OPEN ACCESS

University of BRISTOL

Afonso, M., Katsenou, A., Zhang, A., Agrafiotis, D., & Bull, D. (2017). Video texture analysis based on HEVC encoding statistics. In Picture Coding Symposium (PCS), 2016. Institute of Electrical and Electronics Engineers (IEEE). DOI: 10.1109/PCS.2016.7906312

Peer reviewed version

Link to published version (if available):
10.1109/PCS.2016.7906312

Link to publication record in Explore Bristol Research
PDF-document

This is the author accepted manuscript (AAM). The final published version (version of record) is available online via IEEE at http://ieeexplore.ieee.org/document/7906312/. Please refer to any applicable terms of use of the publisher.

University of Bristol - Explore Bristol Research

General rights

This document is made available in accordance with publisher policies. Please cite only the published version using the reference above. Full terms of use are available:
http://www.bristol.ac.uk/pure/about/ebr-terms.html

Video Texture Analysis based on HEVC Encoding Statistics

Mariana Afonso, Angeliki Katsenou, Fan Zhang, Dimitris Agrafiotis, David Bull
Department of Electrical and Electronic Engineering, University of Bristol, BS8 1UB, UK
{Mariana.Afonso,Angeliki.Katsenou,Fan.Zhang,D.Agrafiotis,Dave.Bull}@bristol.ac.uk

***Abstract*—In this paper, we present an extensive study on encoding statistics of videos with different texture types. These statistics were extracted from HEVC test model (HM), and include, among others, mode selection, partitioning, motion vectors and bitrate allocation. For this study, a new dataset of homogeneous static and dynamic video textures, HomTex, is proposed. A comprehensive analysis of the results revealed a significant variability of coding statistics within dynamic textures, suggesting that this category should be further split into two relevant subcategories, continuous dynamic textures and discrete dynamic textures. This case was then supported by an unsupervised learning approach on the statistics extracted. Finally, following the results obtained, some suggestions of improvements in texture coding are presented.**

I. Introduction

The latest video coding standard, High Efficiency Video Coding (HEVC) has achieved significant gains compared to its predecessor thanks to a number of incremental improvements to the hybrid video coding model including larger block sizes and more flexible partitioning [1]. While this model already considers some perceptual characteristics of human vision, namely in the selection of the transform for residual coding, increased gains could be achieved by further exploiting texture masking. In the context of video compression, textures are typically categorized into two different classes: static and dynamic [2–6]. However, the classification of dynamic textures lacks consistency and is usually very broad, with a large range of diverse content being included in the same class, such as water, leaves, smoke and trees. This might not be efficient when trying to apply and optimize coding strategies, such as texture synthesis [3–6], which to the best of the authors' knowledge, treat all dynamic textures equally. This motivates us to examine more closely how HEVC encoding performs with different texture types as well as to investigate and propose more robust definitions of texture classes.

In particular, this paper makes the case for classifying textures in three types for coding purposes, namely: static textures, continuous dynamic textures and discrete dynamic textures. Multiple contributions concerning the classification of dynamic textures have been recently published [7, 8]. However, their aim is to classify textures based on semantic content. That is not always useful for coding. The distinction between continuous and discrete dynamic textures has already been made by the authors of [9] when annotating the DynTex database. In order to characterize these texture types and to support our case, we performed a detailed analysis of a large number of HEVC encoding statistics using a homogeneous video texture dataset. A homogeneous dataset allows the statistics to be directly associated to a single texture without the need for segmentation. A number of papers have recently included analysis of encoding statistics [10–13]. However, the number of statistics examined was very limited and the objective was not specifically texture coding.

The first contribution of this work is a dataset of homogeneous video textures that is fully available online[1]. An important feature of this dataset is that it is representative of broadcasting video content, as suggested by the methodology proposed by [14] (Section II). The second contribution of this paper is to offer a better understanding of the performance of HEVC on different texture classes. This assists in identifying potential improvements that could be introduced in upcoming standards. The final contribution is a characterization of different types of textures based on their encoding statistics, helping to build a better texture taxonomy.

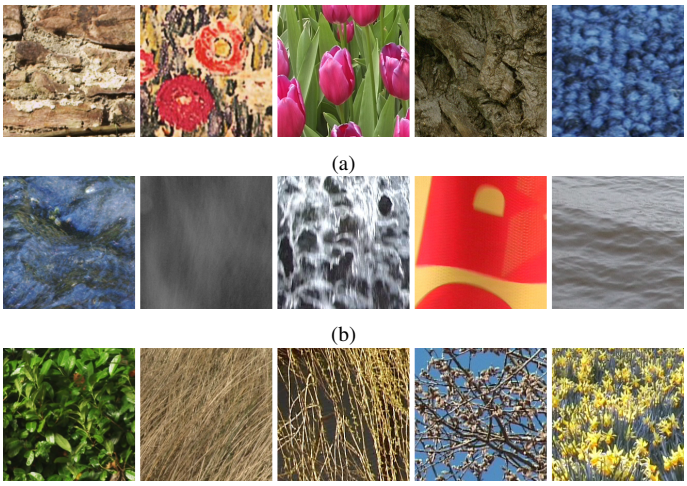
The rest of the paper is organized as follows. Section II describes the proposed dataset, including how it was annotated. Section III presents the experimental design and the statistics extracted from coding the proposed dataset with HEVCs test model (HM). Section IV describes the analysis performed and discusses the results obtained. Finally, conclusions are presented in Section IV-D.

II. A Homogeneous Texture Dataset

Due to the lack of a dataset for both static and dynamic homogeneous textures, a new dataset, the **Hom**ogeneous Video **Tex**ture Dataset (HomTex), has been developed. In the context of this dataset, homogeneity means that the textures are spatially and temporally consistent. It comprises 120 sequences with a spatial resolution of 256×256 pixels. The low resolution was chosen considering the size of the dataset to keep the encoding time manageable. Most of the sequences were obtained by cropping those from two existing datasets: DynTex [9] and the BVI Texture dataset [2]. DynTex contains a large variety of dynamic texture content in over 650 PAL resolution sequences and has been extensively used by the research community as a benchmark for dynamic texture classification and for synthesis. The BVI texture dataset was recently developed and contains both static and dynamic textures at HD resolution.

The developed HomTex dataset was manually annotated by experts considering three different visual characteristics:

[1] https://data.bris.ac.uk/data/dataset/1h2kpxmxdhccf1gbi2pmvga6qp



(a) (b) (c)

Fig. 1: Examples of sequences from HomTex, classified as: a) static textures, b) continuous dynamic textures, c) discrete dynamic textures.

TABLE I: Annotation of the HomTex based on three characteristics.

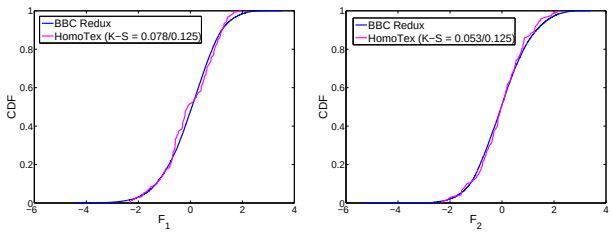
Dynamics	Structure	Granularity	Num. of sequences	Total
static	continuous	high	0	25
		medium	1	
		low	4	
	discrete	high	2	
		medium	6	
		low	12	
dynamic	continuous	high	4	45
		medium	18	
		low	23	
	discrete	high	9	50
		medium	32	
		low	9	

dynamics, structure and granularity. A brief description of each characteristic is presented below:

- **Dynamics:** whether the texture has inherent motion (dynamic) or the texture is still and the only motion present is due to a moving camera (static).
- **Structure:** whether the nature of the texture is one of continuous deformable media (continuous) or the texture is composed of a collection of structured discernible parts (discrete) [9].
- **Granularity:** related to the size of the smallest recognizable repetitive object observed, known as a texture primitive. A texture with smaller size primitives has a high granularity level, while a texture with a larger size primitives has a low granularity level [15].

The number of sequences grouped according to these characteristics is presented in Table I. Moreover, Fig. 1 shows the first frame of some examples of sequences from HomTex under the three major categories defined.

Further characterisation of the HomTex dataset was performed using the methodology of [14], which offers a way of measuring how well a given dataset reflects the characteristics of broadcast consumer video. The dataset was parameterised using low level features, which were then transformed into orthogonal factors. The frequency distribution for the most relevant factors, Naturalness and Movement, was then compared with that of a large-scale database containing modern broadcast content, BBC Redux [16]. These distributions are shown as Cumulative Distribution Functions (CDF) in Fig. 2. Then, the hypothesis that both distributions could come from the same continuous distribution was vallidated using the two-sample Kolmogorov-Smirnov test [17] (shown in Fig. reffig:distribution). These results indicate that HomTex is indeed representative of the characteristics found in broadcast video content.



(a) Factor 1 - Naturalness (b) Factor 2 - Movement

Fig. 2: Distribution comparison between BBC Redux and HomTex.

III. Encoding Statistics

HEVC encoding statistics were extracted using the test model version HM16.2. All the sequences from the HomTex dataset were encoded using the Main profile and three configurations: Random Access, Low Delay and All Intra. The initial quantization parameter (QP) was set to five commonly used values 22, 25, 27, 32 and 37. A total of 37 statistics were obtained from the encoding process at the Coding Tree Unit (CTU) level. These were then post-processed to obtain features per sequence, for various QP values and frame types (I, B and P). For the purpose of this work, P frames are defined as using only past frames for reference and B frames as using both past and future frames. The encoding analyser proposed in [18] was used as the basis for the code that was written to extract the HM statistics.

Table II shows a summary of all the statistics that were extracted grouped into different categories along with a short description of each. For the measure of correlation between the original and the residual frames, only the luminance component of the frames (L_1, L_2) was considered. The 2-D Pearson product-moment correlation coefficient was used, which is computed by the following equation [17]:

$$r = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n}(L_{1i,j} - \overline{L_1})(L_{2i,j} - \overline{L_2})}{\sqrt{(\sum_{i=1}^{m}\sum_{j=1}^{n}(L_{1i,j} - \overline{L_1})^2)(\sum_{i=1}^{m}\sum_{j=1}^{n}(L_{2i,j} - \overline{L_2})^2)}}$$

where (i, j) are the pixels coordinates of the frames, $m \times n$ is the spatial resolution, $\overline{L_1} = \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n} L_{1i,j}$ and $\overline{L_2} = \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n} L_{2i,j}$.

IV. Analysis and Results

The analysis performed and the results obtained are presented in four sub-sections. Section IV-A describes the observations made through examination of the distribution of the statistics of the different classes. Section IV-B analyses the effect that the different levels of granularity have on the encoding

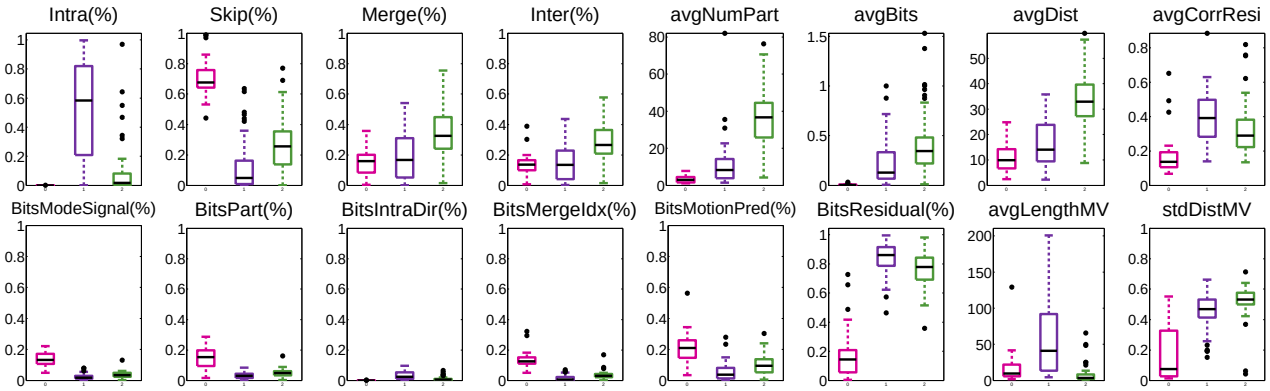


Fig. 3: Distribution of several encoding statistics for the static (magenta), continuous dynamic (purple) and discrete dynamic (green) textures. These results were obtained using Random Access configuration and a QP value of 25.

TABLE II: Statistics extracted from HM during the encoding process.

Category	Statistics	Description
Prediction modes	intra (%) stdIntra Skip (%) stdSkip merge (%) stdMerge inter (%) stdInter	Percentage and standard deviation of the percentage of Coding Unit (CU), scaled by size and predicted by each mode
Reference indexes	ref0 (%) ref1 (%) ref2 (%) ref3 (%)	Percentage of CU (scaled by size) that use each order of image from the reference picture list
Partitioning	avgPart stdPart	Average and standard deviation of the number of partitions per CTU
Bits	avgBits stdBits	Average and standard deviation of bits per pixel
Distortion	avgDist stdDist	Average and standard deviation of distortion (SAD) per pixel
Bit allocation	bitsModeSignal (%) bitsPart (%) bitsIntraDir (%) bitsMergeIdx (%) bitsMotionPred (%) bitsResidual (%) bitsOthers (%)	Percentage of bits spent to encode mode selection, partitioning, intra modes, merge indexes, residual coding and others
Residual Statistics	avgMSEresi stdMSEresi avgMSERecError stdMSERecError avgCorrResi stdCorrResi avgCorrCodedResi stdCorrCodedResi	Average and standard deviation of the MSE of the residual (original - predicted frame), the MSE of the reconstructed frame, correlation between original and residual frames, correlation between residual and coded residual (reconstructed - predicted frame)
Intra mode	DCIntra PlanarIntra avgIntraDir stdIntraDir	Percentage of Intra predicted CU that use DC and planar mode. Average and standard deviation of the Intra mode direction
Motion Vectors	avgLengthMV stdDistMV	Average length of the motion vectors. Standard deviation of the distribution of motion vector's directions

behaviour. Section IV-C presents a clustering analysis that was employed with the aim of finding the inherent structure of the data. Finally, Section IV-D makes some suggestions for areas of further work that have the potential to offer compression performance improvements.

A. *Effect of dynamics and structure on the encoding behaviour*

Herein we split our data into the three classes mentioned previously (static, continuous dynamic and discrete dynamic), which are based on the dynamics and structure characteristics of the textures present. In particular, we compute texture-class specific distributions for each of the extracted statistics with the aim of identifying texture-class related patterns in the behaviour of the encoder. Figure 3 depicts the distributions of a sub-sample of the statistics for the B frames of the Random Access configuration, using a QP value of 25. Having examined the results fully, a number of observations were made. These are described in detail bellow:

The prediction modes selected per CU vary significantly for different texture types. Unsurprisingly, static textures are associated with a high percentage of Skip mode due to the simplicity of the motion present (camera panning or zooming). Continuous dynamic textures are mostly coded using Intra mode, implying that the motion compensation fails to produce a good predictor for these textures. Discrete dynamic textures, on the other hand, exhibit distinct motion and are mainly coded using motion compensation, i.e. using all modes except Intra.

The Rate-Distortion (RD) performance of the encoder varies with texture type. Again, as expected, static textures require a smaller amount of bits to encode compared to dynamic textures and exhibit less distortion (SAD of residual) for the same QP. Of more interest is the fact that discrete dynamic textures, on average, require a higher bitrate compared to continuous textures and result in higher distortion.

The number of CTU partitions varies with texture type, with the lowest number of partitions being observed for static textures (median of 4 partitions per CTU) and the highest for discrete dynamic textures (median of 35 partitions per CTU). As for continuous dynamic textures, they require few partitions (median of 9 partitions per CTU). In general, if the CTUs are highly split, then the content is more likely to be finely textured and have fewer spatially regular regions.

The bit allocation for both types of dynamic textures shows that the majority of bits are spent on residual coding (more than 80% of the bits generated). This implies that the bits used for coding additional information such as motion vectors and mode signalling are nearly irrelevant to the final bitrate of the encoded sequence. This leads to the conclusion that the residual for dynamic textures typically exhibits very high energy, which is further confirmed by the high distortion and

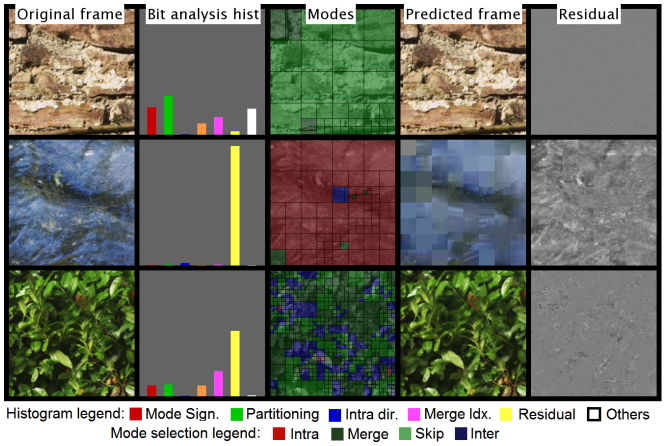


Fig. 4: Visualization of some encoding statistics of three different sequences.

high correlation between the original frame and the residual. In contrast, for static textures, the bit allocation is more evenly distributed among the control data, the motion data and the residual, reflecting the ease in achieving better predictions.

Motion vectors exhibit different characteristics for different texture types. Static textures are associated with small magnitude motion vectors that show directional consistency. Discrete dynamic textures have generally small magnitude motion vectors with high directional irregularity. In contrast, continuous dynamic textures are associated with large magnitude motion vectors with slightly less irregularity in directions.

Figure 4 presents a graphical visualization of a portion of the extracted statistics for three sequences, corresponding to a different texture class: BricksTiling (static), CalmingWater (continuous dynamic) and LampLeaves (discrete dynamic). The results shown are representative of the patterns identified for each texture type, including the dominance of Skip mode for static textures and Intra mode for continuous dynamic textures, the high partitioning of discrete dynamic textures and the high energy in the residual of dynamic textures.

B. *Effect of granularity on the encoding behaviour*

In order to determine how granularity affects encoding, a study of pairs of sequences with different granularities was conducted. For intra frame coding, it was found that higher texture granularity leads to higher average bitrates. This is explained by the fact that textures with higher granularity result in a large number of high frequency coefficients in the residual. For inter frame coding, where the temporal aspect of the texture plays a significant role, changes in texture granularity have a more distinct effect on encoding behaviour. For static and discrete dynamic textures, higher granularities result in decreased bitrates, whereas for continuous dynamic textures, the bitrate tends to increase slightly for higher granularities. The former result (static and discrete dynamic textures) is down to the lower amplitude of the motions in the scene, which results in a more efficient motion estimation, thus decreasing the percentage of CUs coded using Intra mode. In the case of continuous dynamic textures, the prevalence of the intra coding mode leads to similar observations being made as with the case of intra frame coding.

TABLE III: Clustering performance of several clustering algorithms on all the statistics extracted from HM for a QP of 25 and Random Access configuration.

Config.	Clustering method	Silhouette	Purity	NMI	ARI
Random Access	K-Means	0.45	0.84	0.56	0.57
	Hierarchical Clustering	0.36	0.80	0.51	0.48
	Spectral Clustering	0.44	0.84	0.57	0.56
	PCA + K-Means	0.45	0.84	0.56	0.57

C. *Validation of class descriptions through clustering*

The previous section demonstrated that there are significant differences between the encoding statistics of static, continuous dynamic and discrete dynamic textures. However, this outcome alone is insufficient proof that these three classes represent the optimal classification for coding. For this reason, an unsupervised learning technique, clustering, was applied to all the extracted statistics. It is important to note that in clustering analysis, the sequences are not labelled with a class. This ensures that the clusters are obtained solely based on the features themselves, i.e. the encoding statistics.

Three widely used clustering algorithms were employed, K-Means, Hierarchical Agglomerative clustering (using complete linkage) and Spectral clustering [19]. Additionally, Principal Component Analysis (PCA) was also applied for feature extraction, followed by K-Means using the 10 first principal components (80% of the total variance). The input features were 33 statistics of the Random Access B frames at a QP value of 25. Other settings were also tested, and have achieved similar findings. Since the majority of the algorithms used require the number of clusters to be pre-determined, a simple analysis was conducted using a plot of the sum of within-cluster distances. The number of clusters is often decided by the ”elbow” of this plot, which in this case, hints at the existence of three clusters.

The clustering performance evaluation was conducted using four popular cluster validity indices; one internal index (the average Silhouette) and three external indexes (Purity, Normalized Mutual Information (NMI) and the Adjusted Rand Index (ARI)) [19]. The later methods used the manual class annotations mentioned in previous sections with a values of 0 and 1 representing a random and a perfect clustering, respectively. The clustering performance is presented in Table III and indicates that using three clusters is a good representation of the data, since most indices have high values. Additionally, it can be noted that the performance is consistent among the different algorithms employed.

Given the large number of features, visualization of the data is a challenge. One way to do this is to represent the data as an undirected graph where each sequence is a node. The edges that connect two nodes of the graph consist of the n nearest neighbors of each sequence, based on the Euclidean distance (the value of n was chosen to be 15). This representation is depicted in Fig. 5, where static, continuous dynamic and discrete dynamic textures are distinguished by the colors magenta, purple and green, respectively. Looking at the representation, it is possible to identify three distinct clusters in the data in addition to a number of centrally located nodes. These nodes correspond to textures that do not fit into

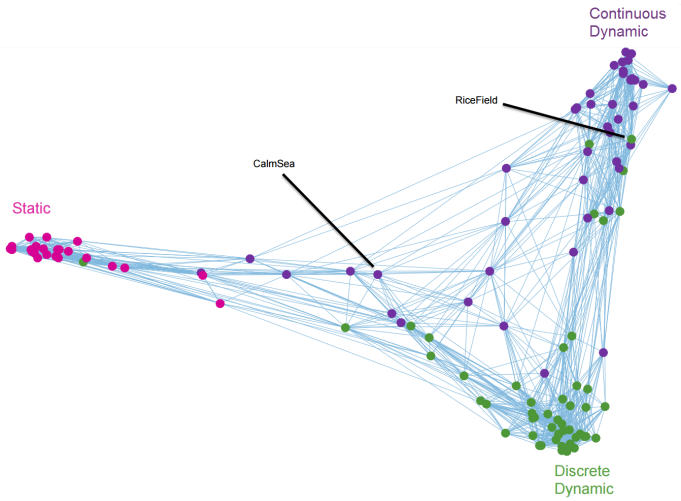


Fig. 5: Undirected graph representation of the sequences.

only one class due to its characteristics. An example is the *CalmSea* sequence, which depicts water flowing very slowly, making it easier to apply motion compensation and bringing it closer to a static texture for some of the statistics such as *avgBits* and *Skip (%)*. Additionally, for some sequences the manual class annotations do not agree with the ones of its nearest neighbours. One example is the *RiceField* sequence, which although discrete in nature, is treated as a continuous dynamic texture by the encoder due to its fine structure and motion patterns.

D. Suggested areas of improvements in texture coding

Based on the conducted analysis and the results presented in the previous sections, it is clear that dynamic textures represent a challenging problem for the compression model of HEVC. In the case of discrete dynamic textures, better local motion estimation could lead to more accurate temporal predictions, reducing the energy of the residual. On the other hand, for continuous dynamic textures, a more robust motion compensation technique may not be sufficient due to the random nature of this type of texture. However, approaches such as texture synthesis may prove helpful in reducing the resulting bitrate by allowing the generation of content without the need to code any residual. Alternatively, given that for dynamic textures the majority of bits are spent on residual coding, future work could focus on developing texture-adaptive residual coding techniques.

V. Conclusion

This paper has presented an extensive and detailed study of different texture types from the perspective of the encoding statistics of HEVC. Since textures represent an important part of video content, knowledge about how HEVC handles textures is key to creating a solid foundation for the development of innovative tools to be added to upcoming standards.

Additionally, the proposed dataset of homogeneous video textures, HomTex, will aid future research by providing a large variety of homogeneous textures for testing new coding approaches. This dataset was proven to be representative of broadcast video content and is annotated considering three different characteristics: dynamics, structure and granularity.

Finally, a clustering analysis of the encoding statistics revealed a clear structure in the data, with a separation between static, continuous dynamic textures and discrete dynamic textures. Following this analysis, some suggestions of improvement were identified for each class, which will serve as a basis for future work.

Acknowledgement

This work was supported by the Marie Skodowska-Curie Actions - FP7 EU programme, project PROVISION ITN.

References

[1] 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, 2012.

[2] M. A. Papadopoulos, F. Zhang, D. Agrafiotis, and D. Bull, "A video texture database for perceptual compression and quality assessment," in *IEEE Inter. Conf. on Image Processing*. IEEE, 2015, pp. 2781–2785.

[3] 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, pp. 1378–1392, 2011.

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

[5] 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, 2011.

[6] P. Ndjiki-Nya, T. Hinz, and T. Wiegand, "Generic and robust video coding with texture analysis and synthesis," in *IEEE Inter. Conf. on Multimedia and Expo*. IEEE, 2007, pp. 1447–1450.

[7] F. Yang, G.-S. Xia, G. Liu, L. Zhang, and X. Huang, "Dynamic texture recognition by aggregating spatial and temporal features via ensemble svms," *Neurocomputing*, vol. 173, pp. 1310–1321, 2016.

[8] Y. Sun, Y. Xu, and Y. Quan, "Characterizing dynamic textures with space-time lacunarity analysis," in *IEEE Inter. Conf. on Multimedia and Expo*. IEEE, 2015, pp. 1–6.

[9] 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.

[10] M. A. Papadopoulos, F. Zhang, D. Agrafiotis, and D. Bull, "An adaptive qp offset determination method for HEVC," in *IEEE Inter. Conf. on Image Processing*. IEEE, 2016.

[11] F. Zhang and D. R. Bull, "An adaptive lagrange multiplier determination method for rate-distortion optimisation in hybrid video codecs," in *2015 IEEE Inter. Conf. on Image Processing*. IEEE, 2015, pp. 671–675.

[12] J. Lei, S. Li, C. Zhu, M.-T. Sun, and C. Hou, "Depth coding based on depth-texture motion and structure similarities," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 25, no. 2, pp. 275–286, 2015.

[13] J. Stankowski, T. Grajek, D. Karwowski, K. Klimaszewski, O. Stankiewicz, K. Wegner, and M. Domański, "Analysis of frame partitioning in HEVC," in *Computer Vision and Graphics*. Springer, 2014, pp. 602–609.

[14] F. M. Moss, F. Zhang, R. Baddeley, and D. Bull, "What's on TV: A large scale quantitative characterisation of modern broadcast video content," in *IEEE Inter. Conf. on Image Processing*. IEEE, 2016.

[15] M. M. Subedar and L. J. Karam, "A no reference texture granularity index and application to visual media compression," in *IEEE Inter. Conf. on Image Processing*. IEEE, 2015, pp. 760–764.

[16] B. Butterworth, "History of the BBC redux project," *BBC Internet Blog*, 2008.

[17] F. J. Massey Jr, "The Kolmogorov-Smirnov test for goodness of fit," *Journal of the American statistical Association*, vol. 46, no. 253, pp. 68–78, 1951.

[18] D. Springer, W. Schnurrer, A. Weinlich, A. Heindel, J. Seiler, and A. Kaup, "Open source HEVC analyzer for rapid prototyping (HARP)," in *IEEE Inter. Conf. on Image Processing*, Paris, France, October 2014.

[19] S. Theodoridis and K. Koutroumbas, *Pattern Recognition, Third Edition*. Orlando, FL, USA: Academic Press, Inc., 2006.