A NEW OBJECTIVE QUALITY METRIC FOR FRAME INTERPOLATION USING IN VIDEO COMPRESSION

Kai-Chieh Yang, Ai-Mei Huang, Truong Nguyen, Clark C. Guest and Pankaj K. Das

Department of Electrical and Computer Engineering
University of California, San Diego
La Jolla, CA 92093
E-mail: kcyang@ucsd.edu, aihuang@ucsd.edu, nguyent@ucsd.edu, cguest@ucsd.eud, das@cwc.ucsd.edu

ABSTRACT

This paper discusses the disadvantages of several existing objective quality evaluation methods for frame interpolation techniques. Samples show that these disadvantages lead the objective quality measurement inconsistent with what humans perceive. Based on these observations, a new metric designed to evaluate the performance of frame interpolation techniques is proposed. This metric combines the severity of interpolation artifacts and several human visual factors into a single quality score. Final implementation shows that the proposed metric out-performs other commonly used metrics.

Index Terms— Objective quality assessment, Frame interpolation, Visual attention

1. INTRODUCTION

In order to meet low bandwidth requirements, video applications such as video telephony usually need to reduce temporal resolution by skipping frames. However, low frame rate video may result in motion jerkiness, especially when the scenes have fast or complex motion. In such a case, motion compensated frame interpolation (MCFI) is often adopted at the decoder to improve temporal video quality.

MCFI interpolates the skipped frames by averaging forward and backward motion compensated predictions using the received motion vectors (MVs). At the encoder, these MVs are generated using block matching algorithm to maximize coding efficiency, rather than finding true motion. As a result, MCFI that directly uses the received MVs often suffers from annoying artifacts such as blockiness, ghost effect, and discontinuous edges.

To solve this problem, a number of MV processing techniques have been proposed to obtain a better motion vector field (MVF) for MCFI. The work in [1] presented an adaptively weighted vector median filter based on prediction residues to obtain a smoother motion field at the encoder. To eliminate blockiness in the interpolated frame, resampling the MVF into finer field with smoothness measurement is presented in [2]. In [3], instead of using high complexity motion re-estimation at the decoder, the authors proposed MV selection that selects the best MV for each merged group from the neighboring MVs based on minimizing the difference between forward and backward motion compensations.

This work is supported in part by Conexant Inc. and matching found from UC Discovery Program.

2. PROBLEM STATEMENT

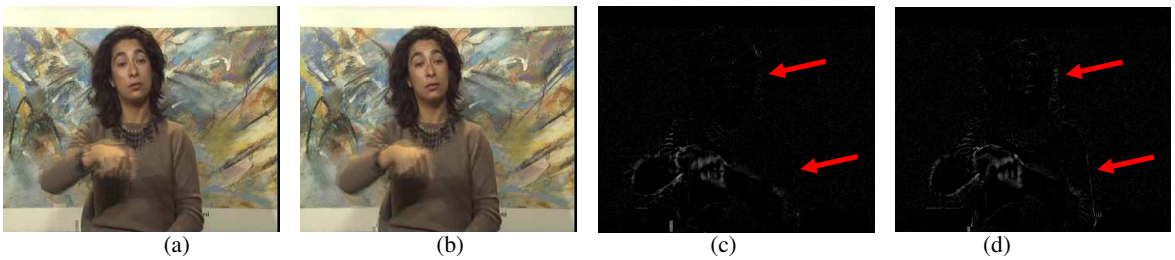
Accurate quality assessment is very essential to understand the performance of different MCFI techniques. Subjective evaluation [1, 2] is the most convincing approach because it collects direct responses from end users. However, it is inconvenient, expensive, and time consuming. Objective methods provide an alternative feasible solution. Most researches evaluate the interpolated frame quality using fidelity metrics. Reference [3] uses Peak-Signal-to-Noise-Ratio (PSNR), a normalized Mean-Square-Error (MSE) between original and processed images, to measure the interpolation quality. Reference [4] measures the quality by Structure Similarity (SSIM) metric from [5]. SSIM uses a combination of three components, luminance, contrast, and structure similarity comparisons, as the quality index. However, the fidelity-wise measurement could fail because of the following three reasons:

1 *Pixel shift*: The moving region is the most challenging part for MCFI. Some approaches excel in this region but it also changes the pixels in the static region. Nevertheless, this pixel shift is hard to notice and human perceived quality is good, but the fidelity metrics usually yield a low quality score. Some examples are shown in Fig. 1. Visual observation shows that Fig. 1(a) is worse than Fig. 1(b), but PSNR gives a higher quality score to Fig. 1(a). From Fig. 1(c) and (d), we observe that both approaches have similar amount of distortion on the moving hand, but Fig. 1(c) has less distortion than Fig. 1(d) on the non-moving region. Therefore, we can reasonably infer that the lower PSNR value of Fig. 1(b) is resulted from the imperceivable distortion in the static region.

2 *Artifacts dominance*: Fig. 1(a) and (b) show some examples of blocking and ghost artifacts. Different sensitivity to each artifact affects the judgment of final quality. Hence, using fidelity measurement alone is not sufficient to capture human perceived quality.

3 *Moving region dominance*: The moving region introduces the most artifacts during frame interpolating. In addition, humans tend to pay more attention to the moving region. Hence, higher weight should be applied to this region when pooling the local spatial quality measurement.

In order to overcome the problems stated above, a novel metric for MCFI is proposed. This metric uses the degree of both blocking and ghost artifacts as the basic measurements. These measurements are adjusted by local motion activity. The final quality score is calculated by an artifacts dominance adaptive integrator.



(a) (b) (c) (d)

Fig. 1. (a) and (b) are the interpolated frame produced by direct MCFI and MV smoothing technique from [2], and the corresponding PSNR values are 31.77dB and 31.55dB respectively, (c)and (d) are the MSE map compared against original frame.

The rest of this paper is organized as follows. Section 3 describes the proposed metric in detail. Section 4 shows some comparisons of quality predicted from different metrics. The conclusion is summarized in Section 5.

3. PROPOSED METRIC

Figure 2 shows the system diagram of the proposed metric. A blockiness metric is used to estimate the amount of blocking artifact, whereas a modified SSIM is used for estimating the ghost artifact. A moving region extractor outputs a moving region map to emulate various sensitivity to the moving region. Finally, all measurements are integrated with a weighted sum. The weights are determined by the dominance of each artifact.

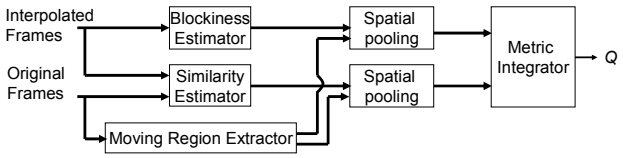


Fig. 2. System diagram of frame interpolation metric

3.1. Blockiness Estimator

A well known blockiness metric from [6] is adopted to measure the amount of blockiness artifacts. This metric first calculates the pixel value discontinuity at each 8×8 boundary. Because blocking artifact cannot be recognized in too dark or too bright lighting condition, the discontinuity is weighted by a luminance masking function. The adjusted pixel discontinuity is normalized by the inter block pixel difference to avoid false alarms from real edges. Finally, a blockiness map is calculated. It will be further adjusted based on its perceptual importance later on.

Consider an image I that is composed of $\{I_{c1}, I_{c2} \dots I_{cN_c}\}$, where I_{cj} indicates the jth column of the image. The core of estimating horizontal blockiness artifacts is

$$B_h = \|w_k(I_{c(8\times k)} - I_{c(8\times k+1)})\|^2, \tag{1}$$

where $\|\cdot\|$ is the l_2 norm and w_k is the output of a luminance masking function used to adjust the perceptual importance of each boundary discontinuity. Let w_k be $w_{i,j}$, where $i = 1, 2\cdots, N_R$, $j = 8\times k$ for $k = 1, 2, \cdots, N_C/8-1$, and N_C , N_R are the width and height of the image respectively. The luminance masking function is defined as

$$w_{i,j} = \begin{cases} \tau ln(1+\frac{\sqrt{\mu_{i,j}}}{1+\sigma_{i,j}}) & \text{if } \mu_{i,j} \le \zeta \\ \tau ln(1+\frac{\sqrt{255-\mu_{i,j}}}{1+\sigma_{i,j}}) & \text{otherwise} \end{cases} \tag{2}$$

where

$$\tau = \frac{ln(1+\sqrt{255-\zeta})}{ln(1+\sqrt{\zeta})} \tag{3}$$

and ζ is set as 81, $\mu_{i,j}$ and $\sigma_{i,j}$ are the mean value and standard deviation of the pixels on the same row within two adjacent blocks respectively, which can be calculated by

$$\mu_{i,j} = \frac{1}{16}\sum_{n=-7}^{8} I(i, j+n) \tag{4}$$

and

$$\sigma_{i,j} = \{\frac{1}{16}\sum_{n=-7}^{8}[I(i, j+n) - \mu_{i,j}]^2\}^{1/2}. \tag{5}$$

The final horizontal blockiness map, B'_h, is obtained after normalizing the discontinuity with the average inter-pixel difference of the non-boundary pixels as $B'_h = B_h/E_h$, where

$$E_h = \frac{1}{7}\sum_{n=1}^{7}\Psi_n, \tag{6}$$

and

$$\Psi_n = \sum_{k=1}^{N_c/8-1}\|w_k(I_{c(8\times k+n)} - I_{c(8\times k+n+1)})\|^2. \tag{7}$$

The vertical blockiness map, B'_v, can also be obtained similarly with horizontal blockiness map.

In order to combine the blockiness map with the moving region map, the blockiness map must be transformed from a boundary basis to a block basis. Therefore, the final blockiness map is

$$\mathcal{B}(i', j') = 1 - \frac{1}{16}\sum_{n=0}^{7} B'_h(8i'+n-7, j'+7) + B'_v(i'+7, 8j'+n-7), \tag{8}$$

where (i', j') is the index of each 8×8 block, $i' = 1, 2, \cdots, N_R/8-1$ and $j' = 1, 2, \cdots, N_C/8-1$. Higher $\mathcal{B}$ implies less blockiness and better quality.

3.2. Similarity Estimator

This estimator is used to measure the severity of the ghost effect using structure similarity measurement. Let I_x and I_y be the original and interpolated images respectively. The SSIM from [5] is modified to avoid the pixel shift problem indicated in Section 2. Because the first two components described at Section 2 are used for pixel-wise fidelity and the last component is a more structure similarity oriented measurement, we only use the last component. The equation for the similarity estimator is

$$s(x,y) = \frac{\sigma_{xy}}{\sigma_x + \sigma_y}, \tag{9}$$

where σ_{xy}, σ_x and σ_y are the correlation and standard deviations of I_x and I_y respectively. The similarity map is further processed into block base by

$$\mathcal{S}(i',j') = \frac{1}{64}\sum_{n=1}^{8}\sum_{n'=1}^{8} s[8(i'-1)+n, 8(j'-1)+n']. \tag{10}$$

Higher S implies higher structure similarity and fewer ghost artifacts.

3.3. Moving Region Extractor

A background subtraction method from [7] is adopted to separate the moving region from non-moving region. The movement of each pixel is modeled by a mixture Gaussian kernel along the temporal axis of R,G and B color channels.

Each pixel of each frame for all three color channels is compared against the pixels at the same spatial location in previous T frames and input into a mixture Gaussian model,

$$Pr(i,j)^{(n)} = \{1 - \frac{1}{T}\sum_{t=1}^{T}\prod_{c}^{3}\frac{1}{\sqrt{2\pi\sigma_c^{(n)2}}}e^{-\frac{1}{2}\frac{[D(i,j)_c^{(n)}]^2}{\sigma_c^{(n)2}}}\}^{\alpha_m}, \tag{11}$$

where n is the frame index, $D(i,j)_c^{(n)} = I(i,j)_c^{(n)} - I(i,j)_c^{(n-t)}$, and $\sigma_c^{(n)2} = m_c^2/0.9248$, the variance of cth color channel for a given pixel, which can also be thought of the bandwidth of the mixture Gaussian kernel where m_c is the median of $|I^n - I^{n+1}|$ for each consecutive pair (I^n, I^{n+1}), and α_m is a constant used to emphasize the moving region. Higher value in Equation(11) means that the pixel is more likely to be considered as part of the moving-region. The final moving region map is

$$\mathcal{M}(i',j') = \frac{1}{64}\sum_{n=1}^{8}\sum_{n'=1}^{8} Pr[8(i'-1)+n, 8(j'-1)+n']. \tag{12}$$

3.4. Spatial Pooling

The quality scores from both blockiness and similarity metrics of each block are pooled together with different spatial importance given by moving region map. The final blockiness and similarity measurements are

$$Q_B = \frac{1}{(\frac{N_C}{8}-1)(\frac{N_R}{8}-1)}[\sum_{i',j'}\mathcal{M}(i',j')\cdot\mathcal{B}(i',j')], \tag{13}$$

and

$$Q_S = \frac{1}{(\frac{N_C}{8}-1)(\frac{N_R}{8}-1)}[\sum_{i',j'}\mathcal{M}(i',j')\cdot\mathcal{S}(i',j')]. \tag{14}$$

3.5. Metric Integrator

The final quality Q is produced by using a weighted sum of Q_B and Q_S as

$$Q = \omega_B(\beta_B Q_B + \Phi_B) + \omega_S(\beta_S Q_S + \Phi_S), \tag{15}$$

where Q ranges from 0 to 1 and higher value implies better quality. The parameters β_B, Φ_B, β_S, Φ_S are factors used to normalize Q_B and Q_S within 0 to 1, and ω_B, ω_S are the weights for combining Q_B and Q_S. Since the blockiness metric is only good in measuring blocking artifacts, it will give a high score (good quality) if ghost artifact is more pronounced than blocking artifact. Hence, an evenly weighted sum of the scores from the blockiness and similarity measurements will inevitably underestimate the amount of distortion in the video under test. Therefore, we have designed a mechanism to select the weights based on the dominance of each artifact. The factor governing the weights selection is γ, where $\gamma = (\beta_B Q_B + \Phi_B)/(\beta_B Q_B + \beta_S Q_S + \Phi_B + \Phi_S)$. If γ is larger than some threshold, then it means that similarity is low and the final quality should be decided by Q_S mainly. Otherwise, the blockiness is considered significant, and higher weight will be applied on normalized Q_B. Detail of the selection threshold and the predefined values of ω_B and ω_S are given in Table 1.

	$0.53 < \gamma$	$0.47 \leq \gamma \leq 0.53$	$\gamma < 0.46$
ω_B	0.3	0.5	0.7
ω_S	0.7	0.5	0.3

Table 1. Values of ω_B and ω_S and the applicable scenarios

4. SIMULATION

Two video sequences, FOREMAN and SILENT, of CIF frame resolution are used with original frame rate of 30 frame per second (fps). They are encoded using H.263, but even frames are skipped to generate video bitstreams of 15 fps. The rate control function is disabled by fixing quantization parameter (QP) values. The averaged bit rates of these two test sequences are 395.77 Kbps and 372.70 Kbps for FOREMAN and SILENT, respectively. The skipped frames are interpolated at the decoder for evaluation of the proposed method using direct MCFI, vector median filter [1], MV smoothing as described in [2], and MV selection in [3] but with fixed block size.

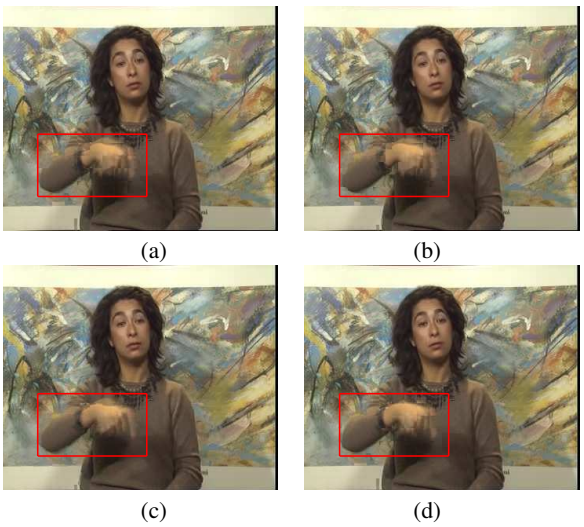
Figures 3 and 4 show the 94th and 64th frame of FOREMAN and SILENT produced by different interpolation approaches, respectively. The corresponding quality measurements from the proposed metric, SSIM and PSNR are shown in Tables 2 and 3.

In Fig. 3, we observe that the face structure is barely preserved and many blocking artifacts are introduced. Fig. 3(a) has the greatest blocking and ghost artifacts. Fig. 3(b) has fewer blockiness but the face structure is still highly distorted. Fig. 3(c) has the least blocking artifact but still fails to preserve the face structure. Although Fig. 3(d) has more blockiness than Fig. 3(c), it preserves the face structure relatively better. Subjectively, we conclude that Fig. 3(d) has the best quality, and the rest from high to low are (c), (b), and (a). From Table 2, the proposed metric is consistent with our visual evaluation (i.e., Fig. 3(d) has the highest score). In the other hand, SSIM metric yields the same value for Fig. 3(b) and (c), whereas the PSNR values for Fig. 3(b) is higher than that of Fig. 3(c). These metrics do not match our subjective observation.

In Fig. 4, most artifacts occur around the moving hand and the amount of blockiness from high to low is (a) $>$ (b) $>$ (d) $>$ (c).

(a) (b) (c) (d)

Fig. 3. The interpolated results of frame 94 of FOREMAN sequence. (a), (b), (c), and (d) are the results produced by Direct MCFI, vector median filter in [1], MV smoothing in [2], and MV selection described in [3] respectively



(a) (b) (c) (d)

Fig. 4. The interpolated results of frame 64 of SILENT sequence. (a), (b), (c), and (d) are the results produced by Direct MCFI, vector median filter in [1], MV smoothing in [2], and MV selection described in [3] respectively

Figure 4(a), (b), and (c) have similar degree of ghost effect and Fig. 4(d) has the least. Although Fig. 4(d) has slightly more blockiness than Fig. 4(c), Fig. 4(d) preserves the hand structure relatively better than Fig. 4(c). Overall, the subjective quality ranking from high to low is Fig. 4(d), (c), (b) and (a). In Table 3, the proposed metric gives Fig. 4(a) and (b) almost equivalent low quality score and Fig. 4(c), (d) higher score. This measurement matches our visual observation closely. SSIM and PSNR give the lowest quality score to Fig. 4(c), which is contradictory to our subjective evaluation. This is the result of pixel shift in the non-moving-region as discussed in Section 2.

	Proposed metric	SSIM	PSNR(dB)
Direct MCFI	0.26	0.79	26.59
Median filtering [1]	0.32	0.80	26.90
MV smoothing [2]	0.51	0.80	26.64
MV selection [3]	0.61	0.82	27.24

Table 2. Quality comparison for FOREMAN 94th frame

	Proposed metric	SSIM	PSNR(dB)
Direct MCFI	0.48	0.84	31.77
Median filtering [1]	0.40	0.84	31.70
MV smoothing [2]	0.67	0.83	31.55
MV selection [3]	0.73	0.84	32.12

Table 3. Quality comparison for SILENT 64th frame

5. CONCLUSION

This paper investigates the performance of widely used quality metrics for frame interpolation. Based on the investigation, a new metric for measuring the quality of interpolated frames is proposed. This metric is designed based on several prior knowledge about frame interpolation, such as type of artifacts, possible region of quality degradation, and the various dominance of different artifacts. The proposed metric also has been implemented and compared against other metrics, such as PSNR and SSIM. The results show that the proposed metric is able to provide more accurate quality assessment.

The proposed metric will be further improved by adding more human visual factors, such as texture and temporal masking. This will be investigated in future work.

6. REFERENCES

[1] L. Alparone, M. Barni, F. Bartolini, and V. Cappellini, "Adaptively weighted vector-median filters for motion-fields smoothing," *Proc. ICASSP'96*, vol. 4, pp. 2267–2270, May 1996.

[2] G. Dane and T. Q. Nguyen, "Smooth motion vector resampling for standard compatible video post-processing," *Proc. Asilomar Conf. Signals, Systems and Computers*, 2004.

[3] A.-M. Huang and T. Q. Nguyen, "A novel motion compensated frame interpolation based on block-merging and residual energy," *Proc. IEEE MMSP'06*, 2006.

[4] J. Wang, N. Patel, and W.I. Grosky, "A fast block-based motion compensation video frame interpolation approach," *Proc. Asilomar Conf. Signals, Systems and Computers*, pp. 1740–1743, 2004.

[5] Z. Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Transaction on Image Processing*, vol. 13, pp. 600–612, April 2004.

[6] H.R. Wu and M. Yuen, "A generalized block-edge impairment metric for video coding," *IEEE Signal Processing Letters*, vol. 4, pp. 317–320, Nov. 1997.

[7] A. Elgammal, R. Duraiswami, D. Harwood, and L. S. Davis, "Background and foreground modeling using non-parametric kernel density estimation for visual surveillance," *Proc. IEEE*, vol. 90, pp. 1151 – 1163, 2002.