# A NON-REFERENCE PERCEPTUAL QUALITY METRIC BASED ON VISUAL ATTENTION MODEL FOR VIDEOS

*Fahad Fazal Elahi Guraya*[1], *Ali Shariq Imran*[1], *Yubing Tong*[2], *Faouzi Alaya Cheikh*[1]

Gjovik University College, Gjovik, Norway[1], Universite de Saint Etienne, France[2]

## ABSTRACT

The Human Visual System (HVS) tends to focus on specific regions of viewed images or video frames, this is done effortlessly, instantly and unconsciously. These are called salient regions and form a saliency map, which could be used to improve a number of image and video processing techniques. In this paper, we propose a novel non-reference objective video quality metric based on the saliency map to improve the estimation of the perceived video quality. This metric estimates the degree of blur and blockiness in each video frame from the impaired video only, and uses it with the saliency map to derive a weighting function. The latter is used to modulate the contribution of the pixel differences to the final quality score. The salient regions of the videos are automatically computed using our video saliency model. A psychophysical experiment is conducted to estimate the perceived quality of the impaired videos. The results of this subjective test are compared to the scores obtained with the proposed objective metric. The objective and subjective scores are found to be highly correlated, which shows that our metric correctly estimates the perceived quality of a video.

## 1. INTRODUCTION

Recent improvements in imaging and video technology allowed us to capture and record large collections of videos. When we need to share these videos on internet, it is hard to do so because of their big size. Therefore, it is always preferred to compress video for storage and transmission. During the compression the estimation of the resulting video quality is an important factor to determine its usefulness for a specific application. For several applications one may want to estimate the perceptual quality of the compressed video. For instance for video communication one would need a model to estimate the perceived video quality at the receiving end to tune the parameters of the encoder. Similarly, in real time video surveillance system, a number of cameras may need to be controlled for proper functioning to ensure a certain level of quality of the recorded videos. This may be useful to account for camera malfunction or to adjust it to the changes in the visual scene, such as changes in the illumination or weather conditions etc. It may be of crucial importance to the surveillance application to have a certain quality of the recorded video for person identification or license plate reading for example. In [11] it was shown that several video surveillance systems record videos which are of no real value for the reliable identification of the faces. Therefore such systems are basically legally blind since they offer no help in court.

There are three types of quality matrices i-e, a full-reference quality metric that takes original and degraded video and computes the difference or quality degradation; the second type of metric is non-reference quality metric, it computes the quality degradation based only on the impaired video and third type is called reduced reference quality metric, that computes certain features from the original and de-graded videos, and finds correlation/match between them.

Under normal viewing conditions human eye movements are tightly coupled to human visual attention [1]. It is known that humans direct attention to the important objects in a scene (image/video frame) using bottom-up and top-down cues [2, 5]. Bottom-up cues use low-level features such as color, orientation, and intensity to compute the maps called conspicuity maps. However top-down models use high level features such as face-detection, object/people detection etc. A saliency maps could be computed automatically using top-down and bottom-up approaches [2]. In [2], authors used high level feature such as face detection with low level features such as color, intensity, orientation to compute the saliency maps for images and improved the saliency maps by 33 percent. Similarly saliency maps for videos can be computed by considering the temporal changes such as object motion along with stationary saliency maps. A full referenced quality metric based on visual attention modeling for images and videos has been proposed in [4]. The authors have used saliency detection to improve PSNR and SSIM.

The rest of this paper is organized as follows: first we discuss our perceptual model for saliency detection, in Section 2. Our no-reference video quality metric is proposed in Section 3. Section 4 presents the subjective psychophysical test and its results; we also compare these results with the proposed quality metric and PSNR. The last section concludes the paper with some future directions.

## 2. MULTI-FEATURE PERCEPTION MODEL FOR SALIENCY DETECTION

A region in a video sequence frame may be considered salient when it stands out from its spatial (stationary saliency) or temporal surrounding regions (motion saliency). Stationary saliency is performed using multi feature conspicuities

978-1-4244-7167-6/10/$26.00 ©2010 IEEE

including high level features such as face and low level features such as color intensity and orientation. Motion saliency is calculated based on motion analysis and distance effect on visual perception. Both stationary saliency map and motion saliency map are used to create a frame saliency map. In the following three sections we present the algorithm for computing the: stationary saliency model with face as a high level feature and intensity, color, orientation as low level features, motion saliency model with motion vector field measurements and distance weights in Gaussian model and fusion method for stationary and motion saliency maps.

### 2.1. Stationary Saliency Model

Stationary saliency map (SSM) is computed in two steps, using low level feature such as color, intensity and orientation and high level features such as face detection. Itti's bottom-up attention model [5] is used to compute low level features (color $C_c$, intensity $C_i$, and orientation $C_o$) represented as conspicuity maps. Seven conspicuity maps are computed, one for intensity, four for orientations 0, 45, 90 and 135 degrees, and two for color combinations Red-Green & Blue-Yellow. These conspicuity maps are combined after normalization step as shown in equation 1.

$$C_{itti} = \frac{1}{7}(C_i + 2C_c + 4C_o) \tag{1}$$

Psychlogical studies show that faces, heads, and hands attracts human attention [7]. Faces and text attracts human gaze independent of the task [8]. Itti's model does not consider high level features such as faces. So face conspicuity map can also be added with itti's stationary saliency map. In this paper, we have used Walther et al face detection model [6] to compute face conspicuity map $C_{face}$. This face detection algorithm compute Gaussian modell for skin hue color distribution. Itti's low level feature's conspicuity maps can be combined with face conspicuity maps as in equation 2.

$$SSM = f(C_{itti}, C_{face}) \tag{2}$$

The $f$ function can be defined impirically. In our case we used a linear combination as shown in the equation 3.

$$SSM = \frac{1}{8}(2C_i + 2C_c + C_o + 3C_{face}) \tag{3}$$

### 2.2. Motion Saliency model

It was shown in [9] that motion dominates other low level features while watching videos. Therefore, motion saliency information is added to our proposed saliency model. Motion attention model based on spatial-temporal entropy proposed by [10] is used to compute the motion saleincy map. Motion saliency map is computed for each frame of a video from motion vectors. The motion vectors are computed using Motion Vector Block matching algorithm between reference and target frames. The reference video frame is divided into macro blocks of size 16x16 pixels, and motion vectors are computed for each are independently. Motion saliency map (MSM) is computed using three inductors from motion vectors,i-e, intensity of the motion I, spatial coherence $C_s$ and temporal phase coherency $C_t$, as given in equation 4 taken from [10].

$$MSM = I * C_t * (1 - I * C_s) \tag{4}$$

### 2.3. Fusion of stationary and motion saliency map

$SSM$ and $MSM$ can be combined to obtain the final saliency map of every frame in a video. Since we usually are more susceptible to those objects in the centre of the frame than those that are far away from the center $(x_c, y_c)$, we propose to use the following distance weighting fusing model with $\alpha = 0.5$.

$$S_{VG} = \alpha * MSM + (1 - \alpha) * SSM \tag{5}$$

**Fig. 1**. Video Sequence 1(a) Original video frame (b) blur in full frame (c) blur in non-salient regions (d) blur in salient regions.

The proposed saliency detection model is used to detect the salient regions in the video frames. The salient regions are used for two purposes. One is to add various kinds of artifacts, as explained in the next section, to the salient /non-salient regions of the video frame. Second purpose is to use saliency maps to compute the quality score for each video, where the salient regions are assigned higher weights than the non-salient regions. The original and impaired video frames of video 1 are shown in Figure 1. The salient regions are highlighted in Figure 2. Figure 3 shows a frame from a video sequence used in the subjective experiment.

## 3. THE PROPOSED QUALITY METRIC

Most of today's compression standard such as MPEG use $8 \times 8$ block size for DCT compression. This causes various artifacts to appear in the compressed videos. Two of such most common artifacts are blocking and blurring, both degrade the video quality drastically. We have proposed

**Fig. 2**. Salient region in a video frame.

a quality metric that detects these blocking and blurring artifacts of video frame and computes the quality score for the overall video. The Saliency maps computed in the last section are used as weighting function to the distorted (blurred or blocky) video sequences. The pixels which are more salient are given higher weights than the less salient neighboring pixels.

**Fig. 3**. Sample video sequence 2 (a) original video frame (b) blur in non-salient region.

The value of the blocking artifact $Blc_v$ across two horizontally adjacent blocks, represents a measure of the discontinuity at the vertical boundary between the two blocks. This value is computed in the following way, first the vertical discontinuity is evaluated for each line across the two blocks. This vertical discontinuity is computed as the absolute difference of the two extrapolated values, $(E_l)$ and $(E_r)$, across the boundaries of two adjacent blocks using first order extrapolator as:

$$E_l = \frac{3}{2} * y_1 - \frac{1}{2} * x_1 \tag{6}$$

$$E_r = \frac{3}{2} * y_2 - \frac{1}{2} * x_2 \tag{7}$$

Where $x1$, $x2$ is the $7^{th}$ and $8^{th}$ column value of the first $8 \times 8$ block and $y1$, $y2$ is the $1^{st}$ and $2^{nd}$ column value of the the second adjacent block in horizontal direction.

The vertical artifact value is the mean of the eight discontinuities within a single block. Where $(E_r)_j$ is the $i^{th}$ row extrapolated values.

$$Blc_v = \frac{1}{8} \sum_{j=0}^{7} \left| (E_r)_j - (E_l)_j \right| \tag{8}$$

The values for the horizontal artifacts can be calculated in similar fashion. A blockiness score can be estimated by summing up the vertical and horizontal blockiness artifacts.

$$B_S = Blc_v + Blc_h \tag{9}$$

Blur on the other hand is hard to compute. It is usually caused by the quantization process and often by the deblocking filter.

Blur can be calculated across the horizontal and vertical boundaries of the $8 \times 8$ adjacent blocks. Local variance [3] may be used to estimate the blurriness in an image constituting the salient blocks. We first compute the local variance across the vertical blocks and then the horizontal ones. The local variance is given by:

$$\sigma = \frac{\sqrt{\sum_{i=1}^{n} |x_i - y_i|}}{n-1} \tag{10}$$

Where $n = 2$, and $x_i$ and $y_i$ are the adjacent pixel values. Next we compute the average of these local variances along row $j$, as follows:

$$\overline{\Delta\sigma_j} = mean\{\sigma_{ji} - \sigma_{j(i+1)} | i \in \{1, 2, ..., K\}\}, \tag{11}$$

Where $K$ is the number of $8 \times 8$ blocks in the horizontal direction of the image. The sum of total blur across vertical direction is given by

$$Blr_v = \sum_{j=1}^{N} \overline{\Delta\sigma_j} \tag{12}$$

Where $N$ is the total number of rows in the image.

The value for the horizontal blur is calculated in similar fashion.

An overall video quality value is obtained by combining the features extracted from the dataset. First the average blocking and blurring values are obtained by combining the vertical and horizontal artifacts.

$$Blc = \left( \frac{Blc_v + Blc_h}{2} \right) \tag{13}$$

$$Blr = \left( \frac{Blr_v + Blr_h}{2} \right) \tag{14}$$

Then the following prediction model is used to combine the artifacts

$$QPM = 10 \times \left( \beta + \delta \times Blc^a \times Blr^b \right) \times T^c \tag{15}$$

Where $\beta$ and $\delta$ are adjusted based on the mean opinion score from subjective tests. The values of $a = -0.24$, $b = -0.16$, and $c = 0.06$ are estimated from the image dataset used to train the algorithm using a non-linear regression routine. While $T$ is a perceptual threshold obtained via the subjective test questionnaire and it represents the acceptable amount of blocking and blurring artifacts in an image. $T$ is used to fine-tune the parameters $\beta$ and $\delta$ by omitting any contradictory mean opinion score value from subjective tests.

## 4. EXPERIMENTAL TESTS AND RESULTS

A total of 90 impaired videos are created by adding blur, compression, and blur plus compression artifacts in salient regions only, in non salient regions only and in the full

frame of the two video sequences. These impaired videos along with the two originals were shown to the subjects. The original videos were shown at the start while the impaired videos were shown in random fashion. Sixteen non-expert subjects participated in the subjective experiment. At the end of each impaired video they are asked to rate the quality on a scale of 1 to 5. Where 1 corresponds to really annoying and 5 corresponds to the imperceptible video quality. Mean opinion score (MOS) was then obtained which was used to correlate with the objective score obtained from the quality prediction metric.

The results are show in Table 1. The table shows Pearson, Spearman and Kendall correlation coefficients between PSNR and MOS & QPM and MOS for video sequence 1,2 and for both 1 & 2. In case of video sequence 1, QPM has higher correlation with MOS than PSNR correlation with MOS. The correlation coefficients for video sequence 2 and correlation coefficient for both video sequences 1 & 2 combined show a little higher correlation in case of QPM than PSNR. Overall results depict that QPM always perform better than PSNR. Figure 4 shows the scatter plots of PSNR vs MOS and QPM vs MOS. The less scattered the data values are the better correlated they are. That's why it is clear from the graph that QPM is better correlated with MOS than PSNR to MOS.

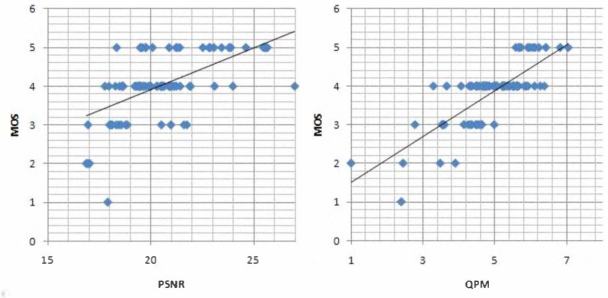


**Fig. 4**. Scatter plots between MOS and PSNR & QPM.

**Table 1**. Correlation coefficient for PSNR and QPM, with MOS.

| Correlation | Video 1 | Video 2 | Both |
|---|---|---|---|
| Pearson(PSNR;MOS) | 0.405 | 0.787 | 0.568 |
| Pearson(QPM;MOS) | 0.872 | 0.927 | 0.782 |
| Spearman(PSNR;MOS) | 0.374 | 0.805 | 0.546 |
| Spearman(QPM;MOS) | 0.811 | 0.834 | 0.75 |
| Kendall(PSNR;MOS) | 0.305 | 0.692 | 0.447 |
| Kendall(QPM;MOS) | 0.701 | 0.727 | 0.636 |

## 5. CONCLUSIONS AND FUTURE DIRECTIONS

We have proposed an objective non-reference metric for perceptual quality evaluation for video. This metric estimates blockiness and blur artifacts strength in the video and gives higher weights to their contribution when they are in the salient region. Saliency has been introduced to incorporate the HVS. The results obtained with our quality metric show high correlation with the subjective MOS obtained from the psychophysical experiment. The results are also shown to be better than those of PSNR. Our proposed metric(QPM) is non-reference, which makes it suitable for applications such as video streaming or surveillance videos quality evaluation. More tests with different types of videos and impairments will be performed to make the metric more generic.

## 6. REFERENCES


[1] T. Jost, N. Ouerhani, R. V. Wartburg, R. Muri, and H. Hugli, Computer Vision and Image Understanding, Elsevier 100-107, 2005.

[2] P. Sharma, F. A. Cheikh, and J. Y. Hardeberg, in Sixteenth Color Imaging Conference (The Society for Imaging Science and Technology, 2008), vol. 16, pp. 332-337, 2008.

[3] Liu Debing; Chen Zhibo; Ma Huadong; Xu Feng; Gu Xiaodong, "No Reference Block Based Blur Detection," Quality of Multimedia Experience, 2009., vol., no., pp.75-80, 29-31 July 2009.

[4] You, J., Perkis, A., Hannuksela, M. M., and Gabbouj, Perceptual quality assessment based on visual attention analysis. In Proceedings of the Seventeen ACM international Conference on Multimedia, China, 561-564, 2009.

[5] L. Itti, Ph.D. thesis, Models of Bottom-Up and Top-Down Visual Attention, California Institute of Technology, Pasadena, California, 2000.

[6] Walther, D., Koch, "Modeling Attention to Salient Proto-objects", Neural Networks 19, 1395-1407, 2006.

[7] R Desimone, TD Albright, CG Gross and C Bruce. " Stimulus selective properties of inferior temporal neurons in the macaque", Journal of Neuroscience, vol4, 2051-2062, 1984.

[8] Cerf, M., Frady, E. P., and Koch, C., "Faces and text attract gaze independent of the task: Experimental data and computer model". Journal of Vision, 9(12):10, 1-15, 2009.

[9] Dwarikanath Mahapatra, Stefan Winkler, and Shih-Cheng Yen, "Motion saliency outweighs other low-level features while watching videos". Proc. SPIE 6806, 2008.

[10] Yu-Fei Ma; Hong-Jiang Zhang, "A model of motion attention for video skimming," Image Processing. 2002. Proceedings. 2002 International Conference on , vol.1, no., pp. I-129-I-132 vol.1, 2002.

[11] Kovesi, P., "Video Surveillance: Legally Blind?", DICTA09, 204-211, 2009.