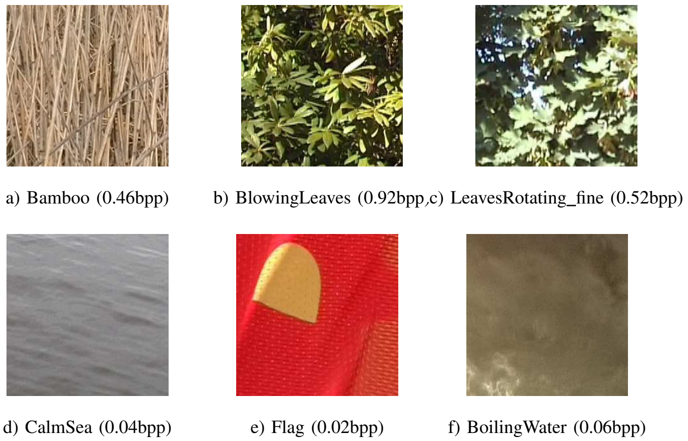
Predicting Video Rate-Distortion Curves using Textural Features

Angeliki V. Katsenou, Mariana Afonso, Dimitris Agrafiotis and David R. Bull
Department of Electrical & Electronic Engineering, University of Bristol, BS1 8UB, UK
{Angeliki.Katsenou, Mariana.Afonso, D.Agrafiotis, Dave.Bull}@bristol.ac.uk

***Abstract*—This work addresses the problem of predicting the compression efficiency of a video codec solely from features extracted from uncompressed content. Towards this goal, we have used a database of videos of homogeneous texture and extracted both spatial and frequency domain features. The videos are encoded using *High Efficiency Video Coding* (HEVC) reference codec at different quantization scales and their *Rate-Distortion* (RD) curves are modelled using linear regression. Using the extracted features and the fitted parameters of the RD model, a *Support Vector Regression Model* (SVRM) is trained to learn the relationship of the textural features with the RD curves. The SVRM is tested using iterative five-fold cross-validation. The presented experimental results demonstrate that RD curve characteristics can be predicted based on the textural features of the uncompressed videos, which offers potential benefits for encoder optimization.**



a) Bamboo (0.46bpp) b) BlowingLeaves (0.92bpp,c) LeavesRotating_fine (0.52bpp)

d) CalmSea (0.04bpp) e) Flag (0.02bpp) f) BoilingWater (0.06bpp)

Fig. 1: Average required *bits per pixel* (bpp) for the example sequences [4].

I. Introduction

Texture perception is a well-studied topic in psychophysics, neurobiology and signal processing. In images, texture is an important visual primitive, defined as a spatially repetitive pattern [1, 2]. In video sequences, texture becomes stochastic with variations in both its spatial and temporal characteristics. Analysing and understanding texture is important for video compression, as how a codec deals with texture content has a significant effect on its *Rate-Distortion* (RD) performance. Videos that contain dynamic textures in particular (e.g. water, moving tree leaves) pose a significant challenge even to the most recent video coding standard, *High Efficiency Video Coding reference codec* (HEVC HM) [3], requiring many bits for a good quality reconstruction. Moreover different types of dynamic texture exhibit different bit rate requirements. Dynamic textures with irregular local motion, for example, generate a higher bit rate compared to other types of dynamic texture when coded using the same *Quantization Parameter* (QP). This is illustrated in Fig. 1, where a random frame from six example sequences [4] is shown along with the average number of bits generated per pixel, when each sequence is coded with HEVC HM at a QP of 27.

This paper presents a method of predicting the encoding difficulty of texturally homogeneous video sequences based on their textural features. First, texture-related features and their statistics in the spatial, temporal and frequency domain are are defined and extracted from uncompressed videos. For the same videos, the RD curves are obtained using HEVC HM for various QPs. A big part of these RD points, along with the extracted features of the videos, are used to train *Support Vector Regression Models* (SVRMs). Based on these trained models, the RD curve of the video may be predicted prior to encoding. This offers the potential for efficient parameter selection in the video encoder. Furthermore, the extracted features, along with the predicted RD curves, could be used for texture classification and segmentation. Finally, the proposed approach could be useful for the annotation of a dynamic texture video database with numerical descriptors instead of textual annotations that are currently used (e.g. [5]).

Textural features are conventionally defined with the purpose of facilitating similarity, browsing, retrieval and classification applications [2, 6–12]. Additionally, most works have only considered static textures, namely images [2, 6–9]. Hence, most textural features do not capture the dynamic characteristics that texture obtains in videos. Some of these features have previously been used for spatial segmentation in video synthesis and coding [10, 11]. An effort to synergise spatial and temporal texture features in video (but only for classification purposes) is reported in [12]. In the present paper, features that capture both spatial and temporal characteristics of different types of texture in videos are extracted and correlated with their compression efficiency. Furthermore, a feature that captures the temporal coherence of successive frames is introduced.

There exists some previous work that relates textural features to video compression efficiency [9, 13, 14]. In [9], Subedar et al. define a no-reference metric of granularity in static textured images and discuss its relation to compression efficiency, but with no clear association to RD curves. In [14], elementary statistics of prediction error for texture and motion vectors for motion are obtained from H.264/AVC encoder and used to build variability-distortion models. In [13], a block-based spatial correlation model is defined and used to predict the RD bounds within an H.264/AVC encoder. This work was extended in [15] to consider the block-based spatial correlation among two successive frames within HEVC HM. The present work moves beyond the state of the art by predicting the RD curves of video sequences with homogeneous texture based on textural features that are extracted from the uncompressed sequences.

The remainder of the paper is organised as follows. Section II describes the selected features of texture in videos and the new dataset. In Section III, the RD curves are modelled and their basic characteristics are discussed. The proposed SVRM to predict the compression efficiency based on the video content is detailed in Section IV, where also the experimental results are presented and discussed. Finally, the conclusions are drawn in Section V.

II. Video Texture Features

The selected features are designed to capture the basic characteristics of video texture: coarseness, directionality, regularity, and temporal stationarity.

A. Gray Level Co-occurrence Matrix (GLCM)

The *Gray Level Co-occurrence Matrix* (GLCM)[6] is a commonly used spatial textural feature [10]. It expresses the intensity contrast of neighbouring pixels in a frame, thus capturing the degree of coarseness and directionality of the texture.

For the present frame I_t, let G be the GLCM, whose element G_{ij} is the number of occurrences for pixel pair ij with intensity values Y_i, Y_j, with $Y \in \{0, 255\}$. The probability that a pixel pair ij assumes Y_i, Y_j values is $p_{ij} = G_{ij}/K$, where K is the number of occurrences. GLCM has five main descriptors: contrast, correlation, energy (or uniformity), homogeneity and entropy that are formally defined in the equations below:

$$GLCM_{\text{contrast}} = \sum_{i=1}^{M}\sum_{j=1}^{N}(i-j)^2 p_{ij}, \tag{1}$$

$$GLCM_{\text{correlation}} = \sum_{i=1}^{M}\sum_{j=1}^{N}\frac{(i-m_r)(j-m_c)p_{ij}}{\sigma_r\sigma_c}, \tag{2}$$

$$GLCM_{\text{energy}} = \sum_{i=1}^{M}\sum_{j=1}^{N}p_{ij}^2, \tag{3}$$

$$GLCM_{\text{homogeneity}} = \frac{p_{ij}}{1+|i-j|}, \tag{4}$$

$$GLCM_{\text{entropy}} = -\sum_{i=1}^{M}\sum_{j=1}^{N}p_{ij}\log_2 p_{ij}, \tag{5}$$

where M, N are the rows and columns dimensions respectively, m_r, m_c the mean and σ_r, σ_c the standard deviation along rows and columns of both I_t and G. All GLCM descriptors are computed at a frame level per sequence. Then, all features are averaged over the number of frames of a sequence.

B. Normalized Cross-Correlation (NCC)

The Normalized Cross-Correlation (NCC) is commonly used in image processing applications for spatial similarity purposes [16]. It assumes values within the range [-1,1] with its maximum value indicating the maximum correlation and vice versa. In this paper, NCC is used as a spatio-temporal feature, as it examines the spatial similarity of two successive frames, I_{t-1} and I_t, using a sliding matching template window T of $w \times w$ size from the reference frame I_{t-1}:

$$NCC = \frac{\sum_{i=1}^{M}\sum_{j=1}^{N}|I_t(i,j) - \bar{I}_t(u,v)||T(i-u,i-v) - \bar{T}|)}{\sqrt{\left(\sum_{i=1}^{M}\sum_{j=1}^{N}|I_t(i,j) - \bar{I}_t(u,v)|\right)^2\left(\sum_{i=1}^{M}\sum_{j=1}^{N}|T(i-u,i-v) - \bar{T}|\right)^2}} \tag{6}$$

where u, v define the area covered by the window T. NCC is recorded for every pair of successive frames and its statistics (standard deviation, skewness, kurtosis and entropy) are computed at a sequence level.

C. Temporal Coherence (TC)

In order to express how easy or difficult one frame can be predicted from its previous temporal neighbour, the spectral magnitude coherence among two successive frames $I_{t-1}I_t$, is employed. It is computed using the *Fast Fourier Transform* (FFT) [17] and is defined as follows:

$$TC = \frac{|P_{I_{t-1}I_t}|^2}{P_{I_{t-1}I_{t-1}}P_{I_tI_t}}, \tag{7}$$

where $P_{I_{t-1}}$ is the auto-spectral density of I_{t-1} and $P_{I_{t-1}I_t}$ the cross-spectral density of frames $I_{t-1}I_t$. TC is normalized within the range [0,1] and assumes its maximum value for static or purely translational motion among two successive frames. Due to its dependence on the frequency response, it is evident that the TC for textures with high density of high frequencies and low motion will have higher TC values (e.g. Bamboo) compared to textures that are less dense in terms of high frequencies (e.g. CalmSea). TC is computed for every pair of successive frames and its statistics (standard deviation, skewness, kurtosis and entropy) are computed at a sequence level.

D. Dataset and Feature Extraction

In order to extract our features, the new dataset, ***Hom**ogeneous Video **Tex**ture Dataset* (HomTex) [4], has been created. This comprises 256 × 256 cropped homogeneous regions from the DynTex [5] and the BVI video dataset [18]. DynTex is a database with annotated videos (original resolution is 720 × 576 at 25 fps) of different types of dynamic textures (with a variation form weakly to strongly dynamic). The BVI video dataset contains textured natural scenes (originally with HD resolution and 60 fps). The 120 selected videos contain spatially homogeneous texture of different types. Particularly, 25 static, 45 dynamic continuous and 50 dynamic discrete. The type of the texture has been manually annotated by experts. The term “dynamic continuous” refers to scenes of moving deformable surfaces (e.g. water, flag), while “dynamic discrete” refers to perspectively moving structures (e.g. straws, leaves). As “static” we characterize the texture videos that have almost exclusively global motion. The granularity level in the videos varies in three levels, from fine to coarse. Moreover, in some sequences, except from the local motion, global motion also occurs due to the camera motion.

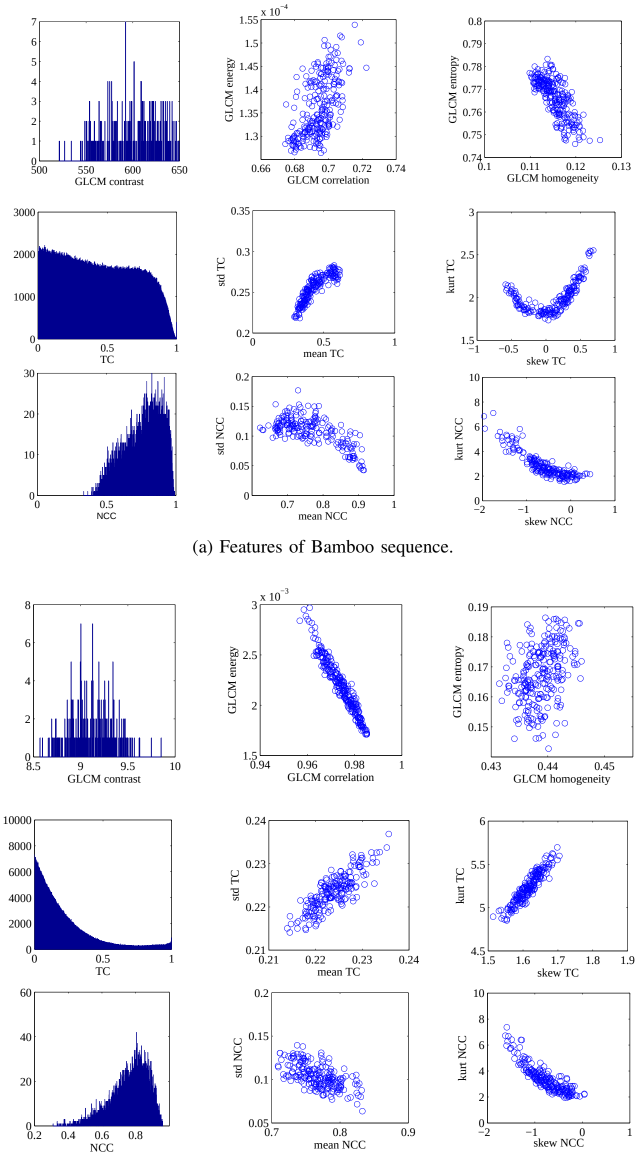
Figure 2 depicts the selected features for two of the example sequences, Bamboo and CalmSea. Bamboo is a dynamic texture with fast and irregular moving structures, while CalmSea is dynamic surface with slow irregular motion. As expected, their features are quite different. For example, Bamboo has a high range of GLCM values and higher TC and NCC, compared to CalmSea. Also, the scattered features reveal different relations for these two different types of texture. It is important to mention that there many variations of these plots depending mainly on the texture granularity and the motion variation that affects the temporal coherence.

III. Rate-Distortion Curves

In order to plot the RD curves, the same video sequences used for feature extraction were encoded for five different quantization scales, $QP = \{20, 25, 27, 32, 37\}$, using the Random Access configuration of HM16.2. The RD curves characterize the relationship of the mean bit rate and the video quality. The RD curves are highly dependent on the video content. Particularly, motion and texture are important aspects that influence video compression efficiency. These factors result in RD curves of different characteristics, as illustrated in Fig. 3 (a), where the RD curves for five different quantization parameters of the example sequences from HomTex are drawn. Particularly, “CalmSea”, “Flag”, and “BoilingWater” represent dynamic continuous texture, while the rest example sequences represent dynamic discrete texture. As it can be observed, these different types of textures result in RD curves with different characteristics. For example, the “CalmSea” sequence, which is annotated as dynamic continuous texture, represents slow motion water and its curve shows that using a lower QP value results in a high *Peak Signal to Noise Ratio* (PSNR) value for a rather small bit rate increase. On the other hand, the “BlowingLeaves” sequence, which is annotated as dynamic discrete texture, depicts irregularly moving tree leaves, and its RD curve shows that many bits are required to provide high video quality.

A. RD Curve Modelling

In this paragraph, it is shown that the RD curves can be approximated linearly, if the logarithm of the rate is used. In Fig. 3 (b), the horizontal axis represents the logarithm of the bit rate R, $\log_{10}(R)$. Since *Pearson linear Correlation Coefficient* (PCC) over the test sequences has a mean value of 0.9845, this shows a strong linear correlation of PSNR and $\log_{10}(R)$. Hence, the relation of PSNR and $\log_{10}(R)$ can be modelled as a linear function. Thus,



(a) Features of Bamboo sequence.

(b) Features of CalmSea sequence.

Fig. 2: Examples of extracted features for two example sequences, (a) Bamboo and (b) CalmSea. The first row illustrates GLCM with their descriptors; contrast in a histogram and correlation, energy, homogeneity and entropy in scatter plots. Rows two and three, depict TC and NCC in histograms with their statistics in scatter plots, respectively.

$$\text{PSNR} = \alpha \log_{10}(R) + \beta \ , \tag{8}$$

where $\alpha \in \mathbb{R}^+$ and $\beta \in \mathbb{R}$. Ordinary a least squares fit is used for the computation of parameters α, β.

To assess the validity of the linear model, the *Bjontegaard delta PSNR* (BDPSNR) measurement for all sequences was used [19]. Figure 3 (c) depicts a histogram of the computed BDPSNR values for all HomTex video sequences. The average BDPSNR value equals to 0.0873 dB. Also, as it is obvious from Fig. 3 (c), the distribution around the mean value is narrow. This means that there is a small deviation between the

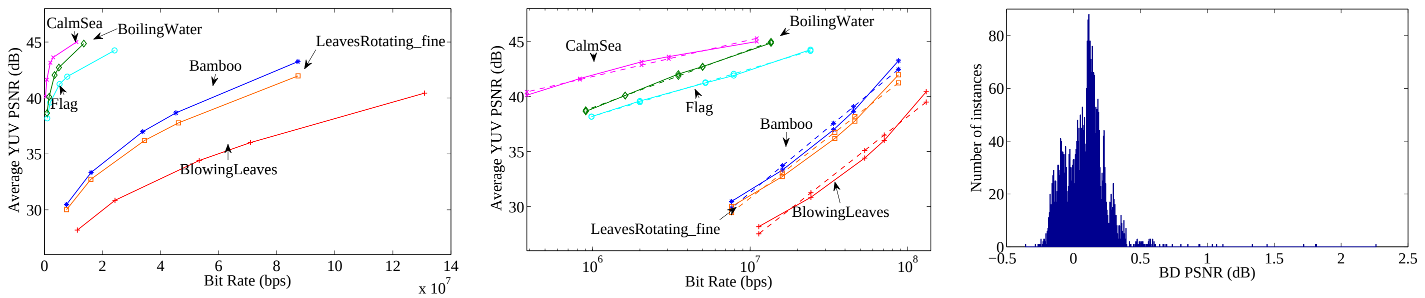


Fig. 3: RD Curves of the example sequences. (a) RD curves based on measurements in HEVC HM. (b) The continuous curves are the same as in (a), but with $log_{10}(R)$ x-axis. The dashed lines depict the linearly fitted curves. (c) Histogram of the BDPSNR for all linearly modelled RD curves.

estimated PSNR values using Eq. (8) and the measured PSNR values. All the aforementioned justify the linear approximation of the curves.

IV. Prediction of the RD curves using SVRM

A. Training and Testing

120 test sequences were used, each with 250 frames and *Group of Pictures* (GOP) length equal to 8. The RD curves were built per sequence and per GOP resulting in 3600 RD curves. The RD curves were obtained by encoding the test sequences for five different quantization scales, QP $=\{20, 25, 27, 32, 37\}$. For all the RD curves, parameters α and β were fitted.

To ensure the validity of the training and testing a repetitive randomized process was followed. A random split of the data (including both extracted features and fitted parameters α, β per sequence and per GOP) was performed with 70% of the data being used for model configuration and training and the remaining 30% of the data for the final prediction and performance evaluation. Regarding the training process, a five-fold cross-validation was used. Parameters α and β were consodered independent and two different regression models were built. Each time, four groups were used to train the SVRM to predict parameter α and the other SVRM to predict parameter β. Then, the respective SVRM is used to predict parameters $\hat{\alpha}$ and $\hat{\beta}$ of the RD curves in the remaining fifth group. The partitioning, training, and final prediction are conducted iteratively 100 times to ensure the validity of the SVRM training and the accuracy of predictions. The LIBSVM ToolBox was used to build the regression models and the radial basis function was adopted as the SVRM kernel [20].

B. Results and Discussion

The evaluation of the performance of the proposed approach takes place in two steps. First, the accuracy of predicting parameters $\hat{\alpha}$ and $\hat{\beta}$ is assessed. Next, the RD curves based on the predicted $\hat{\alpha}$ and $\hat{\beta}$ are validated using BDPSNR.

In Table I, the mean values and standard deviations of *Normalised Root MSE* (NRMSE), PCC and *Spearman's Rank Correlation Coefficient* (SROCC) for the predicted parameters $\hat{\alpha}$ and $\hat{\beta}$ are reported. The prediction accuracy for both parameters is similar. Also, the predicted values of parameters $\hat{\alpha}$ and $\hat{\beta}$ are scattered over the fitted values of α and β in Fig. 4.

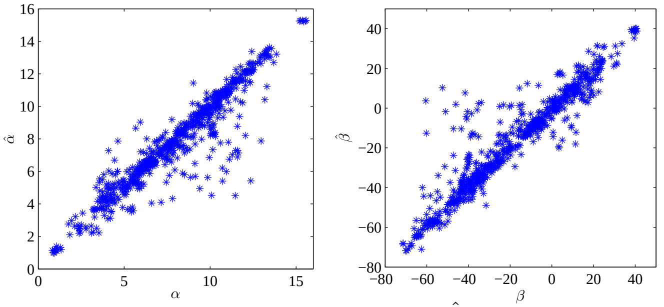


Fig. 4: Scattering of predicted parameters $\hat{\alpha}$ and $\hat{\beta}$ versus the fitted ones.

Both parameters show that most of the predicted values that lie on or very close to the diagonal are predicted with high accuracy, which is also inferred by the mean PCC values in Table I. However, some predicted $\hat{\alpha}$ and $\hat{\beta}$ deviate from the diagonal in Fig. 4 and consequently lead to deviations in the predicted RD curves.

TABLE I: Goodness of prediction of parameters $\hat{\alpha}, \hat{\beta}$ using mean±standard deviation values over all tested sequences.

Parameter	NRMSE	PCC	SROCC
$\hat{\alpha}$	0.3335±0.0198	0.7939±0.0194	0.7700±0.0266
$\hat{\beta}$	0.3094±0.0202	0.8154±0.0180	0.7984±0.0251

A visual example of the predicted curves for the six example sequences is provided in Fig. 5 (a). From this example, it is clear that for all example sequences, the predicted RD curves are close to the encoder-derived RD curves for the used range of QP values. As it can be seen, for some sequences the predicted curves are more accurate, as for example for CalmSea. This can be explained by the fact that the spatial diversity of this sequence is low, as also indicated by its GLCM statistics in Fig. 2 (b). On the other hand, Bamboo has irregularly moving objects in the foreground and a still background, which is also reflected by its GLCM statistics in Fig. 2 (a).

The accuracy of the predicted RD curves compared to the measured RD curves is measured using BDPSNR for the tested data and a random instance of the testing is depicted in the histogram of Fig. 5 (b). The mean BDPSNR value over all predictions is -0.1187 dB with a standard deviation of 3.7012 dB. Regarding the outliers in the prediction, one reason is

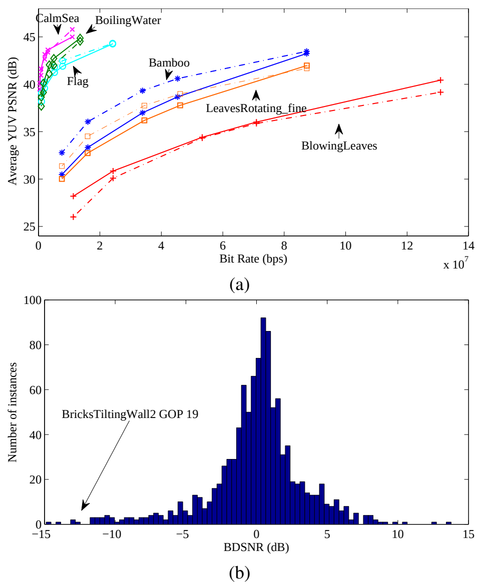


Fig. 5: Predicted RD curves for the example sequences. (a) Predicted (dashed lines) and measured curves (continuous lines). (b) Histogram of the BDPSNR of the predicted RD curves over the measured ones.

that, as in all regression methods, the finite number of test sequences does not cover the infinite number of RD curves and all possible values of the different textural features. Also, the accuracy of the predicted curves depends on two predicted parameters, $\hat{\alpha}$ and $\hat{\beta}$. This is the case for the annotated point in Fig. 5. For example, for the RD curve of GOP19 of the sequence BricksTilting_wall2 [1], although the prediction of $\hat{\alpha}$ is accurate, the prediction of $\hat{\beta}$ deviates from its fitted value, resulting in a RD curve that has a very low BDPSNR value compared to the measured one. This means that a deviated prediction of both parameters might result in outliers. Also, for some outliers, although their textural content was perceived as homogeneous by experts, it is spatially quite variable. The spatial variability affects the values of the extracted features, thus the accuracy of the predicted RD curves.

V. Conclusion

A novel method of predicting the RD curves based on features extracted from uncompressed video sequences with homogeneous texture has been proposed. Two SVRMs have been trained to learn the underlying relationship between the textural features and the RD curve parameters. Based on experimental results, the proposed features are related to the difficulty to encode different types of texture in videos and the regression models perform well in predicting the RD curve parameters. The predicted RD curves are close to the measured, with a mean BDPSNR value equal to -0.1187 dB. The proposed method offers the benefit of a means of prediction of video compression performance and could be used to support encoder configuration and prior to encoding adaptive rate-quality optimization. Future work will focus on extending the dataset and by including more textural features related to directionality and regularity of the texture. Also, the extension in the prediction of the RD curves for videos with content with mixed textures is another interesting direction for future research.

Acknowledgement

The work presented was supported by the “Marie Skłodowska-Curie Actions - The EU Framework Programme for Research” project PROVISION.

References

[1] B. S. Manjunath, J. R. Ohm, V. V. Vasudevan, and A. Yamada, “Color and texture descriptors,” *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 11, no. 6, pp. 703–715, Jun 2001.

[2] P. Salembier and T. Sikora, *Introduction to MPEG-7: Multimedia Content Description Interface*, John Wiley and Sons, Inc., New York, NY, USA, 2002.

[3] G. J. Sullivan, J. R. Ohm, W. J. Han, and T. Wiegand, “Overview of the High Efficiency Video Coding (HEVC) Standard,” *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 22, no. 12, pp. 1649–1668, Dec 2012.

[4] M. Afonso, A. Katsenou, F. Zhang, D. Agrafiotis, and D. R. Bull, “Homogeneous Video Texture Dataset (HomTex),” 2016, https://data.bris.ac.uk/data/datasets/1h2kpxmxdhccf1gbi2pmvga6qp/.

[5] P. Renaud, S. Fazekas, and M. J. Huiskes, “DynTex : a Comprehensive Database of Dynamic Textures,” *Pattern Recognition Letters*, http://projects.cwi.nl/dyntex/.

[6] R. M. Haralick, K. Shanmugam, and I. Dinstein, “Textural features for image classification,” *IEEE Trans. on Systems, Man, and Cybernetics*, vol. SMC-3, no. 6, pp. 610–621, Nov 1973.

[7] R. M. Haralick, “Statistical and structural approaches to texture,” *Proceedings of the IEEE*, vol. 67, no. 5, pp. 786–804, May 1979.

[8] J. Zujovic, T. N. Pappas, and D. L. Neuhoff, “Structural Texture Similarity Metrics for Image Analysis and Retrieval,” *IEEE Trans. on Image Processing*, vol. 22, no. 7, pp. 2545–2558, July 2013.

[9] M. M. Subedar and L. J. Karam, “A no reference texture granularity index and application to visual media compression,” in *2015 IEEE Intern. Conf. on Image Processing (ICIP)*, Sept 2015, pp. 760–764.

[10] M. Bosch, F. Zhu, and E. J. Delp, “Segmentation-Based Video Compression Using Texture and Motion Models,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 5, no. 7, pp. 1366–1377, Nov 2011.

[11] F. Zhang and D. R. Bull, “A Parametric Framework for Video Compression Using Region-Based Texture Models,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 5, no. 7, Nov 2011.

[12] C. H. Peh and L. F. Cheong, “Synergizing spatial and temporal texture,” *IEEE Trans. on Image Processing*, vol. 11, no. 10, pp. 1179–1191, Oct 2002.

[13] J. Hu and J. D. Gibson, “New rate distortion bounds for natural videos based on a texture-dependent correlation model,” *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 19, no. 8, pp. 1081–1094, Aug 2009.

[14] G. Van der Auwera, M. Reisslein, and L. J. Karam, “Video texture and motion based modeling of rate variability-distortion (vd) curves,” *IEEE Trans. on Broadcasting*, vol. 53, no. 3, pp. 637–648, Sept 2007.

[15] J. Hu, M. Bhaskaranand, and J. D. Gibson, “Rate distortion lower bounds for video sources and the HEVC standard,” in *Information Theory and Applications Workshop (ITA), 2013*, Feb 2013, pp. 1–10.

[16] J. P. Lewis, “Fast template matching,” in *Vision interface*, 1995, vol. 95, pp. 15–19.

[17] G. Carter, C. Knapp, and A. Nuttall, “Statistics of the estimate of the magnitute-coherence function,” *IEEE Trans. on Audio and Electroacoustics*, vol. 21, no. 4, pp. 388–389, Aug 1973.

[18] M. A. Papadopoulos, F. Zhang, D. Agrafiotis, and D. R. Bull, “BVI video texture database,” http://data.bris.ac.uk/data/dataset/1if54ya4xpph81fbo1gkpk5kk4, 2015.

[19] G. Bjontegaard, “Calculation of average PSNR differences between RD-curves,” Tech. Rep., 13th VCEGM33 Meeting, Austin. TX, 2001.

[20] C. C. Chang and C. J. Lin, “LIBSVM: A library for support vector machines,” *ACM Trans. on Intelligent Systems and Technology*, vol. 2, pp. 27:1–27:27, 2011.

[1] BricksTilting_wall2 is a static sequence with the camera tilting over a brick wall.