Estimation of Perceptual Redundancies of HEVC Encoded Dynamic Textures

Karam Naser, Vincent Ricordel, Patrick Le Callet

Estimation of Perceptual Redundancies of HEVC Encoded Dynamic Textures

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*—Statistical redundancies have been the dominant target in the image/video compression standards. Perceptually, there exists further redundancies that can be removed to further enhance the compression efficiency.**

In this paper, we considered short term homogeneous patches that fall into the foveal vision as dynamic textures, for which a psychophysical test was used to estimate their amount of perceptual redundancies. We demonstrated the possible rate saving by utilizing these redundancies. We further designed a learning model that can precisely predict the amount of redundancies and accordingly proposed a generalized perceptual optimization framework.

***Index Terms*—Suprathreshold JND; Perceptual Video Coding; Beyond HEVC**

I. Introduction

The latest MPEG video compression standard, known as high efficiency video coding (HEVC) [1], is a hybrid video coding that utilizes both signal prediction and transform in order to provide a compact representation of the video sequences. An entropy based binary coding (CABAC [1]) is used to achieve the minimum amount of information to be stored or transmitted over channels. These mechanisms (prediction, transform and entropy coding) rely on the statistical redundancies of the input signals, such as spatial and temporal correlation. However, beside this, there are also perceptual redundancies that can be further exploited to enhance the coding performance.

It is well known that the human visual system can detect differences when a certain threshold is crossed. The just noticeable difference/distortion (JND), is the threshold at which the change of certain physical quantity causes a perceptual difference. It is of huge importance in many applications involving perceptual optimization. An example of this, in the scope of this paper, is permitting the coding system to further compress the input signal, while assuring an equivalent perceptual quality. In other words, exploiting the presented perceptual redundancies in the input signal.

Typically, JND threshold is estimated based on low level mechanisms of human vision, namely contrast sensitivity [2][3]. Such methods are of limited scope, and can poorly perform in the region of apparent distortion (suprathreshold region) as indicated in [4]. According to this, we argue that the threshold can be properly estimated from the natural image sequences themselves, taking into account computational features describing them.

Dynamic textures are specific components of the visual signals, that are characterized by high spatial and temporal homogeneity. They have been often suitable candidates for perceptual optimization of video coding as they generally possess details irrelevancies. Thus, replacing them by an equivalent stochastic signal results in significant bitrate saving [5] [6].

In this paper, we subjectively estimated the suprathreshold JND profiles of a set of 25 dynamic textures, and provided a regression model to estimate it. We demonstrate the amount of bitrate saving that can be achieved by utilizing this model.

The rest of the paper is organized as follows: In Sec. II, the details and results of the psychophysical test for redundancy estimation are described. In Sec. III, a machine learning approach for predicting the psychophysical test outcome is presented. The discussion of the obtained results with their relevance is given in Sec. IV, while the conclusion and future work are given in Sec. V.

II. Perceptual Redundancy Estimation

A. Method and Apparatus

The perceptual redundancy, as discussed in Sec. I, refers to the amount of possible further compression without altering the visual quality of the decoded video. It can be estimated using well-known psychophysical tests of threshold estimation. In contrast to the classical tests like the method of limits, method of constant stimuli or the method of adjustment [7], there exist also adaptive methods that usually converge with less number of trials. Examples of them are the famous methods like PEST [8] and QUEST [9] methods. In this work, we opted to use the state of the art method, known as Updated Maximum-Likelihood (UML) Procedure [10], which can precisely determine threshold, along with the other parameters of the psychometric function (slope and lapse rate [11]).

The UML method was used to measure the subjective preference probabilities. Given a sequence encoded by HEVC at two compression levels, the subjects would prefer one of the decoded pairs against the other with a certain probability. This probability is dependent on the relative compression

level, which is monotonically related to the relative rate (R_r) between the two levels (R_1 and R_2), where the relative rates is computed as follows:

$$R_r = (R_2 - R_1)/(R_1) \quad (1)$$

An example of the preference psychometric function is given in Fig. 1. In this figure, we can see that sequences having a large negative relative rate is not preferred (and vice-versa), which reflects the fact that negative relative rate represents lower bitrate due to higher compression, which would necessarily results in lower visual quality. An interesting and most informative point on the curve is the point of 50% preference probability. This point is the one at which no clear preference towards any of the compared pairs is present. It is commonly known as the point of subjective equality. Ideally, this point should correspond to the points where the compared videos are exactly the same, which is the point of zero relative rate, but it appears at relative rate of approximately -10% (Fig. 1). This can be interpreted in the way that the given video is perceptually equivalent to the same video being compressed at a higher compression level, namely the given video possesses a certain amount of perceptual redundancy which can be exploited to produce 10% bitrate saving.

The UML test was conducted with 25 naive observers, with normal or corrected (to normal) vision. They received written instructions on using the software as well as the task they have to perform. A screen shot of the used software is shown in Fig. 2, in which two videos are simultaneously shown, and the observer task is to select the sequence with better perceived quality.

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. The test duration was less than half an hour for all of the observers.

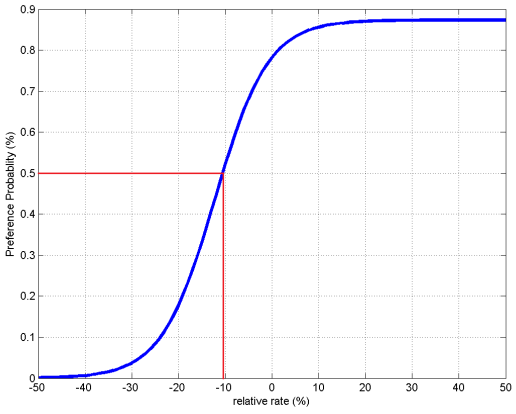

Fig. 1. An example of the measured subjective preference pychometric function. Red line represents the point of subjective equality.

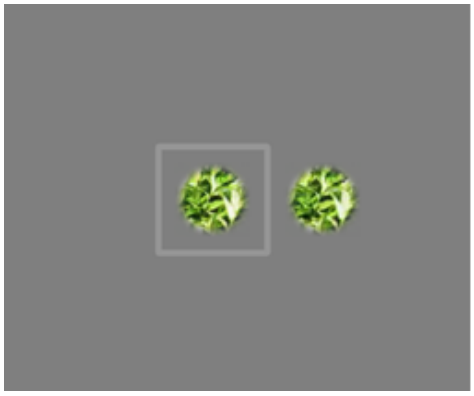

Fig. 2. Screen shot of the software used for psychophysical experiment.

B. Material

In most of the tasks involving visual quality assessment, the test materials are sequences having divergent contents, extending from 5 to 10 seconds. However, other studies of psychophysical threshold estimation uses simplistic signals with controlled properties (bars, Wavelet-Gaussian patches). In our work, which covers both perceptual quality assessment and video compression, we believe that a combination of both is required. In other words, we need to focus on natural videos, having homogeneous properties.

The main goal of the video compression standard (HEVC), is to provide the best trade-off between rate and distortion. Thus, HEVC encoder selects the best prediction mode, splitting depth and etc according to the instantaneous rate and distortion measure. The distortion computed with a limited knowledge about the spatial and temporal part of the signal. For this reason, this work concentrates on spatially small, short term stimulus, having homogeneous properties, which is referred to us as dynamic textures.

Accordingly, we collected 25 sequences from two dynamic texture datasets, namely DynTex dataset [13], and BVI dataset [14]. DynTex is a comprehensive dataset of 650 dynamic textures that have been extensively used for research purpose, while BVI is a new dataset designed mainly for subjective testing. For both datasets, the 25 collected are of resolution 128x128, with temporal extent of 500 ms, which is a duration of perceptual significance.

A circular windowed version of the sequences (as shown in Fig. 3) is used in the experiment. The window radius was chosen as 32 pixels, and the rest of the video were gradually faded to the background level using gaussian filter. This is done such that the signal falls within the foveal vision. Temporally, as the initial signal is quite short, it was repeated upon the end of the sequence with time reversal to avoid the temporal discontinuity artifact.

The sequences were compressed to 3 quality levels (good, medium, and bad) using the HEVC reference encoder (HM 16.2 [15]). This resulted in 75 source materials (SRCs), which were compared to other compression levels (HRCs) to obtain the preference probability using the UML procedure Sec. II-A

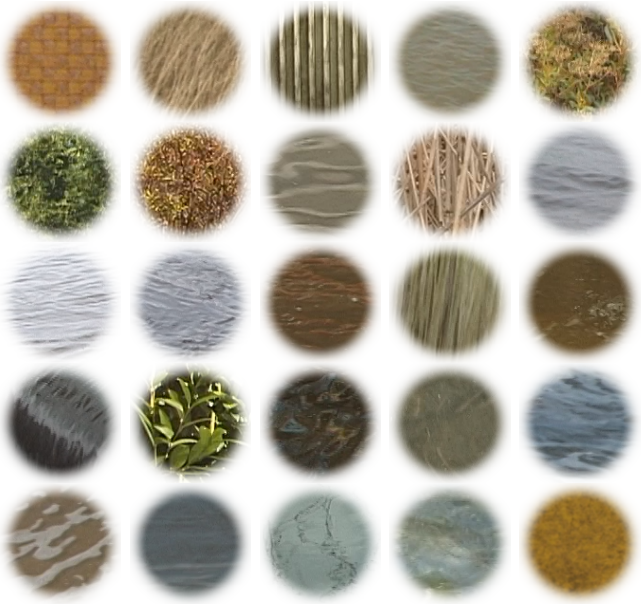

Fig. 3. Data set used in this work.

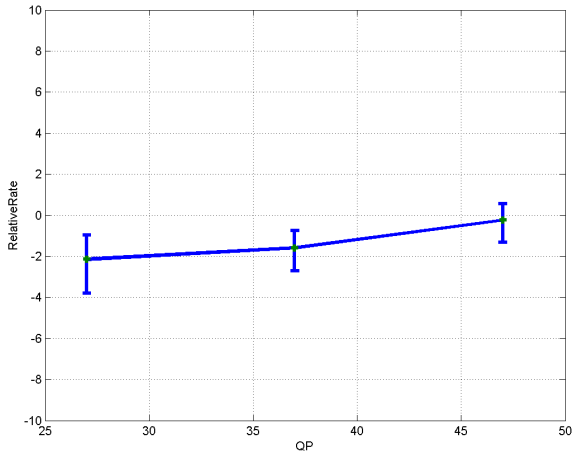

Fig. 4. An example of relative rate at equivalent subjective quality. Error bars correspond to 95% confidence interval.

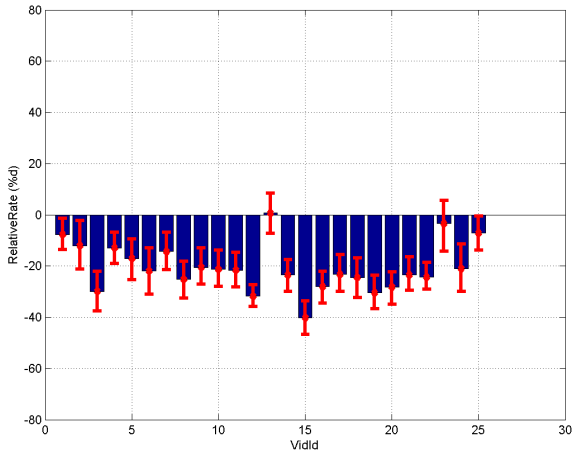

Fig. 5. Overall average relative rate of all videos. Error bars correspond to 95% confidence interval.

C. Results

The results of the psychophysical test are 75 psychometric preference functions (Fig. 1), each representing the probability of preferring a given HRC over another HRC. For each function, the threshold of 50% probability of preference is retained, which represents the point of subjective equality (Sec. II-A). An example of one sequence is shown in Fig. 4, where we can see that the redundancies to high quality region (low QP) is higher than for low quality region (large QP).

The overall average relative rate from the three quality points of all the sequences is shown in Fig. 5. We can see clearly that for most of the sequences, the corresponding subjective equality doesn't appear at the same bitrate. This clearly indicates that there are high perceptual redundancies, that can be exploited to reduce the bitrate, while maintaining an equivalent subjective equality.

III. Perceptual Redundancy Estimation via Features Analysis

In this section, we discuss the possibility of predicting the perceptual redundancy profile of a given dynamic texture. Looking again at Fig 4, we can see that two things need to be predicted. First is to predict whether there is a significant gain when the perceptual redundancies are utilized, and second is to estimate the amount of these redundancies. The first one is a binary classification problem, while the second is regression problem.

We aimed at using computationally simple features in both the classification and regression problem. Accordingly, we used the following set of descriptors: the standard Spatial Information (SI) and Temporal Information [17], The Colorfullness (CF) [18], Gray Level Cooccurrence Matrix (GLCM) [16], and the set of dynamic texture descriptors defined in [19]. The GLCM descriptor combines 4 features, that are contrast, correlation, energy and homogeneity. Similarly, the following descriptors are defined in [19] for normal flow vectors: Divergence, Curl, Peakness and Orientations. For the frame based features, such as SI and TI, we experimented different temporal pooling strategies, such as temporal mean and standard deviation.

For the binary classification problem, we found that the following set of features, accompanied by the compression level (QP) of the tested quality points, provided the required trade-of between number of features and the classification accuracy. The selected features were: the temporal mean of GLCM correlation and homogeneity, the Curl of normal flow, temporal standard deviation and temporal minimum of the Colorfulness, and finally temporal mean of GLCM homogeneity. Support Vector Machine (SVM) was used as a classification tool with

the 6 defined features, as well as QP. To test the learning performance of SVM, we performed a leave-one-out cross validation test and measured the classification accuracy, which is found to be 0.916. This indicates that binary classification works quite well. The other performance metrics results in Table I support this as well.

The other part, which is estimating the amount of perceptual redundancies, is performed using a linear regression approach. The target property of the regression model is to estimate the maximum value of distortion, measured in Peak Signal to Noise Ratio (PSNR), that the encoder can reach, without causing a perceptual difference compared to a given level of quality associated with the considered QP. This value is denoted as Max_PSNR. Using the set of features mentioned earlier, we experimentally found that the following subset of features are the most significant ones: QP, maximum of SI, mean of TI, standard deviation of GLCM Homogeneity, Energy, Contrast and Correlation, mean of GLCM Homogeneity and Contrast. Once more, the leave-one-out cross validation process was employed, and obtained R-Squared value of 0.96, which also indicates also goodness of the trained model (see Table II).

IV. Discussion

Dynamic textures, possess a certain amount of perceptual redundancies that depend on both the signal characteristics and the compression level. Utilizing these redundancies, a large supplementary coding gain can be obtained. Predicting whether a significant gain can be produced, as well as estimating the amount of perceptual fidelity, can be easily done using computationally simple features.

According to the achieved results, we proposed a general framework for perceptual optimization of the video compression standard (HEVC). The block diagram of the framework is given in Fig. 6. In this framework, the input video signal is analyzed and spatio-temporal features are computed (Sec. III), next, the binary classifier decides whether a significant coding gain can be obtained. If the condition is true, the linear regression module estimates the maximum possible distortion level, and finally the encoder increases the compression ratio accordingly.

V. Conclusion

In this paper, the suprathreshold perceptual distortion artifacts were estimated for 25 homogeneous spatio-temporal patches, referred as dynamic textures. For a given sequence, it was shown that subjective equality between two patches can occur at different bitrates. Exploiting this difference, we can achieve significant bitrate saving.

The threshold can be precisely estimated using a linear regression model. The model combines low level computationally simple set of features. According to this, a perceptual optimization framework of HEVC encoding has been proposed, as shown in Fig. 6.

The possible future outcome of this work is to apply the proposed approach on compound scenes, such that the perceptual redundancies of each spatio-temporal patch is well exploited. Higher overall bitrate saving, compared to the base line encoder (HEVC), is expected.

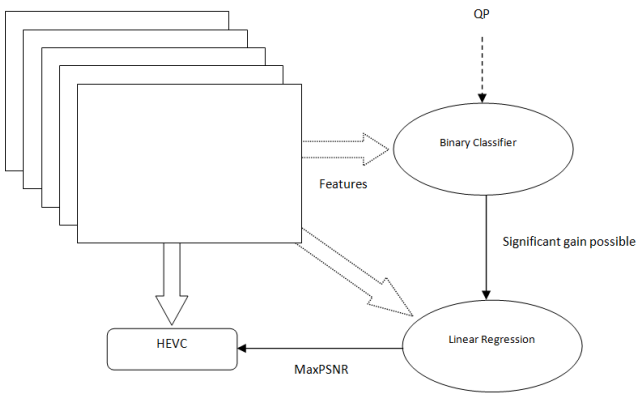

Fig. 6. Proposed perceptual optimization framework of HEVC Encoding.

Acknowledgment

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.

References

[1] G. J. Sullivan, J. 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.

[2] A. B. Watson, "Dctune: A technique for visual optimization of dct quantization matrices for individual images," in *Sid International Symposium Digest of Technical Papers*, vol. 24. SOCIETY FOR INFORMATION DISPLAY, 1993, pp. 946–946.

[3] Z. Wei and K. N. Ngan, "Spatio-temporal just noticeable distortion profile for grey scale image/video in dct domain," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 19, no. 3, pp. 337–346, 2009.

[4] H. Wu and D. Tan, "Subjective and objective picture assessment at suprathreshold levels," in *Picture Coding Symposium (PCS), 2015*. IEEE, 2015, pp. 312–316.

[5] J. Balle, 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.

[6] 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.

[7] P. G. Engeldrum, *Psychometric scaling: a toolkit for imaging systems development*. Imcotek press, 2000.

[8] M. Taylor and C. D. Creelman, "Pest: Efficient estimates on probability functions," *The Journal of the Acoustical Society of America*, vol. 41, no. 4A, pp. 782–787, 1967.

[9] A. B. Watson and D. G. Pelli, "Quest: A bayesian adaptive psychometric method," *Perception & psychophysics*, vol. 33, no. 2, pp. 113–120, 1983.

[10] Y. Shen, W. Dai, and V. 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.

[11] F. A. Wichmann and N. J. Hill, "The psychometric function: I. fitting, sampling, and goodness of fit," *Perception & psychophysics*, vol. 63, no. 8, pp. 1293–1313, 2001.

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

Classification Accuracy	Sensitivity	Specificity	Area Under Curve (ROC)	Precision
0.92	1.00	0.83	0.87	0.86

TABLE I

PERFORMANCE METRICS OF THE SVM BINARY CLASSIFIER, USING LEAVE-ONE-OUT VALIDATION PROCEDURE

Mean Squared Error	Mean Absolute Error	R2
0.002	0.038	0.961

TABLE II

PERFORMANCE METRICS OF THE LINEAR REGRESSION PROCESS, USING LEAVE-ONE-OUT VALIDATION PROCEDURE

[13] 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.

[14] 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*. IEEE, 2015, pp. 2781–2785.

[15] 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.

[16] R. M. Haralick, K. Shanmugam, and I. H. Dinstein, "Textural features for image classification," *Systems, Man and Cybernetics, IEEE Transactions on*, no. 6, pp. 610–621, 1973.

[17] T. Installations and L. Line, "Subjective video quality assessment methods for multimedia applications," *Networks*, vol. 910, p. 37, 1999.

[18] S. 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.

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