PERCEPTUALLY-ALIGNED FRAME RATE SELECTION USING SPATIO-TEMPORAL FEATURES

Angeliki V. Katsenou, Di Ma, and David R. Bull

Department of Electrical and Electronic Engineering, University of Bristol, Bristol BS8 1UB, UK
{Angeliki.Katsenou, Di.Ma, Dave.Bull}@bristol.ac.uk

ABSTRACT

During recent years, the standardisation committees on video compression and broadcast formats have worked on extending practical video frame rates up to 120 frames per second. Generally, increased video frame rates have been shown to improve immersion, but at the cost of higher bit rates. Taking into consideration that the benefits of high frame rates are content dependent, a decision mechanism that recommends the appropriate frame rate for the specific content would provide benefits prior to compression and transmission. Furthermore, this decision mechanism must take account of the perceived video quality. The proposed method extracts and selects suitable spatio-temporal features and uses a supervised machine learning technique to build a model that is able to predict, with high accuracy, the lowest frame rate for which the perceived video quality is indistinguishable from that of video at the acquisition frame rate. The results show that it is a promising tool for prior to compression and delivery processing of videos, such as content-aware frame rate adaptation.

***Index Terms*—** High frame rate, frame rate selection, perceptual video quality, spatio-temporal features.

I. INTRODUCTION

Apart from the semantics of the content, the other key factors of visual experiences that influence immersion are the spatial resolution, the dynamic range and the frame rate of the video. While significant work has been reported on dynamic range extension, less has been reported on the influence of frame rate, which is the focus of this paper. Although there has been an increase in the availability of 4K video at 60 frames per second (fps), the demand of higher resolutions up to 8K adds pressure to further increase frame rate [1, 2]. High frame rates are important as they improve perceptual video quality and reduce the visibility of motion artefacts (e.g. motion blur, aliasing) [3–6]. However, increasing the frame rate, significantly raises bandwidth demands and makes the task of video delivery even more challenging for the service providers. Taking into consideration this challenge on one hand and the content-dependent nature of perceived video quality at different frame rates on the other hand, a content-driven perceptually-aware frame rate selection mechanism is required.

Several efforts have been made recently to explore the relationship between content, frame rate and perceptual video quality [6–12]. Nasiri et al. [7] built a database with compressed videos at various frame rates up to 60 fps and based on a subjective study they investigated the impact of frame rate on perceived video quality and its relationship with quantization level, spatial resolution, and spatial and motion complexities. Based on this study, Nasiri et al. [8] formulated a descriptor of the perceptual aliasing factor. Both of these works however focused on rather low frame rates, up to 60 fps. Mackin et al. [6] demonstrated the relationship between dynamic range and frame rate for frequencies including and beyond 120 fps.

Besides the studies on the impact of frame rate on perceptual quality, some work has also reported new video quality metrics, predictive models [9–11] or frame rate selection mechanisms [12]. Ou et al. explored the impact of frame rate and quantization [9] and later proposed a quality model [10] as a function of spatial resolution, temporal resolution, and quantization stepsize, where each of these parameters was defined by a mathematical expression based on nine spatio-temporal features. However, the number of features was relatively high, the range of considered frame rates was limited to a maximum of 50 fps. Another interesting recent approach was the video quality metric proposed by Zhang et al. [11] that extracts features in the wavelet domain from a video sequence at different temporal resolutions using spatio-temporal pooling. The metric considers videos at high frame rates (up to 120 fps) but requires, as inputs, the video sequences at all potential frame rates. Lastly, Huang et al. [12] proposed a frame rate selection mechanism with the aim to meet the "satisfied user ratio". A feature-based machine learning approach was proposed and tested on a dataset with sequences at various frame rates but only up to 60 fps.

This paper presents a relatively low complexity frame rate selection process that is based on a supervised machine learning technique that uses only a few spatio-temporal features extracted from the original HFR sequence and the outcomes of the one-way Analysis of Variance (ANOVA) on

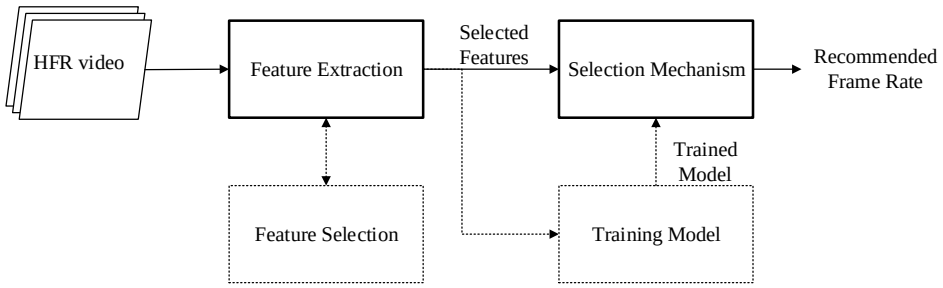

Fig. 1: Diagrammatic illustration of the proposed method. The dashed lines and blocks are only used in the offline phase of the proposed method, while the solid lines and blocks are used in both phases.

the differential mean opinion scores (DMOS) to benchmark the ground truth. Compared to the recent literature, the proposed method has the advantage of employing only a small number of spatio-temporal features extracted from the original video sequence (only at the acquisition frame rate) resulting in reduced complexity. Furthermore, it has been tested with a data set up to 120 fps. After an offline feature selection process, the selected spatio-temporal features are used along, with the DMOS-based ground truth, to train a supervised machine learning model. This approach offers the potential to minimise acquisition bit rates and hence storage or transmission requirements prior to compression.

The remainder of the paper is organised as follows. Section II describes all steps of the proposed perceptual-aware frame selection method. Section III explains the testing setup, the evaluation method and discusses the results. Finally, the conclusions are drawn in Section IV.

II. PROPOSED METHOD

In this section, we describe a method that is able to recommend a frame rate for a short video sequence that has a consistent spatial and temporal behaviour (i.e. similar spatio-temporal features across all frames). The aim is to select the minimum frame rate while maintaining the perceptual quality of the source sequence. A diagrammatic overview of the proposed method is illustrated in Fig. 1. The method is structured in two phases; offline and online. During the offline phase (all blocks are used), spatio-temporal features are extracted from the source sequences and after following a feature selection process, the selected features, along with the ground truth, are fed into the training model. After the training phase is complete, the system can be used online to make a perceptually-aligned frame rate recommendation based on features selected during training. More details are provided in the following subsections.

II-A. Feature Extraction

Motion characteristics are critical in relation to temporal artefacts in video sequences because when the frame rate is above the critical flicker frequency of about 80fps [4], the two most significant motion artefacts are blur and strobing. The amount of motion blur depends on the speed of the motion, leading to an unnatural change between sharp and blurred pictures as objects change speed. Strobing is caused by temporal undersampling. In addition to the temporal characteristics, spatial attributes of the content are also very important. In particular, the combination of temporal and spatial characteristics is related to motion artifacts [3]. Therefore, we employ spatial and temporal low level features that can effectively identify the impact of these artefacts. Also, we assume that it is meaningful to extract the spatio-temporal features per shot as this is typically related to spatial and temporal homogeneity. A total number of 31 features with their statistics were extracted from the videos at the acquisition frame rate as explained below and summarized in Table I.

Since temporal characteristics are very important, many features that can be defined that relate to them. First, a simple measure of the temporal variation within a video sequence can be obtained by the frame difference (FD) between successive frames. FD tends to give high values not only for large pixel displacements, but also for very fine and sharp textures. Thus, in blurry areas it is expected to have lower values. A similar related feature is the normalised frame difference (NFD) [9], that is defined by the ratio of FD to the average standard deviation of the pixel values within a frame. NFD attempts to connect the intensity contrast to the motion between successive frames. A more sophisticated temporal feature is the optical flow (OF), which is computed based on Farneback's method [13] and we can extract the mean and standard deviation of magnitude (mag) and orientation (or) of the OF vectors.

For the representation of spatial information, and particularly the intensity contrast between neighboring pixels, we selected the gray level co-occurrence matrix (GLCM) [14] and extracted the statistics as described in [15]. GLCM can also capture the degree of coarseness and directionality of the neighbouring pixels. GLCM is expected to be significantly affected by motion blur or strobing artifacts.

In an attempt to include a feature that combines both spatial and temporal characteristics, we used the temporal coherence (TC) with its statistics [15]. We also extracted statistics from the histograms of oriented gradients

Table I: List of features and notations.

Feature	Keywords
FD [9]	meanFD
NFD [9]	meanNFD
HoG [16]	meanHoG, stdHoG, skewHoG, kurHoG, entrHoG
OF [13, 15]	$\text{meanOF}_{\text{mag}}$, $\text{stdOF}_{\text{mag}}$, $\text{meanOF}_{\text{or}}$, stdOF_{or}
TC [15]	$\text{meanTC}_{\text{mean}}$, $\text{stdTC}_{\text{mean}}$, $\text{meanTC}_{\text{std}}$, $\text{stdTC}_{\text{std}}$, $\text{meanTC}_{\text{skew}}$, $\text{stdTC}_{\text{skew}}$, $\text{meanTC}_{\text{kur}}$, $\text{stdTC}_{\text{kur}}$, $\text{meanTC}_{\text{entr}}$, $\text{stdTC}_{\text{entr}}$
GLCM [14, 15]	$\text{meanGLCM}_{\text{con}}$, $\text{stdGLCM}_{\text{con}}$, $\text{meanGLCM}_{\text{cor}}$, $\text{stdGLCM}_{\text{cor}}$, $\text{meanGLCM}_{\text{hom}}$, $\text{stdGLCM}_{\text{hom}}$, $\text{meanGLCM}_{\text{enrg}}$, $\text{stdGLCM}_{\text{enrg}}$, $\text{meanGLCM}_{\text{entr}}$, $\text{stdGLCM}_{\text{entr}}$

(HoG) [16], such as the mean, the standard deviation, the skewness, the kurtosis and the entropy.

II-B. Feature Selection

In order to build a robust classifier and avoid overfitting, we need to reduce the dimensionality of the feature space by selecting a suitable subset. The feature selection process takes place during the offline phase. It should be noted that since many of the features are correlated, different subsets of features can lead to the same results. However, in order to achieve a low complexity solution we are interested in finding the subset of features with the lowest cardinality that achieves the best results. Therefore, a feature selection and elimination process based on Random Forest (RF) models, called Recursive Feature Elimination (RFE) [17], is employed. RFs are a popular type of machine learning technique that are robust even with high dimensional data and that also capture both linear and non-linear relationships. RFs use feature ranking techniques, which can be used for feature selection. RFE uses feature ranking and iteratively selects subsets of features with different cardinality, computes the classification accuracy and returns the optimal feature subset.

II-C. Frame Rate Selection Mechanism

Although the frame rate selection problem could be considered as a regression problem, it is treated as a classification problem. This is because the set of frame rates specified in the standardisation activities is discrete and consists of a few values. Therefore it is more efficient to treat it as a classification problem. During the offline phase and after the feature selection process, the training of the classification model takes place based on the selected features that are extracted from the data set used only for training purposes. During the training phase several machine learning techniques (mostly based on random forests and decision trees) are tested and only the most accurate was used in the online phase to make a perceptually-driven frame rate recommendation.

Fig. 2: Examples of frames of the BVI-HFR video sequences [18], from left to right and top to bottom: typing, water-splashing, bobble-head, cyclist, flowers, hamster, joggers, sparkler, martial-arts.

III. EVALUATION OF THE PROPOSED METHOD

III-A. Description of the Data Set

We employ BVI-HFR [18] that contains 22 source sequences at four different frame rates $\{15, 30, 60, 120\}$ fps. As explained in [18], all sequences were natively captured with a RED Epic-X video camera at 4K UHD spatial resolution and a frame rate of 120 fps using a fully open (360°) shutter angle. These sequences were spatially downsampled to HD resolution into YUV 4:2:0 format. All sequences are 10 sec in duration without any shot transitions. The three temporally down-sampled versions were generated by averaging frames. A few examples of frames from the BVI-HFR sequences are depicted in Fig. 2. Furthermore, due to the fact that a fully open shutter angle was used for the acquisition of the sequences, motion blur artifacts are expected [5], especially in sequences with objects moving relative to the camera, as for example the video sequences cyclist and joggers in Fig. 2. This characteristic of the data set is important for the feature selection process, since it affects the spatio-temporal characteristics and thus the features. It should be noted that we expect feature combinations to vary with shutter angle and this is the topic of future work.

III-B. Forming the Ground Truth

Mackin et al. [18] have conducted an ANOVA on the participants opinion scores in BVI-HFR comparing the video sequences at 60 fps and 120 fps to test the significant difference between the examined frame rates. The ANOVA analysis showed that some sequences will have clear perceptual quality benefits at 120 fps compared to 60 fps, while other sequences have no significant difference. Based on these results (see Table 2 in [18]), we set the ground truth for the recommended frame rates of the BVI-HFR sequences as reported in Table II. For 10 out of 22 BVI-HFR video sequences the 60 fps is the subjectively selected lowest frame rate (perceptually no significant difference from the acquisition frame rate), while for the other 12 video sequences the acquisition frame rate of 120 fps is selected. A

common characteristic of most of the sequences that benefit from the highest available frame rate is that the camera is moving.

Table II: Ground truth as derived from the ANOVA on the subjective tests in [18].

Optimal Frame Rate	Video Sequences
120 fps	1. books, 2. catch, 3. catch-track, 4. cyclist, 5. golf-side, 6. hamster, 7. joggers, 8. library, 9. plasma, 10. pour, 11. typing, 12. water-splashing
60 fps	1. bobblehead, 2. bouncyball, 3. flowers, 4. guitar-focus, 5. lamppost, 6. leaves-wall, 7. martial-arts, 8. pond, 9. sparkler, 10. water-ripples

III-C. Test Setup

One of the limitations of the evaluation presented here is that BVI-HFR is the only publicly available high frame rate data set that has been subjectively evaluated and it is rather small, comprising only 22 video sequences. These sequences are uniform but of course cannot cover the full space of spatio-temporal features. The size of the training set and its coverage of the parameter space is critical for the training of any machine learning classification model in order to achieve good performance. To overcome this, we use a five-fold cross-validation method with a random selection of the folds. Particularly, in every iteration, 80% of the data set (18 video sequences) is used for training and 20% (4 video sequences) for testing. After the fifth iteration, the classification accuracy is averaged over all five iterations and an overall confusion matrix is given.

III-D. Results and Discussion

Feature Selection: The results from the feature selection are summarised in Fig. 3 and Fig. 4. Figure 3 illustrates the ranking of the features using RFs with the Mean Decrease Gini index [17]. As it can be seen, the features with the higher ranking are those that capture the temporal behaviour, namely the temporal standard deviation of the features. The feature that is ranked with the highest score is $\text{stdGLCM}_{\text{hom}}$. It is interesting that the second feature is $\text{stdGLCM}_{\text{con}}$, which is linearly correlated to $\text{stdGLCM}_{\text{hom}}$. The third is $\text{stdOF}_{\text{mag}}$ which is a better candidate feature since it is not highly correlated to $\text{stdGLCM}_{\text{hom}}$ and could therefore capture different characteristics.

Regarding the optimal minimum subset of features, the results of the RFE method following a five-fold cross-validation are depicted in Fig. 4. The plot shows the performance of different subsets of features with different cardinalities. As the solid marker indicates in the plot, the suggested minimum number of features is two. The selected features are $\text{stdGLCM}_{\text{hom}}$ and $\text{stdOF}_{\text{mag}}$ (as expected from the ranking). $\text{stdGLCM}_{\text{hom}}$ denotes the temporal standard deviation in the GLCM homogeneity, which describes how diagonal the GLCM matrix is. The more blurry a sequence, the less diagonal the GLCM is. $\text{stdOF}_{\text{mag}}$ is related to the motion blur and expresses the unnatural effect of change between sharp and blurred frames. It is important to note that the selected subset of features highly depends on the acquisition parameters of the data set. For instance, had the video sequences been captured with a different shutter angle, different temporal artefacts might have been better represented by different features.

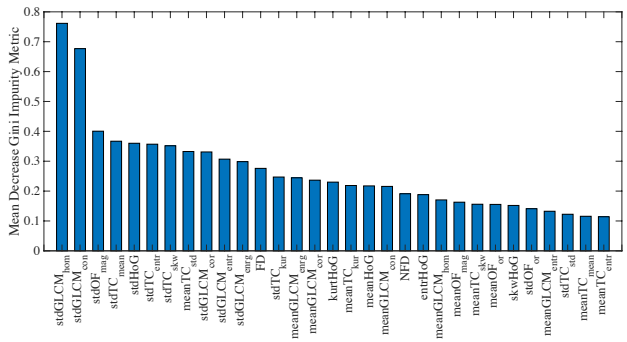

Fig. 3: RF-based Feature Ranking using the Mean Decrease Gini index.

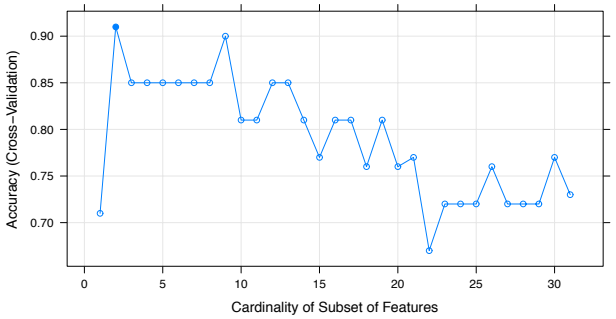

Fig. 4: Results of RFE using a five-fold cross-validation. The solid marker denotes the optimal cardinality of the subset of features.

Table III: Confusion Matrix and AUC values of the proposed frame rate selection method.

Using Bagged Decision Trees		Predicted Frame Rate		
		60 fps	120 fps	AUC
Ground Truth	60 fps	9	1	0.99
	120 fps	0	12	0.99

Frame Rate Selection: Several different classification methods (with varying kernels) were employed such as K-Nearest Neighbours, Support Vector Machines, Decision Trees, Random Forests, etc. Most of the classification methods led to similar results (over 90% classification accuracy and similar confusion matrices), but here we present the results from the method that achieved the highest accuracy using the subset of the two selected spatio-temporal features. The proposed frame rate selection method reached 95.5% accuracy by using a five-fold cross-validation with Bagged Decision Trees [17]. The confusion matrix along with the Area Under

the Curve (AUC) values per class of frame rate are reported in Table III. As can be observed, the AUC values are very high (maximum value equals to 1) and equal for the two classes showing a high accuracy in the prediction per class.

From Table III, we point out two important observations for the accuracy of the recommended frame rate. First, it is important to note that for the critical class of the 120 fps, the accuracy of the selected frame rate is 100%, since for all 12 sequences of the class the optimal frame rate is predicted. The confusion matrix also shows that there is only one misclassified video sequence from the class of the 60 fps; "flowers". For this sequence, instead of selecting 60 fps, the highest frame rate is predicted as the optimal. Although this is not a correct frame rate selection (according to subjective opinions), the decision will not adversely impact the perceived video quality since the higher frame rate is selected. The reasons behind this misclassification can be explained by the content features of this sequence. Despite the fact that the sequence is static, its spatial characteristics have misled the classification model. The sequence has a dense high frequency content that can be interpreted as prone to temporal aliasing artifacts. Furthermore, for the "flowers" sequence, a big part of the frame is unfocused, causing blurriness. This means that the spatial features are comparable to those of other sequences that have blurry areas caused my motion artifacts.

IV. CONCLUSION

In this paper, we investigated the relationship between high frame rate and spatio-temporal features in video. This resulted in a two-phase method that can recommend an optimal frame rate for a shot at which the perceived video quality is not significantly different from that of the original video. In the offline phase, the feature extraction and selection processes take place as well as the training of the machine learning model. In the online phase, the selected spatio-temporal features are extracted from the test sequences and the trained model is employed to recommend the lowest available frame rate without any perceptual quality degradation. The proposed method has the advantage of requiring only the original video sequences to make a frame rate recommendation. The results are promising with the benefit of no perceptual quality degradation nevertheless the limited available high frame rate video data set.

Future work includes the subjective evaluation of the perceived video quality of new video sequences at various frame rates for further validation of the proposed method. Finally, the proposed method will be further extended to make decisions on the adaptation of other video parameters related to immersion (e.g. spatial resolution, shutter angle).

ACKNOWLEDGEMENTS

The work presented was supported by the Engineering and Physical Sciences Research Council (EPSRC), EP/M000885/1 and EP/L016656/1.

REFERENCES

[1] M. Sugawara, S.-Y. Choi, and D. Wood, "Ultra-High-Definition Television (Rec. ITU-R BT.2020): A Generational Leap in the Evolution of Television [Standards in a Nutshell]," *IEEE Magazine on Signal Processing*, vol. 31, no. 3, 2014.

[2] Recommendation ITU-R BT.2020-1, "Parameter Values for Ultra-High Definition Television Systems for Production and Intern. Programme Exchange," Tech. Rep., Geneva:Intern. Telecommunications Union, 2014.

[3] Watson A., "High Frame Rates and Human Vision: A View through the Window of Visibility," *SMPTE Motion Imaging Journal*, vol. 122, no. 2, pp. 18–32, 2013.

[4] Recommendation ITU-R BT.2246-5, "The present state of ultra-high definition television," Tech. Rep., Geneva:Intern. Telecommunications Union, 2015.

[5] R. Selfridge, K. C. Noland, and M. Hansard, "Visibility of motion blur and strobing artefacts in video at 100 frames per second," in *13th European Conf. on Visual Media Production*, 2016, pp. 3:1–3:10.

[6] A. Mackin, K. C. Noland, and D. R. Bull, "High Frame Rates and the Visibility of Motion Artifacts," *SMPTE Motion Imaging Journal*, vol. 126, no. 5, pp. 41–51, 2017.

[7] R. M. Nasiri, J. Wang, A. Rehman, S. Wang, and Z. Wang, "Perceptual quality assessment of high frame rate video," in *IEEE 17th Intern. Workshop on Multimedia Signal Processing*, 2015.

[8] R. M. Nasiri and Z. Wang, "Perceptual Aliasing Factors and the Impact of Frame Rate on Video Quality," in *IEEE Intern. Conf. on Image Processing*, 2017.

[9] Y. F. Ou, Z. Ma, T. Liu, and Y. Wang, "Perceptual Quality Assessment of Video Considering Both Frame Rate and Quantization Artifacts," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 21, no. 3, pp. 286 – 298, 2011.

[10] Y. F. Ou, Y. Xue, and Y. Wang, "Q-STAR: A Perceptual Video Quality Model Considering Impact of Spatial, Temporal, and Amplitude Resolutions," *IEEE Trans. on Image Processing*, vol. 23, no. 6, pp. 2473–2486, 2014.

[11] F. Zhang, A. Mackin, and D. R. Bull, "A Frame Rate Dependent Video Quality Metric based on Temporal Wavelet Decomposition and Spatiotemporal Pooling," in *IEEE Intern. Conf. on Image Processing*, 2017.

[12] Q. Huang, S. Y. Jeong, S. Yang, D. Zhang, S. Hu, H. Y. Kim, J. S. Choi, and C. C. J. Kuo, "Perceptual Quality Driven Frame-Rate Selection (PQD-FRS) for High-Frame-Rate Video," *IEEE Trans. on Broadcasting*, vol. 62, no. 3, pp. 640–653, 2016.

[13] G. Farnebäck, "Two-frame motion estimation based on polynomial expansion," in *Scandinavian Conf. on Image analysis*. Springer, 2003, pp. 363–370.

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

[15] A. V. Katsenou, T. Ntasios, M. Afonso, D. Agrafiotis, and D. R. Bull, "Understanding video texture - a basis for video compression," in *IEEE 19th Intern. Workshop on Multimedia Signal Processing*, 2017.

[16] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *IEEE Computer Society Conf. on Computer Vision and Pattern Recognition*, 2005.

[17] M. Kuhn and K. Johnson, *Applied Predictive Modeling, First Edition*, Springer, New York, USA, 2013.

[18] A. Mackin, F. Zhang, and D. R. Bull, "A study of subjective video quality at various frame rates," in *IEEE Intern. Conf. on Image Processing*, 2015, pp. 3407–3411.