A FOVEATED SHORT TERM DISTORTION MODEL FOR PERCEPTUALLY OPTIMIZED DYNAMIC TEXTURES COMPRESSION IN HEVC

Karam Naser, Vincent Ricordel, Patrick Le Callet

A FOVEATED SHORT TERM DISTORTION MODEL FOR PERCEPTUALLY OPTIMIZED DYNAMIC TEXTURES COMPRESSION IN HEVC

Karam Naser, Vincent Ricordel and Patrick Le Callet

University of Nantes, IRCCyN UMR CNRS 6597
Polytech Nantes, Rue Christian Pauc BP 50609 44306 Nantes Cedex 3, France
karam.naser; vincent.ricordel; patrick.le-callet
@univ-nantes.fr

ABSTRACT

Due to their rapid change over time, dynamic textures represent challenging contents for the video compression standards. While several studies proposed different coding techniques, such as synthesis based coding, they mostly lack the perceptual studies on the perceived distortions.

In this paper, a framework of perceptually optimized dynamic texture compression is presented. It is based on a psychophysically driven distortion model, which is utilized inside the rate distortion loop of HEVC. The model is tested for three compression levels, and showed a significant rate saving for both training and validation sequences set.

I. INTRODUCTION

The demand for high quality video services is continuously expanding. Modern technologies enable the end-user to record, view, stream and share videos with diverse quality and resolution. Due to the limited storage and transmission capacity, these videos are often compressed to match the available rate budget. This compression would necessarily degrade the quality and may introduced undesired decrease in the user quality of experience.

The state of the art video compression standard, known as HEVC [1], has shown significant improvement over the previous (AVC), in which it can provide up to 50% bitrate saving for the same subjective quality [2]. This achievement is mainly due to better prediction mechanism as well as flexibility in the block partitioning.

Despite the high performance of HEVC, it shows weaknesses when it deals with dynamic textures. This prediction tools are not suited for this contents. It is especially true for inter prediction, in which high residual, after motion compensation, is to be encoded . In contrast, for ordinary signals, a small residual signal is yielded, and many blocks are skipped.

For this reason, coding dynamic textures represents a challenge. There has been lots of effort put to provide a better coding strategy. In particular, texture synthesis has been the prosperous alternative of the conventional coding. One of the first approaches was introduced by NdjikiNya et. al in [3]. In this approach the textured areas are synthesized, and only the synthesis parameters are sent to the decoder. more recent approaches, such as [4] and [5], are also following the same methodology.

The backbone of all of the synthesis based approach is the quality of the synthesized textures. In which, a proper metric is needed to decided precisely whether to switch to synthesis or conventional coding. The metric is also desired to work on block level, which makes it difficult to be designed, and thus, it is still unsolved problem.

On the other hand, the coding efficiency can be improved by utilizing the knowledge about the human visual perception. A large body of research was put into developing several ways to compress the videos into the class of perceptual video compression [6] [7]. An example of this, in [8] [9], is to consider the sensitivity of each region of the scene in distributing the bitrate. Most of these kind of approaches consider the static texture properties, whereas the dynamic texture are still not fully explored.

In the foveal vision, as compared to peripheral one, we are more sensitive to visual distortions. This is because of the high concentration of photo-receptors in the fovea. For this reason, the foveal distortion model considers the lower case of the allowed distortions. In other words, by optimizing the distortion for foveal vision, it is guaranteed that it would not result in negative impact if the stimuli are perceived in the peripheral vision. For this reason, the proposed model is designed as a foveated one.

Our recent investigation on dynamic textures showed that there is a large amount of perceptual redundancies, that can be exploited to provide significant saving in HEVC [10]. On the other hand, the perceived distortion of dynamic texture contents can be modeled to derive the rate-distortion mechanism in HEVC to provide a significant bitrate saving [11]. In this paper, we present a general framework for perceptually optimized dynamic texture compression. The model is parameterized via psychophysical measurement, and used inside the rate distortion loop of HEVC. It showed a significant bitrate saving for different categories of contents. Further, the model parameters were able to be predicted

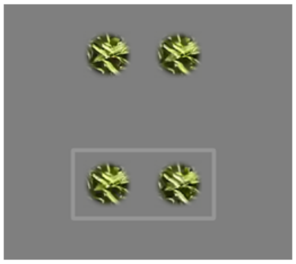

Fig. 1. Screen shot of the software used for MLDS

by linear regression using limited set of dynamic texture features.

The rest of the paper is organized as follows: Sec. II provides the details of the psychophysical test for estimating the perceived distortion. Sec. III describes the implementation of the model inside HEVC, accompanied by the bitrate saving resutls. In Sec. IV, the generalization test and results are explained. Finally, the concluding remarks as well as the future work is summarized in V.

II. HEVC PERCEPTUAL DISTORTION MODEL ESTIMATION

II-A. Method

Subjective Assessment of the perceived distortions is the most accurate way as compared to different available distortion metrics. There exist several standard methods for measuring the perceived distortion/quality which are recommended by ITU-T are listed in [12]. Beside them, there are also some other methodologies that are base on binary decision, such as pair comparison test. Generally, binary tests are more accurate and require less effort from the observers.

One of the promising methodology is known as the Maximum Likelihood Difference Scaling (MLDS [13]). It is based on comparing supra-threshold distortions by comparing two pairs of stimuli, and selecting the pair that shows a higher difference. The method have shown a good performance in the task of estimating the perceived distortions of compressed images [14], thus, we opted to use it in this work.

Adapting MLDS in this work is straightforward. The observers were presented 4 sequences, that are horizontally 1 degree of visual angle apart, and 3 degrees vertically. The observers were asked to select the pair that shows more visual difference, as compared to the other pair. The selection was done via the keyboard arrows, and by pressing ”enter” to validate the selection. A screenshot of the used software is shown in Fig. 1.

The subjective test was conducted in a professional room specifically designed for subjective testing. It complies with the ITU recommendations regarding the room lighting and screen brightness [12]. The used screen was a TVLogic LVM401 with a resolution of 1920x1080 at 60Hz. The viewing distance was 3H, where H is the screen height.

II-B. Material

The distortion model that we are seeking is to be used for video coding purpose. Knowing that the video coding standards work on small blocks, with limited access to the past and future frames, the optimal model is the one that has a very short spatio-temporal extent. However, we are limited by two factors: first, we can’t go below the foveal vision (one degree of visual angle), and second, subjects would need around 100-200 ms to fixate, which is known as the minimum fixation time.

Accordingly, we manually cropped sequences from two dynamic texture datasets, namely DynTex dataset [15] and BVI textures [16]. Overall, we collected homogeneous 43 sequences These sequences have 128x128 spatial extent (4 Coding Tree Units) and 500 ms temporal extent.

For the subjective evaluation, it is inconvenient to use all the 43 sequences, but rather a representative subset that covers the original feature space. This is to reduce the effort required to obtain the subjective results. We considered the rate-distortion behavior of the sequences as a distinguishing feature. Using HEVC reference software (HM 16.2 [17]), the sequences where encoded to 10 levels of Quantization Parameter (QP), and the Bjontegaard delta PSNR (BD-PSNR [18]) was computed between all sequences. The sequence which has the minimum sum of BD-PSNR compared with all the other sequences is considered as the reference one, and the BD-PSNR with respect to this sequence is consider as the sequence feature. Accordingly, 8 sequences were retrieved using k-means clustering algorithm ($k = 8$).

Finally, in the subjective test, the viewers were shown only the inner circle of 91 pixels diameter. Upon the end of each sequence, it was repeated with time reversal in order to avoid temporal flickering artifacts. The sequences are shown in Fig. 2. For clarity, each video was assigned to a SeqId from 1 to 8, which follows the same order as shown in the figure (from left to right, and top to bottom).

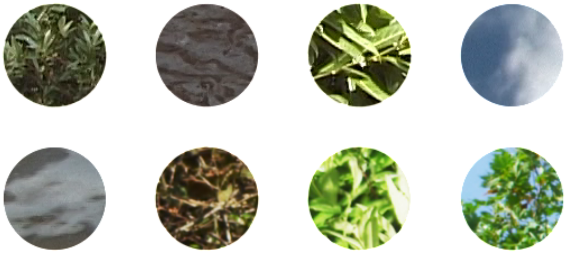

Fig. 2. Sequences used for subjective test

II-C. Results

The binary response of the observers was converted to a perceptual difference scale as described in [13]. It can be interpreted as effect of the change of a physical quantity on the perceived one. In our case, this corresponds to the average MSE and perceived difference. Examples of the results are shown in Fig. 3. The x-axes represents the overall average MSE of all the frames, whereas the y-axes represents the perceived difference. The confidence intervals are computing by learning the observers probability and repeat 10000 simulations using a boot-strapping procedure as explained in [14].

The two curves shown in Fig. 3 represent two different trends in the MSE vs perceptual difference relationship. The first trend, as for SeqId 2, shows that there is a big deviation between the measure distortion (MSE) and the perceived one. On the other hand, the second trend, which is shown for SeqId 7, indicates that MSE is directly proportional to the perceived value of distortion.

The perceptual distortion model is thus a piece-wise linear function that maps the measured distortion in MSE to the perceived distortion. This model is content dependent, but it can be predicted based on feature analysis (Sec. IV-A) and applied in HEVC to enhance the coding efficiency III.

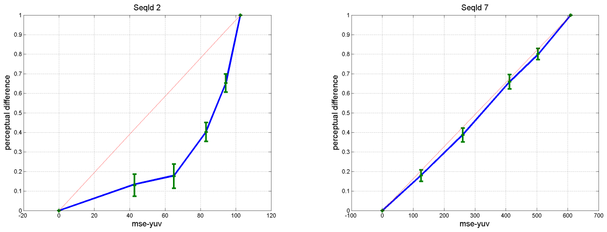

Fig. 3. Subjective test results of 2 sequences representing 2 different trends

III. MODEL APPLICATION IN VIDEO CODING

III-A. Perceptual Rate Distortion Optimization

Looking at the relationship between the measured and perceived distortions in Fig. 3, we can see that the sequences belonging to the first trend are suitable candidates for perceptual optimization. This is because the high deviation between the perceived distortion and the measured distortion can lead to wrong decisions inside the coding loop and thus would not lead to the optimal rate quality compromise.

A straightforward way to utilize the subjective test result in the video compression scenario (HEVC) is to map the distortion measure in HEVC to its perceptual value. This is achieved by using the piece-wise linear mapping functions derived from the subjective test. As HEVC encoder is using the sum of square difference (SSD), the measured MSE is mapped into a perceptual value (SSD_p) as follows:

$$SSD_p = (\alpha MSE + \beta) \times N \tag{1}$$

Where N is the number of pixels belonging to the given block. Similarly, the new Lagrangian multiplier λ can be mapped to the perceptual one λ_p as follows:

$$\begin{aligned}\lambda_p &= -\frac{\partial SSD_p}{\partial r} \\ &= (\frac{\partial SSD_p}{\partial SSD}) \times (-\frac{\partial SSD}{\partial r}) \\ &= \alpha \times \lambda\end{aligned} \tag{2}$$

where r is the bit-rate. Thus, the new λ is a scaled version of the old one.

III-B. Estimating the Bitrate Saving

Instead of only testing whether the proposed framework is more preferred, we were interested in measuring the amount of bitrate saving that it can provide. For this purpose, one needs to compare the bitrate at equal subjective quality. However, finding the subjective equality is not straightforward, but rather requires a specific psychophysical test known as threshold estimation.

For this purpose, we designed a specific psychophysical test to estimate the iso-quality points. It is simply a yes/no forced choice test, in which the observers are asked to select one of the two pairs, that is of better quality One of the pairs is the reference one with a fixed compression level, and the other is tested with different bitrates. The responses are then converted to a psychometric function of preference. The threshold of 50% preference is considered as a point of subjective equality, and the corresponding relative rate is considered as the amount of rate saving. The point of subjective equality is considered as the rate at which 50% or preference occurs. This choice of 50% is because that the subjects are not sure which pair is preferred and they simply make a random selection.

Estimating the psychometric function is a classical task in psychophysics. However, we sought an advanced test, that adaptively selects the stimuli level and requires less number of trails to converge, and thus limits the overall test duration and cost. For this reason, we used the recent psychometric estimation toolbox in [19], known as Updated Maximum Likelihood test.

III-C. Results

The optimization process in Sec. III-A was used for the specific sequences, which showed a large deviation between the measured distortion (in terms of MSE) and perceived one. These sequences, which belong to the first trend as explained in Sec. II-C, are namely defined by their sequence id: SeqId1, SeqId2, SeqId3 and SeqId8 (Fig. 2). The sequences were encoded to three quality levels: high quality (Q1), middle quality (Q2) and low quality (Q3). Using the procedure described in Sec. II, the bitrate saving at the same subjective quality is shown in Table I. One can clearly realize

SeqId	Q1	Q2	Q3	**Ave(row)**
1	12.2+-7.4	6.8+-2.2	19.2+-1.4	12.7+-3,7
2	40.4+-1.3	34.9+-1.0	20.7+-0.9	32.0+-1.1
3	36.9+-4.6	**37.3**+-5.5	33.5+-6.02	35.9+-6.0
8	13.3+-5.9	26.9+-6.3	3.8-7.4	14.6+-6.5
Ave(col)	25.7+-5.3	26.5+-3.7	19.3+-3.9	**23.8+-4.3**

Table I. Bitrate saving (%) due to perceptual optimization. +- refers to 95% confidence interval

that the proposed perceptual optimization algorithm provides a significant bitrate saving, up to 37%.

IV. GENERALIZATION OF THE PROPOSED MODEL

IV-A. Model Prediction via Feature Analysis

The results presented so far are based on optimizing sequences utilizing their measured perceived distortion profile. To generalize the perceptual optimization algorithm, the distortion model parameters need to be estimated from the sequences features. In other words, the piece-wise linear function parameters (values of α and β in Sec. II-C), must be learned from each sequence.

We used a set of few, computationally simple, features in the machine learning approach. First, we selected the spatial and temporal information (SI and TI) [20] and the colorfullness (CF) [21]. These features are often used to categorize contents in datasets. For image analysis, the gray-level co-occurrence matrix is one of the highly used features for different classification/recognition problems [22]. However, we used only its homogeneity property. On the other side, we used also some dynamic texture features, namely the curl and peakness of normal flow as defined in [23].

This set of features were used in the form of linear regression. The performance has been evaluated by the mean squared error (normalized) of leave one out cross-validation test, which has a value of 0.087. This indicates that model prediction is reasonably well.

IV-B. Generalization Test

The trained linear regression model has been used to predict the perceptual distortion model parameters of novel sequences. As explained in Sec. II-B, we have an overall of 43 dynamic texture sequences, 8 of them where used in the first experiment. For the rest of sequences (33 sequences), The trained linear regression model was used to predict their perceptual distortion model.

Among these sequences, we selected the top 24 sequences having the highest deviation between the measured and the perceived (predicted) distortion.

Using the same perceptual optimization algorithm described in Sec. III-A, we encoded these sequences also for 3 quality points (Q1,Q2 and Q3). The bitrate saving was measured subjectively, using the same psychophysical procedure in Sec. III-B. The results are shown in Fig. 4, in which the average saving of the three quality point is plotted for each sequence, defined by its sequence id (SeqId). We can clearly see that the model can provide significant bitrate saving for the majority of the sequences. However, some exceptions are also present.

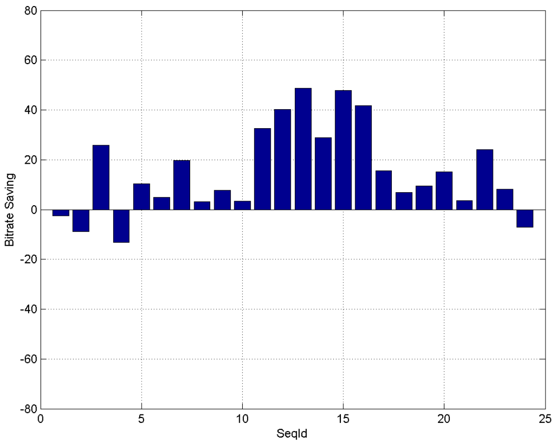

Fig. 4. Average bitrate saving for three quality points of the generalization test

V. DISCUSSION AND CONCLUSION

In this paper, a framework for perceptually optimized dynamic texture compression is presented. It is based on measuring the perceived distortions due to HEVC compression, where a specialized subjective test methodology (MLDS) was used to estimate it.

For a certain category of dynamic textures, a straightforward perceptual optimization was possible, achieving a bitrate saving up to 37%.

The distortion model parameters were shown to be predictable via a linear regression model using computationally simple features. This leaded to show the applicability of the model on novel sequences.

The advantages of the proposed optimization algorithm are the simplicity and compatibility. That is, no need for a complicated quality metric and only linear mapping of the used distortion measure is needed. In terms of compatibility, there is no change in the reference decoder, so the sequences can be directly decoded by the HEVC standard.

The future work would be to deploy the framework in a complete scene. In this case, a proper segmentation is first to be done base on the content features. For each content, the model parameters will be estimated, and used for perceptual optimization. An overall significant bitrate saving is expected.

ACKNOWLEDGEMENT

This work was supported by the Marie Sklodowska-Curie under the PROVISION (PeRceptually Optimized VIdeo

CompresSION) project bearing Grant Number 608231 and Call Identifier: FP7-PEOPLE-2013-ITN.

VI. REFERENCES

[1] Gary J Sullivan, Jens Ohm, Woo-Jin Han, and Thomas 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.

[2] J-R Ohm, Gary J Sullivan, Holger Schwarz, Thiow Keng Tan, and Thomas Wiegand, "Comparison of the coding efficiency of video coding standardsincluding high efficiency video coding (HEVC)," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 22, no. 12, pp. 1669–1684, 2012.

[3] Patrick Ndjiki-Nya, Bela Makai, Gabi Blattermann, Aljoscha Smolic, Heiko Schwarz, and Thomas 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.

[4] Marc Bosch, Fengqing Zhu, and Edward J Delp, "An overviewof texture and motion based video coding at purdue university," in *Picture Coding Symposium, 2009. PCS 2009*. IEEE, 2009, pp. 1–4.

[5] Fan Zhang and David 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] Hong Ren Wu, Amy R Reibman, Weisi Lin, Fernando Pereira, and Sheila S Hemami, "Perceptual visual signal compression and transmission," *Proceedings of the IEEE*, vol. 101, no. 9, pp. 2025–2043, 2013.

[7] Jong-Seok Lee and Touradj Ebrahimi, "Perceptual video compression: A survey," *Selected Topics in Signal Processing, IEEE Journal of*, vol. 6, no. 6, pp. 684–697, 2012.

[8] Chang Sun, Hong-Jun Wang, Hua Li, and Tai-hoon 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.

[9] Meifeng Liu and Ling 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.

[10] Karam Naser, Vincent Ricordel, and Patrick 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.

[11] Karam Naser, Vincent Ricordel, and Patrick Le Callet, "Modeling the perceptual distortion of dynamic textures and its application in hevc," *Image Processing (ICIP), 2016 IEEE International Conference on*, p. to appear, 2015.

[12] ITUR Rec, "Bt. 500-11,," *Methodology for the subjective assessment of the quality of television pictures*, vol. 22, pp. 25–34, 2002.

[13] Laurence T Maloney and Joong Nam Yang, "Maximum likelihood difference scaling," *Journal of Vision*, vol. 3, no. 8, pp. 5, 2003.

[14] Kenneth Knoblauch, Laurence T Maloney, et al., "MLDS: Maximum likelihood difference scaling in R," *Journal of Statistical Software*, vol. 25, no. 2, pp. 1–26, 2008.

[15] Renaud Péteri, Sándor Fazekas, and Mark J Huiskes, "DynTex: A comprehensive database of dynamic textures," *Pattern Recognition Letters*, vol. 31, no. 12, pp. 1627–1632, 2010.

[16] Miltiadis Alexios Papadopoulos, Fan Zhang, Dimitris Agrafiotis, and David Bull, "A video texture database for perceptual compression and quality assessment," in *Image Processing (ICIP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 2781–2785.

[17] Joint Collaborative Team on Video Coding (JCT-VC) of ITU-T SG 16 WP 3 and ISO/IEC JTC 1/SC 29/WG, " High Efficiency Video Coding (HEVC) Test Model 16 (HM 16) Encoder Description, year = 2014," Tech. Rep.

[18] Gisle Bjontegaard, "Calculation of average PSNR differences between RD-curves," *Doc. VCEG-M33 ITU-T Q6/16, Austin, TX, USA, 2-4 April 2001*, 2001.

[19] Yi Shen, Wei Dai, and Virginia M Richards, "A matlab toolbox for the efficient estimation of the psychometric function using the updated maximum-likelihood adaptive procedure," *Behavior research methods*, vol. 47, no. 1, pp. 13–26, 2015.

[20] ITU-T RECOMMENDATION, "Subjective video quality assessment methods for multimedia applications," 1999.

[21] Stefan Winkler, "Analysis of public image and video databases for quality assessment," *Selected Topics in Signal Processing, IEEE Journal of*, vol. 6, no. 6, pp. 616–625, 2012.

[22] Robert M Haralick, Karthikeyan Shanmugam, and Its' Hak Dinstein, "Textural features for image classification," *Systems, Man and Cybernetics, IEEE Transactions on*, , no. 6, pp. 610–621, 1973.

[23] Renaud Péteri and Dmitry Chetverikov, "Dynamic texture recognition using normal flow and texture regularity," in *Pattern Recognition and Image Analysis*, pp. 223–230. Springer, 2005.