band and highpass signals.

In the final stage the motion vector field is interpolated (bilinear) to the size of signal to be compensated. This means that for every pixel there is a motion vector. These motion vectors are further rounded off to the nearest integer before actual shift is started. These motion vectors now actually compensate the band and highpass signal corresponding to $f(x, y)$, to the new locations that would approximately correspond to bandpass and highpass of $g(x, y)$ as shown in Fig. 4. The compensated band and highpass signals are then added to $g(x, y)$ using inverse pyramid reconstruction, shown in Fig. 3b.

It should be noted that the estimated reconstruction is a post processing step and only applied after the loop filter. There is a possibility that the current lowpass $g(x, y)$ might be a reference picture for any other future picture. So, the previous status of the lowpass signal i.e. without post processing is retained in a buffer.

3. EXPERIMENTAL RESULTS

This section provides video synthesis results and potential video coding gains. Although quality metrics have received increasing attention over recent years, no satisfactory tool is able to assess the perceived quality of dynamic textures. For our experiment we chose 5 different UHD sequences named Balls Under Water, Lamp Leaves, Calming Water, Camp Fire Party and Fountain from the well known databases given in [21, 22]. For evaluation purposes, we cropped patches of size 1024×1024 consisting of spatially similar dynamic texture content from a true UHD source. In total, 7 such cropped sequences were used as testing material, shown in Fig. 8. The test set comprises both continuous and discrete dynamic texture patches where Lamp Leaves (Lamp Leaves-One and Lampleaves-Two) are discrete type and the rest are continuous. Temporally, the sequences have an extent of 5 seconds and consist of both 30 and 60Hz content. Tests were performed with JEM 2.0 reference software [23]. RA configuration is used with GOP size 16 and intra period half a second.

3.1. Bitrate Saving

In order to demonstrate the potential saving in bitrate, we compare the size of bitstreams between JEM 2.0 reference software using RA configuration and the proposed synthesis coding scheme at the same QPs i.e. $22, 27, 32, 37, 42$ and 47. In Fig. 5 each bar indicates the average bitrate saving on the 1024×1024 cropped patches for a specific QP using our proposed method. It is clearly visible in Fig. 5 that rate saving is highest at QP$= 22$ i.e. 31% and gradually decreases with the decreasing rate points, minimum at QP$= 47$ i.e.18%. We

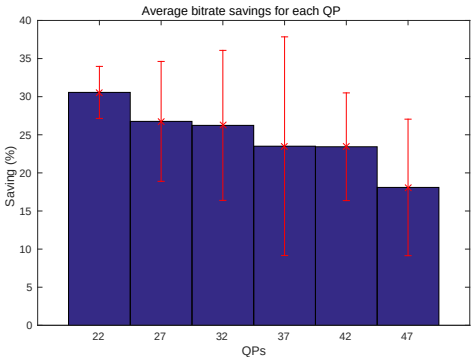

Fig. 5: Average bitrate saving on the BVI and SJTU data set for six different QPs using proposed scheme

also extended the proposed scheme to small patches of resolution 256×256 containing homogeneous dynamic texture as provided in the Hometex database [5]. We selected 86 sequences out of 120 provided in the database (some were discarded as they were static or didn't fit our motion criteria well). Rate savings for each sequence at four different QP levels is shown in Fig. 6

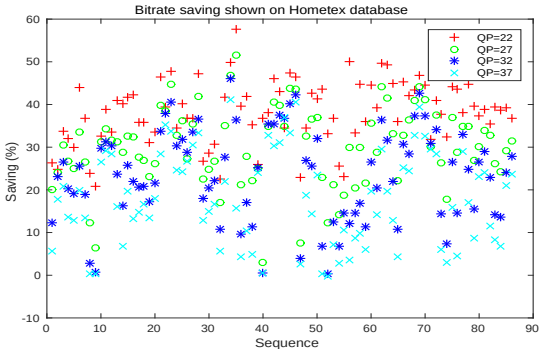

Fig. 6: Bitrate saving over four different compression levels for Hometex database using proposed scheme.

3.2. Subjective evaluation

For evaluation of the proposed model, we designed a subjective test to verify its usefulness. The subjective testing material consisted of seven source sequences. The sequences were encoded at four compression levels (q1; q2; q3 and q4) for QPs ranging in value between 22 to 37. The test employed a pairwise comparison in which the observers saw two sequences side by side and were asked to compare and select the one that they preferred.

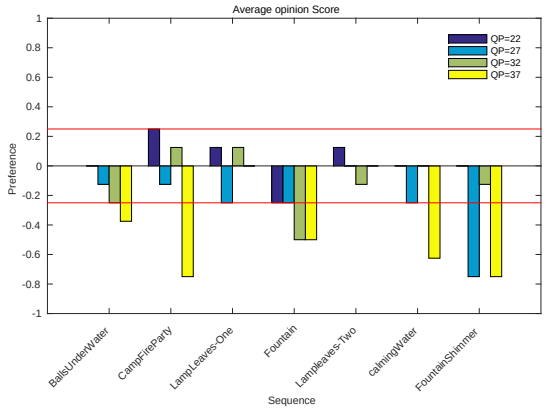

Fig. 7: Average opinion score from the pairwise test. Average preference on 21 out total 28 pairs is in the first quarter both positive and negative, showing that proposed method's quality is comparable to the standard JEM 2.0 reference software at much lower bitrate.

The observers were also allowed to select a no preference option, in-order to reduce the number of random selections. The observers could repeat and play the sequences at their choice. The proposed scheme was compared to conventional JEM RA configuration at same QPs. In total eight subjects participated in the test, their respective average responses for a particular pair is plotted in Fig. 7. Positive refers to when the proposed model is preferred and negative points shows when default JEM is preferred and 0 refers to no preference. Overall, we can see that majority of the time scores are in the first quarter either positive or negative, which means that majority of the subjects were unable to perceive quality differences between proposed versus JEM. It can be concluded from the test data that proposed method provides comparable quality on average at lower bitrate for 21 pairs out total 28 viewed.

Fig. 8: Subjective evaluation material consisting of seven different dynamic texture patches

4. CONCLUSION AND FUTURE WORK

The paper presents a novel scheme for synthesis of details in the context of dynamic texture coding. In the proposed scheme, pictures with $t_{id} > 2$ are downsampled using SP, during encoding. At the decoder these pictures are upsampled using an inverse SP transform. The fine details i.e the band and the highpass are warped from the nearest key picture using motion compensation. The paper demonstrate benefits when coding dynamic textures. The method has great potential for transmission of high frame rate videos as more frames can be sent at very low bitrate using a downsampled picture format. In our future work, we focus on synthesis of bandpass signal by analyzing neighborhood statistics of the compensated subband. Especially, where motion compensation introduces noticeable distortion such as, double contouring in the highly contrast varying regions etc. Fixing such artefacts can greatly enhance the reconstruction quality of upsampled pictures.

5. 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] R. Péteri, S. Fazekas, and M. J. Huiskes, "Dyntex: A comprehensive database of dynamic textures," *Pattern Recognition Letters*, vol. 31, no. 12, pp. 1627–1632, 2010.

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

[4] 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.

[5] M. Afonso, A. Katsenou, F. Zhang, D. Agrafiotis, and D. Bull, "Video texture analysis based on hevc encoding statistics," *Proc. of International Picture Coding Symposium PCS '16*, Dec. 2016.

[6] U. S. Thakur, K. Naser, and M. Wien, "Dynamic texture synthesis using linear phase shift interpolation," in *2016 Picture Coding Symposium (PCS)*, Dec 2016, pp. 1–5.

[7] J. Portilla and E. P. Simoncelli, "A parametric texture model based on joint statistics of complex wavelet coefficients," *International Journal of Computer Vision*, vol. 40, no. 1, pp. 49–70, 2000.

[8] 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.

[9] 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.

[10] U. S. Thakur and B. Ray, "Image coding using parametric texture synthesis," in *2016 IEEE 18th International Workshop on Multimedia Signal Processing (MMSP)*, Sept 2016, pp. 1–6.

[11] 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.

[12] G. Doretto, A. Chiuso, Y-N Wu, and S. Soatto, "Dynamic textures," *International Journal of Computer Vision*, vol. 51, no. 2, pp. 91–109, 2003.

[13] 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.

[14] F. Racape, D. Doshkov, M. Köppel, and P. Ndjiki-Nya, "2d+ t autoregressive framework for video texture completion," in *Image Processing (ICIP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 4657–4661.

[15] O. Chubach, P. Garus, and M. Wien, "Motion-based analysis and synthesis of dynamic textures," in *2016 Picture Coding Symposium (PCS)*, Dec 2016, pp. 1–5.

[16] 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.

[17] 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.

[18] M. A. Papadopoulos, F. Zhang, D. Agrafiotis, and D. Bull, "An adaptive qp offset determination method for hevc," in *2016 IEEE International Conference on Image Processing (ICIP)*, Sept 2016, pp. 4220–4224.

[19] W. T. Freeman and E. H. Adelson, "The design and use of steerable filters," *IEEE Transactions on Pattern analysis and machine intelligence*, vol. 13, no. 9, pp. 891–906, 1991.

[20] S.H. Chan, T. Q. Nguyen, et al., "Subpixel motion estimation without interpolation," in *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2010, pp. 722–725.

[21] 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.

[22] 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.

[23] Joint Video Exploration Team (JVET) of ITU-T SG 16 WP 3 and ISO/IEC JTC 1/SC 29/WG 11, "Joint exploration test model (JEM) 2.0," .