PERCEPTUAL SHARPNESS METRIC (PSM) FOR COMPRESSED VIDEO

Kai-Chieh Yang, Clark C. Guest and Pankaj Das

Department of Electrical and Computer Engineering
University of California at San Diego
9500 Gilman Drive
La Jolla, CA 92093

ABSTRACT

Sharpness, one of the most effective factors in video quality assessment, usually dominates the first impression of the representation of the compressed video or image signals. In this paper, a new sharpness metric is presented. Without the original video sequence, this metric evaluates the level of sharpness of a compressed video sequence based on the presence of high frequency signals components. Also, an attention module and several human visual factors are included in order to make the measurement results more correlated to human perception. Finally, psychovisual experiments show high correlation between the metric prediction and subjective ranking of video sharpness.

1. INTRODUCTION

In video compression, spatial information is quantized in order to reduce the amount of data needed to be encoded based on the assumption of low visibility of high texture information. However, if the compression ratio is too high, many spatial details are disappear and perceived quality is degraded. Among various factors for video quality assessment, sharpness has been reported as one of most important attributes. Therefore, how to accurately measurement of the users' perceived sharpness is a very critical task for balancing the represented quality with the compression ratio. Several related research has been done and can be categorized into spatial and frequency domain approaches.

In a spatial approach, Zhang and Marziliano [1, 2] developed several metrics that measure the sharpness by estimating spatial activity around edges. The assumption is that a sharp image has larger local pixel differences and edge slope but smaller edge width than a blurred image. With the guidance of the edge profile, informative regions for sharpness quantification can be defined. These algorithms rely heavily on accurate edge detection and are designed for still images. In a frequency domain approach, Caviedes and N. Zhang[3, 4] designed a sharpness metric based on measuring the skewness of local Discrete Cosine Transform (DCT) coefficients around each edge location.

In these related works, edge information supplies a very important clue for locating the regions of noticeable sharpness. However, edge detection can fail if some compression artifacts are too strong; i.e., blurriness, blockiness and accurate edge location might not be available. Also, most of the previous works assume that human perception has the same sharpness sensitivity at all spatial locations. However, the fact is that viewers care about the sharpness in the foreground much more than in background. In addition, most of the previous research focuses on the application to individual images and treats a video sequence as a set of independent images. However, as video is being played, some dependency between frames can make subjective observation far different from simply averaging sharpness output through all frames. Therefore, a new approach - *Perceptual Sharpness Metric (PSM)* is presented in this paper The PSM can effectively extract the informative regions based on human sensitivity to sharpness in different frequency bands and spatial locations. Moreover, several important Human Visual System (HVS) phenomena, such as luminance and motion masking are emulated, and the predicted sharpness results show high correlation with subjective ranking. The main idea behind the proposed metric is measuring the sharpness of each frame by estimating the energy of high frequency signals in a block based DCT transformation. The DCT coefficients are weighted based on their corresponding sharpness perceptibility and each block is adjusted by an attention module and lighting condition. Afterward, the sharpness of each frame is adjusted with respect to its global motion activity. Finally, a perceptual sharpness score is obtained by taking an average through all frames with various weights.

This paper is organized as followings. Section 2 gives a detailed explanation of the proposed approach. Section 3 is the experiment set up and simulation results, conclusions are summarized in Section 4.

2. THE PROPOSED METRIC

From a spatial quality point of view, sharp images usually have more acute edges than blurred images. As a result, in the frequency domain, sharp images contain more high frequency energy than blurred images. Based on this property, the correlation between energy of the AC coefficients in the Discrete Fourier Transform(DCT) and sharpness estimation will be employed as basis of the proposed metric. Afterward, several important human visual factors will be considered. The block diagram is shown in Fig. 1.

2.1. Channel Decomposition

Let the n_{th} frame of one video sequence to be denoted as $f^{(n)}(x, y)$ where $x \in 1, 2, ...Width$, $y \in 1, 2, ...Height$ and $Width$, $Height$ are the horizontal and vertical length of this image. After non-overlapped 8×8 block-based DCT, each block is denoted as $B_{ij}^{(n)}$ where i, j are the indices of each block. Fig. 2 shows the DCT basis function for an 8×8 block. The first column and row represent the frequency response to horizontal and vertical edge structures. With some investigation, we found these two orientations are sufficient for sharpness estimation. Hence, only these two sets of coefficients will be taken into account, and are denoted as $H_{i,j}^{(n)}$, $V_{i,j}^{(n)}$ respectively.

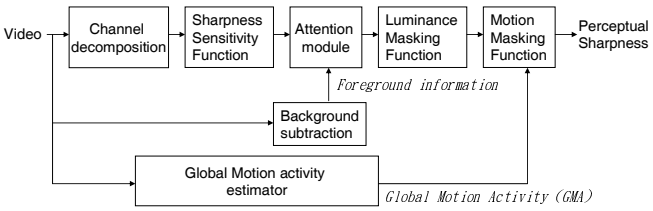


Fig. 1. Block diagram of PSM.

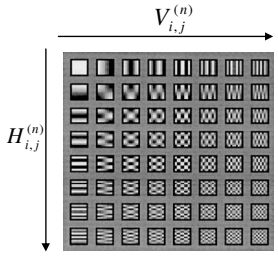


Fig. 2. Basis function of DCT.

2.2. Sharpness Sensitivity Function

Not all DCT coefficients are informative for perceptual sharpness. The energy of some coefficients may come from compression artifacts or uncorrelated signal, i.e. spurious edges from compression artifacts or flat regions. Hence, different weightings should be applied for each frequency channel, several related works have been detail investigated at [5]. A similar approach-*Sharpness Sensitivity Function (SSF)* is adopted and revised to fit the system designing. The SSF is described by Equation(1) and its response is shown in Fig. 4(a)

$$SSF(d) = (a_1 + b_1 \cdot d) \cdot exp(-c_1 \cdot d) \cdot d^{e_1} \tag{1}$$

where d is denoted as index of DCT coefficients and $d \in 1, 2...8$. In Fig. 4(a), the third and forth coefficients are given the higher weights - because most of the noticeable sharpness information is located in these channels. For the same reason, very low weights are assigned to the lowest frequency channel(i.e. DC) and the very high frequency coefficients to reduce the influence of flat regions and noise. Associated parameters, $a_1 = -3.533$, $b_1 = 3.533$, $c_1 = 0.548$ and $e_1 = 0.269$, have been obtained experimentally. The sharpness of block (i, j) of the n_{th} frame is obtained by summing up the products of the SSF with the ratio of horizontal and vertical DCT coefficients to the DC, as in Equation(2).

$$S_{ij}^{(n)} = \sum_{d=1}^{8} \frac{SSF(d)}{B_{ij}^{(n)}(1,1)} \cdot [H_{ij}^{(n)}(d) + V_{ij}^{(n)}(d)] \tag{2}$$

2.3. Attention Model and Background Subtraction

When evaluating the sharpness of an image, observers are usually more sensitive to foreground than background. Therefore, the sharpness of an image should be estimated by taking the average through all the blocks that belong to foreground only. In order to know which part of an image belongs to the foreground, a well known background subtraction approach[6] has been adopted. It is robust against moving backgrounds and has low computational complexity. In this approach, every background pixel in a time series is modeled by a Gaussian distribution. However, a single Gaussian model is not sufficient to model the background accurately since it might be confused by some small movement or the scene changing. In order to make the model more generic and adaptive to various video content, the mixture Gaussian model at Equation(3) is used to compensate the shortcoming of the single Gaussian model.

(a) (b)

Fig. 3. Examples of background subtraction of (a) *Foreman* and (b) *Carphone*.

$$P(x^{(n)}) = \frac{1}{T} \sum_{t=1}^{T} \prod_{c}^{C} \frac{1}{\sqrt{2\pi\sigma_c^2}} e^{-\frac{1}{2} \frac{(x_c^{(n)} - x_c^{(n-t)})^2}{\sigma_c^2}} \tag{3}$$

where $x^{(n)}$ is a pixel value of n_{th} frame and T is the number of look-back frames. The variable C is the dimension of feature vector of each pixel point. The feature vector here is comprised by three-dimension chromatic values. When the output of Equation(3) for a pixel at frame n is larger than a threshold, then this pixel will be classified as background. By this mechanism, we can separate foreground and background blocks and construct an attention mask. Some examples are shown in Fig. 3.

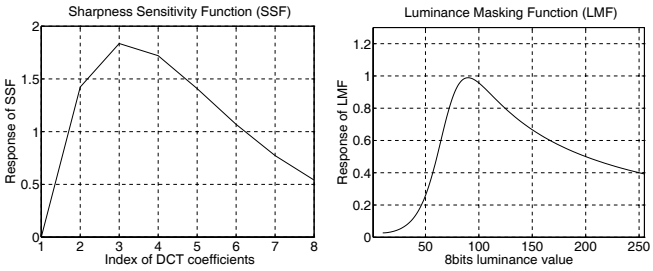
2.4. Luminance Masking Function

Luminance masking is a phenomenon by which human eyes have less discriminating ability under too bright or too dark lighting conditions. As masking occurs, sharpness will be less noticeable or incorrectly judged. In order to make the proposed metric closer to perceptual estimation, this phenomenon is emulated by the *Luminance Masking Function (LMF)* in Equation(4), and Fig. 4(b) is its response.

$$LMF(l) = \frac{1}{10 \cdot l \cdot (a_2 + exp(-b_2 \cdot l))} \tag{4}$$

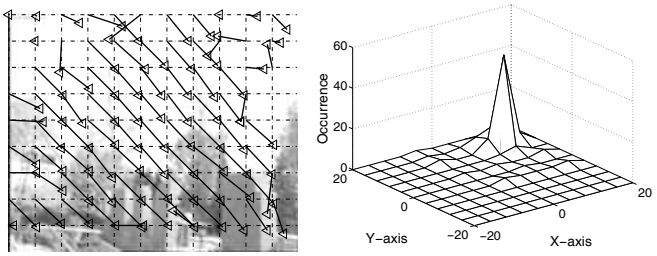
where l is the average luminance value that ranges from 10 to 255 and parameters $a_2 = 0.001$, $b_2 = 0.1$ are obtained experimentally. In Fig. 4(b), it is interesting to note that the LMF provides larger weights when the average luminance value lies between 80 and 90. Based on some subjective observations and previous research[7], image content can be most correctly distinguished under those lighting conditions. Therefore, the LMF gives higher weight to 8×8 blocks whose average luminance lies within this interval. The adjusted sharpness score is the product of the LMF and the sharpness output for the foreground at n_{th} frame, as expressed in Equation(5).

$$S_1^{(n)} = \frac{1}{B'_W B'_H} \sum_{i'=1}^{M'/8} \sum_{j'=1}^{N'/8} S_{i'j'}^{(n)} \cdot LMF(\overline{f_{i'j'}^{(n)}}) \tag{5}$$



(a) SSF (b) LMF

Fig. 4. Response of (a)Sharpness Sensitivity Function(SSF), (b)Luminance Masking Function(LMF).



(b) Motion Vectors (b) Distribution of motion vectors

Fig. 5. Example of distribution of motion vectors.

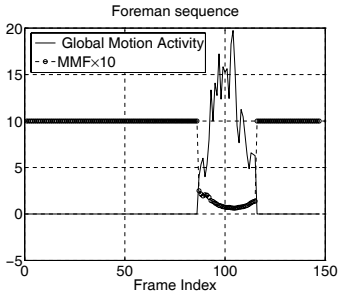


Fig. 6. Example of Global motion activity vs. Motion masking function(MMF) of Foreman sequences at frame rate 15fps.

where $i', j' \in foreground$, $B'_W B'_H$ is the total number of foreground blocks and $\overline{f_{i'j'}^{(n)}}$ is the average luminance value for 8×8 block at location (i', j') of n_{th} frame.

2.5. Motion Masking Function

Video playing is a process of showing a set of images with a very short time interval. The motion activity of the shown objects can be estimated by the amplitude of their corresponding motion vectors. In addition, motion activity can be separated into three types, which are global, local and ambiguous motion activities. Global motion is usually introduced by camera shifting and most of the motion vectors are pointed in a similar direction. Local motion is the related motion between objects and the camera that usually happens as camera is static but object is moving. Ambiguous motion usually occurs in flat regions, which causes the motion estimator to fail to find the right prediction; information of this type of motion is less reliable.

As global motion activity increases, spatial content will be less visible, equivalent to a low pass spatial filter mechanism[8]. Since sharpness features belong to the high frequency bands, their visibility will be decreased as global motion increases. To imitate this human visual effect, the *Motion Masking Function (MMF)* in Equation(6)is introduced, where $gma(n)$ is the global motion activity between the n_{th} and $n_{th} + 1$ frame, $w(i)$ is the weight of i_{th} frame and fr is the playing frame rate. Finally, $MMF(n)$ is truncated so that $MMF(n) \in [0, 1]$. In order to distinguish reliable motion activities from unreliable ones, only the motion vectors belonging to the blocks with high texture and small Sum of Absolute Differences (SAD) will be taken into account. Since most reliable motion vectors are pointed in a similar direction when global motion activity occurs, the distribution of all reliable motion vectors concentrates around some interval as shown in Fig. 5. Therefore, the Global Motion Activity (GMA) is the summation of the amplitude of motion vectors that have the highest frequency of occurrence in reliable motion vectors.

$$MMF(n) = \frac{\sum_{i=1}^{fr/2} w(i)}{\sum_{i=1}^{round(fr/2)} w(i) \cdot gma(n-i)} \tag{6}$$

Since video playing is a continuous and causal process, motion masking of the current frame is only related to previous frames. Thus MMF looks backward a certain number of frames and $fr/2$ is chosen as an appropriate number. Moreover, because the video sequence is synthesized from a group of still images, some dependent relationship exists between each pair of consecutive frames. This relationship fades as temporal distance increases because of the cognitive limitation. Therefore, $w(i)$ decreases monotonically as temporal distance between the look-back frame and the current frame increases. Fig. 6 shows one example of global motion activity vs. MMF output for the *Foreman* sequence. The output of MMF is multiplied by 10 for illustration purpose. In *Foreman*, as strong global motion activity occurred within frame index $90 \rightarrow 120$, response of the MMF decreased to a very low value, meaning the sharpness is least visible. On the contrary, for scenes at rest, the camera motion is more static, and the output of the MMF increases to a large value which means sharpness is most visible.

The final perceptual sharpness of a whole sequence is obtained by taking the average of the weighted sharpness output through each frame as expressed in Equation(7), where N is the number of frames.

$$S_{PSM} = \sum_{n=1}^{N} S_1^{(n)} \cdot MMF(n) \tag{7}$$

3. EXPERIMENTAL RESULTS

3.1. Experiment Setup

In our experiments, two standard video sequences, *Foreman* and *Carphone* have been used for evaluating the performance of our sharpness estimation. The video sequences are sampled in 4:2:0, 176×144 per frame and 15 frames per second.

When video is compressed, the blurriness increases and sharpness decreases as the quantization parameter (QP) becomes larger. In order to simulate different levels of sharpness, the original video sequences have been encoded under various QP values ranging from 2 to 30 with the MPEG4 reference encoder.

The subjective experiment was carried out by *Double Stimulus Continuous Quality Evaluation (DSCQE)* method. Twenty-three viewers comprised by 5 females an 18 males participated in this experiment, age ranges from 25 to 33.

3.2. Simulation Results

Fig. 7 and Table 1 provide a comparison of *carphone* sequence with various QP values and the corresponding sharpness measurement from different metrics . Higher scoring means a sharper image. All the predicted results are mapped to $1 \rightarrow 10$ by a linear transformation. As shown in Fig. 7, the perceived sharpness decreases as QP increases. With shown in Table 1, only PSM and Kurtosis match this trend. In addition, Table 2 and 3 show the correlation between the MOS data and the output of different metrics of *foreman* and *carphone* sequences. Higher correlation means better prediction performance. In both sequences, we can see that the PSM has better performance than other metrics.

	QP = 2	QP = 12	QP = 17	QP = 20
PSM	9.50	5.97	4.45	3.45
Zhang-AETS [1]	7.96	9.24	6.57	6.70
Zhang-AETW [1]	9.27	8.20	4.52	5.32
Zhang-DSS [1]	8.30	8.70	8.00	7.70
Caviedes-Kurtosis [3]	9.00	5.72	4.37	4.00

Table 1. Sharpness from different metrics for *Carphone* sequence.

	Pearson	Spearman	Kendall
PSM	0.98	0.98	0.93
Zhang-AETS [1]	0.86	0.88	0.71
Zhang-AETW [1]	0.96	0.95	0.87
Zhang-DSS [1]	0.85	0.90	0.77
Caviedes-Kurtosis [3]	0.98	0.98	0.93

Table 2. Correlation between the output of different metrics and the MOS data of *Carphone* sequence.

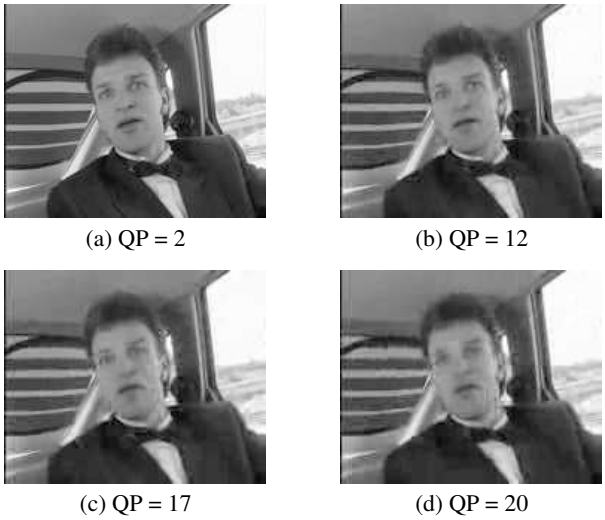
	Pearson	Spearman	Kendall
PSM	0.98	0.99	0.96
Zhang-AETS [1]	0.97	0.98	0.96
Zhang-AETW [1]	0.96	0.98	0.96
Zhang-DSS [1]	0.96	0.98	0.96
Caviedes-Kurtosis [3]	0.97	0.99	0.96

Table 3. Correlation between the output of different metrics and the MOS data of *Foreman* sequence.

4. CONCLUSION

A non-reference HVS-based sharpness metric for compressed video has been proposed in this paper. It employs a sharpness sensitivity function and attention module to extract useful information for image-based sharpness prediction. Several important human visual factors have been included to ensure agreement of the measurement with subjective experiments.

Without guidance from any edge information, the metric provides results consistent with the subjective data and has robust performance when artifacts are apparent. A close-loop video quality enhancement system based on this metric will be included in the future development.



(a) QP = 2 (b) QP = 12

(c) QP = 17 (d) QP = 20

Fig. 7. Compressed *Carphone* with different QP.

5. REFERENCES

[1] B. Zhang, J. P. Allebach, and Z. Pizlo, "An investigation of perceived sharpness and sharpness metrics," in *Proc. SPIE The International Society for Optical Engineering*, 2005, vol. 5668, p. 98.

[2] P. Marziliano, F. Dufaux, S. Winkler, and T. Ebrahini, "Perceptual blur and ringing metrics: Applications to JPEG2000," in *Signal Proccessing: Image Communication*, 2004, pp. 163–172.

[3] J. Caviedes and F. Oberti, "A new sharpness metric based on local kurtosis, edge and energy information," in *Signal Processing: Image Communication*, 2004, vol. 19, pp. 147–161.

[4] N. Zhang, A. E. Vladar, M. T. Postek, and B. Larrabee, "A kurtosis-based statistitcal measure for two-dimensional processes and its application to image sharpness," in *Proc. of Section of Physical and Engineering Sciences of American Statistical Society*, 2003, pp. 4730–4736.

[5] J.A. Movshon and L. Kiorpes, "Analysis of the development of spatial contrast sensitivity in monkey and human infants," in *J. of the Optical Society of America A*, 1988, number 12, pp. 2166–2172.

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

[7] B. Birod, "The information theoretical significance of spatial and temporal masking in video signals," in *Proc. SPIE Conference of Human Vision, Visual Processing and Digital Display*, 1989, vol. 1077, pp. 178–187.

[8] Z. Lu, W. Lin, X. Yang, E. Ong, and S. Yao, "Modeling visual attention's modulatory aftereffects on visual sensitivity and quality evaluation," in *IEEE Transac. Image Processing*, 2005, pp. 1928–1942.