QUALITY ASSESSMENT OF LOW FRAMERATE VIDEOS

Abhinau Kumar Venkataramanan

Department of Electrical and Computer Engineering, UT Austin

ABSTRACT

Framerate conversion is one of the simplest methods of introducing temporal distortion to videos. Videos having low temporal complexity can be transmitted/viewed at lower framerates without affecting subjective quality. Even within a video, different regions in space or time may be affected differently by reduction in framerate. For example, static regions in frames remain unaffected while moving objects may result in motion blur or stuttering. In this work, we propose a quality metric based on three factors - content diversity, perceived motion blur, and motion smoothness. We evaluate it on the BVI-HFR database and show that it outperforms the baseline model, which measures aliasing.

***Index Terms*—** Temporal Downsampling, Motion Blur, Motion Smoothness

1. INTRODUCTION

This work aims to estimate the effect of temporal downsampling on the quality of videos. Widely used Video Quality Assessment (VQA) algorithms such as SSIM [1] involve the temporal aggregation of frame-wise quality scores computed using Image Quality Metrics. This temporal aggregation is usually the arithmetic mean, but may also be other pythagorean means, median, Minkowski sums, etc. However, in all these cases, the quality of the video is dominated, if not solely determined, by the spatial quality of frames in the video.

In case of temporal downsampling, i.e., reduction in the frame rate of the video, the spatial quality of the individual frames may not adequately capture the loss in subjective quality. Suppose we wish to downsample the video by an integral factor K. This can be achieved in two ways - simply skipping $K-1$ frames, or averaging non-overlapping sections of K frames. In the first case, since the retained frames are unchanged, there is no real change in spatial quality, but the absence of intermediate frames leads to stutter. In the second case, motion blur is introduced in regions containing moving objects, leading to a loss in both spatial and temporal quality. This is shown in Figure 1.

Fig. 1. Motion blur due to downsampling

1.1. Previous Work

Oh et. al. [2] proposed a metric for measuring the judder in videos using Difference of Position in Retinal Image (DPRI) and motion vectors (MVs). However, this method was only evaluated on synthetic data since MVs were difficult to obtain for arbitrary videos. Chapiro et. al. [3] quantified judder as a function of the luminance, frame rate and speed of a moving object. However, this work is again restricted to synthetic data of a single object moving at a constant speed from left to right against a constant background.

Nasiri et. al. [4] quantify the aliasing introduced by downsampling as the ratio of energy in the high frequency region to the energy in the low frequency region of the spectrum. Consider a video $f(x,y,t)$. At a coordinate (x_0,y_0), the 1-D pixel signal is $s_{x_0,y_0}(t) = f(x_0,y_0,t)$. When this signal is sampled at a frequency f_s, the temporal aliasing factor (TAF) is defined in terms of the spectrum $\hat{s}_{x_0,y_0}(f)$ as

$$A_{T,x_0,y_0}(f_s) = \frac{\int_{f_s/2}^{f_s} |\hat{s}_{x_0,y_0}(f)|^2 df}{\int_{0}^{f_s/2} |\hat{s}_{x_0,y_0}(f)|^2 df} \tag{1}$$

To obtain a value for the entire video, N pixel locations (x_i,y_i) are sampled and the average $A_T(f_s) = \frac{1}{N}\sum_{i=1}^{N} A_{T,x_i,y_i}(f_s)$. We use this as a baseline against which we evaluate the performance of our models. We later show

that the performance of this model is significantly worse on this database compared to the reported performance in [4], although on a different database.

[5] is another work by Nasiri et. al. which proposes a temporal motion smoothness measure using the complex wavelet transform. We refer to this method as Complex Wavelet Temporal Motion Smoothness (CW-TMS). While the authors report their performance on the same database we use, we were unable to reproduce their results. Nevertheless, we include their reported results for completeness.

1.2. Database

The database used in this work is the BVI-HFR database [6]. This database contains 22 video sequences at 4 different framerates - 15, 30, 60, and 120 fps. Treating the 120 fps video as the reference "pristine" video, we have 22 reference videos and 66 distorted videos.

1.3. Proposed Method

We define our quality model as a fusion of three quality metrics. The temporal diversity factor (TDF) quantifies the diversity of the sets of frames being averaged to create the low framerate video. The perceptual blur factor (PBF) is used to localize spatial motion blur artefacts created due to the downsampling process. Finally, the motion smoothness factor (MSF) is a measure of the difference in motion smoothness between the pristine and distorted videos.

These three metrics are then used as features to train a support vector regressor (SVR) with a radial basis function (RBF) kernel to predict visual quality.

1.3.1. Temporal Diversity Factor

The performance of the baseline model on this database, shown in Table 1, is much lower than the performance reported by the authors on their database. This drop in performance might be because temporal downsampling is carried out by frame-averaging. The resulting motion blur in these regions might "hide" aliasing. Examining the spectra, it was observed that most pixel signals had very low tail energies, due to which the TAF values did not show strong trends across framerates.

However, we can still compare energies. Let $d_{f_s,x_0,y_0}(t)$ denote a pixel signal from the "distorted" video having framerate $f_s = 120/K$ and duration T. We can view this as a vector $\mathbf{d_{f_s,x_0,y_0}}$ having f_sT entries. The pixel signal from the pristine video of the same duration $s_{x_0,y_0}(t)$ is a vector $\mathbf{s_{x_0,y_0}}$ of $120T$ entries.

We can view s_{x_0,y_0} as an interleaved version of K vectors s_{k,x_0,y_0} for $k = 0, 1, \ldots, K-1$, where

$$s_{k,x_0,y_0}(t) = s_{x_0,y_0}(Kt + k) \tag{2}$$

Since the downsampled video is created by frame-averaging, we can express the distorted pixel signal as

$$d_{f_s,x_0,y_0}(t) = \frac{1}{K}\sum_{k=0}^{K-1} s_{k,x_0,y_0}(t) \tag{3}$$

Then, by convexity of norm,

$$\|\mathbf{d_{f_s,x_0,y_0}}\| = \left\| \frac{f_s}{120}\sum_{k=0}^{K-1} \mathbf{s_{k,x_0,y_0}} \right\| \leq \frac{f_s}{120}\sum_{k=0}^{K-1} \|\mathbf{s_{k,x_0,y_0}}\| \tag{4}$$

Since each vector $\mathbf{s_{k,x_0,y_0}}$ has one frame from each set of K in $\mathbf{s_{x_0,y_0}}$, it is reasonable to assume that they have approximately equal energies. Then,

$$\|\mathbf{s_{k,x_0,y_0}}\|^2 \approx \frac{1}{K}\|\mathbf{s_{x_0,y_0}}\|^2 \tag{5}$$

Substituting, we have the inequality

$$\|\mathbf{d_{f_s,x_0,y_0}}\|^2 \leq \frac{f_s}{120}\|\mathbf{s_{x_0,y_0}}\|^2 \tag{6}$$

The only source of the inequality is the convexity of the norm. The "gap" between the two quantities is zero if all frames in each set of K are identical. The more diverse these sets, the larger the gap. Using this relationship, we can define the Temporal Diversity Factor (TDF) at a location x_0, y_0 as

$$D_{x_0,y_0}(f_s) = \frac{120}{f_s}\frac{\sum_{k=0}^{f_sT-1}|d_{f_s,x_0,y_0}(k)|^2}{\sum_{k=0}^{120T-1}|s_{x_0,y_0}(k)|^2} \tag{7}$$

As in TAF, instead of computing this over the entire video, we sample N spatial locations at random. We can then define the TDF for a video as $D(f_s) = \frac{1}{N}\sum_{i=1}^{N} D_{x_i,y_i}(f_s)$.

1.3.2. Perceptual Blur Factor

To identify motion blur (this applies to general blur artefacts also) we propose a DCT-based approach which compares local frequency responses. Let $R(x, y, t)$ and $T(x, y, t)$ be the reference and test frames corresponding to the same time instant t. We compute the DCT of $K \times K$ non-overlapping blocks. The intuition is that in regions of blur, the $K \times K$ DCT spectrum $\hat{\mathbf{T}}_{x_0,y_0}$ has a response more concentrated at low frequencies compared to the reference DCT $\hat{\mathbf{R}}_{x_0,y_0}$.

Let us define the logarithmic difference between the reference and test DCTs computed at spatial index x_0, y_0, at frequency indices k_1, k_2 to be

$$\delta_{x_0,y_0}(k_1,k_2) = \log\left(1 + \left|\hat{T}_{x_0,y_0}(k_1,k_2)\right|\right) - \log\left(1 + \left|\hat{R}_{x_0,y_0}(k_1,k_2)\right|\right) \tag{8}$$

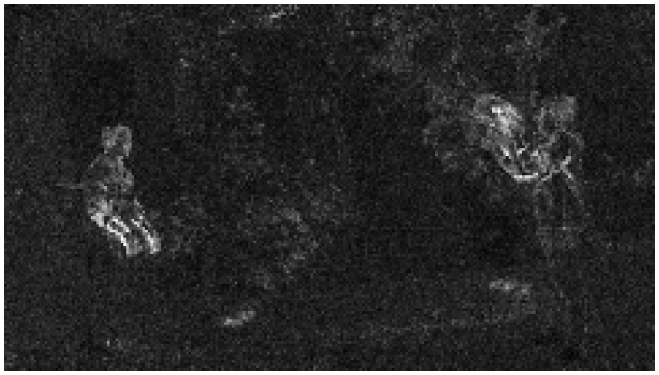

Fig. 2. Blur Index

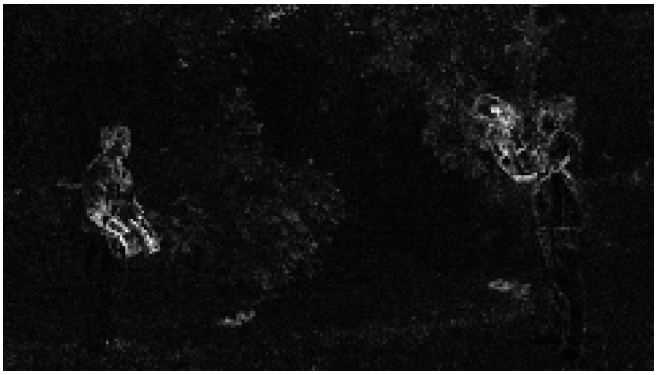

Fig. 3. Perceptual Blur Index

The 1 is added as a saturation constant to avoid negative values of logarithms. Since we want to quantify blur, we expect positive differences at low frequencies and negative differences at higher frequencies. To capture this, define a sign function

$$\sigma(k_1, k_2) = \begin{cases} +1 & \text{if } k_1 + k_2 < K \\ -1 & \text{if } k_1 + k_2 \geq K \end{cases} \tag{9}$$

We can then quantify the blur at index x_0, y_0 as

$$\delta(x_0, y_0) = \frac{1}{K^2} \sum_{k_1=0}^{K-1} \sum_{k_2=0}^{K-1} \{\sigma(k_1, k_2)\delta_{x_0,y_0}(k_1, k_2)\}^+ \tag{10}$$

where $\{x\}^+ = \max(x, 0)$ clips negative values to 0. For the frame in Figure 1, the map of this blur index is shown in Figure 2.

However, due to the contrast sensitivity function of the eye, not all frequencies are created equal. In practice, this, along with the statistical nature of video frames, has been used in the design of quantization matrices in compression algorithms like JPEG. So, choosing $K = 8$ allows us to borrow the 8×8 quantization matrix $\mathbf{Q}$ to use as a weighting function to re-weight DCT coefficients. Frequencies having lower perceptual relevance are weighted more, decreasing their effect on the difference. We define the logarithmic difference between perceptually weighted DCTs computed at spatial index x_0, y_0, at frequency indices k_1, k_2 to be

$$\delta^Q_{x_0,y_0}(k_1, k_2) = \log\left(1 + \left|\frac{\hat{T}_{x_0,y_0}(k_1, k_2)}{Q}\right|\right) \tag{11}$$

$$- \log\left(1 + \left|\frac{\hat{R}_{x_0,y_0}(k_1, k_2)}{Q}\right|\right) \tag{12}$$

We can then quantify the perceived blur at index x_0, y_0 as

$$\delta^Q(x_0, y_0) = \frac{1}{K^2} \sum_{k_1=0}^{K-1} \sum_{k_2=0}^{K-1} \left\{\sigma(k_1, k_2)\delta^Q_{x_0,y_0}(k_1, k_2)\right\}^+ \tag{13}$$

For the frame in Figure 1, the map of this perceptual blur index is shown in Figure 3. We observe that the perceptual weighting has lead to a significant noise reduction in motion blur estimates.

Finally, since differences in logarithmic domain correspond to ratios in the linear domain, we can define the blur factor and perceptual blur factor respectively at a spatial index x_0, y_0 to be

$$\rho(x_0, y_0) = \exp\left(-C\delta(x_0, y_0)\right) \tag{14}$$

$$\rho^Q(x_0, y_0) = \exp\left(-C\delta^Q(x_0, y_0)\right) \tag{15}$$

where the constant C can be used to control the range of values. Values close to 1 indicate no blur, while values close to 0 indicate severe blur.

1.3.3. Motion Smoothness Factor

Since low frame-rate videos are sub-sampled versions of the true video, we expect to observe larger motion magnitudes and larger changes in motion between successive frames. To investigate this, we use Farneback's Optical Flow algorithm to calculate flow fields using adjacent frames.

Let $u_x(i, j, k)$ and $u_y(i, j, k)$ be the motion vectors in the x and y directions computed using frames k and $k + 1$ of a video. We observe that the shapes of the histograms / empirical probability distribution function (EPDFs) of these flow values are content-dependent and vary between videos and between frames in a video. This is shown in Figure 4

Motion smoothness is better described by the discrepancy between motion in adjacent frames. To capture this, we consider the statistics of differences between flow fields. Let us define $d_x(i, j, k) = u_x(i, j, k) - u_x(i, j, k + 1)$ and $d_y(i, j, k) = u_y(i, j, k) - u_y(i, j, k + 1)$. The EPDFs of motion vector differences are typically unimodal and have more consistent shapes across contents. This is shown in Figure 5. To capture these histograms, we calculate means, variances and kurtoses, denoted by $(\mu_{d,x}(k), \mu_{d,y}(k))$, $(\sigma^2_{d,x}(k), \sigma^2_{d,y}(k))$, $(\kappa_{d,x}(k), \kappa_{d,y}(k))$.

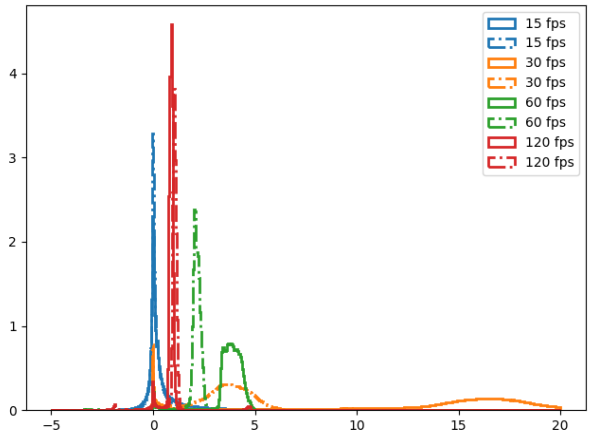

Fig. 4. Flow values

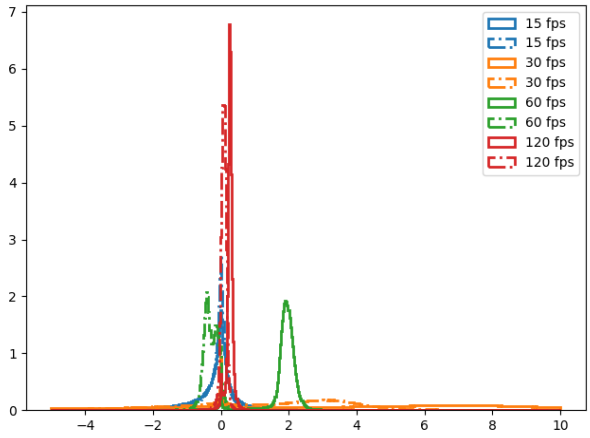

Fig. 5. Flwo Difference values

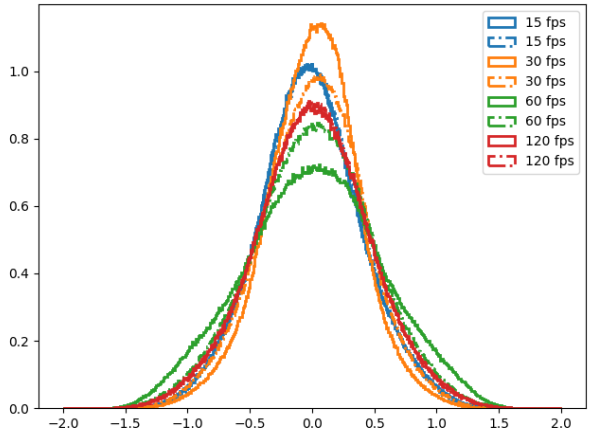

Fig. 6. Flow DNTs

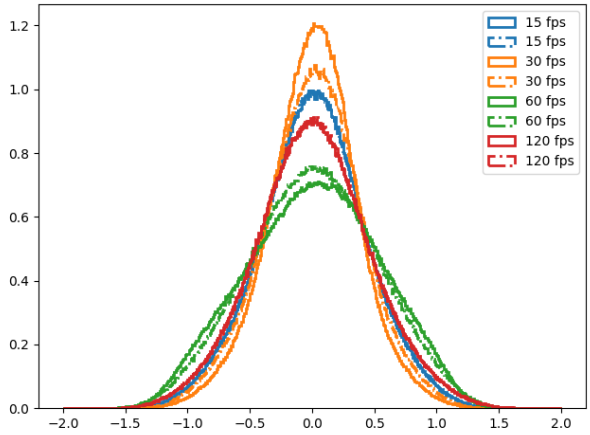

Fig. 7. Flow Difference DNTs

To extract more statistical regularities, we apply the divisive normalization transform (DNT) to these optical flow fields. We subtract the local mean and divide by the local standard deviation, i.e.,

$$z_x(i,j,k) = \frac{u_x(i,j,k) - \mu_x(i,j,k)}{\sigma_x(i,j,k) + c} \tag{16}$$

$$z_y(i,j,k) = \frac{u_y(i,j,k) - \mu_y(i,j,k)}{\sigma_y(i,j,k) + c} \tag{17}$$

Finally, we also apply the DNT to the differenced flow-fields, giving us $z_{d,x}(i,j,k)$ and $z_{d,y}(i,j,k)$. We record the mean, variance and kurtoses of these DNT coefficients.

Figures 4 and 5 show the histograms of optical flow field and differenced flow field respectively for one video (catch-track) at various framerates from the BVI-HFR database, using $c = 0.0001$. For the same video, we have two histograms per framerate corresponding to motion in the x and y directions. For the same video, figures 6 and 7 show the statistics after applying DNT to the flow field and differenced flow field respectively.

Using these flow and flow-difference statistics, I tried to derive quality metrics. When we compare frame-wise statistics across frame-rates, we resample the values obtained for the high framerate sequence. The best performing metrics differed in both databases.

On the BVI-HFR database, the best performing metric was

$$M(f_s) = \frac{\sum_{k=0}^{f_s T-2} |\mu_{d,x}^{(f_s)}(k) - \mu_{d,x}^{(120)}(k)|}{\sum_{k=0}^{f_s T-2} |\mu_{d,x}^{(120)}(k)|} \tag{18}$$

$$+ \frac{\sum_{k=0}^{f_s T-2} |\mu_{d,y}^{(f_s)}(k) - \mu_{d,y}^{(120)}(k)|}{\sum_{k=0}^{f_s T-2} |\mu_{d,y}^{(120)}(k)|} \tag{19}$$

where the superscripts denote the frame-rate of the videos.

Normalized differences are used to account for masking effects where small variations in motion statistics are masked

when the content itself has a large change of motion, i.e., higher means $(\mu_{d,x}^{(120)}, \mu_{d,y}^{(120)})$.

It is notable that even though DNT appears to "standardize" the histograms of flow fields and flow differences, the best performing metric does not use the DNT. In fact, on the BVI-HFR database, correponding metrics operating on DNT coeffecients consistently performed worse than those operating on untransformed data.

1.4. Results

We compare the performance of our methods with the baseline TAF model and SSIM. We empirically choose $C = 1$ for the blur factor and $C = 2$ for the perceptual blur factor. We report the Spearman Rank Order Correlation Coefficient (SROCC) between predicted quality and subjective scores for various methods in Table 1.

Method	SROCC
SSIM	0.4041
Temporal Aliasing Factor	0.4096
Perceptual Blur Factor	0.4281
Temporal Diversity Factor	0.4626
Motion Smoothness Factor	0.7105
Combined	**0.7232**
CW-TMS (Reported)	**0.78**

Table 1. Performance of various models

1.5. Conclusion

In this work, we propose a new quality measure for low framerate videos. We propose a simple energy-based measure of the reduction in diversity of frames due to downsampling. We also propose a perceptually-motivated DCT-based method of quantifying the blur in a region, using which we localize the motion blur caused by downsampling. Finally, we measure the loss in motion-smoothness by comparing the statistical properties of optical flow vectors across framerates. Using these, we showed a significant improvement in predicting subjective quality over the baselines.

In the future, wavelet-based methods may improve performance of blur localization since they provide simultaneous spatial and frequency resolution and allow for multi-scale analysis. Statistical modelling of motion in videos is still an open problem and incorporating more sophisticated statistical models of may also lead to better quality prediction.

2. REFERENCES

[1] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

[2] Se Ri Oh, Dongchan Kim, Pyeong Gang Heo, and Hyun-Wook Park, "A new metric for judder in high frame-rate video," in *2016 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2016, pp. 3802–3806.

[3] Alexandre Chapiro, Robin Atkins, and Scott Daly, "A luminance-aware model of judder perception," *ACM Transactions on Graphics (TOG)*, vol. 38, no. 5, pp. 1–10, 2019.

[4] Rasoul Mohammadi Nasiri and Zhou Wang, "Perceptual aliasing factors and the impact of frame rate on video quality," in *2017 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2017, pp. 3475–3479.

[5] R. M. Nasiri, Z. Duanmu, and Z. Wang, "Temporal motion smoothness and the impact of frame rate variation on video quality," in *2018 25th IEEE International Conference on Image Processing (ICIP)*, 2018, pp. 1418–1422.

[6] Alex Mackin, Fan Zhang, and David R Bull, "A study of subjective video quality at various frame rates," in *2015 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2015, pp. 3407–3411.