COMPARISON OF BLOCKING AND BLURRING METRICS FOR VIDEO COMPRESSION

Athanasios Leontaris
University of California, San Diego
aleontar@code.ucsd.edu

Amy R. Reibman
AT&T Labs - Research
amy@research.att.com

ABSTRACT

We evaluate the performance on compressed video of a number of available similarity, blocking, and blurring quality metrics. Using a systematic, objective framework based on simple subjective comparisons, we evaluate the ability of each metric to correctly rank order images according to the subjective impact of different (1) spatial content, (2) quantization parameters, (3) amounts of filtering, (4) distances from the most recent I-frame, and (5) long-term frame prediction strategies. The indicated weaknesses of available metrics can be used as guides in the development of future quality metrics.

1. INTRODUCTION

Many methods to measure the quality of an image or video sequence have been proposed in recent years. Although the Mean Squared Error (MSE) and its popular variant, the Peak-Signal-to-Noise Ratio (PSNR), are widely used, it is clear that neither metric takes into account the perceptual functions of the human visual system (HVS). Video quality metrics are useful in many applications, including assessing the quality of video transported over networks and guiding the design of video codecs.

Humans perceive quality on multiple dimensions simultaneously. One video may be simultaneously more blocky and less blurry than another video. One approach to assess video quality is to use HVS principles to determine a single value that characterizes the overall video quality [1, 2]. An alternate approach is to design quality metrics that assess a single impairment type in isolation, such that the impact of multiple impairments can be subsequently combined into a single quality value [3, 4].

In this paper, we concentrate on evaluating single-impairment quality metrics in the context of video compression. We are interested primarily in how well blockiness, blurriness, and similarity metrics can rank order images according to the amount of impairment added by common video compression components. In addition, we are primarily interested in no-reference (NR) quality metrics, since these are applicable in an environment where the original video is not available. The ability of a metric to correctly predict an ordered ranking of images given different amounts of transmission errors is outside the scope of this work.

Some earlier reviews of image and video quality metrics are available [5, 6, 7, 8, 9]. Each considers a different set of metrics and a different framework for evaluation. A comprehensive overview of monochrome image quality metrics dating from 1974 to 1999 was presented in [5], which focuses on describing HVS models without evaluation. Several monochrome full-reference (FR) image quality metrics were evaluated in [6], using the correlation between the metric output and a subjective evaluation. In [7], four quality metrics were evaluated in an error-prone environment: the SNR, PSNR, ITS [10] (an objective metric based on subjective tests) and MPQM [1] (an objective metric based on HVS). Three still FR image metrics that use HVS criteria are compared in [8] using subjective tests. Three quality metrics [11, 12, 13] were evaluated for video streaming in [9], which indicates the metrics had a difficult time correctly ranking the quality produced by different codecs.

In this paper, we present a framework to systematically explore the impact of multiple parameters that affect compressed video quality. Our goal is to test the ability of available similarity, blocking, and blurring metrics to correctly order the subjective impact of (1) different spatial content, (2) different quantization parameters, (3) different amounts of filtering, (4) different distances from the most recent I-frame, and (5) different long-term frame-prediction strategies. Because many metrics *appear* to work well when averaged over an entire video sequence, we explore their performance on a per-frame basis. Our evaluation, based on simple subjective comparisons, exposes several inadequacies in the performance of most metrics. However, any metrics which pass our criteria may still require further evaluation using more exhaustive subjective tests.

The paper is organized as follows: Section 2 discusses the multi-dimensional aspects of quality. The metrics we evaluate are summarized in Section 3. Our comprehensive comparison framework is described in detail in Section 4. Section 5 compares the metrics, while the paper concludes in Section 6.

2. MULTI-DIMENSIONAL ASPECTS OF QUALITY

Viewers perceive quality on many axes simultaneously. For example, an image may be blurry, blocky, have ringing artifacts, or have added high frequency energy. Temporally, a video may have jerky motion, or added "mosquito" noise. In this paper, while we are interested in the quality of videos, we assess only the quality of the individual still images that comprise the video.

Blockiness arises from the vertical and horizontal edges along a regular blocking grid that result from the block-based processing in many image (JPEG) and video (MPEG) codecs. Coarse quantization yields more blockiness, while edge-attenuating filters reduce its perceptual effect.

Blurriness is caused by the removal of high-frequency content from the original image/video signal. Increased blurriness can be caused by coarser quantization, edge-attenuating filters, or overlapped block motion compensation (OBMC).

Ringing artifacts are caused by the absence of high frequency terms during coarse quantization. They are most evident near high contrast edges, and most prevalent in wavelet coders.

This work was performed while the first author was a summer intern at AT&T Labs - Research.

Added high-frequency (HF) content is typical in video codecs that use block-based motion compensation (MC). When coarse quantization is combined with MC, blocking artifacts propagate from I-frames into subsequent frames, causing structured HF noise that is no longer aligned with block boundaries. Fractional-pel MC and edge-attenuating in-the-loop filters help minimize this artifact.

Farias et al. [14] indicates that when artifacts are perceived to have equal strength, blurriness is more annoying than either added high-frequency noise or blockiness. Furthermore, interactions among two artifacts types affect the perceived strength of each.

3. SUMMARY OF QUALITY METRICS

In this paper, we consider three similarity, nine blocking, and three blurring metrics.

Three similarity metrics. The similarity measures all require both the original and degraded images. MSE and, equivalently, PSNR use a pixel-by-pixel comparison between two images. The structural similarity index (SSIM) [15] uses means, variances, and correlations of both images. The psychovisual image quality evaluator (PIQE) [16] is a FR method consisting of two parts: a blockiness component and a similarity component, denoted PIQE-S, which counts the number of edges common to both the original and degraded image.

Nine blocking metrics. We consider one FR and eight NR blocking metrics. The blockiness component PIQE-B of PIQE [16] uses the DCT DC coefficients to compute a FR blocking measure. Two slope-based NR blocking measures, boundary discontinuity (BD) [17] and MSDS [18], were designed to identify areas to apply de-blocking algorithms. Both assume that the slopes internal to a block are preserved during encoding. Gao et al. [19] present a metric based on differences of averages of eight-pel rows (columns) across vertical (horizontal) block edges. In the Phase Correlation method [11], the denominator of the metric measures inter-block similarity while the numerator measures intra-block similarity.

The remaining four blocking metrics all incorporate some form of HVS modeling. In the generalized block impairment metric (GBIM), described in [13], HVS masking is incorporated by means of weights derived from localized averages and standard deviations across block boundaries. In [12], the Power Spectrum of the 1-D absolute difference signal is calculated using FFT. Luminance and texture masking are exploited to scale the signal before the frequency analysis. The DCT-Step metric [20] models blocking artifacts as 2-D step functions, and weighs results using local background luminance and activity masking measures. The perceptual block impairment metric (PSBIM) [21], which modifies GBIM to include more comprehensive luminance masking, has a similar structure to the metric in [11]. In PSBIM, the numerator represents edge strength, while the denominator represents the inner-block spatial similarity. None of these blocking metrics uses temporal masking, or is parameterized with respect to the viewing distance.

Three blurring metrics. All three blurring metrics are NR. The first blurring metric computes a global blur for an image using a histogram of DCT coefficients gathered from the compressed bitstream [22]. The second computes the spatial extent of each edge in an image using inflection points in the luminance to demark the start and end of an edge [23]. The third blurring metric actually computes sharpness, the inverse of blurring, by calculating local edge Kurtosis [24].

4. COMPARISON FRAMEWORK

In [9], video sequences were encoded at a single rate and the sole parameter explored was the frame number. Our intention is to explore the multi-dimensional parameter space that characterizes quality in a systematic fashion. Quality perception is affected by a number of parameters, including the six we consider:

1. The quantization parameter (QP) of the I-frames, Q_I.
2. The quantization parameter of the P-frames, Q_P.
3. The distance d between the current frame and the most recent I-frame.
4. The video *content*. Static vs. high motion activity; contrast, luminance and temporal masking influence quality perception.
5. The presence of edge attenuating *filtering*. Filters can be applied before compression (pre-filtering), during compression (in-the-loop filtering), after reconstruction (post-processing), or implicitly by using overlapped block motion compensation (OBMC) and fractional-pel MC.
6. The video *codec*. We consider H.263 and H.264/AVC. These differ, in part, in the size of their block partitions and the degree of fractional-pel MC prediction.

To evaluate the metrics, we use simple subjective comparisons to create a list of expectations that a well-designed metric should satisfy. Expectation (A) considers the effect of spatial content, (B,C) consider quantization, (D) considers filtering, (E) considers distance d, and (F) considers the influence of a high-quality long-term frame memory with H.264 [25].

Expectations:

(A) For the same QP and no filtering, *coastguard* is less blocky than *foreman*. Similarly, the quantized *coastguard* is more similar to its original than is the quantized *foreman*.

(B) Without filtering, blockiness increases and similarity decreases as QP increases. This is always valid for I-frames, where blocking artifacts are easy to define and locate. For P-frames, artifacts due to heavy quantization may no longer align with block boundaries, so many metrics may fail.

(C) With filtering, blurriness increases and similarity decreases as QP increases.

(D) For fixed QP and d, more filtering decreases both blockiness and similarity but increases blurriness.

(E) For fixed QP and filtering and for a single reference frame, blurriness increases, blockiness increases (because artifacts accumulate), and similarity decreases as distance from the most recent I-frame, d, increases.

(F) For fixed QP, fixed filtering, and fixed d, using a long-term (LT) high-quality reference frame improves quality; namely, it increases similarity and reduces blurring and blocking.

We use the following framework to explore how well the metrics satisfy the above expectations. We consider three case studies. The first, **I0**, consists only of I-frames with varying Q_I. We explore Expectations (A-D) for H.263 and H.264 with and without loop filtering across different spatial content. The second, **P1**, consists of an I-frame followed by multiple P-frames, where we vary the quantizer while keeping $Q_I = Q_P$. For this case, we explore Expectations (A-E), focusing on three values of d: $d = 1, 6, 14$. Finally, the third case, **P2**, is designed to examine Expectation (F), the visual impact of using a high-quality LT prediction frame in H.264 [25]. **P2** fixes both $Q_I = 18$ (nearly lossless) and $Q_P = 32$ (medium) and varies d across all values from 1 to 14.

To isolate the impact of spatial content, we choose identical frames for the comparison, regardless of the prediction structure. Thus, for frame number 21 to use $d = 1$, we set frame 20 to be an I-frame, while to achieve $d = 14$, frame 7 is an I-frame. We examine two sequences, *foreman* and *coastguard* (frames no. 21 and 141, respectively). For **I0** and **P1**, we vary the quantization parameter using constant increments starting from nearly lossless (2 for H.263 and 18 for H.264) to nearly unwatchable (30 for H.263 and 45 for H.264). Our H.263 test sequences are computed using the H.263+ codec in MPlayer/MEncoder [26]. For H.263, we considered the case of no filtering, and the use of 8×8 block motion vectors and OBMC. Our H.264/AVC test sequences, with and without loop filtering, are generated with the JVT reference software version JM 8.2 using CABAC.

Expectation (A) should hold across all codecs and frame-types, and Expectation (D) should hold across all possible codecs, frame-types and spatial content. Expectation (F) is based on observations using *coastguard* with and without LT prediction. The section of *coastguard* we considered particularly benefits from LT prediction because occluded areas were uncovered. Expectation (E) is derived from a small-scale subjective test with 6 viewers, who were shown *coastguard* using H.263 at a distance of 6 times the picture height. Comparing **I0** to **P1** with $d = 1$ and 14, for $Q_I \geq 22$, all observers found I-frames to be significantly less blurry than P-frames for $d = 14$ while four of six observers found I-frames to be less blocky than P-frames with $d = 1$.

In addition to holding across all possible codecs and spatial content, Expectations (B,C) should also be monotonic. For example, as QP increases, blockiness in (B) should increase monotonically. To characterize how well a blocking or blurring metric is able to capture this monotonic increase, we use Kendall's tau, τ_a [27], which is an estimate of the probability that a pair of variables is more likely to be correctly ordered than incorrectly ordered. For a set of data $\{x_i\}, i = 1, ..., N$, which should always increase, Kendall's tau is defined to be $\tau_a = (\gamma - \delta)/(\gamma + \delta - \epsilon)$, where γ is the number of possible pairs $(x_i, x_j), i < j$ for which $x_i < x_j$ (i.e., the number of pairs correctly ordered), δ is the number of pairs incorrectly ordered, and ϵ is the number of pairs for which $x_i = x_j$. Note that $\gamma + \delta + \epsilon = N(N-1)/2$. For completely monotonic data, $\tau_a = 1$. As pairs become incorrectly ordered, τ_a decreases.

5. COMPARATIVE RESULTS

Table 1 summarizes how well each metric described in Section 3 is able to satisfy the expectations listed in Section 4. Ability to satisfy Expectations (A,D,E,F) is indicated by "Y", while inability is indicated with "x". Results for Expectations (B,C) show the minimum value of Kendall's τ_a achieved across the set of situations considered. Recall that $\tau_a = 1$ means a metric completely satisfies this expectation, while negative τ_a clearly indicates an inability to satisfy this expectation. For Expectation (B), we consider separately the different frame types **I0** and **P1**.

As shown in Table 1(a), the FR similarity metrics perform well for most expectations. In particular, PSNR and SSIM similarly pass all tests except that they each predict *foreman* is better visually than *coastguard*. Since neither include any HVS masking, they are unable to recognize that spatial masking conceals the perceptual impact of blocking in *coastguard*. While PIQE-S is able to meet Expectations (A-C), it fails for (D-F) because it is heavily dependent on extracting edges from an image. As a result, when the image is compressed using an edge-attenuating filter, the edge detector may have difficulty extracting sufficient edges.

method	A	B(I0)	B(P1)	C	D	E	F
PSNR	x	.995	1.0	.995	Y	Y	Y
SSIM [15]	x	.995	.995	.995	Y	Y	Y
PIQE-S [16]	Y	.947	.921	.946	x	x	x

(a)

method	A	B(I0)	B(P1)	D	E	F
PIQE-B [16]	x	.759	.980	x	Y	x
BD [17]	x	.725	-.994	Y	x	x
MSDS [18]	x	.815	-.994	Y	x	x
[19]	x	.963	.894	Y	x	x
[11]	x	.606	.090	x	x	x
GBIM [13]	Y	.995	.741	Y	x	x
[12]	x	.968	-.798	Y	x	x
DCT-Step [20]	x	.852	.651	Y	x	x
PSBIM [21]	Y	.980	.906	x	x	Y

(b)

method	C	D	E	F
[22]	.892	Y	x	x
[23]	.956	Y	x	Y
Kurtosis [24]	.591	Y	x	x

(c)

Table 1. Ability of each metric to satisfy Expectations (A-F). A "x" in Columns (A,D,E,F) means metric failed expectation; "Y" denotes satisfaction. Columns (B,C) denote Kendall's τ_a characterizing monotonicity. (a) Similarity, (b) Blocking, and (c) Blurring metrics.

Results for blocking metrics are given in Table 1(b). All metrics were designed to quantify the situation measured with Expectation (B) using **I0**; however, [11], BD, PIQE-B, and MSDS are clearly weakest in this regard. Because motion compensation moves blocking artifacts off block boundaries, most metrics have not been designed to address Expectation (B) using **P1**; however, PIQE-B, PSBIM, [19], and GBIM still perform reasonably well in this situation. Regarding Expectation (D), it is interesting to note that three metrics are unable to consistently measure across all QP and d that additional filtering reduces blockiness. Only PIQE-B is able to show (E) that blockiness increases for P-frames over I-frames, only GBIM and PSBIM are able to show (A) *coastguard* is visually less blocky than *foreman*, and only PSBIM is able to show (F) that using a high-quality LT prediction reduces blockiness for *coastguard* frame 141.

The blocking metrics share a number of weaknesses. First, only four incorporate some form of HVS modeling. Second, BD and MSDS both assume that the encoding process preserves the inner-block structure and especially the pixel slopes. Unfortunately, the experimental results indicate this assumption does not hold for the codecs we considered. Third, many metrics appear to discard useful information when they compute blockiness. For example, PIQE-B and [19] use only DC coefficients on a block and row/column level, respectively. Further, [19] employs a very strong cutoff threshold for measuring an edge. In DCT-Step, the simple 2-D step function model discards many coefficients. Similarly, the Phase Correlation method [11] discards vital information during spatial sub-sampling. Fourth, as pointed out in [9], many

blocking metrics assume block artifacts appear only on 8×8 block boundaries. However, this is generally not true in P-frames, where blocking artifacts have also propagated from previous frames. Finally, most of these metrics average the blockiness across the image. Thus, a very strong edge will be averaged with weaker edges. On the other hand, humans are likely to perceive blockiness using only the most visible blockiness. HVS masking is often used to give more weight to stronger edges, but often it is not enough.

As shown by the results for blurring metrics in Table 1(c), all are effective at showing (D), that filtering increases blurriness, while none were able to show (E), that increasing d increases blurriness. Only [23] was successful with (F), the LT prediction, while Kurtosis was the weakest with regard to Expectation (C).

None of the blurring metrics incorporates any HVS modeling. The Kurtosis metric and [23] are similar to PIQE-S, in that they are heavily dependent on obtaining reliable edge information. Filtering during compression reduces the number of available edges in their sample space, which decreases their statistical reliability. On the other hand, [22] discards potentially significant information when forming its histogram of received DCT coefficients. Since it is based on received DCT coefficients, its performance for P-frames degrades with heavier quantization.

6. CONCLUSIONS

Our systematic evaluation shows that the recently proposed quality metrics we consider all have some weakness in measuring the quality of still frames from compressed video. We derive our expectations using simple subjective comparisons, and each metric is unable to correctly rank order images for at least one of our expectations. Several metrics also prove inadequate when applied on H.264 video, since they are designed for 8×8 DCT blocks. Kurtosis and Power Spectrum are particularly weak for H.264 due to its rich blocking structure that involves blocks as irregular as 4×8.

The most challenging expectations for the metrics to satisfy, as a whole, are to correctly characterize (A) the impact of spatial content, (B) the impact of blockiness in P-frames, (E) the increased blurriness as the distance d from the most recent I-frame increases, and (E) the increased blockiness with increasing d. Among these, the inability to characterize the second and fourth of these are due to the fact that these metrics have been designed for images, and are then applied to stills taken from video. However, because MC propagates blocking artifacts into subsequent frames such that they are no longer aligned on block boundaries, these impairments are not observed by most metrics.

7. REFERENCES

[1] C. J. Van den Branden and O. Verscheure, "Perceptual quality measure using a spatio-temporal model of human video system," in *Proc. SPIE EI, vol. 2668*, 1996, pp. 451–461.

[2] M. A. Masry et al., "A metric for continuous quality evaluation of compressed video with severe distortions," *Sig. Proc.: Im. Comm.*, vol. 19, pp. 133–146, Feb. 2004.

[3] Z. Yu et al., "Vision-model-based impairment metric to evaluate blocking artifacts in digital video," *Proc. of the IEEE*, vol. 90, no. 1, pp. 154–169, Jan. 2002.

[4] K. T. Tan et al., "A multi-metric objective picture-quality measurement model for MPEG video," *IEEE Trans. on CSVT*, vol. 10, no. 7, pp. 1208–1213, Oct. 2000.

[5] A. M. Eskicioglu, "Quality measurement for monochrome compressed images in the past 25 years," in *ICASSP*, 2000.

[6] A. M. Eskicioglu et al., "Image quality measures and their performance," *IEEE Trans. on Comm.*, vol. 43, Dec. 1995.

[7] P. Cuenca et al., "Study of video quality metrics for MPEG-2 based video communications," in *PACRIM CCSP*, 1999.

[8] A. Mayache et al., "A comparison of image quality models and metrics based on human visual sensitivity," in *Proc. IEEE ICIP*, 1998.

[9] S. Winkler et al., "Perceptual video quality and blockiness metrics for multimedia streaming applications," in *Proc. ISWPMC*, Sept. 2001.

[10] A. Webster et al., "An objective video quality assessment system based on human perception," in *Proc SPIE, vol 1913*, 1993, pp. 15–26.

[11] T. Vlachos, "Detection of blocking artifacts in compressed video," *IEE El. Let.*, vol. 36, no. 13, June 2000.

[12] Z. Wang et al., "Blind measurement of blocking artifacts in images," in *Proc. IEEE ICIP*, 2000, vol. 3, pp. 981–984.

[13] H. R. Wu et al., "A generalized block-edge impairment metric for video coding," *IEEE Sig. Proc. Let.*, Nov. 1997.

[14] M. C. Q. Farias et al., "Perceptual contributions of blocky, blurry and noisy artifacts to overall annoyance," in *Proc. IEEE ICME*, 2003.

[15] Z. Wang et al., "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. on Im. Proc.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[16] R. W. Chan et al., "A psychovisually-based image quality evaluator for JPEG images," in *Proc. IEEE ICSMC*, 2000.

[17] J. Yang et al., "Noise estimation for blocking artifacts reduction in DCT coded images," *IEEE Trans. on CSVT*, vol. 10, no. 7, pp. 1116–1134, Oct. 2000.

[18] S. Minami et al., "An optimization approach for removing blocking effects in transform coding," *IEEE Trans. on CSVT*, vol. 5, no. 2, pp. 74–82, Apr. 1995.

[19] W. Gao et al., "A de-blocking algorithm and a blockiness metric for highly compressed images," *IEEE Trans. on CSVT*, vol. 12, no. 12, pp. 1150–1159, Dec. 2002.

[20] S. Liu et al., "Efficient DCT-domain blind measurement and reduction of blocking artifacts," *IEEE Trans. on CSVT*, vol. 12, no. 12, pp. 1139–1149, Dec. 2002.

[21] S. Suthaharan, "Perceptual quality metric for digital video coding," *IEE El. Let.*, vol. 39, no. 5, pp. 431–433, Mar. 2003.

[22] X. Marichal et al., "Blur determination in the compressed domain using DCT information," in *ICIP.*, Oct. 1999.

[23] P. Marziliano et al., "A no-reference perceptual blur metric," in *Proc. IEEE ICIP*, 2002, vol. 3, pp. 57–60.

[24] J. Caviedes et al., "No-reference sharpness metric based on local edge kurtosis," in *Proc. IEEE ICIP*, 2002.

[25] A. Leontaris et al., "Optimal mode selection for a pulsed-quality dual frame video coder," *IEEE SPL*, to appear.

[26] MPlayer 1.0-pre4 software, http://www.mplayerhq.hu/.

[27] M. G. Kendall, *Rank correlation methods*, Hafner Publishing Co, New York, 1955.